Computational Linguistics and Chinese Language Processing
Vol. 6, No. 2, August 2001, pp. 11-30

A Simple Method for Chinese Video OCR and Its Application to Question Answering

Chuan-Jie Lin[*], Che-Chia Liu[*], Hsin-Hsi Chen[*]

Abstract

Captions in videos contain valuable information for video retrieval. Although texts in captions can be obtained easily in the new image compression formats like MPEG2, there still are many video programs encoded in older formats. Thus, video OCR is indispensable for content-based video retrieval. This paper proposes a simple video OCR method for Chinese captions, including image capturing, caption region deciding, background removing, character segmentation, OCR and post-processing. We employed Discovery Channel films as training and testing corpus. In an outside test, the accuracy of the video OCR was 84.1%. The hardware used in the experiment consisted of a computer with a P4-1.7G CPU, 256MB RAM and a 40G, 7200rpm hard disk. On average, it took 29 minutes and 11 seconds to process a film 495MB in size. We also applied the results of video OCR to video retrieval and question answering.

Keywords: digital library, question answering, Chinese video OCR, video retrieval

1. Introduction

In the new information era, multimedia is widely used, and the amount of existing video data is huge. How to extract the content of video data for further application has become an important issue. The well-known project "Informedia" [Wactlar, 2000] in digital library is a typical example. Captions in videos contain valuable information for video retrieval. Although texts in captions can be easily obtained in the new image compression formats like MPEG2, there still are many video programs encoded in older formats. Thus, video OCR is indispensable for content-based video retrieval. This paper proposes a simple video OCR method for Chinese captions and demonstrates its application in video search and question answering.

[*] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, TAIWAN, R.O.C.
E-mail: {cjlin, jjliu}@nlg2.csie.ntu.edu.tw, hh_chen@csie.ntu.edu.tw

OCR research started very early and has achieved many good results. In a traditional OCR system, textual data is scanned and saved as images, and then transformed into text files [Lee and Chen, 1996]. There have also been many researches on handwriting OCR. In contrast, video OCR is more challenging than traditional OCR because we have to recognize small characters on a colorful background instead of black characters on a white background.

Several approaches have been proposed to video OCR. Wu *et al*. [1997, 1998] tried to find characters in pictures by means of connected components. Their method performs well on pictures but not films because the background of a film is more complicated, and text will also connect with other objects in the film. Lienhart *et al*. [1998, 2000] found text by means of color segmentation, contrast segmentation, geometry analysis, and texture analysis. Li, Doermann and Kia [2000] adopted a neural network to detect strings in images. Li and Doermann [1999] also employed multiple images to enhance resolution. Smith and Kande [1997] used text and object shifting, and facial recognition to reduce the size of images. Sato *et al*. [1998] achieved higher OCR correctness by means of image improving and multi-frame integration.

This paper focuses on Chinese captions in videos. Section 2 introduces several issues concerning video OCR and the architecture of our system. Sections 3 to 8 describe each strategy and each module in detail. The performance was evaluated using films made by the Discovery Channel. Section 9 demonstrates an application for question answering. Section 10 presents conclusion.

2. Architecture

There are two kinds of texts in videos, i.e., captions and image texts. Captions often appear at specific positions, such as a textual line in the lower part of a screen, or a vertical text line in the left or right part of a screen. Image texts consist of characters appearing in an image, such as shop signs, automobile registration numbers, *etc.* They are themselves part of the image, so they change their positions when the camera moves. Captions are narratives or dialogues in a film, so they often carry valuable information. The focus of this paper is how to extract texts in captions.

Complex backgrounds often show up behind captions; thus, the first problem is how to remove backgrounds. After backgrounds are removed, the remaining captions are black characters on a white background. That will make the following OCR task easier. We also apply a post-processing procedure to improve OCR performance. Figure 1 shows the architecture of the whole system.

