# Story Segmentation in News Videos Using Visual and Text Cues

Yun Zhai, Alper Yilmaz, and Mubarak Shah

School of Computer Science,
University of Central Florida,
Orlando, Florida 32816

**Abstract.** In this paper, we present a framework for segmenting the news programs into different story topics. The proposed method utilizes both visual and text information of the video. We represent the news video by a Shot Connectivity Graph (SCG), where the nodes in the graph represent the shots in the video, and the edges between nodes represent the transitions between shots. The cycles in the graph correspond to the story segments in the news program. We first detect the cycles in the graph by finding the anchor persons in the video. This provides us with the coarse segmentation of the news video. The initial segmentation is later refined by the detections of the weather and sporting news, and the merging of similar stories. For the weather detection, the global color information of the images and the motion of the shots are considered. We have used the text obtained from automatic speech recognition (ASR) for detecting the potential sporting shots to form the sport stories. Adjacent stories with similar semantic meanings are further merged based on the visual and text similarities. The proposed framework has been tested on a widely used data set provided by NIST, which contains the ground truth of the story boundaries, and competitive evaluation results have been obtained.

## 1 Introduction

News programs provide instant and comprehensive reporting of what is happening around the world. It usually contains two portions: news stories and miscellaneous stories. One of the standard definitions by the U.S. National Institute of Standards and Technologies (NIST) for the news stories is that a news story is a segment of a news broadcast with a coherent news focus, which may include political issues, finance reporting, weather forecast, sports reporting, etc [4]. On the other hand, non-news stories are called miscellaneous stories, covering commercials, lead-ins, lead-outs, reporter chit-chats, etc. Both types of the stories are composed of one or more shots. The coverage of the news program is very comprehensive, and it is likely that individual viewers maybe interested in only a few stories out of the complete news broadcast. This interest can be summarized based on the type of news, the geographic locations, etc. Automatic story segmentation and indexing techniques provide a convenient way to store, browse and retrieve news stories based on the user preferences.

News segmentation is an emerging problem, and many researchers from various areas, such as multimedia, information retrieval and video processing, are interested in

W.-K. Leow et al. (Eds.): CIVR 2005, LNCS 3568, pp. 92–102, 2005.

this problem. Hoashi *et al.* [3] has proposed an SVM-based news segmentation method. The segmentation process contains the detection of the general story boundaries, in addition of the special type of stories, e.g., finance report and sport news. Finally, the anchor shots are further analyzed based on the audio silence. Hsu *et al.* [6] proposed a statistical approach based on discriminative models. The authors have developed the *BoostME*, which uses the Maximum Entropy classifiers and the associated confidence scores in each boosting iteration. Chaisorn *et al.* [1] used HMM to find the story boundaries. The video shots are first classified into different categories. The HMM contains four states and is trained on three features: type of the shot, whether location changes and whether speaker changes.

In this paper, we propose a two-phase framework for segmenting the news videos. The method first segments the news videos into initial stories. Then, these stories are refined by further detection of special types of news stories and the merging of similar stories. The rest of the paper is organized as follows: Section 2 describes the proposed framework in detail; Section 3 demonstrates our system results; and, Section 4 presents the conclusion and the discussions.

## 2 Proposed Framework

In the news videos, we often observe the following pattern: first, the anchor person appears to introduce some news story. Then, the camera switches to the outside of the studio, e.g., the scene of the airplane crash site. After traversing around the key sites, the camera switches back to the studio, and the anchor person starts another story. It can be summarized in this form: [anchor]→[story1]→[anchor]→[story2]→[· · · ]. This pattern can be represented by the Shot Connectivity Graph (SCG) (Fig.1). In SCG, the nodes represent the shots in the video, and similar shots are represented by a single node. The edges connecting the nodes are the transitions between the shots, as shown in Fig.1. The stories in the video correspond to the large cycles in the SCG that are connected at the node representing the anchor, and our objective is to detect these cycles in the SCG.

