

Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation

Winston Hsu[†], Shih-Fu Chang[†], Chih-Wei Huang[†], Lyndon Kennedy[†], Ching-Yung Lin[‡], and
Giridharan Iyengar[§]

[†]Dept. of Electrical Engineering, Columbia University, New York, NY 10027

[‡]IBM T. J. Watson Research Center, Hawthorne, NY 10532

[§] IBM T. J. Watson Research Center, Yorktown Heights, NY 10598

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 measures up to 0.76 in ABC news and 0.73 in CNN news), present how these multi-level multi-modal features construct the probabilistic framework, and more importantly observe an interesting opportunity for further improvement.

Keywords: story segmentation, Maximum Entropy Model, multi-modal fusion, prosody, face detection, TRECVID, exponential model

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 works addressing news story segmentation. A well-known work is the Infromedia project,¹ in which rules on combining image, audio, transcripts, and closed captions are used to locate story boundaries. However, for international news programs, closed captions, and accurate speech recognizers are usually unavailable. Besides, the production rules vary from different channels or countries. Qi et al. identify the story boundaries by a clustering based algorithm that detects anchorpersons by performing an AND operation on visual and speech anchor segments.² 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 the key frame of each shot only and is sensitive to shot detection errors. A promising statistical framework⁴ based on Hidden Markov Model (HMM) is employed with a decision tree to classify shots into one of the 13 pre-defined categories. These category labels are then fed to HMM to locate story boundaries. In general, these approaches are categorized as heuristic rules or statistical approaches. 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 heuristic algorithms lack the generality in handling diverse video sources with different features and production rules.

In our prior work,⁵ 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

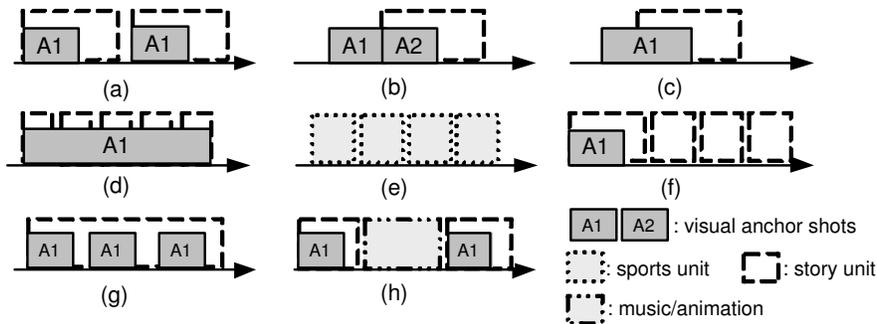


Figure 1. Common story types seen in the ABC/CNN programs, where $A1$ and $A2$ represent two visual segments of different anchor persons: (a) two normal stories start with an anchor person; (b) a story starts after switching to a different visual anchor; (c, d) stories start within anchor shots; (e) 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.

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 switches 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 2, where boundary detection $F1^*$ 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 5.1. In this paper, we will present significant performance gain over the baseline by using statistical multi-modal fusion and demonstrate satisfactory performance with $F1$ measures up to 0.76 in ABC news and 0.73 in CNN news.

In TRECVID 2003, we have to detect all the boundaries at the transition between segments such as those 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 2. The probabilistic framework and the feature wrapper are addressed in Section 3. Relevant features are presented in Section 4. The experiment, evaluation metrics, and discussions are listed in Section 5 and followed by the conclusion and future work in Section 6.

1.1. Data set

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

From the 111 videos in the development set, we found that the story length ranges from 4.05 to 223.95 seconds on CNN and from 7.22 to 429.00 seconds on ABC. The average story length on CNN is 42.59 second and is 71.47 on ABC. Apparently, CNN tends to have shorter and more dynamic stories.

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 as shown in Figure 5.1; the latter models the temporal dynamics within each story. We adopt the first approach in this paper. In such an

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

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

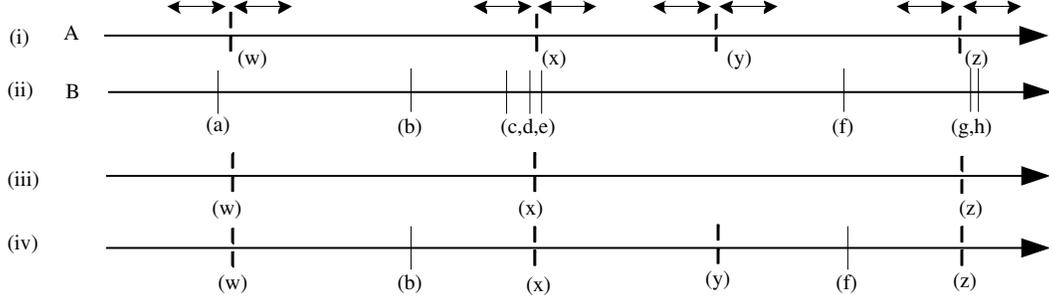


Figure 2. Example of temporal operators: (i) the sequence A is composed of time points $\{w, x, y, z\}$; (ii) B is the time sequence with points $\{a, \dots, h\}$; (iii) the *AND* operation $A \odot_{\epsilon} B$ yields $\{w, x, z\}$; (iv) the *OR* operation $A \oplus_{\epsilon} B$ unifies sequences A and B into $\{w, b, x, y, f, z\}$ by removing duplications within the fuzzy window ϵ , illustrated with " \longleftrightarrow ".

approach, one basic issue is the determination of candidate points, each of which is tested and classified as either a story boundary or a non-boundary.

