COLUMBIA-IBM NEWS VIDEO STORY SEGMENTATION IN TRECVID 2004

Winston H. Hsu†, Lyndon S. Kennedy†, Shih-Fu Chang†, Martin Franz‡, and John R. Smith‡

†Dept. of Electrical Engineering, Columbia University, New York
{winston, lyndon, sfchang}@ee.columbia.edu
‡IBM T. J. Watson Research Center, New York
{franzm, jsmith}@us.ibm.com

Columbia University ADVENT Technical Report #209-2005-3

ABSTRACT

In this technical report, we give an overview of our technical developments in the story segmentation task in TRECVID 2004. Among them, we propose an information-theoretic framework, visual cue cluster construction (VC^3), to automatically discover adequate mid-level features. The problem is posed as mutual information maximization, through which optimal cue clusters are discovered to preserve the highest information about the semantic labels. We extend the Information Bottleneck framework to high-dimensional continuous features and further propose a projection method to map each video into probabilistic memberships over all the cue clusters. The biggest advantage of the proposed approach is to remove the dependence on the manual process in choosing the mid-level features and the huge labor cost involved in annotating the training corpus for training the detector of each mid-level feature. When tested in TRECVID 2004 news video story segmentation, the proposed approach achieves promising performance gain over representations derived from conventional clustering techniques and even the mid-level features selected manually; meanwhile, it achieved one of the top performances, F1=0.65, close to the highest performance, F1=0.69, by other groups. We also experiment with other promising visual features and continue investigating effective prosody features. The introduction of post-processing also provides practical improvements. Furthermore, the fusion from other modalities, such as speech prosody features and ASR-based segmentation scores are significant and have been confirmed again in this experiment.

1. INTRODUCTION

In large video databases, the news story is a basic unit for browsing, summarization, and understanding. Automatic segmentation of continuous video programs into constituent story units is challenging due to the diversity of story types and the complex composition of attributes in various types of stories. We excerpt some of the common story types in Figure 5. Review of existing approaches and definition of news stories can be found in [1]. In this experiment, we try to discover adaptive and automatic visual and audio features and utilize them through effective multi-modal fusion approaches.

News channels are diverse and usually have different visual production events, across channels and over time, which are statistically relevant to story boundaries. These production events might be correlated to some "mid-level" features. Recent research in video analysis has shown a promising direction, in which mid-level features (e.g., people, anchor, indoor) are abstracted from low-level features (e.g., color, texture, motion, etc.) and used for discriminative classification of semantic labels [2, 3].

For years, researchers have tried different ways to manually enumerate all these mid-level features through inspection, and then train the specific classifiers for each. For example, in [2], 17 domain-specific detectors are trained as mid-level features such as *Intro/Highlight, Anchor, 2Anchor, People, Speech/Interview, Live-reporting, Sports, Textscene, Special, Finance, Weather, Commercial, LEDS, TOP, SPORT, PLAY*, and *HEALTH*. [1] The work requires intensive annotations and classifier training, which are usually limited to the designated channel. Another work in [4] uses specific anchor, commercial detectors, and clustering algorithms to determine classes of "visual delimiters," where human intervention is required. In recent work [5], authors manually train "section-"dependent classifiers to determine the story boundaries. The sections are similar to the mid-level features such as *Top Stories, Dollar and Sense*, start and end of *Headline sports*, and long anchor shots, etc. Most of them are detected by specific jingle spotting. Generally, researchers try to derive those mid-level representations, *which deliver semantic meanings or are statistically*

[1] See [2] for more explanations.

relevant to the target event, story boundaries, and then apply a discriminative or Bayesian approach to classify the story boundaries.

These prior approaches are intuitive but not feasible for extension to multiple channels. For example, if we hope to deploy the approach worldwide on 100 channels and for each channel we assume requiring 5 mid-level classifiers, the cost is overwhelming since we would have to totally annotate and train these mid-level classifiers about 500 times. Even worse, before that, human inspection is required to define domain-specific mid-level representations.

Such intensive work can be eased through other rigorous and automatic approaches. Motivated by cue word clustering or selection in information retrieval, we try to automatically and adaptively discover these mid-level representations while avoiding human inspection and time-consuming annotations, given the target or auxiliary labels, i.e. story boundaries.

Most of all, in this story boundary segmentation experiment, we propose an information-theoretic framework, visual cue cluster construction (VC^3), to automatically discover adequate mid-level features that are relevant to story boundary detection. The problem is posed as mutual information maximization, through which optimal cue clusters are discovered to preserve the highest information about the semantic labels. We extend the Information Bottleneck framework [6, 7, 8] to high-dimensional continuous features and further propose a projection method to map each video into probabilistic memberships over all the cue clusters. The biggest advantage of the proposed approach is to remove the dependence on the manual process in choosing the mid-level features and the huge labor cost involved in annotating the training corpora for training the detector of each mid-level feature. The proposed VC^3 framework is general and effective, leading to exciting potential in solving other problems of semantic video analysis. When tested in news video story segmentation, the proposed approach achieves promising performance gain over representations derived from conventional clustering techniques and even our prior manually-selected mid-level features.

In this experiment, we focus on the automatic approach to construct salient mid-level visual cue clusters. To exploit the support from other modalities, we also statistically fuse other features, such as speech prosody features [1] and ASR-based segmentation scores [9], which have been shown highly relevant to story boundary detection in the broadcast news videos. We adopt the Support Vector Machines (SVM) [10] as the major fusion approach.

The computational model and fusion approach is briefed in Section 2. The investigated feature set are presented in Section 3. In Section 4, we proposed two post-processing techniques for TRECVID 2004. The performance and multimodal fusion are discussed in Section 5, followed by conclusions and future work in Section 6.

