Boundary and Feature Recognition at The University of Iowa

David Eichmann,[1,2] and Dong-Jun Park,[2]

[1]School of Library and Information Science
[2]Computer Science Department
The University of Iowa
david-eichmann@uiowa.edu

The University of Iowa participated in the shot boundary detection, story segmentation and feature extraction tracks of TRECVID-2004.

1 – Shot Boundary Detection

Our shot boundary work was based upon three core techniques. The first, histogram similarity, involved construction of color histograms for each frame, where the color space was compressed to RGB values of three bits each, yielding a 512 bin histogram and also represented in HSB space. We have also included an averaging filter that is very useful in reducing the effects of color 'jitter' in noisy video.

The second technique (distance) involves computing for a pair of frames the aggregate color distance for pixel pairs (having the same location in their respective frames) and then normalizing this value by the dimensionality of the frame. This technique has proven quite useful in avoiding false positives due to dramatic, but localized, color shifts between frames.

The third technique (edge) involves edge generation from each frame (using the ImageJ library) and applying the distance metric to pairs of now-monochrome frames. This yields a measure of gross movement occurring between frames that is tunable by how aggressively we erode edges before calculating distances between frames.

Our composite HSB technique first does a histogram-based cut detection and then overlays that with an averaged HSB gradual detection, with graduals trumping any contained cuts.

Official Runs

Our official runs for this year were accidently submitted with only one video in each run having the underlying cut detection data - our development configuration (which was still turned on) flushes all data prior to executing a single video.

Table 1: Shot Boundary Task, Overall Results

Run	Method	All		Cuts		Gradual			
		Rec	Prec	Rec	Prec	Rec	Prec	F-Rec	F-Prec
UIowaSB0401	histogram	0.089	0.118	0.120	0.117	0.024	0.123	0.324	0.649
UIowaSB0402	distance	0.083	0.119	0.121	0.120	0.003	0.109	0.400	0.328
UIowaSB0403	edge	0.015	0.105	0.004	0.088	0.039	0.109	0.311	0.615
UIowaSB0404	composite HSB	0.146	0.073	0.213	0.073	0.004	0.064	0.349	0.740

Future Work on Shots

Last year we reported a recurring source of false alarms in camera flashes [3]. Including camera flash suppression has proved to have noticeable positive effect and no negative effect in performance. We are considering additional localized event detectors for commercial boundaries and recurring graphic animations.

2 – Story Segmentation

Based upon our experience with TDT story segmentation, and the techniques that other TDT participants employed for that task, we focused on two aspects of the LIMSI [5] ASR data, speech pauses longer than a certain threshold and trigger phrases (e.g., "John Smith, ABC News, Atlanta") as indications of story boundaries. We have greatly expanded the pattern set for trigger phrases for this year. For the video/audio data, we concentrated solely on the video, and used our shot boundaries as indications of story boundaries.

Condition 1 (Video only): Our official run entailed a composite measure involving color histogram, aggregate pixel similarity and edge similarity used to define shot boundaries, and inferentially story boundaries (described in proposal 2). This resulted in a high-recall, low-precision result (as you might expect) but with noticeable differences when results are analyzed separately by source. Both precision and recall are better for CNN than for ABC using this simple technique. We considered this configuration a baseline for comparison

Condition 3 (ASR only): Our runs for this condition include: trigger phrase only, speech pause only and a composite measure run at two different threshold values. Trigger phrases prove to be a high precision means of identifying story boundaries, assuming that a proper set of trigger phrases have been identified. Our results show a substantial difference in recall for ABC over CNN with no sacrifice in precision. We speculated last year that, given some additional experimentation, the set of trigger phrases we identified using equal numbers of development videos was not sufficient to identify the full (and larger) set of trigger phrases for CNN. This conjecture will be discussed, with presentation of differential results. Speech pauses prove to be a meager source of story boundaries in ABC videos, and a substantially better source of boundaries in CNN.

Condition 2 (Video & ASR): We submitted a single run for this condition which used the composite measure of condition 1 for video and the composite measure for condition 2 for ASR data. This actually proved useful in improving performance relative to the corresponding condition 3 run. Even though precision is poor for the video composite scheme, using it improves the ASR result. Effects differ for ABC and CNN, with CNN results suffering some degradation in recall but improvement in precision. ABC results little or no recall degradation for a relatively (compared to CNN) greater improvement in precision.