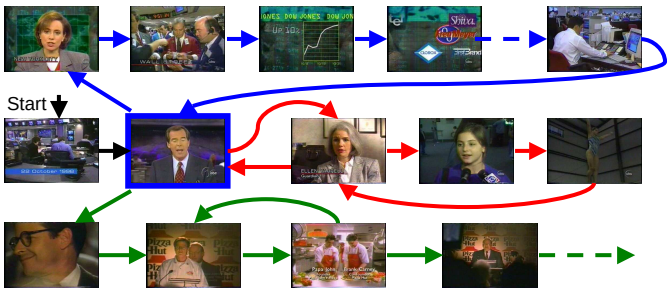


**Fig. 1.** Shot Connectivity Graph. The node with blue bounding box represents the anchor person shots in the news video. In this simple example, the video consists of two news stories and one commercial. The blue cycle and the red cycle are the news stories, and the green cycle represents a miscellaneous story

We have developed an efficient and robust framework to segment the news programs into story topics, motivated by [12]. The framework contains two phases: (1). the initial segmentation based on the anchor person detection, including both the main anchor and the sub-anchor(s); and (2). the refinement based on further detections of the weather and sports stories and the merging of the semantically related stories.

In the first phase, we segment the video by detecting the cycles that are connected at the "anchor" node in SCG. The properties of the extended facial regions in the key-frames of the shots are analyzed for clustering the similar shots into corresponding groups. Note that besides the large cycles, there are also smaller cycles that are embedded in the larger cycles. This can be explained as the appearance of the reporters, interviewers, or the sub-anchors for a specific news story, e.g., finance news. We consider the later case as a type of the news story segments. The detection method for the sub-anchor(s) is the same as the detection of the main anchor.

In the second phase, the initial segmentation is refined by further detecting news stories with special formats and merging the semantically related stories. For some news stories with special formats, there is no anchor involved. These stories are "hidden" in the large cycles in the SCG. Other techniques are used to "discover" them from the initial story segments. There are two kinds of special stories we have incorporated into the system: weather news and sports news. The color pattern of the shots is examined to filter out the candidate weather shots. Then, the candidate weather shots are verified by their motion content. The largest continuous segment of the remaining weather shots forms the weather story. For the detection of the sports story, we used the text correlation of the shots to a sport-related word set. Similar to the weather detection, the adjacent sport shots are grouped into the sports story. It maybe possible that the initial segmentations from the first phase are not semantically independent. For example, for a particular story, the anchor may appear more than once, and this causes multiple cycles in the SCG. Thus, merging of the semantically related stories is needed. Two adjacent stories are merged together if they present similar pattern in either visual appearance or word narration, or both. The visual similarity is computed as the color similarity between the non-anchor shots in the adjacent stories. The narrative similarity is defined as the normalized text similarity based on the automatic speech recognition (ASR) output of the videos. The visual and text similarities are later combined to represent the overall similarity between the stories.

### 2.1 Phase I - Anchor Detection

We construct the SCG by representing the shots of the same person by a single node. There are two common approaches for clustering the similar shots: (1) using similarity measures based on the global features, e.g., the color histograms of the key frames; (2) using similarities based on the face correlation. The problem with the first approach is that if the studio settings change, the global features for the anchor shots possess less similarity. In the later case, the face correlation is sensitive to the face pose, lighting condition, etc. Therefore, it tends to create multiple clusters for the same person. To overcome these problems, we use the "body", an extended face region. In a single news video, the anchor wears the same dress through out the entire program. We take this fact as the cue for this problem. For each shot in the video, we select the middle frame

(a). Original Sample Key Frames with Detected Faces

(b). Corresponding Body Regions Based on the Face Detection

**Fig. 2.** (a). The sample key-frames with the detected faces; (b). The body regions extended from the faces. Global feature comparison or face correlation fails to cluster the same anchor together

as the key frame, detect the face using [11], and find the body region by extending the face regions to cover the upper body of the person. The similarity of two shots $s_i$ and $s_j$ is defined as the histogram intersection of the body patches $f_i$ and $f_j$:

$$HI(f_i, f_j) = \sum_{b \in allbins} min(H_i(b), H_j(b)) \,, \tag{1}$$

where $H_i(b)$ and $H_j(b)$ are the $b$-th bin in the histogram of the "body" patches $f_i$ and $f_j$, respectively. Some example "body" patches are shown in Fig.2. Non-facial shots are considered having zero similarity to others. Then, the shots are clustered into groups using iso-data, and each of those groups corresponds to a particular person. If a shot contains multiple "bodies", the shot is clustered into the existing largest group with the acceptable similarity. Eventually, the shots that contain the main anchor form the largest cluster. Once the anchor shots are detected, the video is segmented into the initial stories by taking every anchor shot as the starting points of the stories.

