Chinese Optical Character Recognition for Information Extraction from Video Images *

Wing Hang Cheung, Ka Fai Pang, Michael R. Lyu, Kam Wing Ng, Irwin King
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong

Abstract *A number of research work on text extraction from videos is conducted in these few years, but not many focus on the Chinese language. Due to different characteristics between Chinese and English, new methods are urgently needed for Chinese character extraction. We are applying optical character recognition (OCR) techniques to video frames, and trying to extract character texts from videos. By doing this, we can automatically convert video contents to text, and then exploit the Chinese subtitles for indexing and searching in a digital video library. This paper covers ways we filter the heavy noise and segment out each Chinese character in video segments. We also describe how we perform the optical character recognition for Chinese and evaluate its performance. Some future work will also be described.*

Keywords: video image processing, Chinese OCR, character segmentation, character recognition

1 Introduction

There are numerous research work on text extraction from videos in recent years [1][2][3][4], but few people focus on a Chinese text extraction system. As Chinese language includes various dialects, quite a lot of Chinese video programs, like news, dramas, and movies, would provide Chinese subtitles for their audience. Hence, OCR techniques on the subtitles can also be widely applied to the extraction of information from videos besides using speech recognition techniques.

The target of our project is to implement a Chinese character recognition tool to automatically extract the subtitles in videos for indexing and searching purposes. Besides giving abstracts to the video contents, the extracted text can also be used to index the video with respect to their playing time. We can simply relate the extracted text with the frame number, and then retrieve a particular frame by using the extracted text as an index later. Many Asian countries, such as Japan and Korea, have been greatly influenced by the Chinese culture, and their written languages have very similar features with the Chinese characters. Consequently, our work can also be applied to subsets of Japanese and Korean languages.

In Section 2, we describe how Chinese characters differ from those in western languages, and discuss the features of subtitles in a video. Then, we present the noise filtering methods and character segmentation methods used in Section 3, and describe the ways of performing Chinese character recognition in Section 4. In Section 5, we evaluate the performance of the system and then present the conclusion and our future work in Section 6.

2 Features of Chinese Subtitles in Video

The Chinese language, which is being used by billions of people in the world, is quite different

*The work described in this paper was supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region (Project No. CUHK4432/99E).

from other western languages in its representation methods. It has the following features:

1. Chinese characters are stand-alone characters.

2. They are square-shaped.

3. Each of the characters has its own meaning.

4. Phrases can be formed by combining separate characters.

For traditional Chinese characters, there are more than 47,000 distinct characters [5], but only about 3,000 to 5,000 characters are frequently used. The large number of Chinese characters greatly increases the complexity in the character recognition process.

Following is a summary of the assumptions we made for the features of subtitles in videos[1]:

1. Characters are in the foreground, and would not be covered by other objects.

2. They are monochrome, and contrast when compared to the background.

3. They are upright and rigid. From frames to frames, their shapes, sizes, or orientations would not change.

4. They have size restriction - they cannot be too large or too small.

5. They usually pop out to the screen, but not fade in or slide in.

6. For Chinese subtitles in videos, they usually appear in clusters and are aligned in horizontal lines, from left to right.

Our noise filtering methods and character segmentation methods are based on these features.

Figure 1: A frame extracted from news video

3 Noise Filtering and Character Segmentation

Extracting characters from videos is very different from extracting characters from printed documents. The background of a printed document is usually in pure color (e.g. white paper), but the background of a video may be full of colorful, complicated, or even moving objects. These complications the difficulty in extracting characters. In order to extract characters from videos, we have to locate the position of the lines of characters. Agnihotri and Dimitrova [4] suggest that we can use the density of edges to locate the text area, as texts are usually rich in edges and contrast with background. We have exploited the property of high edge-density in a text area in a similar way, and developed a similar segmentation method for Chinese characters.