2.1. Temporal sequence operators

During candidate determination, multi-modal fusion, or feature development, we conduct many temporal operations on time instances. To assist with these operations, we define two temporal sequence operators, *OR* \oplus_{ϵ} (Equation 1) and *AND* \odot_{ϵ} (Equation 2), with a fuzzy window ϵ ,

$$A \odot_{\epsilon} B = \{z | z \in A, \exists b \in B, |z - b| < \epsilon\}, \quad (1)$$

$$\begin{aligned} A \oplus_{\epsilon} B &= A \cup B - \{z | z \in B, \exists a \in A, |z - a| < \epsilon\}, \\ &= A \cup B - B \odot_{\epsilon} A. \end{aligned} \quad (2)$$

$A, B = \{z_n \in \mathbb{R}\}_{n \geq 1}$ represent two sequences of time points such as (i) and (ii) in Figure 2. We also need a filter operation to filter those time instances within certain segments with a fuzzy window ϵ . The filter function $\Psi_{\epsilon}(A, S)$ is defined in the following,

$$\Psi_{\epsilon}(A, S) = \{z | z \in A, \exists i, z \in [s_i - \epsilon, e_i + \epsilon], (s_i, e_i) \in S\}, \quad (3)$$

where $S = \{(s_i, e_i)\}$ is composed of segments defined by starting point s_i and end point e_i . The filter operation is to locate those points in A falling within the range of segments S with fuzzy window ϵ .

The "AND" operator locates those points from two time sequences coincide within a fuzzy window and reserves the time points from the first operand. The example is shown in (iii) of Figure 2. The "OR" operator unifies time points from two sequences but removes the duplications from the second operand within the fuzzy window. The example is illustrated in (iv) of Figure 2. Both operators are not commutative.

With these two operators and the filter, we could easily unify or filter the occurrence of features from different modalities. For example, we use \oplus_{ϵ} to unify audio pauses and shot boundaries into candidate points (Section 2.2). We also compactly represent the boundary performance metrics with \odot_{ϵ} in Equations 11 and 12. We later use the filter $\Psi_{\epsilon}(\cdot)$ to locate feature points in a specific region or state (Section 4.7). From our experiments, these operations contribute essential significance to locate story boundaries (Section 5.1).

2.2. Candidate points

A good candidate set should have a very high recall rate on the reference boundaries and indicate the places where salient and effective features occur. Shot boundaries are usually the candidate points used in most news segmentation projects.^{4,6} However, we found that taking the shot boundaries alone is not complete. We evaluate the candidate completeness by detecting reference boundaries with 5-second fuzzy window (defined in Section 5.1). Surprisingly, the recall rate for the shot boundaries on ABC/CNN 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 T_{shot} and audio pauses T_{pas} as candidate points but remove duplications within a 2.5-second fuzzy window by simply taking an "OR" operation $T_{shot} \oplus_{2.5} T_{pas}$. The union candidates yield 100% recall rate.

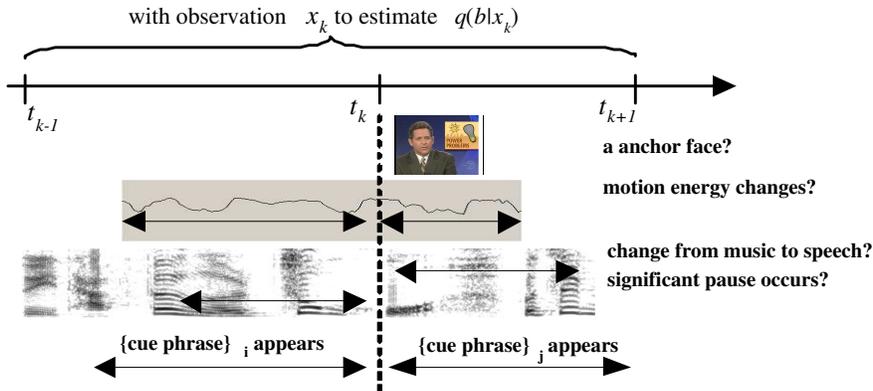


Figure 3. Estimation of the posterior probability $q(b|x_k)$, where $b \in \{0, 1\}$ is the random variable of boundary existence at the candidate point t_k , by fusing multiple mid-level perceptual features extracted from observation x_k .

2.3. 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 is 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.

3. 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 from 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 and leave details in our prior work.⁵

3.1. Maximum Entropy model

The ME model^{5,7} constructs an exponential log-linear function that fuses multiple binary features to approximate the posterior probability of an event (i.e., story boundary) given the audio, visual or text data surrounding the point under examination, as shown in Equation 4. 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 an indication function; g_i is a predictor of story boundary using the i 'th binary feature, generated from the feature wrapper (Section 3.2). f_i equals 1 if the prediction of predictor g_i equals b , and is 0 otherwise. The model is illustrated in Figure 5.1.

