AUDIO-VIDEO SUMMARIZATION OF TV NEWS USING SPEECH RECOGNITION AND SHOT CHANGE DETECTION

Chien-Lin Huang, Chia-Hsin Hsieh and Chung-Hsien Wu

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.
{chicco, ngsnail, chwu}@csie.ncku.edu.tw

ABSTRACT

This paper presents an approach to audio-video summarization of TV news to provide concise information about the content while preserves the essential message of the original. In this study, anchor speech and field report videos are considered separately. First, speech signal is automatically recognized as transcripts and a confidence measure considering syntactic and semantic relations is used to estimate the reliability of words. For video skimming, RGB color histogram difference is adopted to segment video shots and evaluate the smoothness of images concatenation. As a result, the extracted anchor speech and the field report image sequence of TV news are aggregated into a summarization output. The experimental results indicate that the proposed approach effectively extracts important speech segments and gives a concise video sequence.

1. INTRODUCTION

In the age of the information explosion, there are a lot of digital speech records in news, presentations, lectures and entertainment. In order to efficiently search the amount of multimedia, it is important to be able to summarize and retrieve these contents. Summarization can save not only the transmission time but also the browsing time of users. In digital video retrieval applications or mobile device applications, it is appropriate to use concise summarized contents instead of original multimedia documents.

All types of multimedia including speech, audio, text and video are rapidly growing today. In the past years, many efforts have been devoted to multimedia summarization. The main subject in text summarization is to extract important sentences according to the context structure or discourse relation between paragraphs and sentences in an article [1]. The difference between text and speech is the prosody feature in the voice presentation. Speech summarization generally relies on the transcription from a large-vocabulary continuous-speech recognizer (LVCSR). The results of speech summarization are compact speech sequences obtained from the analysis of transcripts [2]. Video summarization analyzes image sequence and segment shots into the compendious and meaningful video stream [3].

This study focuses on TV news summarization according to anchor speech and field report videos. For the purpose of stable environment and rich information of anchor speech, speech summarization technology can be useful for extraction of key information. Based on the scores from speech recognition confidence, word significance, word trigram and semantic dependency, the dynamic programming algorithm is used for speech summarization [2]. On the other hand, the field report is segmented by the color histogram difference method [4]. Then, image sequences of field report are contracted according to the length of the summarized speech and the redundancy of each shot.

2. THE PROSODED SCHEME

As shown in Figure1, a news video is divided into two components: the anchor speech and the field report video. In anchor speech summarization, anchor speech is recognized into transcripts. Four confidence scores for anchor speech summarization are estimated and used to choose the best summarization result. In video skimming of the field report, this study selects the image sequences by minimizing the visual redundancy in the field report videos. Finally, the summarized anchor speech is synchronized with the video summarization result to provide a concise audio-video summarization output.

2.1. Speech Summarization

Given an anchor speech utterance with N words, the corresponding transcription $X = (w_1, w_2, ..., w_N)$ is obtained using an LVCSR. A dynamic programming algorithm is applied to find a speech summary with highest confidence score. Given the compression ratio ∂, a summarized word sequence $Y = (w_1, w_2, ..., w_M)$ with $M (= N \times \partial)$ words which maximizes the following four summarization scores is obtained:

$$S(Y)=\sum_{m=1}^{M}\{\lambda_C C(w_m)+\lambda_R R(w_m)+\lambda_L L(w_m \mid w_{m-2},w_{m-1}) + \lambda_B B_{SDG}(w_{m-1},w_m)\} \quad (1)$$

where $C(w_m)$ denotes the confidence score of word w_m obtained from the LVCSR. $R(w_m)$ denotes the word significance score. $L(w_m \mid w_{m-2},w_{m-1})$ represents the trigram probability and $B_{SDG}(w_{m-1},w_m)$ is the semantic dependency score.