We employ two methods to filter the noise. One is using a static frame, and the other is comparing two frames. For the first method, we get a frame from the video, and extract the red channel for processing which can give a higher contrast in general. A frame from a news video is presented in Figure 1 as an example. Then, an edge filtering to the frame is done by using Sobel filter [6], as shown in

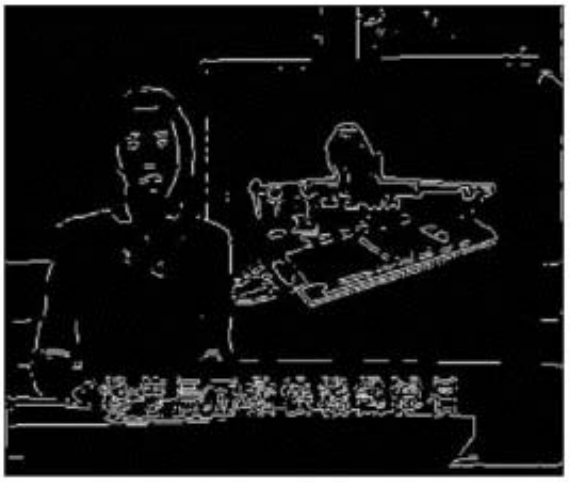

Figure 2: Frame edge filtered by Sobel filter

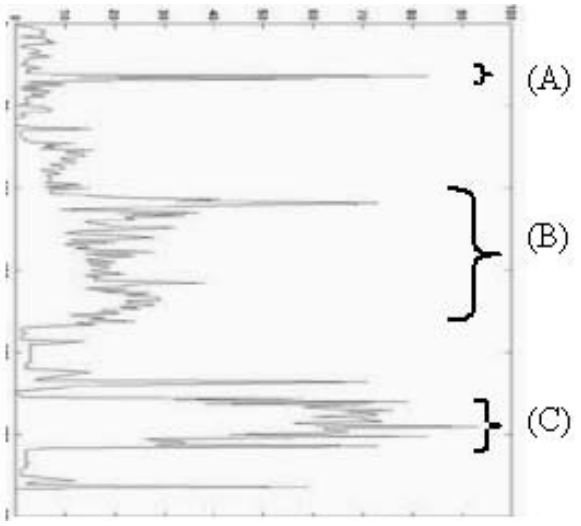

Figure 3: Graph for edge density in horizontal lines

Figure 2.

We can then plot the histogram of edge density in horizontal lines as shown in Figure 3. Range (A) in Figure 3 is so narrow that no text in that area is indicated, as text may be too small to read. Range (B) is too wide to indicate any text in that area, as text may be too large and it will block the screen. Range (C) is within a normal range of text size, so we can conclude that a text line may be located there. If no suitable range of edge density is found from one frame, we can then conclude that the frame contains no text. Now we filter out the range with values lying outside the threshold noise value, and remove all the non-text ranges found. Consequently, we have a text line as shown in Figure 5. This method is very efficient in determining whether there are subtitle-lines in one frame or not, but the disadvantage is that the extracted text lines would still be noisy.

Another method of filtering is by comparing two frames. While processing a video, we will not examine every frame in order to save processing time. We examine one in every five frames so that we will have a better efficiency and will not miss any subtitles. We use the previous method to determine which frames contain a new text. When we notice one frame containing new text, we will perform a subtraction with the previous frame. As the text usually pops out in the screen, we can filter out all the background and figure from the outstanding text. One weakness of this method is that if the subtitle's color is not in great contrast with the background color in some places, this area would be missed after the subtraction. We will combine the results of both methods to proceed further for character segmentation and recognition.

Figure 4: Some Chinese characters would be wrongly regarded as two characters (fig. a) or three characters (fig. b) in vertical projection

Some research work suggest using vertical projection profile for checking the boundaries of English letters. This method, however, is

not suitable for Chinese characters. For example, if we apply vertical projection to some unconnected Chinese characters like the one in Figure 4 (a) (sigh, read as 'shen') or the one in (b) (river, read as 'chuan'), it would incorrectly indicate that there are 2 and 3 characters respectively. To tackle this kind of problem, we have to use the square-shaped features of Chinese characters to perform the segmentation.