Usually, in the news stories with special interests, the main anchor is switched to a sub-anchor. For example, such phenomenon is often found in finance news. The sub-anchor also appears multiple times with different story focuses. Reappearing of sub-anchors result in small cycles in the SCG. Note that some of the stories also cause the small cycles due to other reasons: reporters or interviewers. However, sub-anchor usually appears more times than other miscellaneous persons. Therefore, the true sub-anchor can be classified by examining the size of its corresponding group. Only the groups with sufficient facial shots are declared as the sub-anchor shots. The detections of the main anchor and the sub-anchors provide the initial result of the story segmentation, which is refined during the second phase, and it is discussed in the next section.

### 2.2 Phase II - Refinement

**Weather Detection.** In the news story segmentation, segments related to weather news are considered as separate stories from the general ones. To detect a weather shot, we use both the color and motion information in the video. The weather shots possess certain color patterns, such as greenish or bluish. Some example key-frames are shown in Fig.3. The motion content of the candidate shots is used for the verification purpose.

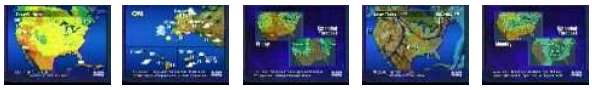

**Fig. 3.** Some key-frames of weather shots used to build the color model for weather detection

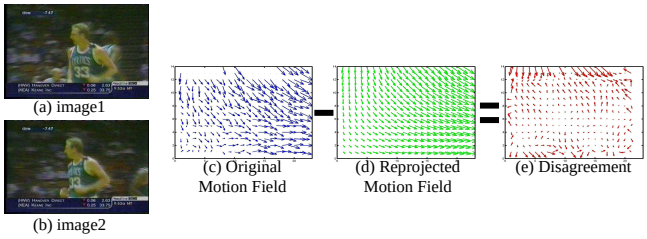


**Fig. 4.** (a,b) Two consecutive images; (c) The motion field with 16x16 blocks; (d) Re-projected motion field by applying the Affine parameters; (e) The difference map between (c) and (d)

From the training data set, we obtain the key-frames of the weather shots. For a key-frame $k_m$, a color histogram $H(k_m)$ in RGB channels is computed. The histograms of all the key-frames then are clustered into distinctive groups using Bhattacharya measures. These groups form the color model $T = \{t_1 ... t_n\}$ for the weather shot detection, where $t_i$ is the average histogram for model group $i$. To test if a shot $s$ is a weather shot, we compute the histogram $H(s)$ of its key-frame and compare it with the color model. If the distance between $H(s)$ and $t_i$ in the color model can be tolerated, then shot $s$ is classified as a weather shot.

The motion content is analyzed for the verification of the initial detected weather shots. To verify if a candidate shot $s$ is a true weather shot or not, we perform the following steps:

1. For each frame $F_i$ in the shot, compute the motion field $U_i$ between $F_i$ and $F_{i+1}$ based on the 16x16 blocks grid $X_i$.
2. Estimate the Affine motion parameters $A_i$ from $U_i$ using the equation $U_i = A_i X_i$.
3. Apply parameters $A_i$ to $X_i$ to generate the re-projected motion field $U_i^p$.
4. Compute motion content $M_i$ as the average magnitude of the "disagreement" between the original motion field $U_i$ and the re-projected field $U_i^p$.
5. The motion content of shot $s$ is the mean of $\{M_1 ... M_{n_s - 1}\}$, where $n_s$ is the number of frames in the shot.
6. If the motion content of the candidate shot $s$ is above some defined threshold, this shot is declared as a non-weather shot.

Fig.4 shows an example of the motion content analysis. Finally, other false detections are eliminated by taking only the largest temporally continuous section as the true weather news story.

**Sport Detection.** We utilize the text similarity measure to detect sporting shots. In sports video, we often hear the particular words related only to the sport games, "quar-

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| NBA | Hawks | Bulls | Lakers | Pistons | Heat | Magic | ..... |
| NFL | Ravens | Broncos | Jaguars | Dophins | Cardinals | Panthers | ..... |
| NHL | Devils | Rangers | Penguins | Sabres | Wings | Senators | ..... |
| MLB | Angles | Indians | Braves | Yankees | Giants | Marlins | ..... |
| MISC | Basketball | Baseball | Football | Olympics | Swimming | Tennis | ..... |