Given a labelled training set, we construct a linear exponential function as below,

$$q_\lambda(b|x) = \frac{1}{Z_\lambda(x)} \exp \left\{ \sum_i \lambda_i f_i(x, b) \right\}, \quad (4)$$

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 4 is a valid conditional probability distribution. Basically, λ_i controls the weighting of i 'th feature in estimating the posterior probability.

3.1.1. 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

$$\lambda^* = \operatorname{argmax}_\lambda D(\tilde{p} \parallel q_\lambda), \quad (5)$$

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

$$D(\tilde{p} \parallel q_\lambda) = \sum_x \tilde{p}(x) \sum_{b \in \{0,1\}} \tilde{p}(b|x) \log \frac{\tilde{p}(b|x)}{q_\lambda(b|x)}. \quad (6)$$

Meanwhile, minimizing the divergence is equivalent to maximizing the log-likelihood defined as

$$L_{\tilde{p}}(q_\lambda) = \sum_x \sum_b \tilde{p}(x, b) \log q_\lambda(b|x). \quad (7)$$

The log-likelihood is used to measure the quality of the estimated model and $L_{\tilde{p}}(q_\lambda) \leq 0$ holds all the time. In the ideal case, $L_{\tilde{p}}(q_\lambda) = 0$ corresponds to a model q_λ , which is "perfect" with respect to \tilde{p} ; that is, $q_\lambda(b|x) = 1$ if and only if $\tilde{p}(x, b) > 0$.

We use an iterative process to update λ_i till divergence is minimized. In each iteration, $\lambda'_i = \lambda_i + \Delta\lambda_i$, where

$$\Delta\lambda_i = \frac{1}{M} \log \left\{ \frac{\sum_{x,b} \tilde{p}(x, b) f_i(x, b)}{\sum_{x,b} \tilde{p}(x) q_\lambda(b|x) f_i(x, b)} \right\} \quad (8)$$

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 empirical 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 if it is measured based on the empirical distribution $\tilde{p}(x, b)$ or the estimated model $\tilde{p}(x)q_\lambda(b|x)$. When the exponential model underestimates the expectation value of feature f_i , its weight λ_i is increased. Conversely, λ_i is decreased when overestimation occurs.

3.1.2. Feature induction

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

$$q_{\alpha, h}(b|x) = \frac{\exp\{\alpha h(x, b)\}q(b|x)}{Z_\alpha(x)}, \quad (9)$$

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, divergence reduction, or likelihood increase. The selected feature h^* in each iteration is represented in Equation 10. h^* is then removed from the candidate pool C . The induction process iterates with the new candidate set $C - \{h^*\}$ till stopping criterion is reached (e.g., upper bound of the number of features or lower bound of the gain).

$$\begin{aligned} h^* &= \operatorname{argmax}_{h \in C} \left\{ \sup_\alpha \{D(\tilde{p} \parallel q) - D(\tilde{p} \parallel q_{\alpha, h})\} \right\} \\ &= \operatorname{argmax}_{h \in C} \left\{ \sup_\alpha \{L_{\tilde{p}}(q_{\alpha, h}) - L_{\tilde{p}}(q)\} \right\} \end{aligned} \quad (10)$$

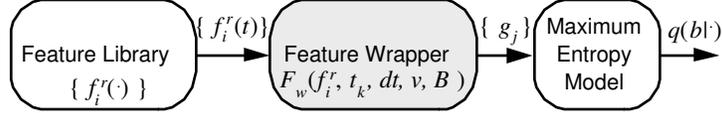


Figure 4. The raw multi-modal features f_i^r are collected in the feature library and indexed by raw feature id i and time t . The raw features are further wrapped in the feature wrapper to generate sets of binary features $\{g_j\}$, in terms of different observation windows, delta operations, and binarization threshold levels. The binary features are further fed into the ME model.

3.2. Feature wrapper

We have described the parameter estimation and feature induction processes from a pool of binary features in the previous sections. However, those available multi-modal multi-level features are usually asynchronous, continuous, or heterogeneous and require a systematic mechanism to integrate them. Meanwhile, the ME approach is based on binary features. For these purposes, we invent a feature wrapper to bridge raw multi-modal features and the ME model.

In Figure 4, 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 locations and 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 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 instance, 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 5) 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 $B = \{B_p, B_n, B_c\}$ around each candidate t_k . The first window is the interval B_p before the candidate point with window size T_w ; another is the same time-span B_n after the candidate; the other is the window B_c 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 sequentially 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.

3.3. 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 short 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.

4. RAW MULTI-MODAL MULTI-LEVEL FEATURES

The raw multi-modal features, repositied in the feature library as shown in Figure 4, are from different feature detectors. We summarize some relevant and novel features in this section. Other features such as motion intensity and music/speech discrimination could be found in our prior work.⁵

The shot boundaries are directly from the common reference shot boundaries of TRECVID 2003. The shots have no durations of less than 2 second (or 60 frames); short shots have been merged with their neighbors. Therefore, many shots (roughly 20%) actually contain several sub-shots. When a shot contains several sub-shots, the corresponding key-frame is always chosen within the longest sub-shot.