Figure 5: First filtering of the non-text area

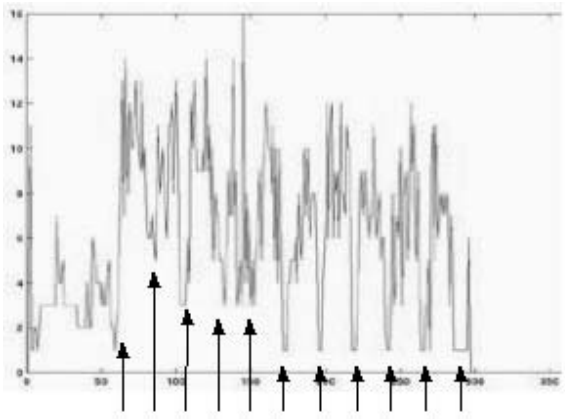

Figure 6: Edge density graph in vertical lines

The density graph of vertical lines, i.e., the vertical projection, can help us to indicate the margins of characters, where troughs can be seen. Chinese characters are mostly square-shaped, and hence periodic troughs are expected. The arrows in Figure 6 indicate the periodic occurrence of troughs, which are also the places where the character margins are located. By exploiting the trough periods in Chinese characters, we can:

1. Avoid the wrong segmentation of unconnected Chinese characters.

2. Help to determine the margins even when there are heavy noise between two characters.

We can then segment each character into a character box and normalize them for character recognition.

4 Character Recognition

In the previous section, we have segmented the lines of Chinese characters into rectangular blocks. Now, we are making use of them to do the character recognition. We have made a Chinese Character image library with 5401 character images. They are in 24×24 binary image matrixes and in the order of Big-5 code. All of them are regarded as 'frequently used characters'.

We have to normalize the segmented characters to 24×24 binary image matrixes before recognition. For each character image we obtained in the last section, we will first do simple pattern match with the images in the library, i.e., by subtracting two character images, and calculating the degree of likeliness. The following formula is used to calculate the score:

Let s be the segmented block and r be a reference image in the library. Without loss of generality, we use the dimension of (m×n) for both matrices. The distance between them is defined as,

$$\text{Pattern Distance} = \sum_{i=1}^{m}\sum_{j=1}^{n} |r(i,j) - s(i,j)|$$

The second method we use is the peripheral features [7]. The outlines of a character pattern can be described by peripheral features, where the distances between the matrix edge and the first (ET1) and second (ET2) black jump in a number of segments are summed. The extraction method of the ET1 feature is shown in Figure 7. Feature ET1 gives the outline structure of the characters. The character block is divided into eight horizontal stripes and eight vertical stripes. Within each stripe, the area between the edge and the first deviation from black pixel to white pixel is calculated and this operation would then create a 32-dimensional feature vector (i.e., 4 edges by 8 stripes). Feature ET2 gives the arrangement of strokes inside the character, which the principle is shown in Figure 8. The character pattern is also divided into eight horizontal stripes and eight vertical stripes, like ET1. The area between the edge and the second deviation from black pixel to white pixel is calculated within each stripe so that again a 32-dimensional vector is produced. By comparing ET1 and ET2 of the segmented characters to the referencing characters in the library, we can calculate the scores. For printed document, these two methods can yield 96% to 98% accuracy in size of 64×64 pixels characters respectively [7].

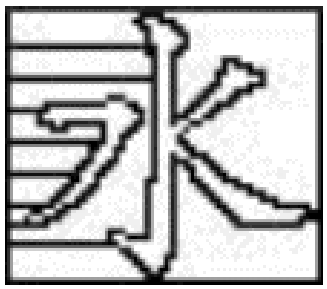

Figure 7: Principle of ET1

By using both ET1 & ET2 principles and the pattern match scheme for calculating the matching scores, we can select the best-matched characters from the library.

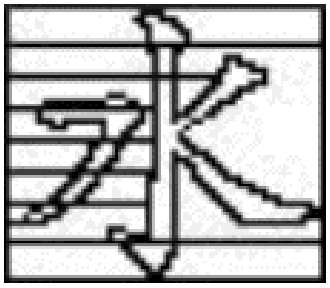

Figure 8: Principle of ET2

5 Evaluation

The videos used for testing our approach are the news videos provided by Hong Kong Asia Television Limited. We converted the videos in videotapes to files in MPEG format, with resolution 352x288 pixels per frame and a frame rate of 25 frames/sec. The video quality is not very good, due to the fact that they have been scaled down for lower bandwidth transmission in the Internet. We have used 10 videos, adding up to 6 minutes, for evaluation. All subtitles in the videos contain Chinese Characters only.