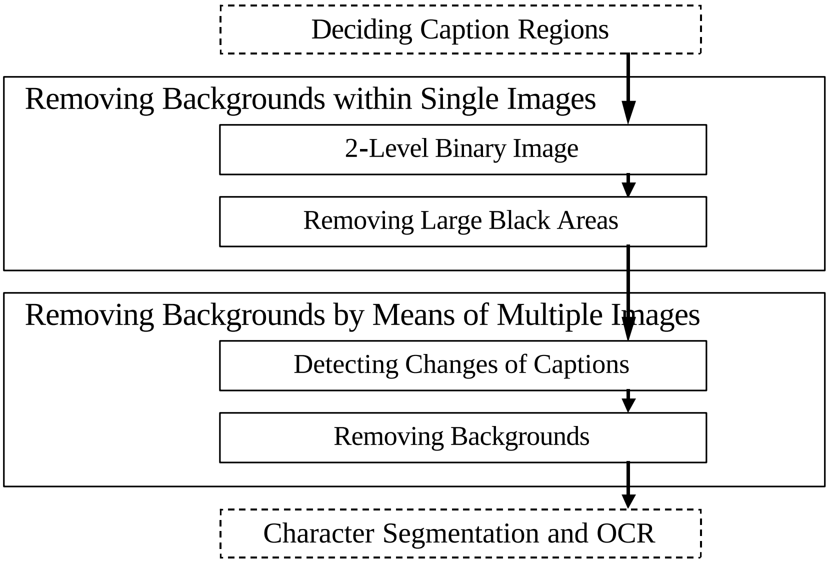

Deciding Caption Regions
Removing Backgrounds within Single Images
2-Level Binary Image
Removing Large Black Areas
Removing Backgrounds by Means of Multiple Images
Detecting Changes of Captions
Removing Backgrounds
Character Segmentation and OCR

Figure 1 *The Architecture of the Video OCR System.*

To evaluate the performance of the system, some films produced by the Discovery Channel were used as experimental materials. Their topics vary widely from natural science to history, military, adventures and human life.

3. Deciding Caption Regions

The characteristics of captions are: (1) they are always in a straight line from left to right or up to down; (2) the characters usually have colors which contrast with the background, and often have perceivable borders; (3) they are always in the foreground of the image; (4) they usually consist of two or more characters; (5) the height of the caption region is not often higher than one third of the height of the image, because characters cannot be too large or too small for reading; (6) they have fixed height, width, and size; (7) they have fixed colors. We employ these characteristics to locate captions.

3.1 Binary Image

Before processing, we first transform the original images into binary images. This technique is often used in video processing. It helps to simplify the background and make the retrieval of captions much easier.

When extracting images from a film, we take 2 pictures in a second and save them in the BMP format. In a BMP file, the color of each point is recorded as its *RGB*-value, (*red-value*, *green-value*, *blue-value*). Each value ranges in brightness from 0 to 255. Here, 0 indicates the darkest value and 255 the brightest value.

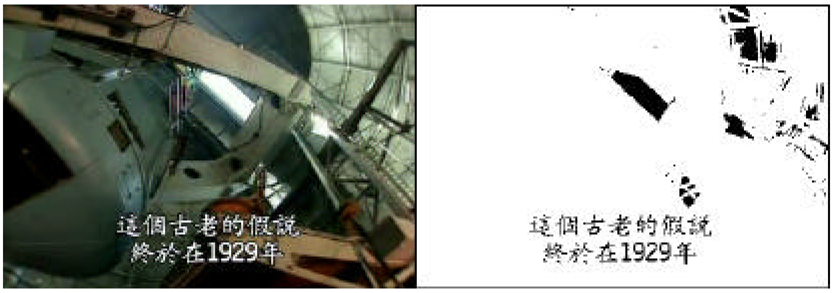


Figure 2 *An Example of Binary Image Transformation.*

Using the *RGB*-values, we can transform an image into a binary image using the following method:

```
Let the binary-threshold be SegColorScore
For each point (red-value, green-value, blue-value) in an image:
    IF   red-value, green-value, and blue-value are larger than SegColorScore
        THEN change the color of this point to black, i.e., (0, 0, 0)
        ELSE change the color of this point to white, i.e., (255, 255, 255).
```

In our experiment, *SegColorScore* was set to 190. Figure 2 shows an example of binary image transformation. The captions are clearly separated from the background. The result is black characters on a white background.

3.2 Deciding Caption Regions

After performing binary image transformation, we decide where the captions are. Here, we employ another characteristic of captions: if we draw a horizontal line across a caption, the line will go through many vertical lines of Chinese characters. As in printed characters, these vertical lines are often of the same width.

Consider every point at the same height $height_i$. A sequence of black points is called a *segment*. In this way, a horizontal line at $height_i$ is composed of a set $SEGMENT_i=(segment_{i1}, segment_{i2}, ..)$ of segments. If the difference between the numbers of black points in two neighboring segments is not larger than a predefined threshold (e.g., 3 in this paper), then we say these two segments belong to the same group. Thus, we have a set $GROUP_i=(group_{i1}, group_{i2}, ..)$ at $height_i$. $Seg(group_{ij})$ is defined as the number of segments in $group_{ij}$. Now we define *Score As Caption Region* (abbreviated as *SACR* hereafter) of $height_i$ as

$$SACR_i = \sum_{j=1}^{|GROUP_i|} Seg(group_{ij}) \times \log_2 Seg(group_{ij}) . \qquad (1)$$