For news typing, we took a very simplistic approach. The first segment in every video was declared as non-news and all other segments were declared as news. Table 2 shows our overall results for all submitted runs.

Table 2: Story Boundary Task, Overall Results

Run	Text Method	Video Method	Threshold (sec.)	Condition	Story Boundary	
					Rec	Prec
UIowaSS0401	–	comp. HSB	–	1 - video	0.818	0.137
UIowaSS0402	trigger phrases	comp. HSB	–	2 - ASR & video	0.319	0.477
UIowaSS0403	speech pauses	comp. HSB	1.25	2 - ASR & video	0.256	0.207
UIowaSS0404	trigger phrases	–	–	3 – ASR	0.441	0.510
UIowaSS0405	speech pauses	–	1.50	3 – ASR	0.285	0.211
UIowaSS0406	speech pauses	–	1.25	3 – ASR	0.354	0.211
UIowaSS0407	speech pauses	–	1.00	3 - ASR	0.460	0.213

3 – Feature Extraction

This year's TREC was our first time in this task, and we extracted 4 features: Boat/ship, Bill Clinton, Beach and Basket scored. Our system for this task was based on the same initial hypothesis from the shot boundary detection task of TRECVID 2003. That is, a relatively small number of basic metrics could be used in fairly simple combination to construct metrics that performed well on complicated data. We selected well-known low-level features and constructed simple combination metrics without using complicated mechanism such as face detection.

Below, we explain what low-level features were used in our system, and then we present how we constructed our submitted runs. Finally, we discuss our official results.

Low-level features

Low-level features are obtained directly from various information sources such as image, audio and text. These features are combined to extract high-level semantic features such as boat, train, airplane take-off and basket scored. We used a set of image-based features extracted from the MPEG videos as well as the text-based feature.

Image-based features

We selected simple and well-known low-level image-based features: color histogram and edge histogram. The color histogram is a simple frequency count measure, with 512 bins, in RGB color space. Each pixel is mapped to a bin by extracting RGB values. The global color histogram is extracted from the entire frame where the local color histogram is from a certain regions of a frame. Another image feature is the edge histogram. Each frame is transformed into a grayscale image and is applied the Sobel operator to detect edges.

Text-based feature

We also used commonly provided speech transcripts (ASR) from the LIMSI system. Mainly, we manually constructed the keyword lists for each feature by examining the training set of LIMSI transcripts.

Runs

We began our experiments by manually collecting a set of sample images from our training data. These images were the key frames that contain a specific feature and were thought to be representative to each feature.

Figure 1: Manually collected key frames for Bill Clinton and Boat/ship features

Then, image-based features were extracted from the representative sample images, and the similarity score was computed by comparing these sample images with the shot key frames. For local color histogram, we took advantage of the fact that the face is usually located at the center in the close-up images, and this face region is similar within the example set. We pondered the idea that our system determines this region dynamically. However, because of the system complexity, we opted to manually predetermine the region from the set of representative images. For typical boat/ship feature, the region was determined to be lower part of the images so that the local color histogram avoided the actual boat and captured the water region.

In addition, we also developed anchor detection classifier. This was because that none of the features to be detected would appear within anchor shots. For this classifier, we closely examined LIMSI ASR data and constructed keywords indicating the first anchor shot. We then compared this shot with the subsequent key frames by calculating color similarity. Since the anchor shots usually shows little movement, consequently these comparisons yielded very high similarity values so we could extract fairly accurate anchor shots from news video data.