2. DISCRIMINATIVE MODEL FOR STORY BOUNDARY DETECTION

News videos from different channels usually have different production rules or dynamics. We choose to construct a model that adapts to each different channel. According to our research in [11], the fusion capability of the discriminative model, such as SVM [10], is more effective than other generative or ensemble models. We train a SVM classifier to classify a candidate point as a story boundary or non-boundary. The features fed to the SVM classifier include the membership probabilities of the induced VC^3 clusters, ASR based segmentation scores [9], and speech prosody features [1], etc.

As our prior work [1], we take the union of shot boundaries and audio pauses as candidate points but remove duplications within 2.5-second fuzzy window. Our study showed these two sets of points account for most of the story boundaries in news. The issues of fusing heterogeneous and asynchronous features are solved with the same approaches addressed in our prior work [1].

2.1. Support Vector Machines

SVM has been shown to be a powerful technique for discriminative learning [10]. It focuses on structural risk minimization by maximizing the decision margin. We applied SVM using the Radial Basis Function (RBF) as the kernel,

$$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2), \gamma > 0.$$

In the training process, it is crucial to find the right parameters C (tradeoff on non-separable samples) and γ in RBF. We apply five fold cross validation with a grid search with varying (C, γ) on the training set to find the best parameters achieving the highest accuracy.

3. FEATURES

The major visual features we adopt in this experiment are the VC^3 features projected from low-level visual features towards those induced cue clusters for CNN and ABC channels (See Section 3.1 and [12]). The required low-level visual features for VC^3 are discussed in Section 3.5. We continue investigating prosody features in Section 3.2. Additionally, there are other features inherited from prior work [1]; for example,commercial detection is used in post-processing (Section 4.2); speech rapidity, speech change, and motions are later fused respectively.

3.1. Visual Cue Cluster Construction (VC^3) Features

In the research of video retrieval and analysis, a new interesting direction is to introduce "mid-level" features that can help bridge the gap between low-level features and semantic concepts. Examples of such mid-level features include location (indoor), people (male), production (anchor), etc., and some promising performance due to such mid-level representations have been shown in recent work in news segmentation and retrieval [2, 13]. It is conjectured that mid-level features are able to abstract the cues from the raw features, typically with much higher dimensions, and provide improved power in discriminating video content of different semantic classes. However, selection of the mid-level features is typically manually done relying on expert knowledge of the application domain. Once the mid-level features are chosen, additional extensive manual efforts are needed to annotate training data for learning the detector of each mid-level feature.

Our goal is to automate the selection process of the mid-level features given defined semantic class labels. Given a collection of data, each consisting of low-level features and associated semantic labels, we want to discover the mid-level features automatically. There is still a need for labeling the semantic label of each data sample, but the large cost associated with annotating the training corpus for each manually chosen mid-level feature is no longer necessary. In addition, dimensionality of the mid-level features will be much lower than that of the low-level features.

Discovery of compact representations of low-level features can be achieved by conventional clustering methods, such as K-means and its variants. However, conventional methods aim at clusters that have high similarities in the low-level feature space but often do not have strong correlation with the semantic labels. Some clustering techniques, such as LVQ [14], take into account the available class labels to influence the construction of the clusters and the associated cluster centers. However, the objective of preserving the maximum information about the semantic class labels was not optimized.

Recently, a promising framework, called Information Bottleneck (IB), has been developed and applied to show significant performance gain in text categorization [6, 7, 8]. The idea is to use the information-theoretic optimization methodology to discover "cue word clusters" which can be used to represent each document at a mid level, from which each document can be classified to distinct categories. The cue clusters are the optimal mid-level clusters that preserve the most of the mutual information between the clusters and the class labels.

In the visual features of this experiment, we propose new algorithms to extend the IB framework to the visual domain, specifically video. Starting with the raw features X such as color, texture, and motion of each shot, our goal is to discover the cue clusters C that have the highest mutual information or the least information loss about the final class labels Y, such as video story boundary or semantic concepts. Note that X and Y are given in advance. The optimization criteria given some constrain R under IB principle can be form as the following:

$$C^* = \underset{C|R}{\operatorname{argmin}}\{I(X;Y) - I(C;Y)\},$$

where $I(X;Y) = \sum_y \sum_x p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$ is the mutual information between two discrete-valued random random variables X and Y as defined in [15]. For continuous high-dimensional feature variables, the approximation is proposed in [12].

Our work addresses several unique challenges. First, the raw visual features are continuous (unlike the word counts in the text domain) and of high dimensions. We propose a method to approximate the joint probability of features and labels using kernel density estimation (KDE) [16]. Second, we propose an efficient sequential method to construct the optimal clusters and a merging method to determine the adequate number of clusters. Finally, we develop a rigorous analytic framework to project new video data to the visual cue clusters by taking into account the cluster prior probabilities and feature likelihood for each of the specific cue clusters. The probabilities of such projections over the cue clusters are then used for the final discriminative classification of the semantic labels or fused with other modalities such as text or speech prosody features.

The fusion results with other modalities and the comparison with our prior work [1], where a set of manually selected mid-level features are used, are presented in Section 5. More details regarding the VC^3 framework are addressed in [12].

3.2. Prosody features

In story segmentation tasks in broadcast news video, the timing and intonation (or *prosody*) of the news anchor's speech have been shown to be powerful cues for the detection of story-change cues. For example, many anchors exhibit a behavior at the end of a story where they lower their pitch, slow their speech and pause. They then indicate the start of the next story by resetting their pitch to a high level and speeding up their speech. This behavior has been leveraged in previous works on story segmentation and has been shown to be independent of language and gender [17, 18, 19, 20, 1].