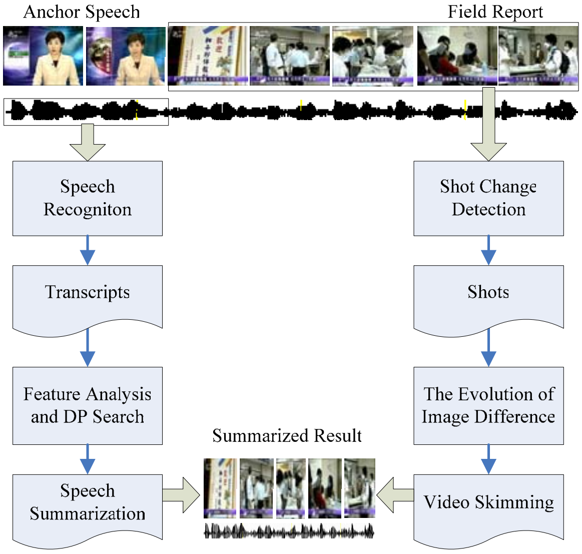

Fig.1: Flowchart of News Video Summarization

1) Speech Recognition Confidence: In order to obtain the text information of the TV news, an LVCSR is used to automatically translate the speech into transcriptions. In order to evaluate the assurance of the recognition result, the confidence measure is estimated to remove unreliable information. In this method, the posterior probability of each transcribed word is used to calculate speech recognition confidence using a linguistic decoder [5].

2) Word Significance: Word significance is used to measure the importance of the words in the speech signal. As in [6], a topic-related corpus consisting of two elements: an article and its corresponding topic words is used. For each word recognized from the LVCSR, a retrieval model is applied to obtain the most relevant document and the corresponding title keywords. After retrieving the most relevant document d^*, the words in the corresponding title t^* contain the most important information related to document d^*. The word significance score $R(w_m)$ of w_m in the transcribed sentence is calculated according to the words in title t^* and the word correlation matrix obtained using latent semantic indexing (LSI) [7]:

$$R(w_m)=\max_b\{P_{LSI}(w_m,w_b^{t^*})\cdot f_{w_m}\cdot \ln(N/df_{w_m})\} \quad (2)$$

where $P_{LSI}(w_m,w_b^{t^*})$ denotes the similarity between word w_m and title keyword $w_b^{t^*}$; f_{w_m} is the term frequency of word w_m in the document. df_{w_m} represents the document frequency of word w_m and N denotes the number of documents in the corpus.

3) Word Trigram: The word trigram score $L(w_m \mid w_{m-2},w_{m-1})$ is used to estimate the trigram probability of a word sequence. The trigram probability is interpolated from trigram, bigram and unigram to smooth the frequencies.

4) Semantic Dependency: This paper proposes a semantic dependency grammar (SDG) to obtain the semantic dependency score $B_{SDG}(w_a,w_b)$ as follows:

$$B_{SDG}(w_a,w_b) = \frac{1}{N_s}\sum_{j=1}^{N_s}\sum_i\sum_r f_{DR_i^r(w_a,w_b)}(T_i,S^j(w_a,w_b))\times f_{T_i}(S^j(w_a,w_b)) \quad (3)$$

where $f_{T_i}(.)$ is the probabilistic context-free grammar (PCFG) used to parse the hierarchical tree structure. $f_{DR_i^r}(.)$ means a probability of SDG. N_s denotes total sentence numbers. $S^j(w_a,w_b)$ denotes sentence S^j containing words w_a and w_b. T_i is the parse tree. r denotes the existing relation index. $D_i=\{DR_i^r(w_a,w_b) \mid 1\le r\le N_w-1\}$ represents the possible dependency relation DR_i^r in the parse tree i with word length N_w. In order to avoid the problem of sparse data, each word is converted into its corresponding hypernym based on HowNet [8], a Chinese knowledgebase.

$$f_{DR_i^r(w_a,w_b)}(T_i,S^j(w_a,w_b))\cong f_{DR_i^r(H(w_a),H(w_b))}(T_i,S^j(w_a,w_b)) \quad (4)$$

where $H(w_a)$ denotes the hypernym of w_a. Furthermore, the score $f_{DR_i^r(H(w_a),H(w_b))}(T_i,S^j(w_a,w_b))$ is estimated using the following equation:

$$f_{DR_i^r(H(w_a),H(w_b))}(T_i,S^j(w_a,w_b)) = F(DR_i^r(H(w_a),H(w_b)))/F(H(w_a),H(w_b)) \quad (5)$$

where $F(DR_i^r, H(w_a), H(w_b))$ denotes the frequency that dependency relation $DR_i^r(H(w_a), H(w_b))$ happens in the training corpus. $F(H(w_a), H(w_b))$ denotes the co-occurrences of $H(w_a)$ and $H(w_b)$ in the training corpus.