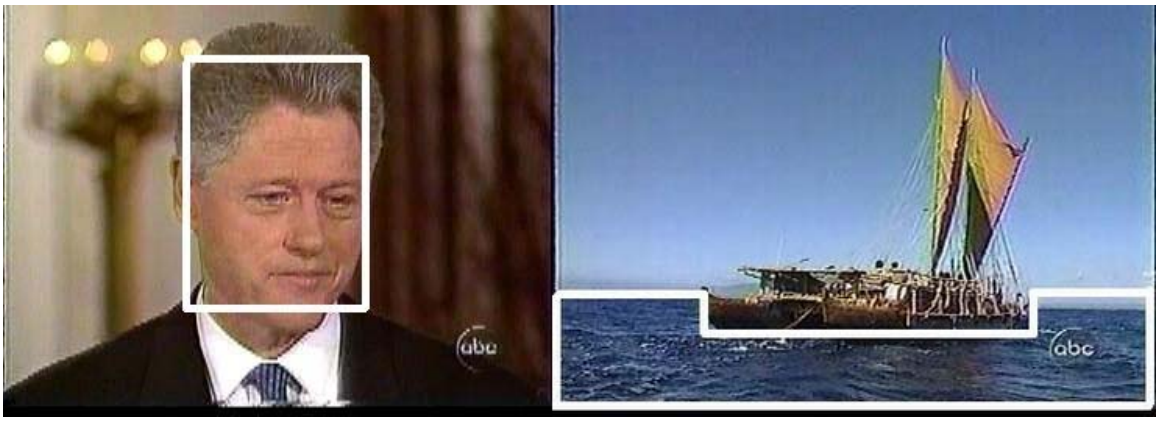

Figure 2: Typical face region for Bill Clinton feature and water region for Boat/ship feature

We submitted a total of 5 runs for each feature (except Basket scored, which we submitted only one), as shown in Table 3. The runs are designed to show the impact of each added classifier, starting from using only text-based features. For adding additional classifiers, we opted to use simple arithmetic sum and product of each metric.

Table 3: Submitted Run Descriptions

Run ID	Description
04FE1	text only (keyword matching)
04FE2	sum (text, image features)
04FE3	product (text, image features)
04FE4	sum (text, image features) with anchor detection
04FE5	product (text, image features) with anchor detection

Results

Table 4 shows our official results for this task. The numbers are not very high but shows a few interesting points.

Table 4: avgP For Each Run

Run ID	Boat/ship	Clinton	Beach	Basket scored
04FE1	0.028	0.177	0.032	0.096
04FE2	0.017	0.191	0.012	
04FE3	0.018	0.193	0.016	
04FE4	0.018	0.222	0.013	
04FE5	0.02	0.22	0.016	

It is interesting to note that simple keyword matching is best performing for both Boat/ship and Beach features, and adding additional image-based classifiers actually diminished the results. It seems that the choice of the region in local color histogram might have bad impact on our system.

It is also noteworthy that the product combination performed better than the sum runs. It seems that the product combinations well suppressed weak text-based classifier while boosting image-based classifiers.

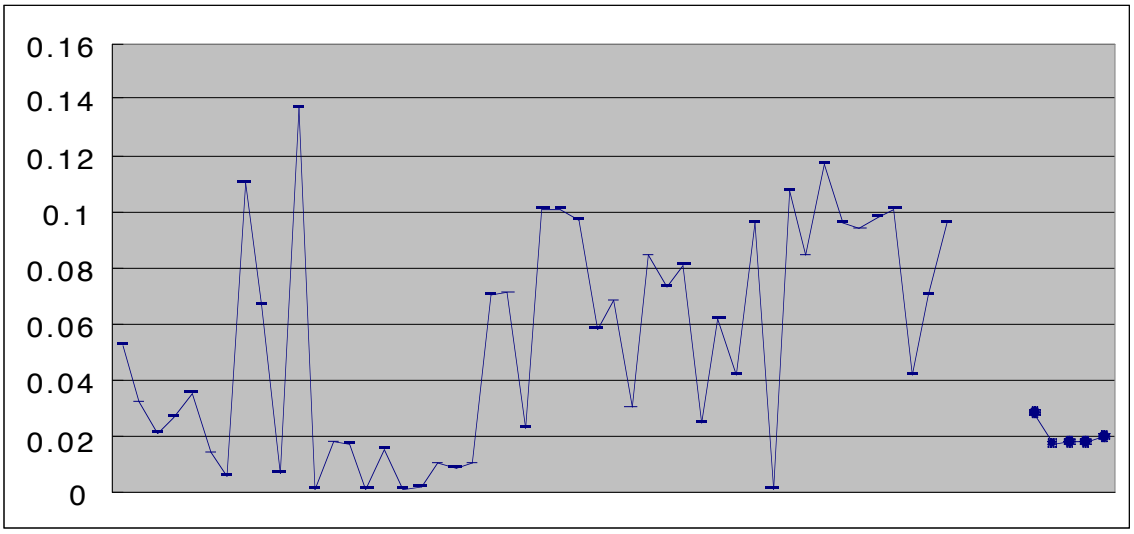

Figure 3: avgP for feature 28 (Boat/ship), all groups. The mark - indicates other groups, and our runs are plotted with dots in the right side. The order of plots coincides with the RunID shown in Table 3.