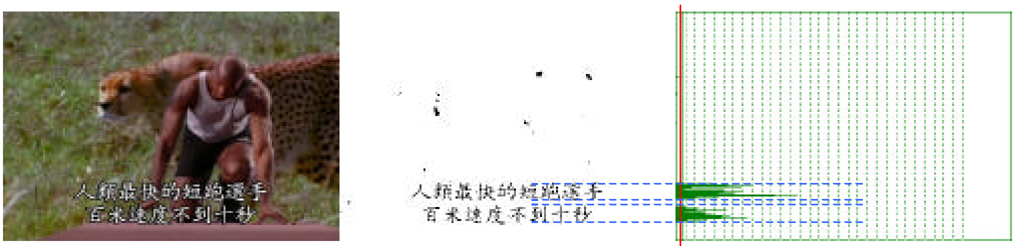


Figure 3 *Examples of Deciding Caption Regions (1).*

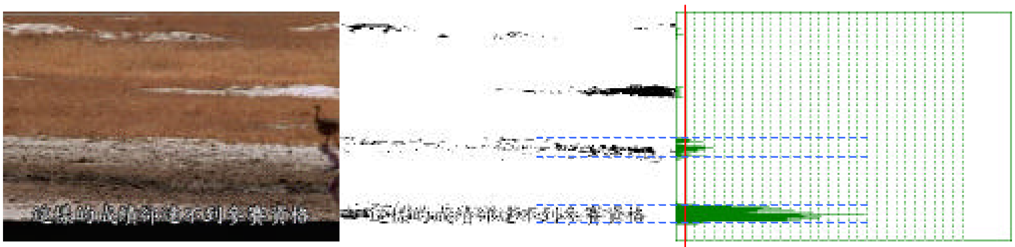


Figure 4 *Examples of Deciding Caption Regions (2).*

Consider the following example. Here, 0 denotes a white point and 1 a black point.

```
points:     00111011111001100011101111111011111111100111101110110111111
segments:   --111-22222--33---444-5555555-666666666--7777-888-99-AAAAAA
groups:     |---------1----------||--------2-------||-----3-----||--4--|
Seg(group):           4                   2             3         1
```

SACR in this example is $4\log_2 4+2\log_2 2+3\log_2 3+1\log_2 1=14.75$.

Assume that the height of an image is m. We calculate m *SACR*'s for the height levels and compute the average $\overline{SACR}$. The height levels that have *SACR*'s higher than the average one are in the caption region. Figures 3 and 4 show two examples. On the left side is the original image; in the middle is its binary image; and on the right side is the corresponding *SACR* of each height level, where the x-axis denotes the height, the y-axis denotes the *SACR* value, the solid vertical line is $\overline{SACR}$, and the horizontal dashed lines denote the caption regions.

3.3 Evaluation

The experiment was performed on three Discovery films: "Lightening," "Animals in the Wild," and "Whales." There were 69, 66, and 41 sentences in captions, respectively. The first 500 images of each film were extracted as experiment data. As shown in Table 1, the precision rates obtained were 76.7%, 39.8% and 82.0%, respectively, but the recall rates were nearly 100%. Errors occurred in cases like the stone road shown in Figure 4. The white stone road in the image had many black segments of the same width, so it was misjudged as a caption region. Such misjudgments can be filtered out in the OCR processing stage. Hence, the recall rate is more important here for retrieving all the captions.

Table 1. *Evaluation of Caption Region Deciding.*

Films	Actual	System Decided	Correct	Precision	Recall
Lightening	69	90	69	76.7%	100.0%
Animals in the Wild	66	161	64	39.8%	97.0%
Whales	41	50	41	82.0%	100.0%

4. Removing Backgrounds within Single Images

When we adjusted the binary image threshold *SegColorScore*, we found an interesting phenomenon: if *SegColorScore* was set too low, the background could not be removed very well; on the other hand, if it was set too high, the background was removed, but the captions were too unclear to do OCR. The value 190 used in the previous module resulted in very unclear images.

To do OCR more precisely, we have to keep the character clear while removing all the background. In this section, we will propose a method for removing backgrounds within single images by employing the difference between the captions and the background. How information from multiple images is used to remove backgrounds will be discussed in the next section.

4.1 2-Level Binary Image

During transformation into binary images, the values of *SegColorScore* will affect the clearness of the remaining images of captions. As shown in Figures 5 and 6, captions are clearly seen when *SegColorScore* is set to 140, but more background parts remain. The situation is reversed when it is set to 180.

Here, we propose a new method, called **2-level binary image transformation**, which employs two different *SegColorScore* values to keep captions clear and to remove backgrounds at the same time. The method is described in the following.