**Fig. 5.** Some example key-frames of the sporting shots, and example sporting key-words

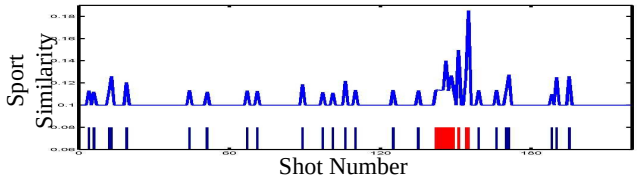


**Fig. 6.** The plot of the sport similarity of shots in a video. Bars in the bottom row represent the potential sport shots, and the red region represents the actual sporting story

terback", "basketball", etc. Given such a database of sporting words, we find the relationship between a shot and the sporting database by computing the correlation between the words spoken in the shot with the words in the database. Some of the key-works are shown in Fig.5, and in total we have over 150 key-words. The text information is provided by the automatic speech recognition (ASR) output of the video [2]. For each candidate shot $s$, we extract the key-words between the time lines by pruning the stop words, such as "is" and "the". The remaining key-words form a *sentence* $Sen_s$ for this shot. The similarity between shot $s$ and the sporting database is defined as:

$$SportSim(s) = \frac{K_s}{L(Sen_s)}\ , \tag{2}$$

where $K_s$ is the number of the key-words in shot $s$ that also appear in the database, and $L(Sen_s)$ is the length of the key-word *sentence* of shot $s$. Our method declares the shots having the strong correlation with the sporting database to be the sporting shots. Similar to the technique used for weather detection, false detections are removed by taking only the largest continuous section of the detected sporting shots as the sporting story. In Fig.6, the upper plot shows the similarity of the shots to the sporting database, while the bars in the bottom row represent the potential sporting shots. The red region represents the true sporting story in the video.

**Story Merging.** The proposed segmentation method over-segments the video in case of an anchor appearing more than once in a single story. To overcome this problem, we merge adjacent segments based on their visual and text similarities. We use the histogram intersection technique to compute the visual similarity of two stories and the Normalized Text Similarity (NTS) as the text similarity measure.

Suppose stories $S_i$ and $S_j$ are the news sections related to the same topic created by phase 1 and have $n_i$ and $n_j$ non-anchor shots, respectively. For each shot in the stories, we extract the middle frame as the key-frame of that shot. The visual similarity $V(i,j)$ between stories $S_i$ and $S_i$ is defined as:

$$V(i,j) = max(HI(s_i^p, s_j^q)),\ p \in [1...n_i],\ q \in [1...n_j], \tag{3}$$

where $HI(s_i^p, s_j^q)$ is the histogram intersection between shots $s_i^p$ and $s_j^q$. This means if there are two visually similar shots in the adjacent stories, these two stories should belong to the same news topic.

Sometimes, the semantic similarity is not always reflected in the visual appearance. For example, a news program related to a taxation plan may first show the interviews with a field expert. Then, after a brief summary by the anchor, the program switches to the congress to show the debate between the political parties on the same plan. In such case, the visual appearances of these two adjacent stories are not similar at all. However, if any two stories are focused on the same story, there is usually a correlation in the narrations of the video. In our approach, this narrative correlation between stories $S_i$ and $S_j$ with *sentences* $Sen_i$ and $Sen_j$ is computed by the Normalized Text Similarity (NTS):

$$NTS(i,j) = \frac{K_{i\to j} + K_{j\to i}}{L(Sen_i) + L(Sen_j)}\ , \tag{4}$$

where $K_{i\to j}$ is the number of words in $Sen_i$ that also appear in $Sen_j$, and similar definition for $K_{i\to j}$. $L(Sen_i)$ and $L(Sen_j)$ are the lengths of $Sen_i$ and $Sen_j$ respectively. One example story *sentence* is shown in Fig.8.

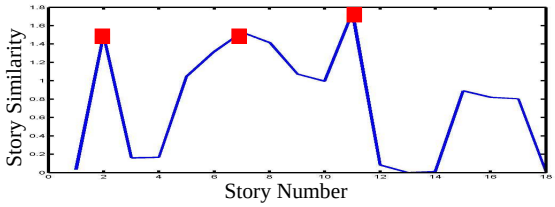


**Fig. 7.** The story similarity plot for the stories created by phase-1. The red peaks correspond to the stories which are merged during the second phase

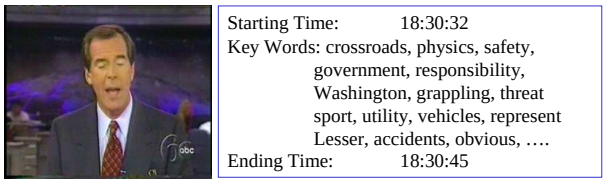


**Fig. 8.** The key-frame of an example shot in a video, accompanied by the key-words extracted from that shot. The starting and ending times are from the analogue version of the video (tape)

The final similarity between stories $S_i$ and $S_j$ is a fusion of the visual similarity $V(i,j)$ and the normalized text similarity $NTS(i,j)$ (Fig.7),

$$Sim(i,j) = \alpha_V \times V(i,j) + \alpha_{NTS} \times NTS(i,j) \ , \tag{5}$$

where $\alpha_V$ and $\alpha_{NTS}$ are the weights to balance the importance of two measures. If $Sim(i,j)$ for the two adjacent stories $S_i$ and $S_j$ is above the defined threshold, these two stories are merged into a single story.

