Automatic Story Segmentation for Spoken Document Retrieval

Pui Yu Hui[1], Xiaoou Tang[2], Helen M. Meng[1], Wai Lam[1] and Xinbo Gao[2]
[1]Human-Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management,
[2]Department of Information Engineering,
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong SAR, China
hmmeng@se.cuhk.edu.hk

ABSTRACT

We have been working on speech retrieval based on Cantonese television news programs. Our video archive contains over 20 hours of news programs provided by a local television station. These programs have been hand-segmented into video clips, where each clip is a self-contained news story. The audio tracks in our archive are indexed by Cantonese speech recognition. This is integrated with a vector-space information retrieval model to achieve speech retrieval. This paper proposes an approach for *automatic story segmentation* from television news programs, intended to replace hand-segmentation as described above. Automatic story segmentation is critical for rapid expansion of our video archive. Our approach relies on the assumption that nearly all the news stories follow the temporal syntax of (begin_story → anchor shots → field shots → end_story). Therefore our algorithm aims to detect field-to-anchor shot boundaries, that should also coincide with the story boundaries. The proposed approach utilizes the video frame information for story boundary detection, and involves such techniques as fuzzy c-means and graph-theoretical clustering. The approach achieved precision and recall values of over 70%, based on a 20-hour video corpus.

1. INTRODUCTION

The explosive growth of the Internet has created a rich source of electronic information in a variety of media – text, audio and video. This creates a demand for multilingual and multimedia information retrieval technologies to enable the user to retrieve personally relevant content on demand. Text-based search engines are widely used, and audio / video searching are active areas of research. We have been working on the problem of Chinese spoken document retrieval [Meng et al., 2000]. In particular, we work with Cantonese, which is a major dialect of the Chinese language, commonly used in Hong Kong, Macau, South China and many overseas Chinese communities. We have developed a video archive of Cantonese television news broadcast, by hand-segmenting the television news reports into individual news stories. Users can type in a textual query through our Web-based interface, to retrieve the video clips of the relevant news stories. This work attempts to apply an *automatic story segmentation technique* to replace hand-segmentation of each news story into an individual video clip. For our speech retrieval system, we combine the technologies of speech recognition and automatic story segmentation for indexing our audio tracks, and applied a vector-space model for information retrieval [Meng et al., 2001]. Previous work in this area include Mandarin (the major dialect of Chinese) spoken document retrieval by [Chien et al., 1999] and [Wang et al., 1999]; and the CMU Informedia project [Wactlar et al., 1996] which uses image and audio information concurrently for digital video access.

2. CORPORA

The video content for our experiments is provided by the Hong Kong Television Broadcasts Ltd. (TVB). It consists of Cantonese news broadcasts from the Jade[1] channel (i.e. the Cantonese channel), with 934 news stories. Table 1 provides a detailed description.

Language	Cantonese Chinese
Source	TVB Jade channel
Number of Stories	934 (~20.01 hours)
Extraction Period	7 Jul 1999 to 16 Aug 1999 5 Oct 2000 to 11 Dec 2000
Average Length of News	1 min 15.62 sec (per story)
Minimum Length of News	5 sec
Maximum Length of News	7 min 7 sec
Digital Video Format	MPEG-1

Table 1. Information about the video content used on our experiments.

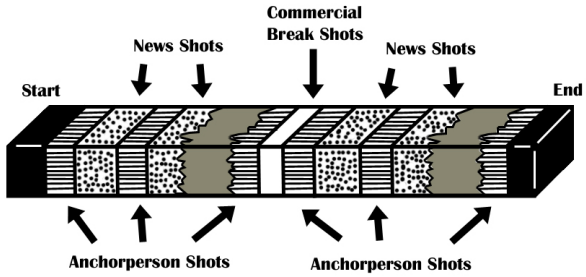


Figure 1. The temporal structure of a news program.

Each MPEG file contains a single news story manually segmented from the news program, which is illustrated in Figure 1. Very often, the news story begins with a report from the anchor(s) in the studio, followed by a live report from the field. The anchor shots are relatively homogenous – there are mainly four patterns of anchor shots, as shown in Figure 2. However, there is no fixed pattern for field shots.