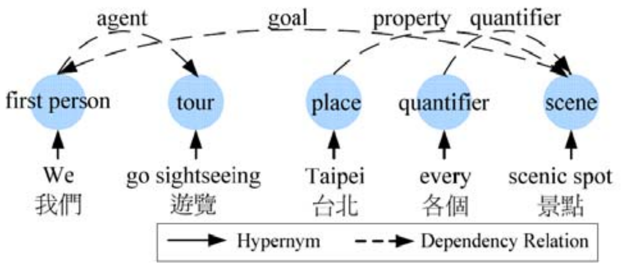

Fig. 2: Example of the semantic dependency graph

Figure 2 shows the example of semantic dependency graph which is constructed from the Chinese sentence "我們 (We) 遊覽 (go sightseeing) 台北 (Taipei) 各個(every) 景點(scenic spot)."

2.2. Video Summarization

For video summarization, this approach extracts the key frames of the field report according to the length of the extracted anchor speech. There are three steps in the video summarization procedure: First, this study employs the color histogram-based shot boundary detection algorithm to segment video shot. Second, we analyze the image sequence of each shot and extract the desired length of the image sequence. Finally, the summarized anchor speech is combined with the field report video to give a concise summarization result.

1) Shot Boundary Detection: In general, there are three kinds of shot changes: hard cuts, fades and dissolves [4]. In the application of news video summarization, the hard cuts method is able to obtain good result for video skimming. In this study, we measure the color histogram difference to detect the shot change boundary. This basic idea is that the color content does not change rapidly within but across shots. In the video decoding procedure, we can get the image sequence with 256 pixels of color BMP format $P_i(R,G,B)$. Each color component is composed of $R(red)$, $G(green)$ and $B(blue)$, and the value is ranged from 0 to 255. The color histogram difference $diff(P_i, P_{i-1})$ is estimated between two images P_i and P_{i-1} as follows:

$$diff(P_i, P_{i-1})_R = \sum_{k=0}^{255} | P_i^R - P_{i-1}^R |$$

$$diff(P_i, P_{i-1})_G = \sum_{k=0}^{255} | P_i^G - P_{i-1}^G | \qquad (6)$$

$$diff(P_i, P_{i-1})_B = \sum_{k=0}^{255} | P_i^B - P_{i-1}^B |$$

$$diff(P_i, P_{i-1}) = \sum\nolimits_k^{R,G,B} diff(P_i, P_{i-1})_k$$

If the difference $diff(P_i, P_{i-1})$ exceeds a threshold θ and the number of successive image frames is larger than a threshold C_N, a hard cut will be detected. According to our preliminary experiment, we chose a setting of $\theta = 5000$ and $C_N = 60$ for a frame size of 160×120 pixels. This setting value is different for different video sizes.

2) Determine the Frame Length of Each Shot: After the shot boundaries have been detected, a suitable length of image sequence is selected according to the anchor speech summarization. A video shot is composed of a series of image contents with similar characteristic. In order to reduce the display length, we analyze the consistency between two images in a shot. In this study, the color histogram is also applied to evaluate the image consistency. Furthermore, the color histogram difference $diff(P_i, P_{i-1})$ of each shot is ranked as $rank(P_i)$ in descending order. The skimming ratio β is decided by the length of anchor speech summarization L_A and the length of field report L_F as follows:

$$\beta = \frac{L_A}{L_F} \qquad (7)$$

To obtain the video skimming content VSC, the ranked image $rank(P_i)$ in each shot S_N is selected based on the video summarization ratio multiplied by the shot length L_S.

$$VSC = \sum_{k=1}^{S_N} \sum_{l=1}^{\beta \times L_S} rank(P_i) \qquad (8)$$

Finally, an MPEG2 encoder is used to combine the summarization result of the field report with the anchor speech summarization. This combination provides a brief overview about this news story.

3. EXPERIMENTS

For evaluation, an HMM-based Mandarin LVCSR was constructed. The character recognition accuracy achieved about 82%. Furthermore, the semantic dependency grammar was constructed from the Sinica Treebank [9] with 36,953 sentences and the HowNet knowledgebase [8]. We extracted 22,025 rules according to the tree structure of Part-of-Speeches (POSs) and their corresponding probabilities estimated from the Treebank were obtained.