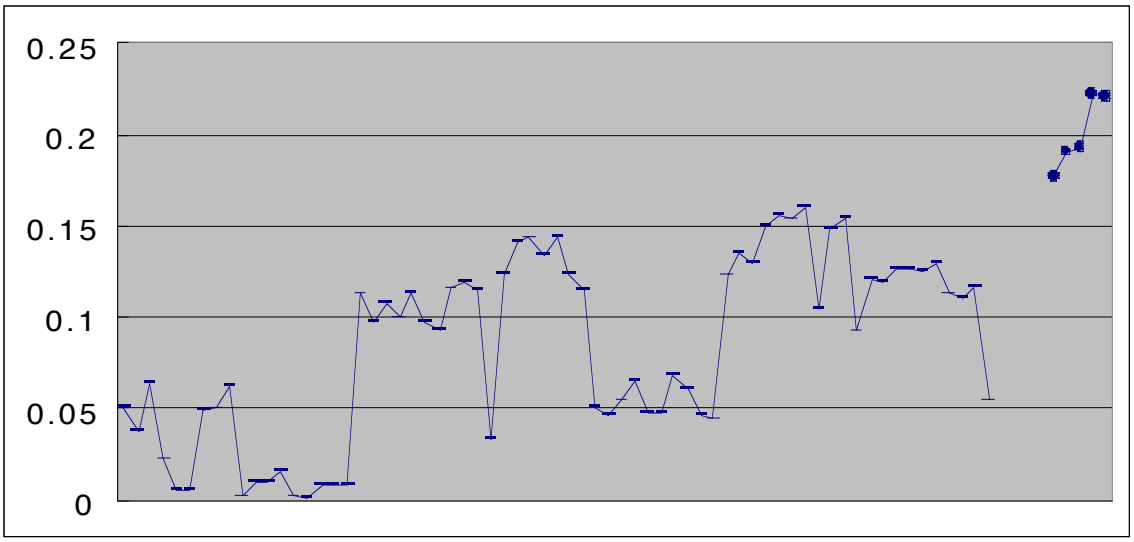

Figure 4: avgP for feature 30 (Bill Clinton) with other groups' runs.

Anchor detection seems to help the overall performance. In news video, the anchor shots are the most frequent shots, and removing these shots proved to be quite useful.

Lastly, Bill Clinton feature performed well with increasing performance for each run. It starts with decent avgP in text only run, and adding image-based classifiers improves the results.

Figures 3, 4, 5 and 6 compare our system runs with other groups' runs. We can see that both Boat/ship and Beach features were hard features, showing the maximum avgP of below 0.15 and 0.08 respectively.

3.4 Future work on Feature extraction

Our system used simple set of low-level features with fairly simple combinations such as sum and product. As a first attempt at the feature extraction task, our system surely has a lot of room to improve further. Underlying all good feature extraction system is a well-balanced system of low-level features, classifiers and special purpose detectors.