3. AUTOMATIC STORY SEGMENTATION

In this section we will describe our method of automatic story segmentation, which detects scene changes from the video frames in order to locate the transition from the studio to the field. With this information, we can segment the audio track for each story into the segment of anchor speech, followed by the segment of live report. Thereafter, these segments can be processed individually for speech retrieval.

We found that the temporal syntax of the news video from our local television station is rather straightforward. The majority of our Cantonese news stories consist of two parts – the anchor shots (studio shots) followed by the field shots. Also, an *entire* news program contains a series of news stories interleaving with commercial breaks (see Figure 1).

[1] http://news.tvb.com.hk/

(1) One anchorperson on the left, news icon in the upper right corner

(2) One anchorperson in the middle

(3) One anchorperson on the right, news icon in the upper left corner

(4) Two anchorpersons

Figure 2. The four typical patterns of anchor shots in our video corpus.

We have manually labeled all the news story boundaries in our 20-hour video corpus. This gave us 934 stories, of which we found that 94.3% have the temporal syntax described above. The few remaining cases have one of the following temporal syntax: (i) anchor shots only; or (ii) anchor-field-anchor combination.

We adopted the story segmentation scheme developed by [Gao & Tang, 2000][2]. It consists of four modules as shown in Figure 3.

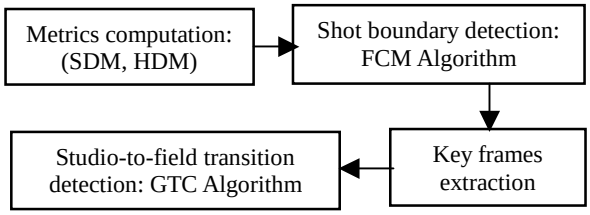


Figure 3. Control flow of our automatic story segmentation scheme.

3.1 Metric Computation

We have extracted the video frames from the MPEG-1 video files of the news programs, and sampled one out of every five frames to obtain a sparser frame sequence. If we compare the gray level or color histograms of a pair of consecutive frames in this sequence, we see that if both frames are anchor shots, they tend to have very similar histograms, as illustrated in Figure 4. However, if the pair of frames crosses a shot boundary, e.g. one belongs to an anchor shot while the other a field shot, their histograms are very different, as illustrated in Figure 5. In order to compare the qualitative difference between a pair of consecutive frames, we use to metrics – the *spatial difference metric* (SDM) and *the histogram difference metric* (HDM), as shown in Equations 1 and 2.

$$SDM = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left| I_t(i, j) - I_{t+1}(i, j) \right| \qquad (1)$$

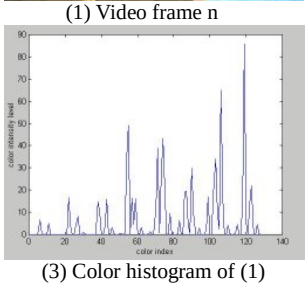
$$HDM = \frac{1}{M \times N} \sum_{k=1}^{L} \left| H_t(k) - H_{t+1}(k) \right| \qquad (2)$$

[2] Co-authors of this paper.

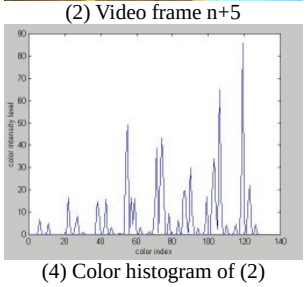
where $M \times N$ is the frame size,
$I_t(i, j)$ denotes the intensity of a pixel at location (i, j),
$H_t(k)$ denotes the number of pixels with color k in the t-th frame pair, and
L is the total number of colors.

(1) Video frame n

(2) Video frame n+5



(3) Color histogram of (1)

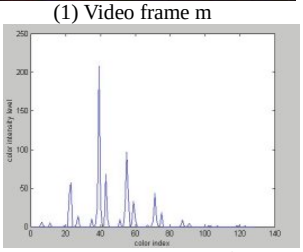


(4) Color histogram of (2)

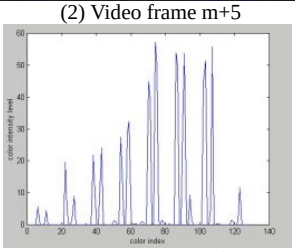
Figure 4. Color histograms of the video frames in the same shot.