## 3 System Performance Evaluation

Different people may have different definitions of a story, e.g., when the story should start, when it should end. This may create argument among different researchers about how their systems should be evaluated. To prevent this problem, we have tested our system on a open-benchmark data set. This data set is provide by the National Institute of Standards and Technologies (NIST). It contains 118 news videos recorded from news networks CNN and ABC. Among these videos, 58 are from ABC's *World News Tonight with Peter Jennings*, and the other 60 are from CNN's *Headline News*. Each video is around 30 minutes long and contains continuous news program. The Language Development Center (LDC) has provided the ground truth for the story boundaries based on the manual annotation.

In the field of information retrieval, two accuracy measures are often used: precision and recall. They are defined as follows:

$$Precision = \frac{X}{A}, \ \ Recall = \frac{X}{B} \ , \tag{6}$$

where $X$ is the number of correct matches between system detections and the ground truth data; $A$ is the total number of the system detections; $B$ is the total number of the ground truth references. Instead of taking the average values of the precision and recall of each video, the evaluation system computes the precision and recall based on the number of total matches over all the videos. For our method, the matching program provided by NIST returned 0.803 and 0.539 for the precision and recall respectively.

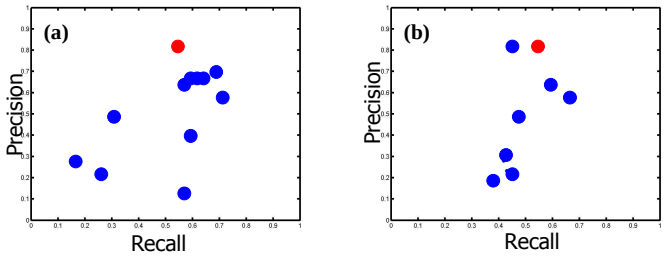


**Fig. 9.** (a). Precision/recall plot of runs using visual and text information; (b). Precision/recall plot of the average performance of the different research groups. The red dots represents the standing of the proposed method

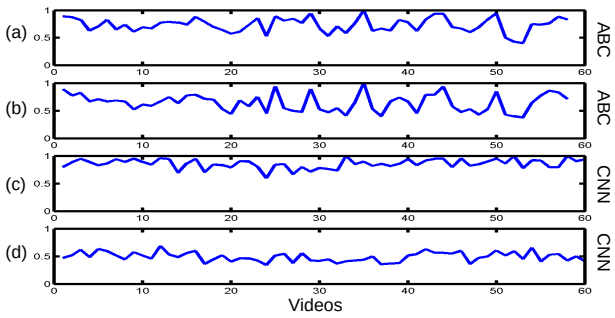


**Fig. 10.** (a) Plot of the precision values for ABC videos; (b) the recall values for ABC videos; (c) precision values for CNN videos; (d) recall values for CNN videos

**Table 1.** Accuracy measure for ABC and CNN videos separately. Precision 1 and Recall 1 are the measurements based on the overall performance, treating every story in all the video equally important. Precision 2 and Recall 2 are the average performance of the system, treating every video equally important. Insertion is the number of the false positives, and deletion is the number of the false negatives

| ***Measures*** | ***ABC*** | ***CNN*** |
|---|---|---|
| Number of Videos | 58 | 60 |
| Total Match | 696 | 1002 |
| Total Insertion | 247 | 169 |
| Total Deletion | 388 | 1043 |
| Precision 1 | 0.7381 | 0.8557 |
| Recall 1 | 0.6421 | 0.4900 |
| Precision 2 | 0.7341 | 0.8585 |
| Recall 2 | 0.6452 | 0.4927 |

It should be noted that the merging technique based on the visual and text similarities reduced average 2-3 false positives every video and increased the overall precision by $5\% \sim 10\%$. We also obtained the results of multiple runs from 7 other research groups in the field. Fig.9 shows the standings of the proposed method comparing with others. The detailed precision/recall values for every video are shown in Fig.10. The overall performance on ABC and CNN videos separately is shown in Table 1.