To further investigate prosody features, we devise a set of low-level speech features which can be extracted automatically from audio signals in fixed windows around candidate topic change points and use SVM to learn the characteristics of story boundary points.

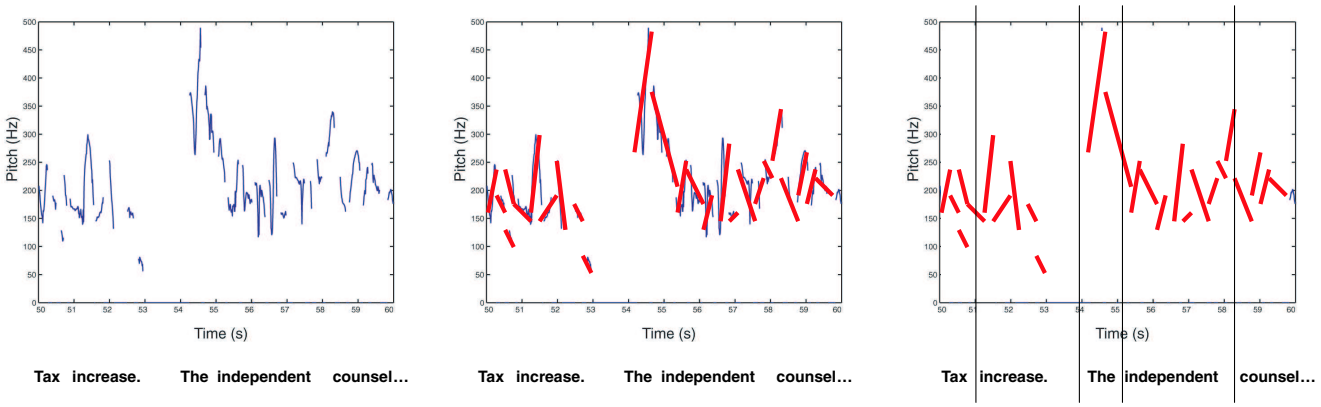


Fig. 1. Preprocessing of pitch features.

The LIMSI ASR transcripts [21] of the news broadcasts are used to determine candidate topic change points. The beginning of each new word is taken to be a candidate topic change point and the features are calculated within a fixed window of each candidate point. An SVM is then learned to distinguish between the story boundary and non-boundary points. The normalized SVM margins at each candidate and its associated time code are later fused with other modalities.

3.2.1. Word Rate

Word rate is determined by counting the number of tokens appearing in the ASR transcript within a 5 second window before and after each candidate point.

3.2.2. Pause Duration

Pause duration is calculated by statistically characterizing the lengths of the pauses within 1 second and 5 second windows before and after each candidate point. The lengths of the pauses are determined by subtracting the end time of each token appearing in the ASR transcript within the window from the start time of the following token. The mean, variance, minimum, and maximum of the pause lengths in each of the two window sizes are then used as the pause duration features.

3.2.3. Pitch

We extract pitch estimates for each 10 ms segment of audio in the data set using the ESPS method in the Snack Sound Toolkit [22]. The pitch estimates are then converted into an octave scale and then normalized according to the mean and variance of the active speaker's pitch. The active speaker is determined by using the speaker turn annotations from the speech recognizer. Each segment is only normalized locally, within each same-speaker turn. Continuous voiced segments (contiguous estimates with accuracy probabilities higher than a certain threshold) are then fit with a first-degree polynomial regression line to smooth out irregularities in the pitch estimates. This preprocessing of the pitch estimates is shown in Figure 1.

- *pitch features*: The smoothed and normalized pitch estimates are then used to determine some pitch features within 1 second and 5 second windows before and after each candidate point. The mean, variance, minimum, and maximum of the smoothed pitch estimates are calculated in each window size before and after each candidate point and taken to be the set of pitch features.
- *pitch slope features*: The slopes of the smoothed and normalized pitch estimates are then used to determine some additional features of the pitch contour. The mean, variance, minimum and maximum of the the slopes of the pitch estimates in the 1 second and 5 second fixed windows are again used to determine the values of the pitch slope features for each candidate point.

3.2.4. Duration of Voiced Segments

The duration of the voiced segments is determined by statistically characterizing the lengths of continuous voiced segments as shown in the output of the pitch estimator. The pitch estimator tends to give pitch estimates for voiced vowel segments of speech and does not give estimates for noise caused by non-vowel phonemes and silence. We take each set of consecutive pitch estimates, which is not interrupted by a non-estimate, to be a voiced segment. The duration of each voiced segment is then just the time in seconds for which it is sustained. The duration features are then determined by examining the durations of voiced segments within 1 second and 5 second windows before and after each

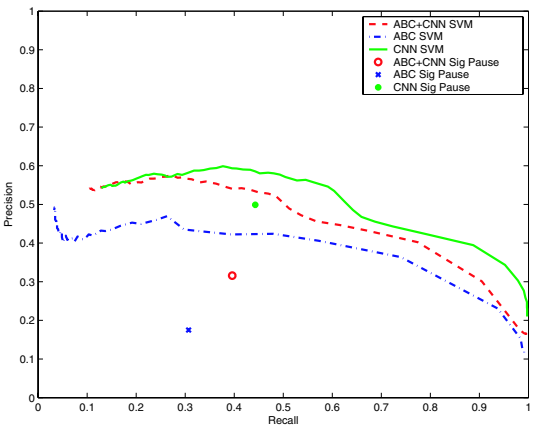


Fig. 2. Performance of SVM using rich prosody features and single "significant pause" feature.

candidate point and taking the mean, variance, minimum and maximum of the voiced segment lengths in each of the two window sizes and the duration of voiced segments features.