4.1. Anchor face

A salient visual feature is the visual anchor segments reoccurring in the news video. Though the TRECVID 2003 videos are from TV news programs, the video quality varies a lot due to different recording resolution and lighting conditions. It is still challenging to locate visual anchor segments from these 218 videos. Our prior work⁵ locates the visual anchor segments in three steps. (1) We first find those prospective face regions at each I-frame in each shot by an efficient face detector.⁸ It locates macro-blocks with possible skin-tone colors and further verifies vertical and size constraints both from DCT coefficients. The detector reports static face regions in each I-frame. (2) Within each shot, we take into account the temporal consistence by counting the face appearance frequency at each macro-block from the same shot to ensure that the faces should appear consistently in the same shot. (3) Regions of interest, extended from detected face regions, of each shot are extracted and are further featured with HSV color histograms. A distance matrix between HSV histograms of regions of interest is later yielded and fed to an unsupervised agglomerative clustering algorithm to locate the dominant cluster which implies the anchor segments in the entire video.

To boost the performance, we add another face detection approach^{9‡} that uses GMM skin-tone model and geometric active contour to locate the possible set of face regions. From them, we repeat steps 1-3 to yield another possible anchor segments. Another unsupervised agglomerative clustering algorithm is applied on these two sets of anchor segments to distill more correct results.

4.2. Commercial

Frame matching based on image templates such as station logos and caption titles is used to discriminate commercial and non-commercial sections since we observe that in CNN and ABC news the non-commercial portions are usually with certain logos or caption titles representing the station identification. From the commercial detector, in the entire video, we label each frame as "1" if it is in the non-commercial portion (matched templates found in this frame) and "0" otherwise. The process yields the binary sequence $A \in \{0, 1\}$ of the entire video. However, the detection process could not avoid the noise due to the dynamic content within the commercial and the variances of production rules. Two morphological operators are further applied and yield a smoothed result A' with temporal consideration as the following,

$$\begin{aligned} A' &= (A \circ M_W) \bullet M_W, \\ M_W &= u[n] - u[n - W]. \end{aligned}$$

[‡]Thanks to Dongqing Zhang of Columbia University for providing the geometric active contour face detection system.

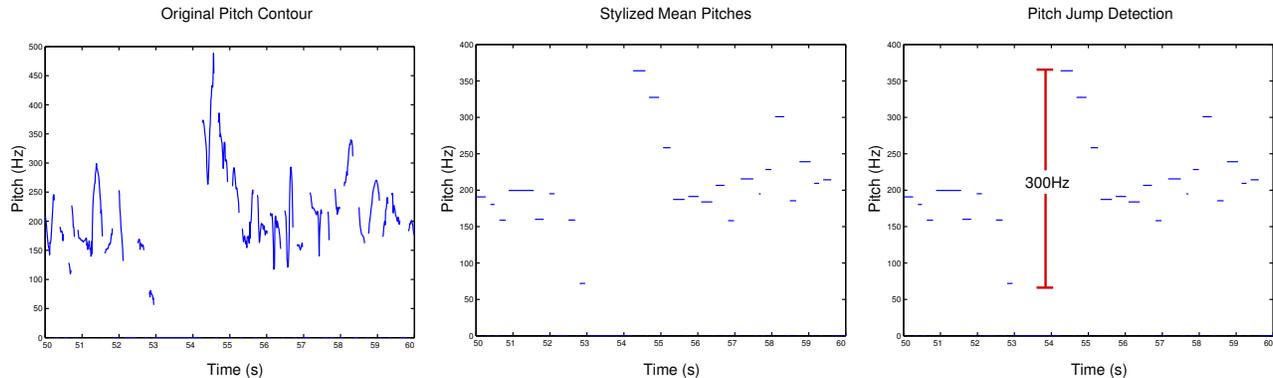


Figure 5. Visualization of original pitch contour, stylized chunk-level mean pitches, and pitch jump detection. For better illustration, the example is not normalized with the mean pitch.

From A' , we yield more correct commercial and non-commercial segments. Here, \circ and \bullet are morphological *OPEN* and *CLOSE* operators; M_W is a mask with W pulses represented by step function $u[n]$. We choose $W = 450$ and hypothesize that the caption titles or station logos in a non-commercial section might disappear but not longer than 450 frames or 15 seconds and there should be no caption logos lasting longer than this duration in the commercials.

4.3. Pitch jump

Pitch contour has been shown to be a salient feature for the detection of syntactically meaningful phrase and topic boundaries^{10, 11} and independent of language and gender.¹¹ A particularly useful behavior in pitch contour has been described as "pitch reset."¹² This behavior is characterized by the tendency of the speaker to lower his or her pitch towards the end of a topic and then to raise it, or reset it, at the beginning of the new topic. Past efforts have tried to characterize and detect this feature as a statistical combination of mean pitch, pitch variance¹³ or stylized pitch contour,¹² where slope of pitch change is relied upon heavily. We hypothesize that the mean pitch change will be sufficient for our task. Since the mean pitch and pitch variances vary between speakers, the pitch distributions used in our pitch jump detection scheme are normalized over same-speaker segments and represented in octaves.