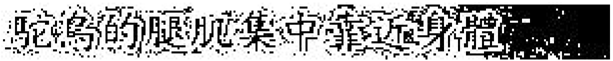


Figure 5 *Binary Image of SegColorScore=140.*

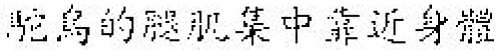


Figure 6 *Binary Image of SegColorScore=180.*

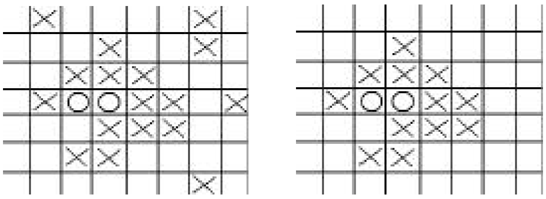


Figure 7 *Illustration of 2-Level Binary Image Transformation.*

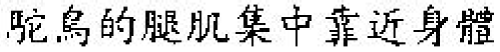


Figure 8 *2-Level Binary Image of Figure 5 and Figure 6.*

Given a picture, we overlap two binary images obtained using two different *SegColorScore* values (let *HiSegColorScore* be the higher one, and *LowSegColorScore* the lower one). Consider the example shown in Figure 7. '○' denotes a black point in both binary images, and '×' a black point only in the binary image obtained using a lower *SegColorScore* value. We keep only those '×' areas adjacent to a '○', because those areas are regarded as black points, and change the other areas into white points. The resulting image is shown on the right side of Figure 7. Figure 8 shows the 2-level binary image result obtained from Figures 5 and 6, which is a clearer caption image.

4.2 Removing Large Black Areas

Consider the image shown in Figure 9, which contains large black areas. It is not easy to remove a background area with a high brightness value using the above method. Thus, another method shown below is proposed to clean such an area if it is large and wide. We will try to deal with small fragments in the next section by using multiple images of the same caption texts.

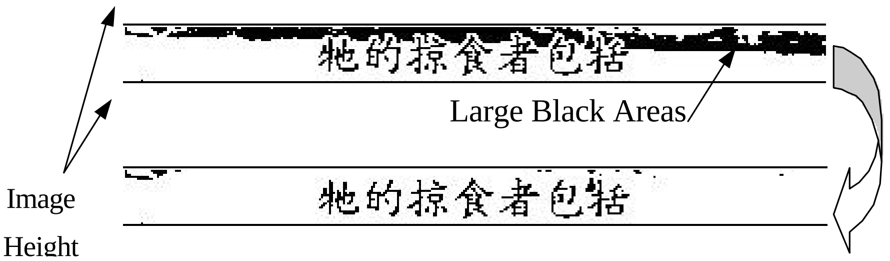


Figure 9 *An Example of Removing Large Black Areas.*

```
Range = (height of the caption region) ÷4;
Total = Range × Range × 0.9;
CHECK each black point in the caption region
    Look at a square with edge of Range and with an upper-left corner at this point
    IF the number of black points in this square >= Total (i.e., 90% of the points are
      black)
      THEN clear all the points in the area adjacent to this point
END
```

5. Removing Backgrounds by Means of Multiple Images

We employ another characteristic of captions to remove small and bright backgrounds; i.e., the positions of the images of captions will not change with the camera, but the background will. We overlap all the images with the same caption texts. Those black points which appear in almost all the images are considered as caption texts. In the next two subsections, we will introduce the method we use to detect the changes of caption texts and the method we use to remove backgrounds by means of multiple images.

5.1 Detecting Changes of Captions

The first task in removing backgrounds with multiple images is to decide which images have the same caption texts. Refer to the example shown in Figure 10. We record the border information of all the black areas. After reading the next image, we compare the border information with that of the previous one. If the difference is larger than a threshold, say,