3.2.5. Prosody performance

To demonstrate the effectiveness of the rich prosody features, we train SVM for classifying each candidate point as a story boundary or a non-story boundary by using the features described in the previous section. We also compare the performance of this system with the "significant pause" features, which are an ad hoc combination of long pauses and jumps in pitch, from our previous work [1]. The results are shown in Figure 2.

We see that our current method shows significant improvement over our previous methods. Some trends, such as the tendency to underperform on the ABC broadcasts by a considerable margin are still present in our current approach. This may be due to the lack of story boundaries which occur without visual cues. These types of story boundaries are abundant on CNN, but less so on ABC. These are the cases where the anchor tends to give extra emphasis to the topic change with his or her pauses and intonation.

Further work should attempt to verify some properties of these prosody features, mainly, their invariance to gender, speaker, and language.

3.3. Text

Like our prior work [1], we adopt the text segmentation scores from the same text segmentation method [9] except that the maximum entropy (ME) and decision tree approaches are delivered separately and not fused as a single score. The performance evaluation in the held-out validation set and the comparison with the text segmentation in our prior work [1] is shown in Table 1. Apparently ME approach has performance gain over decision tree.

Table 1. Boundary evaluation text segmentation only and the comparison with our prior work [1].

	ABC (F1)	CNN (F1)
[1]	0.59	0.59
decision tree	0.51	0.52
maximum entropy	0.61	0.57

3.4. Superimposed text spotting

According to our experiments, one of the most challenging portions is the short dynamic sport stories in CNN channel. The background speech is usually loud and noisy and video is usually high-motion. An interesting observation is that there usually appears superimposed text "*courtesy of ...*" at seconds after switching to a new sports topic. We hypothesize that it might be a good indicator to augmenting our decision making in this challenging portion.

To spot the superimposed text, we can take the video OCR outputs from CMU or Columbia. However, due to the low resolution in recorded videos, this portion is usually not detected or incurs lots of errors. We then turn to another approach to spot the superimposed text by extracting relevant low-level features from the region of interest of each key frame. The first feature we adopted is time index of the key frame since the sports section usually appears in certain time locations. The other is 181-dimensional Zernike moments [23] which have been shown effective in OCR problems.

The processing steps of Zernike moment extraction is presented in Figure 3. We first extract the region of interest from certain locations and convert the region into greyscale, which is then smoothed by a Gaussian kernel. The image binarization is thresholded by the value of mean plus one standard deviation of the smoothed grey region. Before extracting Zernike moments [24], we need centralize the words, crop margins, and square the region.

For measuring the similarity towards the superimposed text, a sample set S of manually annotated features are prepared in advance as templates. The superimposed text similarity to *courtesy* of each key frame is represented by $J_S(x)$, the similarity of its extracted feature x to set S, and expressed with KDE as the following:

$$J_S(x) = \frac{1}{Z(S)} \sum_{x_i \in S} K_\sigma(x - x_i), \qquad (1)$$

where K_σ is the high-dimensional kernel and σ is the bandwidth. The feature representation and the similarity measure is quite promising. By evaluating on a small set (ran-

domly selected), the detection of existence of *courtesy* in the superimposed text has $F1 \approx 0.91$

Similar approach is used to detect the session of "top news of the day" in CNN channel, where anchors are briefing 3-6 top news in the background and a purple logo is shown in the left-bottom area over the news footage. The low-level features used for similarity measure with Equation 1 are time index, color moments, and Gabor texture of the region of interest. The detection result for the top news session is high as well but it does not improve the story boundary detection as prosody is introduced in the A+V fusion.

3.5. Low-level image features

We try to induce the cue clusters from visual raw features, which should provide a good similarity measures in the kernel space. According to rich literatures in content-based image retrieval, we select color autocorrelogram, Gabor texture, and color moments.

3.5.1. Color autocorrelogram

The color correlograms [25] include the spatial correlation of colors and can be used to describe the global distribution of local spatial correlation of colors. We consider the HSV color space with 36 quantized color bins and then calculate the autocorrelogram with the pixel distance 1, 3, 5, and 7. Finally, it results in 144 continuous features in each key frame.

3.5.2. Gabor texture

Gabor functions are Gaussians modulated by complex sinusoids. The Gabor filter masks can be considered as orientation and scale-tunable edge and line detectors. The statistics of these micro-features in a given region can be used to characterize the underlying texture information [26]. We take 4 scales and 6 orientations of Gabor textures and further use their mean and standard deviation to represent the whole key frame and result in 48 textures.

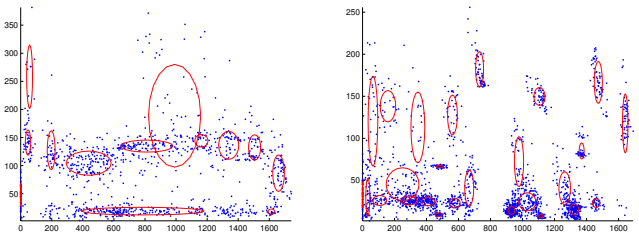
3.5.3. Color moments

We store the first three moments of each color channel of HSV images and result in 9-dimensional features [27].

4. POST-PROCESSING APPROACHES

4.1. Time-dependent Viterbi smoothing

According to our observation, both in CNN and ABC news channels, the story length L depends on the time T where the story starts. We can model the 2-D (T vs. L) data in Gaussian Mixture Model (GMM) which is penalized with minimum description length (MDL) and derive $p(T, L) = \sum \pi_i \mathcal{N}(\mu_i, \Sigma_i)$, where $\mathcal{N}$ is the Gaussian process, μ_i is the mean of the i'th Gaussian component with its diagonal covariance matrix Σ_i. The story length and story starting time from 60 ABC and CNN video clips are exemplified respectively in Figures 4(a) and 4(b). For example, in the CNN videos, the commercials usually start around 800 seconds from the video clip and right followed by top news briefings; the other short stories start around 1300 second with series sports briefings.