In the pitch jump detection scheme, the pitch contour is extracted from audio streams automatically with the *Snack toolkit*[§]. The pitch contour is sampled at 100 estimates per second and is given as a pitch estimate (in Hz). The pitch estimates are converted into octaves by taking the base-2 logarithm of each estimate. To remove absolute pitch variances from different speakers, the octave estimates are then normalized in same-speaker segments by incorporating outputs from the ASR system (Section 4.5). The ASR system identifies segments in which there is a single active speaker. We iterate through all of these segments and normalize the octave pitch estimates according to the the mean in each coherent segment.

The pitch jump points were found by searching for points in the speech where the normalized magnitude of the inter-chunk pitch change was above a certain normalized threshold. We implemented the pitch jump detector by first segmenting the pitch contour into chunks, which are groups of adjacent valid pitch estimates. We then find the mean normalized pitch of each chunk. After that, we find the change in pitch between each chunk by finding the difference or change between the mean normalized pitches of adjacent chunks. A threshold of the mean of the positive pitch change is then applied to all of the chunk boundaries. Chunk boundaries where the normalized pitch change is greater than the threshold are then selected as pitch jump points T_{pcj} . Figure 5 illustrates the processing applied to the pitch contour to find pitch jump points.

[§]The Snack Sound Toolkit: <http://www.speech.kth.se/snack/>

Table 1. Boundary evaluation with significant pauses. The "uniform" column is to generate points uniformly with the same mean interval between significant pause points; it is 40.76 second in ABC and 41.80 second in CNN. The performance is evaluated with two fuzzy windows, 2.5-second and 5.0-second.

Set	ϵ	Significant Pause			Uniform		
		P	R	$F1$	P	R	$F1$
ABC	5.0	0.20	0.38	0.26	0.10	0.22	0.14
	2.5	0.16	0.34	0.22	0.10	0.22	0.14
CNN	5.0	0.40	0.45	0.42	0.20	0.24	0.22
	2.5	0.37	0.43	0.39	0.20	0.24	0.22

4.4. Significant pause

Significant pause is another novel feature that we have developed for the news video segmentation task. It is fairly inspired by the "pitch reset" behavior¹² that we discussed in Section 4.3 and the "significant phrase" feature developed by Sundaram.¹³ Significant pause is essentially an "AND" operation conducted on the pauses and the pitch jump points. We look for coincidences of pitch jump and pause in an attempt to capture the behavior where news anchors may simultaneously take a pause and reset their pitch contour between news stories.

In the significant pause detection scheme, we use the starting point of the pause (in seconds) and the duration of the pause (also in seconds), the pitch jump time (in seconds) and the normalized pitch jump magnitude. The significant pauses T_{sgps} are located by performing an *AND* operation, $T_{pcj} \odot_{0.1} T_{pas}$, on the pitch jump points T_{pcj} and pauses T_{pas} with 0.1-second fuzzy window. The normalized magnitudes and the pause durations associated with significant pauses are continuous raw features and would be further binarized with different thresholds in the feature wrapper.

To gauge the potential significance of this feature before fusing into the framework, we test it on the reference boundaries and show the performance on ABC/CNN in Table 1. Taking the feature in CNN videos, we found its F1 measure to be 0.42, which is quite impressive compared with other features. As a comparison, another salient feature, anchor face on CNN, contributes to the story boundaries with a F1 measure of 0.51. In Table 1, we also compare the feature significance with uniformly generated points of the same mean interval between significant pause points. Apparently, significant pauses deliver more significance than uniformly sampled points both on ABC and CNN news. However, the feature performance on ABC is less significant than that on CNN.

According to our observation, the feature points usually match the boundaries where the anchors tend to raise a new topic. Upon further inspection, we find that the non-significance of significant pause on ABC and the success of that on CNN is due not to an inconsistency in the feature detection, but to a difference in anchor behaviors and production styles on the two stations. The CNN data is from CNN Headline News, which is a newscast dedicated to bringing all of the top stories quickly in short formats. On CNN, the anchors often switch between short stories without any visual cues being conveyed to the viewers. They compensate for this lack of visual cue by emphasizing the change of story with their voice by injecting a pause and resetting their pitch. This pause/reset behavior is very salient and is the behavior that we were trying to capture with our features. Closer examination of the ABC data, however, revealed that the anchor behavior on story changes is rather different from CNN. The ABC news is a nightly newscast that presents fewer stories in 30 minutes than CNN news and dedicates more time to each story. On ABC, the stories rarely change without some sort of visual cue or without cutting back to the anchor from some field reports. This visual obviousness of story change causes the anchors to have less of a need for inserting story-change cues in their speech. We therefore attribute the weakness of the audio features on ABC, when compared with CNN, to a fundamental difference in anchor speech patterns on the two channels.

4.5. 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, separated by a long non-speech segment, a pause or music. The segment boundaries, starting or ending, might imply a story boundary. However, there are still some

boundaries that are within the speech segment; e.g., an anchor person briefs several story units continuously without a break.