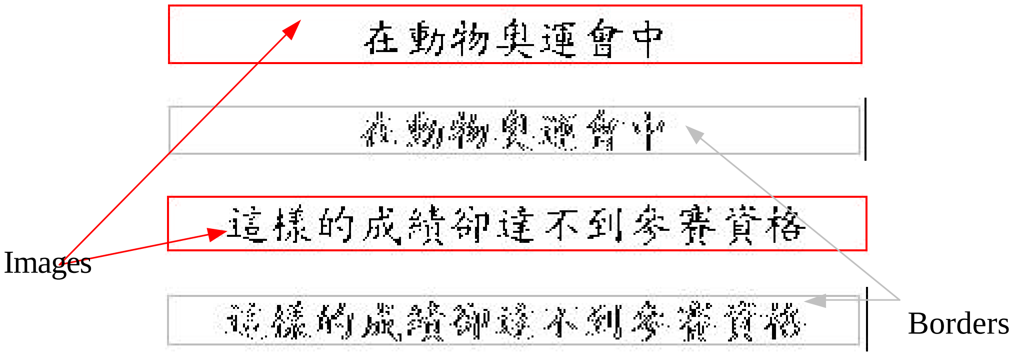


Figure 10 *An Example of Detecting the Change of Captions.*

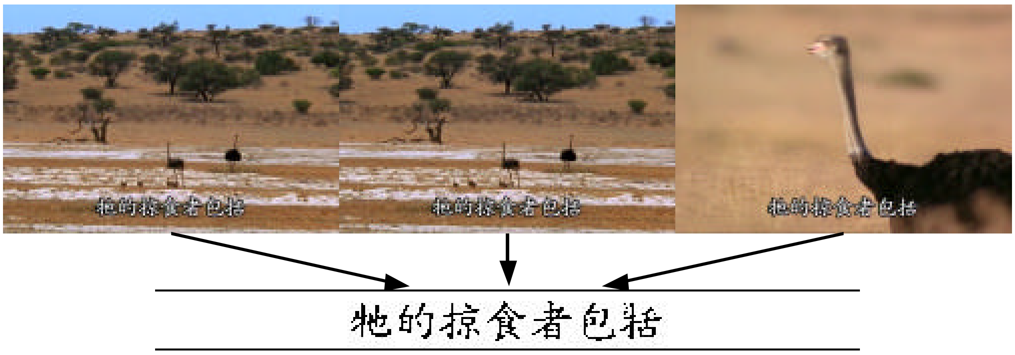


Figure 11 *An Example of Removing Backgrounds by Multiple Images.*

SceneChangeScore, we postulate that the caption texts are different. In the experiment, the value of *SceneChangeScore* was set to 0.6.

The same three films used to evaluate the method used to determine caption regions were also used to evaluate this method. Table 2 shows that the performance was quite good.

Table 2. *Evaluation of Detecting Changes of Subtitles.*

Film	Number of Changes	Number of False Alarms	Correctness
Lightening	69	0	100.0%
Animals in the Wild	66	3	95.5%
Whales	41	0	100.0%

5.2 Removing Backgrounds by Means of Multiple Images

After detecting a sequence of images with the same caption texts, we use the following method to remove the backgrounds. Let *NumFrames* be the total number of sequential images. We consider each point in the caption region. If it is black in 90% of the images (i.e., *NumFrames* × 0.9), then we set the point as black. Otherwise, it is set as a white point. Figure 11 shows an example. The background is removed more clearly than that is in Figure 9.

6. Character Segmentation

At this point, there exists a binary image that has black characters on a white background for each sentence in a caption. We next apply traditional OCR techniques to retrieve caption texts. The first step in performing OCR is to decide the boundaries of each character.

We first decide the left and right boundaries. The most popular way to perform character segmentation is to use projection profiles [Lu, 1995]. As shown in Figure 12, we project every black point onto a horizontal line. Intuitively, the projection for the space between Chinese characters is zero. However, there is also space inside a Chinese character. We employ another cue to resolve this problem. The width of Chinese characters is often

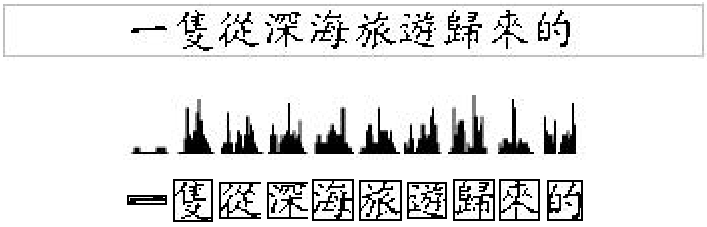


Figure 12 *An Example of Character Segmentation.*

approximately equal to their height. Let the height of a caption region be *ImageHeight*. The gap that is a distance of *ImageHeight*×0.7 ~ *ImageHeight*×1.4 from the previous gap will be regarded as a possible segmentation point.

After deciding the left and right boundaries, we use the same method to decide the upper and lower boundaries of each character.

7. Optical Character Recognition

We adopt a statistical model similar to that of Oka [1982] to perform Chinese OCR. Figure 13 shows an example. Each character image is separated into 16 equal parts. Starting from the center of each part, we observe its up, down, left, and right directions. If there is a black point in a given direction, the corresponding signature value is set to 0. Otherwise, it is set to 1. In this way, we will have 64 (16 parts × 4 directions) values (called a signature) for each character image.

A set of character images that were retrieved from the Discovery Channel films formed a corpus for collecting the signatures of a standard character corpus. When recognizing a new image, we first extract its signature and then compare it with the ones in the standard character corpus. The similarity is measured by counting how many values are matched. Therefore, the similarity score will be between 0 and 64. The higher the score is, the more similar the two patterns are. If the highest score of a new image is less than 16, it is regarded as non-character image.