(a) (b)

Fig. 4. Samples of story length (Y-axis) and story starting time from 60 TRECVID videos respectively from ABC (a) and CNN (b). Their corresponding GMM components are represented in red ellipses and there are 13 components for ABC and 26 for CNN.

The conditional probability $p(L = y|T = \tau)$ can be derived from $p(T, L)$, a GMM form from discrete samples, and represents the probability of story length equals y if the programs starts at τ.

$$P(L = y|T = \tau) = \frac{\sum \pi_i p_T(\tau) p_L(y)}{\sum \pi_i p_T(\tau)}, \qquad (2)$$

where p_T and p_L are probability density function of T and L and can be derived through marginalization of $p(T, L)$. Note that we assume a diagonal covariance matrix in each GMM component. Due to its diagonal simplicity, we can directly use their corresponding variance directly and convert to 1-D GMM model for both random variables T and L. Equation 2, a time-dependent story length distribution, can be used to define the penalty function or transition cost for Viterbi decoding as the same approach mentioned in Section 3 of [28]. The only difference is that the prior work utilizes a time independent story distribution, whereas we extend that in a time-dependent fashion.

We do not finish the result before the test set submission, but it is still a promising post-processing technique according to our observation in the TRECVID data set and the experiments conducted in [28].

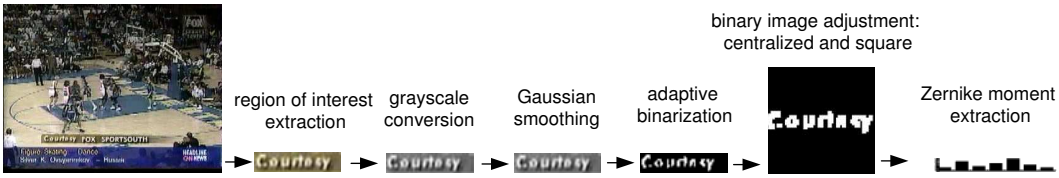


Fig. 3. Processing steps for Zernike moment extraction.

4.2. Best K-score decoding

According to our story segmentation error case analysis from the development set and test sets of TRECVID 2003. A fixed and single decision threshold for the classifier output, e.g., SVM margins on candidate points, does not work well for all video clips since in some video clips the best decision thresholds tend to be higher or lower due to the change of visual or audio production styles; for example, in a few ABC videos, the anchor background is replaced with White House rather than the blue graphics or studio setups. One of the solution to this problem is through the Viterbi decoding approach which determines the best decision path considering accumulated confidence scores within the video clip.

Another observation is that the number of story boundaries within CNN or ABC news is almost fixed with small variances. It is reasonable since these are news videos with predefined production rules. According to this observation, we constrain the number of the hypothesized (positive) story boundaries within some ranges. We then select the top specific number of the candidate points based on their classification confidence. Such constrains are parameterized by mean μ_b and standard deviation σ_b of the number of story boundaries. Within a video clip, if the number of hypothesize story boundaries based on the global threshold is less than $\mu_b - \sigma_b$ or larger than $\mu_b + 4\sigma_b$, we choose the top μ_b candidates of highest classification confidence as story boundaries. We call this approach "best K-score decoding," represented with $\mathcal{F}$. According to the calculations in the development set, the story boundary number pair (μ_b, σ_b) for CNN is $(34.49, 4.37)$ and $(18.2, 2.01)$ for ABC. This approach contributes slightly in the validation set. The performance in the test set is discussed in Section 5.4.

4.3. Commercial filter

Anther post-processing is to apply commercial filter and remove the hypothesized story boundaries within the detected commercial regions. The commercial detection approach is the same in our prior work [1]. We represent the commercial filter with $\mathcal{C}$.

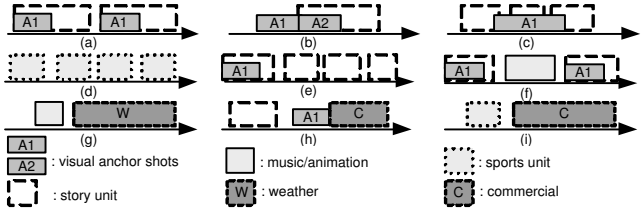


Fig. 5. Common CNN story types seen in the TRECVID 2003 data set. $A1$ and $A2$ represent segments showing visual anchor persons; (a) two stories both starting with the same visual anchor; (b) the second story starts with a different visual anchor; (c) multiple stories reported in a single visual anchor shot; (d) a sports section constitutes series of briefings; (e) series of stories that do not start with anchor shots; (f) two stories that are separated by long music or animation representing the station id; (g) weather report; (h) a non-story section consists of a anchor lead-in followed by a long commercial; (i) a commercial comes right after a sports section.

5. EXPERIMENTS

5.1. 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. In general, a video is roughly 30 minutes long and is composed of news stories and commercials.