In the ASR outputs along with the TRECVID 2003 data, as illustrated in the following paragraph, the XML tags are used to describe the recognized transcripts and associated structures. A speech segment is defined by the tag "`<SpeechSegment>`" and the associated attributes "`stime`" and "`etime`", which represent the starting and ending points of the segment. Transcribed words are described by the tag "`<Word>`." In this example, the speech region starts at 11.80 second and ends at 20.73 second in the video "19980204_CNN.mpg."

```
<Audiofile filename="19980204_CNN">
  <SpeechSegment lang="us" spkr="FS4" stime="11.80" etime="20.73">
    <Word stime="12.35" dur="0.34" conf="0.953"> THREE </Word>
    <Word stime="12.69" dur="0.49" conf="0.965"> EUROPEAN </Word>
    <Word stime="13.18" dur="0.64" conf="0.975"> BALLOONISTS </Word>
    ...
  </SpeechSegment>
</Audiofile>
```

We further measure the speech rapidity by counting words per second in each segment. Usually the speaker tend to speak faster at the start of the speech. For those speech segments with speech rapidity larger than the mean plus a stand deviation measured within the same video are fast speech segments S_{fsp} .

4.6. ASR-based story segmentation

The ASR-based story segmentation scheme in this work is a combination of decision tree and maximum entropy models. It is based on the IBM story segmentation system[¶] used in TDT-1 evaluation.¹⁴ It takes a variety of lexical, semantic and structural features as inputs. These features are calculated from the ASR transcript. The performance in the validation set is show in Table 2.

Prior to feature extraction, the ASR transcript is converted into "sentence" chunks. The preprocessing converts the transcript into chunks consisting of strings of words delimited by non-speech events such as a silence or pause. These chunks are then tagged by an HMM part of speech (POS) tagger and then stemmed by a morphological analyzer which uses the POS information to reduce stemming ambiguities. The feature extractor extracts a variety of features from the input stream, for example, the number of novel nouns in a short lookahead window.¹⁴

The decision tree model is similar to what Franz et al. proposed.¹⁵ It uses three principal groups of input features to model the probability of a story boundary. Our observation has been that the most important group of features related to questions about the duration of non-speech events in the input. The second group of features depends on the presence of key bigrams and key words indicating the presence of story beginning (such as "good morning", "in new-york"). The last group of features compares the distribution of nouns on the two sides of the proposed boundary. The ME models uses three categories of features as well. The first group of features encodes the same information used by the decision tree model. The next group of features looks for n-gram ($n \leq 3$) extracted from windows to the left and right of the current point. The last category of features is structural and detect large-scale regularities in the broadcast news programming such as commercial breaks.¹⁴

In the specific ASR transcript provided with the TRECVID 2003 evaluation data, we observe that a manually provided story boundary in the development data mostly does not correspond to non-speech events in the transcript. This is observed in 15-20% of the boundaries. One possibility is that the ASR system used to generate the transcript was not tuned to report on short pauses. Since our system relies on non-speech events to chunk the text and compute features, in the instances when the manual boundary did not correspond to a non-speech event in the transcript, we inserted a 0-second non-speech event to facilitate feature extraction at such points in the development data. We note that this is a limitation of our system as this pre-processing cannot be achieved for the test data. However, with tuning of the ASR engine to provide information about short-pauses, this limitation becomes less relevant.

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

4.7. Combinatorial features

We observe that some of the features in a specific state or the combination of some features would provide more support toward the story boundaries; e.g., the significant pauses are composed of pitch jump and pause points. With the help of these features, we are able to boost the challenging parts or rate events useful for story segmentation. Some of these features also present significance in the feature induction process as shown in Table 3. We further generate some combinatorial binary features based on previous features, temporal operators, and filters such as:

- *Pitch-jump near the start of the speech segments*: we tend to use this feature to catch the instance when a speech segment starts with a pitch reset. The feature is composed of the pitch jump points T_{pcj} and the start of speech segments T_{ssps} and are computed from $T_{pcj} \odot_{\epsilon} T_{ssps}$.
- *Significant pauses near shot boundaries*: we hope to use this feature to catch the short briefings without leading visual anchor segments at the start of stories but with shot changes such as types (e) or (f) in Figure 1. The feature is calculated with significant pauses T_{sgps} and shot boundaries T_{sht} by taking $T_{sgps} \odot_{\epsilon} T_{sht}$.
- *Fast speech segments within the non-commercial sections*: we hope to catch the starting points of fast speech segments T_{fsp} within the non-commercial segments S_{ncom} by taking $\Psi_{\epsilon}(T_{fsp}, S_{ncom})$.
- *Significant pauses within the fast speech segments or non-commercial segments*: we design these features to boost the detection in short news briefings. Usually the anchor tends to speak faster or has a pitch reset when changing the topic. The features are yielded by simply filtering significant pauses T_{sgps} within fast speech segments S_{fsp} or non-commercial segments S_{ncom} by filters $\Psi_{\epsilon}(T_{sgps}, S_{fsp})$ or $\Psi_{\epsilon}(T_{sgps}, S_{ncom})$.

5. 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.