3.1. Evaluation of Character Accuracy Compared with Manual Summarization Result

The results from automatic speech summarization were compared with the subjective results from manual summarization. We invited five graduate students to summarize target references from correct news articles and the character accuracy is estimated using the following measure:

$$P_{accuracy} = (W - I - D - S) / W \quad (9)$$

where W is the number of characters. I , D and S denote the numbers of insertion, deletion and substitution character errors, respectively.

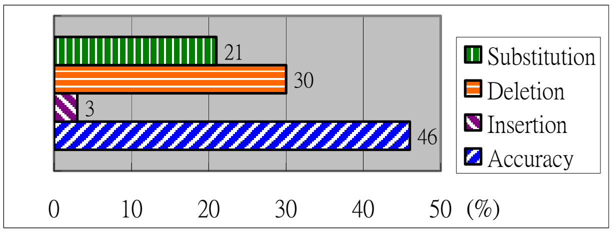

Fig. 3: The accuracy of speech summarization

As shown in Figure 3, the deletion and insertion error are due to subjective variation between different persons. The substitution error is caused by mis-recognition and therefore some important information is missing.

3.2. MOS Evaluation of Video Summarization

The performance of video summarization was evaluated by the Mean Opinion Score (MOS) test. Eighteen graduate students were invited to evaluate the video summarization results. Then, they were asked to subjectively evaluate the results according to the following criterions: fluency (FLU), favorite (FAV) and average (AVG). For each evaluation, evaluators can assign the level from 0.0 to 10.0. The comparison result is shown in Figure 4.

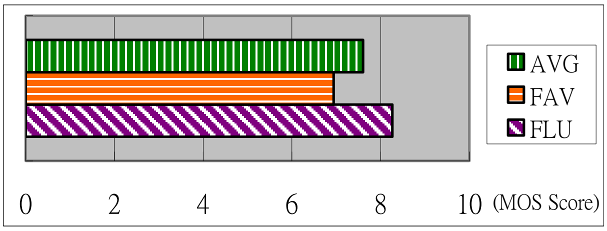

Fig. 4: Experimental results for subjective evaluation

FLU is decided by the smoothness of speech summarization and video skim. FAV depends on personal favorite for video summarization result. Finally, AVG is used to measure the average of fluency and favorite scores.

4. CONCLUSION

This study has presented an approach for audio-video summarization of TV news using speech recognition and shot change detection. The TV news program is chosen as the experimental materials and is divided into two parts: the anchor speech and the field report videos. A speech summarization method is used to extract key speech segments from the anchor speech. Then, a video skimming method is applied to minimize dispensable images. Finally, the extracted anchor speech and the field report image sequence of TV news are aggregated into a summarization output. Experimental results from subjective and objective evaluation demonstrate that the proposed framework achieves a satisfactory performance.

5. REFERENCES

[1] I. Manu and M. Maubury, *Advances in Automatic Summarization*. Cambridge, MA: MIT Press, 1999.

[2] C. Hori and S. Furui, “A new approach to automatic speech summarization,” *IEEE Trans. on Multimedia*, vol. 5, no. 3, pp. 368-378, 2003.

[3] Sangkeun Lee; Hayes, M.H, "An application for interactive video abstraction," *in Proc. Of ICASSP,* vol. 5, pp. 17-21, 2004.

[4] R. Lienhart. "Comparison of automatic shot boundary detection algorithms." *In Image and Video Processing VII 1999*, *Proc. SPIE* 3656-29, January 1999.

[5] T. Kemp and T. Schaaf, “Estimating confidence using word lattices,” *in Proc. 5th Eurospeech*, vol. 2, Rhodes, Greece, 1997, pp. 827-830.

[6] C.H. Hsieh, C.L. Huang and C.H. Wu, “Spoken document summarization using topic-related corpus and semantic dependency grammar,” in *Proc. ISCSLP'04*, Hong Kong, 2004, pp. 333-336.

[7] Christopher D. Manning and Hinrich Schutze, *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[8] HowNet, http://www.keenage.com/

[9] CKIP Treebank http://godel.iis.sinica.edu.tw/CKIP/treebank/