In TRECVID tasks, 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. 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 (See more examples in Figure 5). The evaluation metrics in terms of precision and recall are

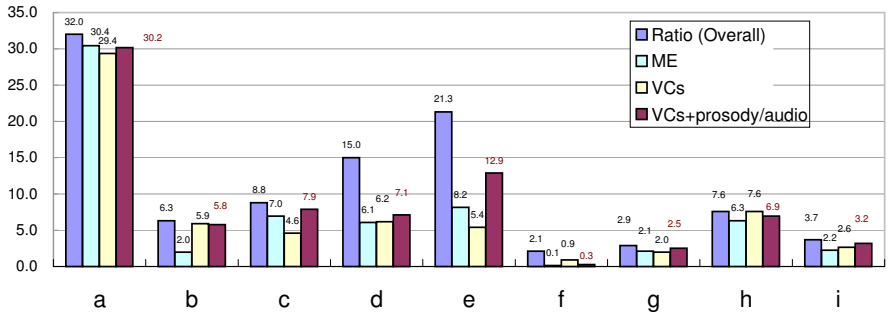


Fig. 6. Story boundary detection performance comparisons of our previous approach (ME) [1], VCs only, and VCs combined with speech prosody features. The performance results over different story types (as defined in Figure 5) are listed separately. The ME approach uses both audio and visual modalities; VCs uses visual cue clusters only. The 'ratio' group is the percentage of stories in each type. All the numbers shown are the recall values, with the precision of each experiment fixed at 0.71.

defined in [1].

For the discovery of VC^3 features, the number of visual cue clusters is determined by observing the break point of accumulated mutual information loss as described in [12] and is 60 both for ABC a CNN videos. To induce the visual cue clusters, 15 videos for each channel are used; 30 videos, with key frames represented in the cue cluster features, for each channel are reserved for SVM training; the validation set is composed of 22 CNN and 22 ABC videos. Another 28 CNN and 23 ABC videos are reserved for training multi-modal feature fusion. They are all from TRECVID 2004 development set.

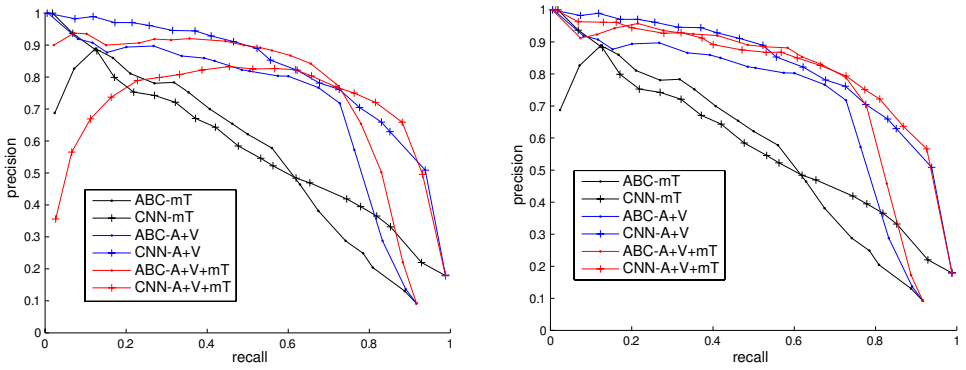
5.2. Boundary detection with VC^3 features

We compare the VC^3 only results with our previous work in Figure 6, where performance breakdowns are listed for different types of video stories (as defined in Figure 5). The left-most bar is the percentage of data in each story type (among the 759 stories of 22 annotated CNN video programs). The second bar is the recall rate achieved by our prior work [1], where specific classifiers of anchors, commercials, and prosody-related features are fused through the Maximum Entropy approach, denoted as ME. Note that the ME approach uses both the audio and visual features. The VC^3 feature is shown in the third bar from the left, and includes visual features only. It is interesting that even with an automatic method to discover the salient visual features, the VC^3 approach can match our previous method (ME) that uses a specific detector like anchor, and uses features from multiple modalities. Meanwhile, the new VC^3 adaptive approach also improves detection performance of some story types by being able to discover unique visual features such as the second anchor in story type (b), specific station animations in type (f), weather stories in type (h), and transition from the end of story to commercials in type (i).

As the case in our prior work [1], we further add other features such as prosody to the visual-only approach to achieve more performance improvement. The prosody features have been shown to be effective for detecting syntactically meaningful phrases and topic boundaries [1, 19]. By adding similar feature set, we can see the improvement for several story types, especially when topic changes are correlated with the appearance of reporters or anchors such as type (a). Those challenging types such as type (c), multiple stories within a single anchor shot, or type (e), continuous short briefings without studio or anchor shots, benefit the most from adding the prosody features. The improvements can be clearly seen in the right-most bar, denoted by "VCs+prosody/audio," in Figure 6.

5.3. Discussion: multi-modal fusion

The performance through the fusion of audio and visual modalities (ABC/CNN-A+V), text segmentation through maximum entropy (ABC/CNN-mT), and the fusion of all modalities (ABC/CNN-A+V+mT) are shown in Figure 7(a). We find that the A+V modalities are more effective than text only. Unlike our prior work [1], CNN-A+V performs better than ABC-A+V. It might be that the introduction of VC^3 discovers those adequate cue clusters which catch the visual dynamics of the CNN channel. The fusion of text in Figure 7(a) is not normal and seems over-fitted. It is due to that we adopted late fusion approach to find an non-linear combination (through SVM) of the confidence scores of A+V, T, and some other audio features again. The original idea is to emphasize the audio features to boost those short briefing portions. However, the result gets worse. That's, taking SVM as function $\Psi(\cdot)$, the fusion approach we originally used is $\Psi(\Psi(\mathrm{A}+\mathrm{V})+\mathrm{A}+\mathrm{T})$. After the formal test set submission, we discover the problem and revise the fusion approach, $\Psi(\mathrm{V}+\mathrm{A}+\mathrm{T})$, as early fusion and derive the



(a) (b)

Fig. 7. Performances from maximum entropy text segmentation (mT), A+V, and A+V+mT on the held-out validation set for ABC and CNN channels. The A+V+mT fusion approach in (a) is $\Psi(\Psi(\mathrm{A}+\mathrm{V})+\mathrm{A}+\mathrm{mT})$, which causes over-fitting problems, and is later revised in (b) with $\Psi(\mathrm{V}+\mathrm{A}+\mathrm{mT})$. Please see Table 2 for the explanations of symbols.

new result in Figure 7(b), showing salient improvement by introducing text segmentation scores.