5.1. Boundary detection performance

The segmentation measure metrics are precision P_{seg} and recall R_{seg} and are defined in the following. According to the TRECVID 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 11 and 12; $|\cdot|$ means the number of boundaries; B_{cpt} and B_{ref} are computed and reference boundaries and formal evaluation fuzzy window ϵ is 5 second.

$$P_{seg} = \frac{|\text{computed boundaries}| - |\text{false alarms}|}{|\text{computed boundaries}|} = \frac{|B_{cpt} \odot_{\epsilon} B_{ref}|}{|B_{cpt}|} \quad (11)$$

$$R_{seg} = \frac{|\text{detected reference boundaries}|}{|\text{reference boundaries}|} = \frac{|B_{ref} \odot_{\epsilon} B_{cpt}|}{|B_{ref}|} \quad (12)$$

The performance in the development set is shown in Table 2, 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 ABC stories being 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 and these short stories are types (e) and (f) in Figure 1. About CNN news, the A+V boosts

Table 2. Boundary detection performance in ABC/CNN news. In "A+V" and "A+V+T", a candidate point is assigned as a boundary if $q(1|\cdot) > b_m$ with $b_m = 0.5$, where q is the estimated posterior probability. While in "A+V (BM)" and "A+V+T (BM)", a boundary movement (BM) is conducted and b_m is shifted and determined in a sperate development set to maximize the F1 measure. Here we set the shifted values b_m 0.25 for CNN and 0.35 for ABC. Generally BM trades a lower precision for a higher recall rate.

Modalities	ABC			CNN		
	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	0.77	0.63	0.69	0.82	0.52	0.63
A+V+T	0.90	0.63	0.74	0.82	0.57	0.67
A+V (BM)	0.75	0.67	0.71	0.70	0.68	0.69
A+V+T (BM)	0.85	0.70	0.76	0.72	0.75	0.73

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.

Since we are estimating posterior probability $q(b|\cdot)$ accounting for the existence of a story boundary, a straightforward boundary decision is just to select those candidate points with $q(1|\cdot) > 0.5$. The results are presented in modalities "A+V" and "A+V+T" of Table 2. However, we found that the story segmentation problem with the ME model also suffers from imbalanced-data learning¹⁶ since the boundary samples are much fewer than non-boundary ones. The best F1 measure of boundary detection does not come from the decision threshold 0.5 but requires a boundary movement¹⁶ meaning that we have to move the posterior probability threshold from 0.5 to a certain shifted value b_m and a candidate point is assigned as a boundary if $q(1|\cdot) > b_m$. The shifted value b_m is determined in a sperate small development set to maximize the boundary detection F1 measure. Here we take the shifted values b_m 0.25 for CNN and 0.35 for ABC. Intuitively, a smaller BM trades a lower precision for a higher recall rate. The results with boundary movements are shown in modalities "A+V (BM)" and "A+V+T (BM)" of Table 2 and improve the most in CNN news.

To understand more how BM affects the boundary detection performance, we plot the precision vs. recall curves of story segmentation performance with modalities "A+V" and "A+V+T" on ABC/CNN news in Figure 6 by ranging b_m from 0.02 to 0.86.

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 for 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 (~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.

To illustrate the binary features induced in the feature selection process of the ME model, we list the first 12 induced features from the CNN A+V model in Table 3. Clearly, the anchor face feature is the most relevant feature according to the training set but the relevance also depends on the location of observation windows. The next induced binary feature is the significant pause within the non-commercial section from combinatorial features. The audio pauses and speech segments are also relevant to the story boundaries since that usually imply a topic change. Interestingly, those features further filtered with non-commercial sections deliver more significance than the original features since the video contains transitional dynamics or states and features in different states have different significance; for example, significant pauses in non-commercial segments are much more relevant to boundaries than those in commercials; also, the rapid speech segments in commercials are less relevant to story boundaries. Another interesting feature is the 7th induced binary feature that a commercial

Table 3. The first 12 induced features from the CNN A+V model. λ is the estimated exponential weight for the selected feature; Gain is the reduction of the divergence as the feature added to the previously constructed model. $\{B_p, B_n, B_c\}$ are three observation windows; B_p is before the candidate point; B_n is after the candidate point; B_c is surrounding the candidate.

Num.	Binary Feature id	Raw Feature Set	λ	Gain	Interpretation
1	160	Anchor face	0.4771	0.3879	An anchor face segment just starts in B_n .
2	142	Significant pause + Combinatorial	0.7471	0.0160	A significant pause within the non-commercial section appears in B_c .
3	91	Pause	0.2434	0.0058	An audio pause with the duration larger than 2.0 second appears in B_c .
4	3	Significant pause	0.7947	0.0024	B_c has a significant pause with the pitch jump intensity larger than the normalized pitch threshold v_{pcj}^0 and the pause duration larger than 0.5 second.
5	113	Speech segment	-0.3566	0.0019	A speech segment starts in B_p .
6	115	Speech segment	0.3734	0.0015	A speech segment starts in B_c .
7	183	Commercial	1.0782	0.0015	A commercial starts in 15 to 20 seconds after the candidate point.
8	117	Speech segment	-0.4127	0.0022	A speech segment ends in B_n .
9	156	Anchor face	0.7251	0.0016	An anchor face segment occupies at least 10% of B_n .
10	85	Pause	0.0939	0.0008	B_c has a pause with the duration larger than 0.25 second.
11	127	Speech rapidity + Combinatorial	0.6196	0.0006	A fast speech segment within the non-commercial section starts in B_c .
12	2	Significant pause	-0.5161	0.0004	B_n has a significant pause with pitch jump intensity larger than the normalized pitch threshold v_{pcj}^0 and pause duration larger than 0.5 second.