The following is an example. A new image 傳 is compared with '**傳**' and '**博**' in the standard character corpus. The corresponding signatures are as follows:

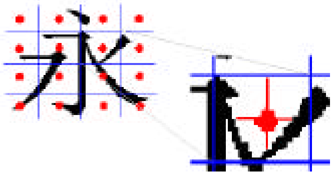


Figure 13 *Signature of Image "永".*

00000003-1-01.bmp: (56)**探** (52)**權抓徹** (51)**撇獵很多育擺**
00000003-1-02.bmp: (58)**孝素** (57)**幸索** (56)**案業** (55)**希考常** (54)**途**
00000003-1-03.bmp: (56)**遠速** (53)**達遺逝遺道** (52)**道情運**
00000003-1-04.bmp: (60)**傳** (52)**博** (51)**偉** (50)**佈搏佛格** (49)**踏彈像**
00000003-1-05.bmp: (59)**學卡層** (51)**餵峰單軍旁革帶**
00000003-1-06.bmp: (59)**的勾鄉** (52)**稀將哺特豹** (51)**均擠**
00000003-1-07.bmp: (60)**奇** (53)**槍青賽考** (52)**希老奔逢旁**
00000003-1-08.bmp: (60)**異** (52)**具其提隻** (51)**週各姿域農**
00000003-1-09.bmp: (59)**世** (53)**親摺甘種** (52)**奇發音實奮**
00000003-1-10.bmp: (63)**界** (55)**善** (53)**輩華** (52)**谷才在像毒舞**

Figure 14 *The First Ten Candidates of OCR.*

1010100010001101110100100000101111101010010001011111111111001111

傳: 1010100000001101010101100000100111101010010001011111111111001111

博: 1110100000011111010001000100100111101100010001011111111111101101

The similarity of the image with '**傳**' and '**博**' is 60 and 50, respectively, so '**傳**' is ranked as the first candidate of this image.

Figure 14 illustrates the first ten candidates of each character image in after OCR is performed. The correct rate of the top one is very high, and most of correct answers appear in the top ten. Table 3 shows the top one performance. The film "Genetics" was used in inside test, and "King of the Pyramids" and "The Real Cleopatra" were used in the outside tests. The results show that the correct rates in the inside test was 91.5%, and that the performance of the outside tests was 78.5% and 81.5% for the two films, respectively.

Table 3. *Experiment Results of OCR.*

Films	TOTAL	CORRECT	ERROR	MISS
Genetics	809	739 (91.5%)	69 (8.5%)	0
King of the Pyramids	684	537 (78.5%)	110 (16.1%)	37 (5.4%)
The Real Cleopatra	750	611 (81.5%)	86 (11.5%)	53 (7.1%)

8. OCR Post-Processing

We found that nearly 95% of the correct answers were in the top ten candidates, and Table 3 shows that the top one achieved 91.5% performance in the inside test. This section will touch on how to promote the correct answer which is not ranked first initially to the first position to improve the overall performance.

8.1 Basic Model

In Figure 14, a value enclosed in parentheses before a candidate denotes its similarity score. First, we filter out those candidates whose scores are lower than the score of the top one candidate by a threshold. The filtered characters are shadowed in Figure 14. Only the characters with larger scores are retained. This will reduce the number of possible candidates. Then, we perform the following steps. Consider three characters denoted *ABC* sequentially. Generate all the possible candidate pairs for *ABC*, e.g., A_iB_j or B_mC_n. Check if a candidate pair is in a dictionary (i.e., a two-character word), or is a part of a three-character word. If it is, we multiply the OCR similarity scores of these two candidates. Otherwise, their score is set to zero. Next, we find the pair with the highest score. If it is A_iB_j, then A_i and B_j are selected, and we start the next iteration from *C* (i.e., *CDE*). If it is B_mC_n, then A_1, i.e., the top one candidate of *A*, is selected, and we start the next iteration from *B* (i.e., *BCD*).

8.2 Strategies Used in Experiments

For the above algorithm, several issues had to be evaluated in the experiments. For example, should we consider all the combinations of characters? Is the top one candidate more important than the others? Are longer words in the dictionary more helpful? We applied 3 strategies to the basic model to examine these factors. The experimental results were compared with those obtained using the Select-First and Longest-First models.

[Strategy 1] All pairs of candidates are considered.

[Strategy 2] Only pairs consisting of at least one ranked first candidate are proposed. In other words, when *AB* are recognized, only A_1B_1, A_1B_2, . .,.A_2B_1, A_3B_1, . . are considered.

[Strategy 3] 4- or 3-character words in the dictionary are proposed first. Then, Strategy 2 is considered.

8.3 Evaluation

The standard character corpus was collected from six Discovery films (i.e., "Natural Born Winners," "Snakes," "Genetics," "The Southern Rockies," "Great Quakes: Kobe, Japan," and "Galapagos: Beyond Darwin"). There were in total 7,818 character images, and only 2,256 signatures of distinct characters were recorded. Tables 4 to 6 show the experimental results for the three different films. Among them, "Genetics" was used in the inside test; "King of the Pyramids" and "The Real Cleopatra" were used in the outside tests. The first 700 images of each film were extracted as experiment data. The notations used in the tables are defined as follows:

TOTAL:	total number of characters in captions;
CORRECT:	number of characters recognized correctly;
ERROR:	number of characters collected in the standard corpus but recognized incorrectly;
MISS:	number of characters which are not collected in the standard character corpus;
Improve:	improvement relative to the baseline;
Select-First:	(baseline) select the top one candidate;
Longest-First:	select the longest candidate combination which is collected in the dictionary.