(1) Video frame m

(2) Video frame m+5



(3) Color histogram of (1)



(4) Color histogram of (2)

Figure 5. Color histograms of the video frames across a shot boundary.

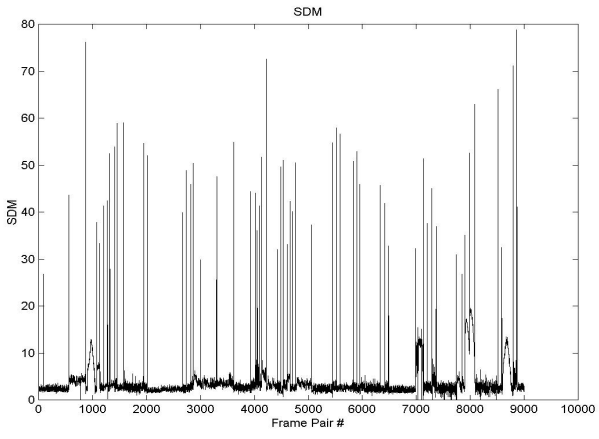


Figure 6. A plot of the spatial difference metric against frame pair number.

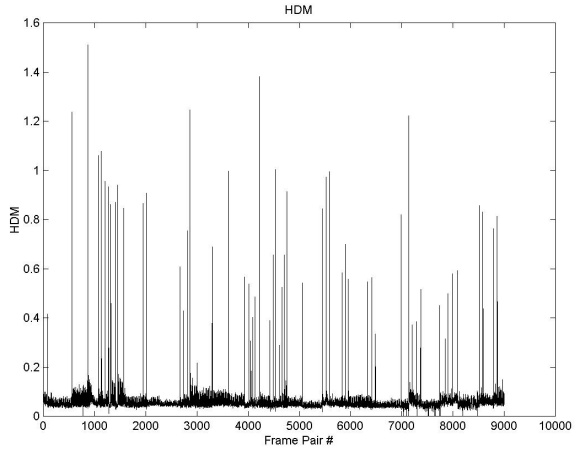


Figure 7. A plot of the histogram difference metric against frame pair number.

If we plot the SDM or HDM values against the frame pair number, we obtain a sequence of pulses, as shown in Figure 6 and 7. A high pulse often results from large content changes in the frames, corresponds to a shot boundary.

3.2 Shot Boundary Detection by the Fuzzy C-Means Algorithm

For each consecutive frame pair in our frame sequence, we plot its normalized SDM and HDM values in a scatter plot as shown in Figure 8. To detect the frame pairs with significant change, we partition this feature space into two subspaces, the significant change (SC) category and non-significant change (NSC) category. Hard clustering assigns each data point (feature vector) to one and only one of the subspaces, with a degree of membership equal to one, assuming well-defined boundaries between the subspaces. This model does not reflect the description of real data, where boundaries between subspaces might be fuzzy – i.e. the new shot may fade in, the previous shot can fade out, or we have a combination of both fade in and fade out. Hence we introduce the *fuzzy c-means* (FCM) clustering algorithm [Bezdek 1981] to classify the feature vectors into SC and NSC categories based on an objective function. By minimizing the defined objective function, we can obtain the optimal fuzzy space partition U^* and the optimal cluster prototype v^*. For our application, the objective function and the fuzzy 2-partition space for feature vectors $F_d(t)$ are defined as

$$J_m(U,v) = \sum_{t=1}^{T}\sum_{i=1}^{2}(u_{it})^m \cdot \|F_D(t) - v(i)\|^2 \qquad (3)$$

$$M_{f2} = \left\{ \begin{array}{l} U \in V_{2T} \mid u_{it} \in [0,1] \forall\, i,t; \\ \sum_{i=1}^{2} u_{it} = 1 \forall\, k; 0 < \sum_{t=1}^{T} u_{it} < T \;\forall\, i \end{array} \right\} \qquad (4)$$

where $U \in M_{f2}$;

$v \in R^{2\times 2}$ is the cluster center or prototype of fuzzy subset u_i, $i = 1,2$;

$m \in [1,\infty]$ is the weight exponent;

u_{it} denotes the degree of the t-th vector $F_D(t)$ belonging to the i-th category;

V_{2T} is the set of real $2 \times T$ matrices.