The previous problem occurs due to limited fusion training time caused by text feature availability but can be avoided if some validation metrics are conducted by observing the PR curves or the average precisions (APs) used in [12] since they will provide performance over all recall areas rather than by looking at a single decision threshold only. For example, in Figure 7(a), the A+V+mT result actually has lower AP than that of A+V only and appears as a worse fusion configuration.

From Figure 7(b), the fusion gain through text is salient in ABC but only slightly in CNN comparing with our prior work [1]. We hypothesize that the error cases remaining in CNN A+V are those short briefing portions which still can not be addressed well by text modality.

5.4. Performance in the TRECVID 2004 test set

The submitted 10 runs of story segmentation are listed in terms of F1 measures in Figure 8, where the abbreviations for different experiment configurations are explained in Table 2. Among the participants, the submitted runs are ranked the second and the best result is F1=0.69 (BOA of Figure 8).

Among 10 submitted runs, dT and mT, text segmentation performances have only minor degradation from those in our validation set (Table 1). However, the degradations in A+V and A+V+T are more significant. One reason is that the visual quality in 2004 test set is much better than those in the development set. Our major visual features come from VC^3 whcih which requires feature projection into the visual cue clusters induced in the development. The inconsistent visual quality might incur low-level feature instability. Such observations are also reported from other TRECVID teams.

The other problems suffers from the late fusion strategy mentioned in Section 5.3. The approach was applied on all A+V and A+V+T runs except in "AV_fc+fc" and causes overfitting problems. That is why in A+V experiments, "AV_fc+fc" performs better than "AV_efc+efc" and "AV_efc+fc." The detailed configurations and fusion steps are described in Table 2.

Generally, the post-processing through commercial filtering $\mathcal{C}$ is effective since "AVmT" is slightly boosted by adding commercial filters. However, best K-score approach degrades the CNN performance by reducing the recall in the test set of CNN, according to the comparison of "AVmT_fc+c and "AVmT_fc+fc" of Figure 8. The effects of best K-score on ABC is not clear according to the test submissions.

Other prospective performance improvement might come from taking temporal dependence of the boundary candidate points which is not utilized in TRECVID 2004 story boundary submissions. The time-dependent Viterbi smoothing (Section 4.1) is worthy of further investigations.

6. CONCLUSION AND FUTURE WORK

We have proposed an information-theoretic VC^3 framework, based on the Information Bottleneck principle, to associate continuous high-dimensional visual features with discrete target class labels. We utilize VC^3 to provide new representation for discriminative classification, conduct feature selection, and prune "non-informative" feature clusters. The proposed techniques are general and effective. Most importantly, the framework avoids the manual procedures to select features and greatly reduces the amount of annotation in the training data.

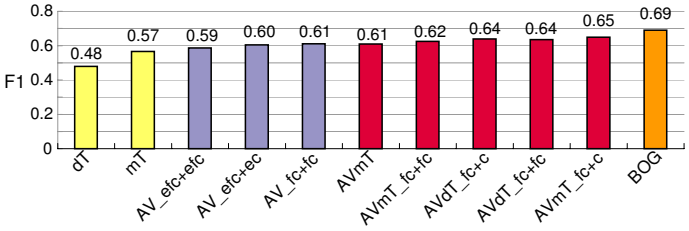


Fig. 8. 10 runs of story segmentation performance (F1 measures) submitted to TRECVID 2004 and the comparison with the best of the other groups (BOG). See explanations in Table 2 for configurations. Color legends: yellow-T, purple-A+V, and red-A+V+T. Please see Table 2 for the explanations of symbols.

Some extensions of VC[3] to induce audio cue clusters, support multi-modal news tracking and search are underway. Other theoretic properties such as automatic bandwidth selection for KDE and performance optimization are also being studied.

We also experiment other promising features and continue investigating effective prosody features. The introduction of post-processing also provides practical improvements. Furthermore, the fusion from other modalities, such as speech prosody features and ASR-based segmentation scores are significant and have been confirmed again in this experiment.

Acknowledgments

We thank Ching-Yung Lin of IBM T. J. Watson Research Center for useful discussions and his kind support of visual features. This material is based upon work funded in part by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Government

Table 2. Explanations of the 10 submitted runs for TRECVID 2004 story segmentation. "dT": decision tree text segmentation, "mT": maximum entropy text segmentation, "A": audio and prosody features, "V": VCs+motions+superimposed text spotting (CNN only), Ψ: SVM fusion, $\mathcal{F}$: K-best scores, and $\mathcal{C}$: commercial filter.