Table 4. *Experimental Results of Post-Processing for the Film "Genetics".*

	TOTAL	CORRECT	ERROR	MISS	Improve
Select-First	809	739 (91.5%)	69 (8.5%)	0	------
Longest-First	809	753 (93.1%)	56 (6.9%)	0	1.6%
Strategy 1	809	751 (92.8%)	58 (7.2%)	0	1.3%
Strategy 2	809	759 (93.8%)	50 (6.2%)	0	2.3%
Strategy 3	809	762 (94.2%)	47 (5.8%)	0	2.7%

Table 5. *Experimental Results of Post-Processing for the Film "King of the Pyramids"*

	TOTAL	CORRECT	ERROR	MISS	Improve
Select-First	684	537 (78.5%)	110 (16.1%)	37 (5.4%)	-------
Longest-First	684	544 (79.5%)	103 (15.1%)	37 (5.4%)	1.0%
Strategy 1	684	546 (79.8%)	101 (14.8%)	37 (5.4%)	1.3%
Strategy 2	684	559 (81.7%)	88 (12.9%)	37 (5.4%)	3.2%
Strategy 3	684	563 (82.3%)	84 (12.3%)	37 (5.4%)	3.8%

Table 6. *Experimental Results of Post-Processing for the Film "The Real Cleopatra".*

	TOTAL	CORRECT	ERROR	MISS	Improve
Select-First	750	611 (81.5%)	86 (11.5%)	53 (7.1%)	------
Longest-First	750	614 (81.9%)	83 (11.1%)	53 (7.1%)	0.4%
Strategy 1	750	635 (84.5%)	62 (8.3%)	53 (7.1%)	3.0%
Strategy 2	750	640 (85.3%)	57 (7.6%)	53 (7.1%)	3.8%
Strategy 3	750	644 (85.9%)	53 (7.1%)	53 (7.1%)	4.4%

Tables 4, 5, and 6 show that Strategy 3 was the best one. The correct rates were 82.3% and 85.9% in the outside tests, and 94.2% in the inside test. 5.4% and 7.1% of the characters could not be found in the dictionary in the outside tests, respectively.

We further compare the experimental results obtained using Strategy 3 and the Select-First Model in Table 7, where “T→F” is the number of characters recognized correctly using Select-First but incorrectly using Strategy 3, and “F→T” is the number of characters recognized correctly using Strategy 3 but incorrectly using Select-First. From Table 7, we can find that “T→F” case was only 0.7%, but that 3.0% to 5.2% of more characters could be recognized correctly. This leads us to the conclusion that post-processing is helpful.

Table 7. *Comparison of Strategy 3 and Select-First.*

Film	Total	Result	T→T	T→F	F→T	F→F
Genetics	809	94.2%	738 (91.2%)	6 (0.7%)	24 (3.0%)	41 (5.1%)
King of the Pyramids	647	87.0%	532 (82.2%)	5 (0.8%)	31 (4.8%)	79 (11.2%)
The Real Cleopatra	697	92.4%	608 (87.2%)	3 (0.4%)	36 (5.2%)	50 (7.2%)

Table 8 shows the experimental results for the three whole films. The main error in the outside test was that about 7~10% of the characters were not collected in the standard character corpus. The signatures of the standard character corpus were collected from the real images of the six films, and only those of 2,256 distinct characters were included.

Table 8. *Experimental Results for the Entire Films Obtained Using Strategy 3.*

Film	Real Answers	Reported by System	Correct (Recall)	Error	Miss
Genetics	9189	8834	8105 (88.2%)	1481(16.1%)	26(0.3%)
King of the Pyramids	7976	7878	6582 (82.5%)	851(10.7%)	543(6.8%)
The Real Cleopatra	8862	8874	7365 (83.1%)	636(7.18%)	861(9.7%)

To solve this problem, we tried to collect the signatures from the existing font types. We experimented on **標楷體** and **華康中楷體**. The experimental results are listed in Table 9. The first experiment was the same the experiment reported in Table 8. In the second and the third experiments, we used 5,401 frequently used Chinese characters as the standard character corpus in **標楷體** and **華康中楷體**, respectively. Comparatively speaking, the results were worse, and using **華康中楷體** was better than using **標楷體**.

In addition, we prepared another standard character corpus for the fourth and the fifth experiments, in which 2,256 signatures came from the original corpus, and the other Chinese characters came from the **華康中楷體** images. The performance was improved, but it was still not as good as that obtained in the inside test. Meanwhile, the whole character set (13,060) did not perform better than the set of frequently used characters.