Using the FCM algorithm, the feature vectors have been classified. We implemented the defuzzifying operation on the classification result by FCM algorithm, we can get the crisp classification result as shown in Figure 9. The defuzzifying operation rule is defined as

$$F_D(t) \in SC, \text{ if } u_{1t} \geq u_{2t}\text{; and}$$
$$F_D(t) \in NSC, \text{ if } u_{1t} > u_{2t} \qquad (5)$$

According to the temporal structure of television new programs, frames between two consecutive boundaries form a shot. Based on these shot boundaries, i.e. the feature vectors in SC category, each video frame can be assigned to either an anchor shot or a field shot.

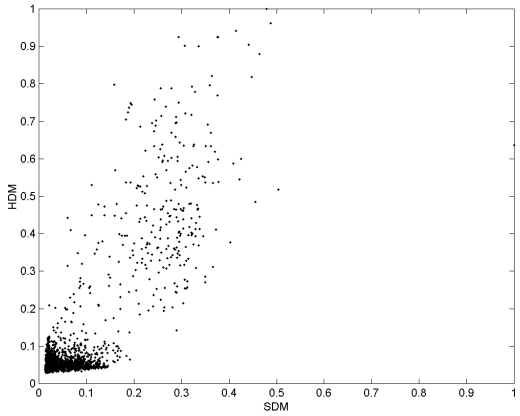


Figure 8. A plot of the two-dimensional feature space — HDM against SDM.

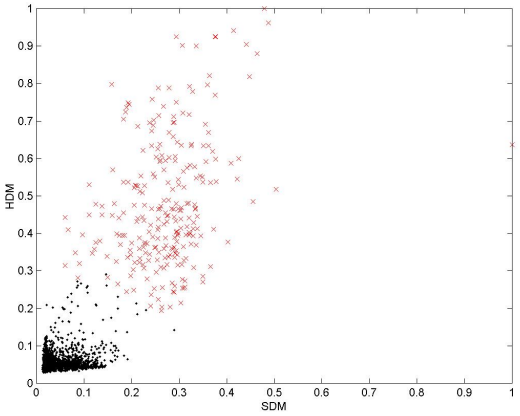


Figure 9. The classification result of Figure 6. Frame pairs with significant change labeled as shot boundaries are labeled as 'x'.

3.3 Key Frames Extraction

There exist two categories of shots – (i) the anchor/studio shot and (ii) the field shot. After automatic shot boundary detection in the previous section, we can partition a news program into individual shots. The next step is to classify the shots into studio shots or field shots. To this end, we first extract the key frame for each shot – we simply take the first frame of each shot to be representative of the shot. These key frames are used for further classification into anchor shots or field shots.

3.4 Studio-to-field Transition Detection by the Graph-Theoretical Clustering Algorithm

Although there are several patterns of studio shots (anchor frames), such as the four typical patterns shown in Figure 2, we observe that each pattern appears no less than twice in a news program. Most key frames of field shots are very different from each other. Because the background region of the anchor key

frames tends to be relatively fixed, the different studio key frames with the same pattern must have a similar color histogram. Using this similarity we can group and detect studio shots of the same pattern in a self-organized fashion through the graph-theoretical cluster (GTC) analysis algorithm [Zahn 1971].

Key frames in the 128-dimensional color histogram space are illustrated as the nodes in Figure 10. We formed a minimal spanning tree (MST) using a single link algorithm to link all the key frames in this space. The path length connecting any two nodes (representing two key frames) in MST is proportional to the HDM difference between the two key frames. We then sever the edges of the tree which have a distance larger than a threshold, and thus four connected clusters remain, as shown in Figure 11. Each of these clusters contains key frames that closely resemble one another. We have four clusters corresponding to the four types of anchor shots shown in Figure 2. Hence the key frames in these clusters are treated as studio shots. The other frames are automatically treated as field shots.

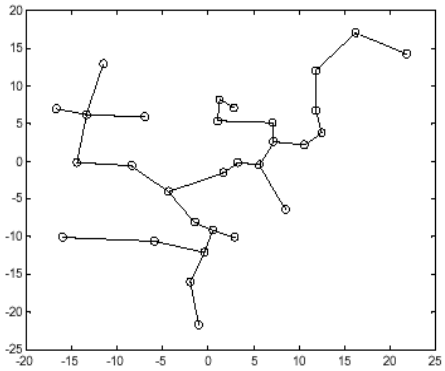


Figure 10. A plot of minimal spanning tree.