starting in 15 to 20 seconds after the candidate point would imply a story boundary at the candidate point. After our inspection on CNN news, it matches the dynamics of CNN news since before turning to commercials the video is finished back to the anchors who would take seconds to shortly introduce coming news. The binarization threshold for the commercial feature is selected by the feature induction process in the training set rather than by heuristic rules.

5.2. 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.

6. CONCLUSION AND FUTURE WORK

Story segmentation in news video remains a challenging issue even after years of research. We believe multi-modality fusion through effective statistical modelling and feature selection are keys to the 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.

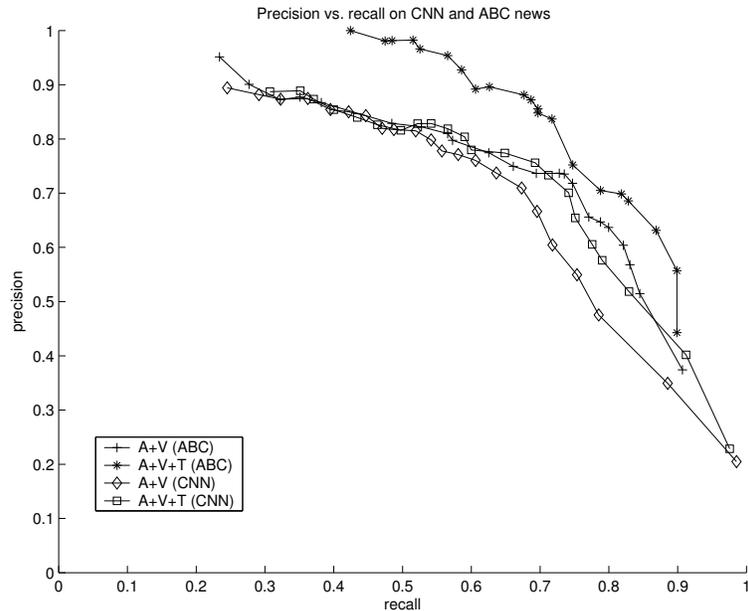


Figure 6. Precision vs. recall curves of story segmentation performance with modalities "A+V" and "A+V+T" on ABC/CNN news.

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.

REFERENCES

1. A. G. Hauptmann and M. J. Witbrock, "Story segmentation and detection of commercials in broadcast news video," in *Advances in Digital Libraries*, pp. 168–179, 1998.
2. W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, "Integrating visual, audio and text analysis for news video," in *7th IEEE Int'l Conference on Image Processing*, 2000.
3. Z. Liu and Q. Huang, "Adaptive anchor detection using on-line trained audio/visual model," in *SPIE Conf. Storage and Retrieval for Media Database*, (San Jose), 2000.
4. L. Chaisorn, T.-S. Chua, and C.-H. Lee, "The segmentation and classification of story boundaries in news video," in *IEEE International Conference on Multimedia and Expo*, 2002.
5. W. H.-M. 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.
6. S. Boykin and A. Merlino, "Machine learning of event segmentation for news on demands," *Communication of the ACM* **43**, February 2000.
7. D. Beeferman, A. Berger, , and J. Lafferty, "Statistical models for text segmentation," *Machine Learning* **34**(special issue on Natural Language Learning), pp. 177–210, 1999.
8. H. Wang and S.-F. Chang, "A highly efficient system for automatic face region detection in mpeg video," *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)* **7**(4), 1997.

9. D. Zhang, "Face detection in news video with gemoetric active contours," tech. rep., IBM T. J. Watson Research Center, 2003.
10. B. Arons, "Pitch-based emphasis detection for segmenting speech recordings," in *International Conference on Spoken Language Processing*, (Yokohama, Japan), 1994.
11. J. Vaissiere, "Language-independent prosodic features," in *Prosody: Models and Measurements*, A. Cutler and D. R. Ladd, eds., pp. 53–66, Springer, Berlin, 1983.
12. E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication* **32**, pp. 127–154, 2000.
13. H. Sundaram, *Segmentation, Structure Detection and Summarization of Multimedia Sequences*. PhD thesis, Columbia University, 2002.
14. 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.
15. S. D. M. Franz, J. S. McCarley, S. Roukos, and T. Ward, "Story segmentation and topic detection in the broadcast news domain," in *1999 DARPA Broadcast News Workshop*, 1999.
16. G. Wu and E. Chang, "Adaptive feature-space conformal transformation for imbalanced-data learning," in *The Twentieth International Conference on Machine Learning (ICML-2003)*, (Washington DC), 2003.