It is very difficult to argue that precision is more important than recall or vice versa. Usually there exists the trade-off between these two measures. For better comparison, we have computed the F-scores of the precision and recall. The F-score is defined as,

$$FScore = \frac{2 \times Precision \times Recall}{Precision + Recall}\ , \tag{7}$$

The proposed approach achieved 0.645 for the F-score, and the relative standing comparing with other groups is shown in Fig.11.

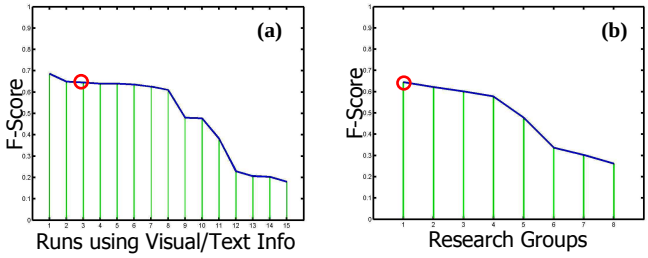


**Fig. 11.** (a). F-scores of the runs using visual and text information; (b). F-scores of the average performance of each research group. Red circles represent the standing of the proposed method

## 4 Conclusions and Discussions

In this paper, we proposed an efficient and robust framework for segmenting the news videos. The method first segments the news videos into initial stories based on the detections of the main anchor and the sub-anchor. Then, the initial segments are refined by further detection of weather news and sports stories, and the merging of adjacent semantically related stories. We have experimented the proposed method on a large scale of data set provided by NIST, and competitive results have been obtained.

The proposed method is biased towards more structured news broadcast. For instance, in ABC videos, since it often follows the pattern we described in Section 2.1, the initial segmentation is able to provide the closed solution to the true segmentations. On the other hand, in CNN videos, sometimes multiple stories exist in a single shot. For example, or the news stories start with a non-anchor shot. These two situations cause the false negatives of the segmentation. This explains why the recall value of the CNN videos is lower than the ABC videos (Table 1). Further research on such issue is needed to solve this under-segmentation problem.

Furthermore, other cues in the video can be exploited in the story segmentation. Audio signal processing and the closed captions (CC) of the videos will be included into our framework in the future.

## References

1. L. Chaisorn, T-S. Chua and C-H. Lee, “The Segmentation of News Video Into Story Units”, *International Conference on Multimedia and Expo*, 2002.
2. J.L. Gauvain, L. Lamel, and G. Adda. “The LIMSI Broadcast News Transcription System”, *Speech Communication*, 37(1-2):89-108, 2002.
3. K. Hoashi, M. Sugano, M. Naito, K. Matsumoto, F. Sugaya and Y. Nakajima, “Shot Boundary Determination on MPEG Compressed Domain and Story Segmentation Experiments for TRECVID 2004”, *TREC Video Retrieval Evaluation Forum*, 2004.
4. http://www-nlpir.nist.gov/projects/tv2004/tv2004.html#2.2
5. A. Hanjalic, R.L. Lagendijk, and J. Biemond, “Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems”, *IEEE Transaction on Circuits and System for Video Technology*, Vol.9, Issue.4, 1999.

6. W. Hsu and S.F. Chang, "Generative, Discriminative, and Ensemble Learning on Multi-Model Perceptual Fusion Toward News Video Story Segmentation", *International Conference on Multimedia and Expo*, 2004.
7. J.R. Kender and B.L. Yeo, "Video Scene Segmentation Via Continuous Video Coherence", *Computer Vision and Pattern Recognition*, 1998.
8. R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Scene Determination Based on Video and Audio Features", *IEEE Conference on Multimedia Computing and Systems*, 1999.
9. C.W. Ngo, H.J. Zhang, R.T. Chin, and T.C. Pong, "Motion-Based Video Representation for Scene Change Detection", *International Journal of Computer Vision*, 2001.
10. H. Sundaram and S.F. Chang, "Video Scene Segmentation Using Video and Audio Features", *International Conference on Multimedia and Expo*, 2000.
11. P. Viola and M. Jones, "Robust Real-Time Object Detection", *International Journal of Computer Vision*, 2001.
12. M. Yeung, B. Yeo, and B. Liu, "Segmentation of Videos by Clustering and Graph Analysis", *Computer Vision and Image Understanding*, vol.71, no.1, 1998.