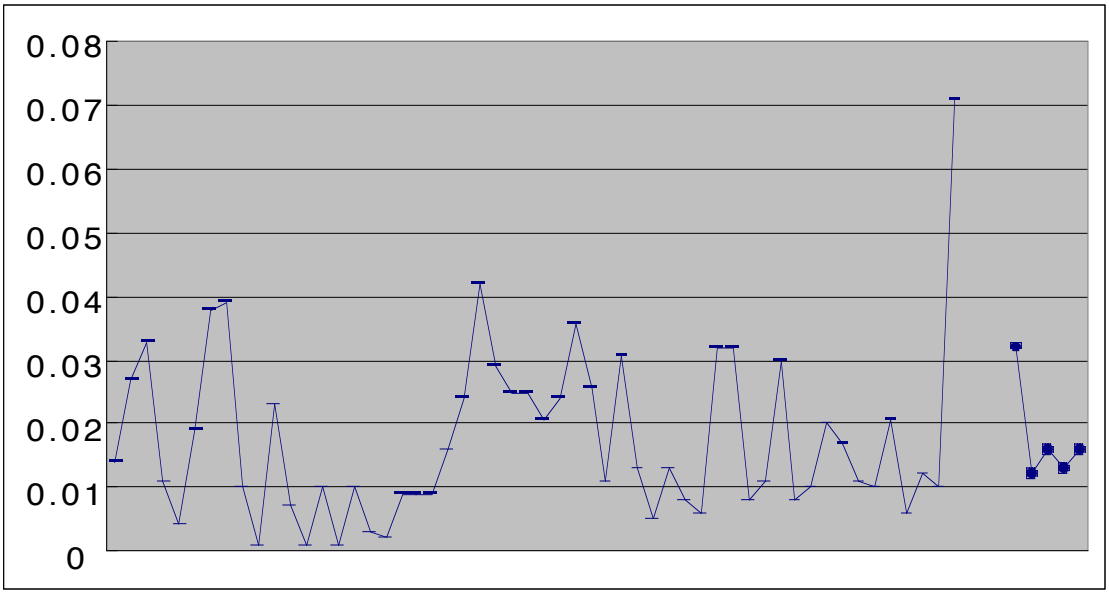

Figure 5: avgP for feature 32 (Beach) with other groups' runs.

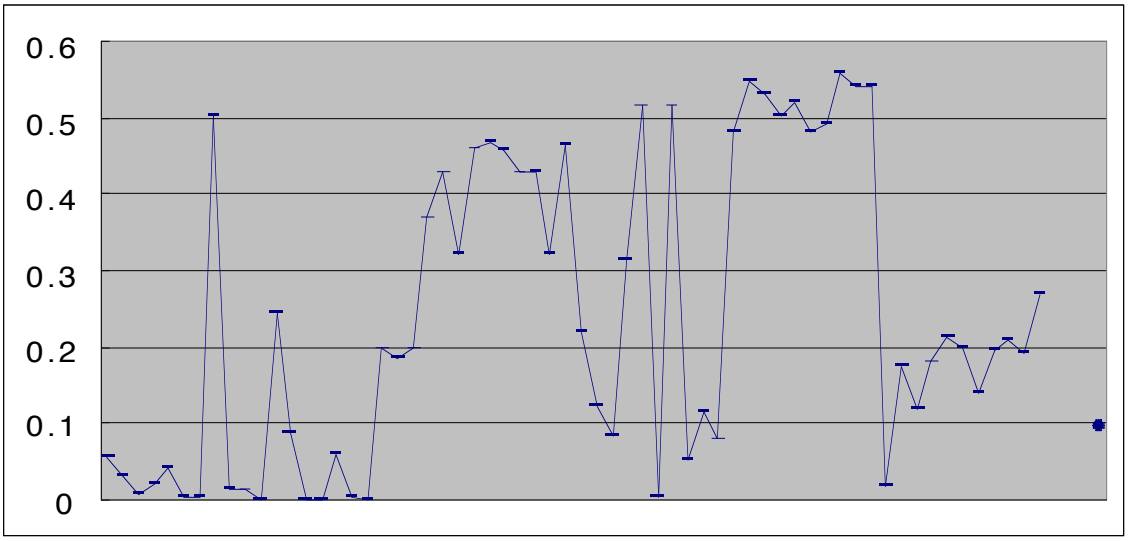

Figure 6: avgP for feature 33 (Basket scored) with other groups' runs.

Low-level features: It seems that audio/sound feature and texture information might be beneficial as they were used frequently in the past TRECVIDs.

Advanced classifiers: Instead of simple combinations, more advanced classifiers may produce improved results. This includes not only classifier itself but also the whole training phase. It seems that machine learning approach such as SVM or neural network can be used for optimum results.

Special purpose detectors: Certain features require object/shape detection. For example, Bill Clinton feature would have benefited by employing face recognizer.

References

[1] Dietterich, T. G., "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization," *Machine Learning,* v. 40, no. 2, 2000, p. 139-157.

[2] Eichmann, D., "Ontology-Based Information Fusion," *Workshop on Real-Time Intelligent User Interfaces for Decision Support and Information Visualization, 1998 International Conference on Intelligent User Interfaces*, San Francisco, CA, January 6-9, 1998.

[3] Eichmann, D and D. J. Park, "Experiments in Boundary Recognition at the University of Iowa," *Proceedings of the 2004 TRECVID Workshop*.

[4] Fiscus, J., "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," *Proc. IEEE ASRU Workshop,* p. 347-352, Santa Barbara, CA, 1997.

[5] Gauvain, J. L., L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication,* 37(1-2):89-108, 2002. ftp://tlp.limsi.fr/public/spcH4_limsi.ps.Z.

[6] Zabih, R., J. Miller, and K. Mai, "A feature-based algorithm for detecting and classifying scene breaks," *Third International Multimedia Conference and Exhibition, Multimedia Systems*, pages 189-200, San Francisco, California 1995