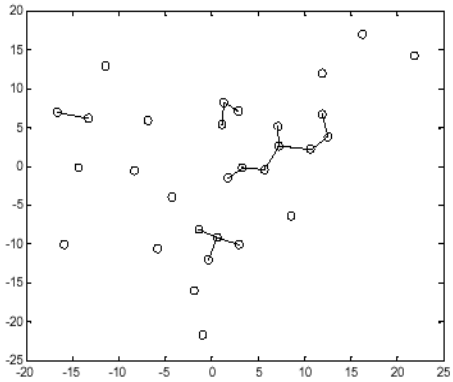


Figure 11. The remaining clusters of Figure 8 after deleting the edges with distance larger than a threshold

4. PERFORMANCE ON AUTOMATIC STORY SEGMENTATION

The previous section describes how we can label each shot (in between two consecutive shot boundaries) into either an anchor shot or a field shot. Recall our assumption that every news story follow the temporal syntax of (begin_story → anchor shots → field shots → end_story), we can now automatically label the field-to-anchor shot boundaries to be our story boundaries. It should be noted that in reality only 94.3% of our news stories (i.e. 881 out of 934) follow the specified temporal syntax, hence our evaluation is based on the 881 stories only.

From our 20-hour video corpus, the automatic story segmentation algorithm labeled 877 story boundaries. Of these 638 are correct. We consider a story boundary to be correctly detected if it lies within 50 frames (i.e. 2 seconds of video) of the manually labeled story boundary. Hence automatic story segmentation achieved a precision of 0.727 and a recall of 0.724.

5. CONCLUSIONS AND FUTURE WORK

This paper reports on our preliminary attempt in automatic story segmentation using video information to extract individual news stories from Cantonese television news programs. Most of our news stories follow the temporal syntax of (begin_story → anchor shots → field shots → end_story). Hence our story segmentation algorithm aims to detect field-to-anchor transitions. Our approach involves shot boundary detection by a fuzzy c-means algorithm and shot classification by a graph-theoretical clustering algorithm.

The automatic story segmentation algorithm can streamline our development process of a video news archive. Our next step involves indexing the audio tracks of these video clips by automatic Cantonese speech recognition. As such users can type in a text query to retrieve relevant video clips. This constitutes a cross-media retrieval task where the query is in Chinese text and the documents are Chinese video clips.

ACKNOWLEDGMENTS

We would like to thank the Television Broadcasts Limited for providing the Cantonese news video in this project. This project is supported by a grant from the Area of Excellence in Information Technology.

REFERENCES

1. Bezdek, J. C., *Pattern Recognition with Fuzzy Objective Function Algorithm*, New York, Plenum, 1981.
2. Chien, L. F. and Wang, H. M., “Exploration of Spoken Access for Chinese Text and Speech Information Retrieval”, Proceedings of the International Symposium on Signal Processing and Intelligent Systems, 1999.
3. Gao X. and X. Tang, “Automatic Parsing of News Video Based on Cluster Analysis”, Proceedings of the 2000 Asia Pacific Conference on Multimedia Technology and Applications, Taiwan, December 2000.
4. Meng, H., W. K. Lo, Y. C. Li and P. C. Ching, “Multi-Scale Audio Indexing for Chinese Spoken Document Retrieval”, Proceedings of ICSLP, 2000.
5. Meng, H., X. Tang, P. Y. Hui, X. Gao and Y. C. Li, “Speech Retrieval with Video Parsing for Television News Programs”, Proceedings of ICASSP, 2001.
6. Wactlar, H., T. Kanade, M. Smith and S. Stevens, “Intelligent Access to Digital Video: Informedia Project”, IEEE Computer, Theme issue on Digital Library Initative, May 1996.
7. Wang, H. M., “Retrieval of Mandarin Spoken Documents Based on Syllable Lattice Matching”, Proceedings of IRAL, 1999.
8. Zahn, C. T., “Graph-theoretical Methods for Detecting and Describing Gestalt Clusters”, IEEE Transactions on Computers, vol. 20, no. 1, pp. 68-86, 1971.