#	symbols	modalities/configurations
1	dT	decision tree text segmentation
2	mT	maximum entropy text segmentation
3	AV_efc+efc	all: $\mathcal{C}(\mathcal{F}(\Psi(\Psi(\mathrm{A}+\mathrm{V})+\mathrm{A})))$
4	AV_efc+ec	abc: $\mathcal{C}(\mathcal{F}(\Psi(\Psi(\mathrm{A}+\mathrm{V})+\mathrm{A})))$ cnn: $\mathcal{C}(\Psi(\Psi(\mathrm{A}+\mathrm{V})+\mathrm{A}))$
5	AV_fc+fc	all: $\mathcal{C}(\mathcal{F}(\Psi(\mathrm{A}+\mathrm{V})))$
6	AVmT	all: $\Psi(\Psi(\mathrm{A}+\mathrm{V})+\mathrm{A}+\mathrm{mT})$
7	AVmT_fc+fc	all: $\mathcal{C}(\mathcal{F}(\Psi(\Psi(\mathrm{A}+\mathrm{V})+\mathrm{A}+\mathrm{mT})))$
8	AVdT_fc+c	abc: $\mathcal{C}(\mathcal{F}(\Psi(\Psi(\mathrm{A}+\mathrm{V})+\mathrm{A}+\mathrm{dT})))$ cnn: $\mathcal{C}(\Psi(\Psi(\mathrm{A}+\mathrm{V})+\mathrm{A}+\mathrm{dT}))$
9	AVdt_fc+fc	all: $\mathcal{C}(\mathcal{F}(\Psi(\Psi(\mathrm{A}+\mathrm{V})+\mathrm{A}+\mathrm{dT})))$
10	AVmt_fc+c	abc: $\mathcal{C}(\mathcal{F}(\Psi(\Psi(\mathrm{A}+\mathrm{V})+\mathrm{A}+\mathrm{mT})))$ cnn: $\mathcal{C}(\Psi(\Psi(\mathrm{A}+\mathrm{V})+\mathrm{A}+\mathrm{mT}))$

7. REFERENCES

[1] Winston Hsu, Shih-Fu Chang, Chih-Wei Huang, Lyndon Kennedy, Ching-Yung Lin, and Giri Iyengar, "Discovery and fusion of salient multi-modal features towards news story segmentation," in *IS&T/SPIE Electronic Imaging*, San Jose, CA, 2004.

[2] Lekha Chaisorn, Tat-Seng Chua, , Chun-Keat Koh, Yunlong Zhao, Huaxin Xu, Huamin Feng, and Qi Tian, "A two-level multi-modal approach for story segmentation of large news video corpus," in *TRECVID Workshop*, Washington DC, 2003.

[3] Arnon Amir, Marco Berg, Shih-Fu Chang, Giridharan Iyengar, Ching-Yung Lin, Apostol Natsev, Chalapathy Neti, Harriet Nock, Milind Naphade, Winston Hsu, John R. Smith, Belle Tseng, Yi Wu, and Dongqing Zhang, "IBM research trecvid 2003 video retrieval system," in *TRECVID 2003 Workshop*, November 2003.

[4] Pinar Duygulu and Alexander Hauptmann, "What's news, what's not? associating news videos with words," in *International Conference on Image and Video Retrieval*, Dublin City University, Ireland, 2004.

[5] Keiichiro Hoashi, Masaru Sugano, Masaki Naito, Kazunori Matsumoto, Fumiaki Sugaya, and Yasuyuki Nakajima, "Shot boundary determination on mpeg compressed domain and story segmentation experiments for trecvid 2004," in *TRECVID Workshop*, Washington DC, 2004.

[6] Naftali Tishby, Fernando C. Pereira, and William Bialek, "The information bottleneck method," in *The 37th Allerton Conference on Communication, Control and Computing*, 1999.

[7] Noam Slonim, Nir Friedman, and Naftali Tishby, "Unsupervised document classification using sequential information maximization," in *25th ACM international Conference on Research and Development of Information Retireval*, 2002.

[8] Ron Bekkerman, Ran El-Yaniv, Naftali Tishby, and Yoad Winter, "On feature distributional clustering for text categorization," in *SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval*, 2001.

[9] 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.

[10] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[11] Winston Hsu and Shih-Fu Chang, "Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation," in *IEEE International Conference on Multimedia and Expo (ICME)*, Taipei, Taiwan, June 2004.

[12] Winston H. Hsu and Shih-Fu Chang, "Visual cue cluster construction via information bottleneck principle and kernel density estimation," in *International Conference on Content-Based Image and Video Retrieval*, Singapore, 2005.

[13] W. H. Adams, Giridharan Iyengar, Ching-Yung Lin, Milind Ramesh Naphade, Chalapathy Neti, Harriet J. Nock, , and John R. Smith, "Semantic indexing of multimedia content using visual, audio, and text cues," *Applied Signal Processing*, vol. 2003, no. 2, 2003.

[14] Teuvo Kohonen, *Self-Organizing Maps*, Springer, Berlin, third edition, 2001.

[15] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[16] David W. Scott, *Multivariate Density Estimation : Theory, Practice, and Visualization*, Wiley-Interscience, 1992.

[17] Barry Arons, "Pitch-based emphasis detection for segmenting speech recordings," in *International Conference on Spoken Language Processing*, Yokohama, Japan, 1994.

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

[19] 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.

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

[21] J. L. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, pp. 89–102, 2002.

[22] Snack Sound Toolkit, "http://www.speech.kth.se/snack/," .

[23] A. Khotanzad and Y. H. Hong, "Invariant image recognition by zernike moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 489 – 497, 1990.

[24] LANS Pattern Recognition Toolbox (MATLAB), "http://www.lans.ece.utexas.edu/," .

[25] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih, "Image indexing using color correlograms," in *Computer Vision and Pattern Recognition*, 1997.

[26] W. Y. Ma and B. S. Manjunath, "Texture features and learning similarity," in *Computer Vision and Pattern Recognition*, 1996.

[27] Markus Stricker and Markus Orengo, "Similarity of color images," in *Storage and Retrieval for Image and Video Databases (SPIE)*, 1995.

[28] Winston H. Hsu and Shih-Fu Chang, "A statistical framework for fusing mid-level perceptual features in news story segmentation," in *IEEE International Conference on Multimedia and Expo*, 2003.