For character segmentation, we have 118 frames containing different subtitle text lines, and all of those frames can be detected. There are 860 characters in those frames, and 831 characters can be segmented out correctly. For the remaining 29 characters, they are unable to be detected. The missed characters have the same features that their colors have very little contrast with the background colors. We have achieved a successful segmentation rate at 96.6%.

For character recognition, we have used all the 831 segmented characters. 295 characters have been correctly recognized, which is equivalent to a character recognition rate at 35.5%. Characters are wrongly recognized due to five main reasons:

1. The character size for recognition is too small (in 24×24 pixels).

2. The edges and skeletons of the characters

become unclear and blurred after being scaled down.

3. The background is still noisy after filtering.

4. Colors of some characters give little contrast with the background colors.

5. The structures of the wrongly-recognized characters are hardly distinguishable from that of the correct ones.

For the correctness, the rate of successful segmentation is satisfactory, but the resulting recognition rate is too low for video indexing. Further improvement is needed.

And for the efficiency, the processing time is proportional to the number of characters in the frames. The processing time for a single frame with 7 to 8 characters is around one minute (in a Sun Ultra 5/270 machine with 128 Mb RAM). We will detect and discard the frames that without characters or that with the same characters as the previous frame, to reduce the time processing time for a whole video clip. The distribution of processing time consumed in different procedures of the application is shown in the following table:

Table 1: Processing time distribution among different procedures

Procedures	% of time
Character Recognition	59%
Reading MPEG Frames	34%
Character Segmentation	5%
Others (e.g. UI)	2%

The long recognition time is due to the retrieval of 5401 character images for character matching in every frame. As the purpose of the appliation is to perform information preprocessing rather than real time processing, the speed is not our prime concern. But the speed can be improved by enhancing the recognition method.

6 Conclusion and Further Work

In this paper, we have presented the features of Chinese characters, and also discussed why some similar work in English character extraction in video cannot be applied to Chinese characters. We have described how we perform Chinese character segmentation and recognition by using our new approaches, and discussed the performance evaluation of our system. The recognition rate is not very satisfactory because the video quality is not very good, as they have been scaled down for lower bandwidth transmission in the Internet. Modification of this project is in progress. One direction is to use some more intelligent character recognition methods in order to increase the character recognition rate. We would try to train and build a decision tree for Chinese character and use it in the recognition process, which can also speed up the recognition time. Moreover, a phrase dictionary would also be used to provide more evidence in recognition, instead of focusing relying solely on OCR. Another direction is to enhance the character segmentation method. Our method is targeted on Chinese character, and we would like to make it more generic for English letters and other symbols as well. This would post a challenge as other languages have quite different features comparing to Chinese. When higher recognition rate can be attained, searching engine for the indexed videos can be applied in the future.

References

[1] Rainer Lienhart. *Automatic Text Recognition for Video Indexing.* ACM Multimedia 96 pp. 11 - 20, Boston, MA, USA, November, 1996.

[2] T. Sato, T. Kanade, E. K. Hughes, M. A. Smith. *Video OCR for Digital News Archive.* Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on, pp. 52-60, 1998.

[3] J. C. Shim, C. Dorai, R. Bolle. *Automatic Text Extraction from Video for Content-Based Annotation and Retrival.* In Proc. of the International Conference on Pattern Recognition, pp. 618-620, 1998.

[4] L. Agnihotri, N. Dimitrova. *Text Detection for Video Analysis.* Content-Based Access of Image and Video Libraries, 1999. (CBAIVL '99). Proceedings. IEEE Workshop, pp 109-113, 1999.

[5] Qing Dynasty. *Kangxi Dictionary.* Commerical Press, 1903.

[6] R. Gonzalez, R. Woods. *Digital Image Processing.* pp 199, 419, Addison-Wesley, 1993.

[7] J. Guo, R. Suchenwirth, I. Hartmann, G. Hincha, M. Krause, Z. Zhang. *Advances in Control Systems and Signal Processing - Optical Recognition of Chinese Characters.* pp 60-64, Braunschweig : F. Vieweg, 1989.