Table 9. *Experimental Results on Different Standard Character Corpora ("King of the Pyramids").*

	Real Answers	Reported by System	Correct (Recall)	Error	Miss
2,256, Original	7976	7878	6582 (82.5%)	851 (10.7%)	543(6.8%)
5,401, 標楷體	7976	7092	2648 (33.2%)	5325 (66.8%)	3(0.0%)
5,401, 華康中楷體	7976	7380	3265 (40.9%)	4708 (59.0%)	3(0.0%)
5,401, Original+華康中楷體	7976	7885	6701 (84.0%)	1272 (15.9%)	3(0.0%)
13060, Original+華康中楷體	7976	7885	6612 (82.9%)	1272 (15.9%)	0(0.0%)

9. Question Answering (QA) System

Since caption texts in video can be extracted successfully using the procedures proposed in the previous sections, we tried to integrate the IR and QA techniques to develop a video question answering system in the next step.

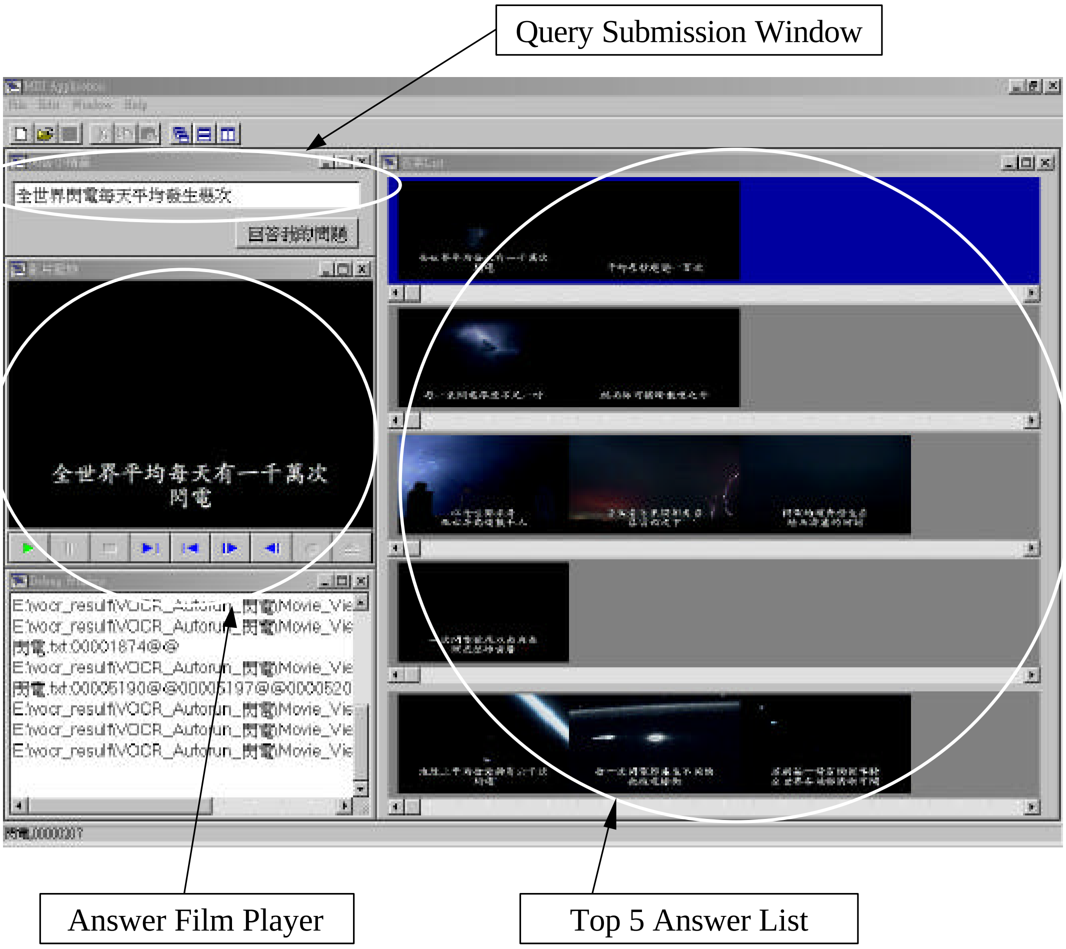
9.1 Video QA System

Figure 15 shows the interface of the Video QA System. Users issue questions in the submission window. The system finds answers in a film corpus and shows them in the answer window with several indicative pictures extracted from the video for each answer. If the user wants to watch the original film for an answer, he can click on that picture, and the system will play the film starting from the answer fragment.

The technique used for QA was proposed by Lin *et al*. [2001]. It implements a question answering system on heterogeneous collections including video. The correctness of Video OCR is not 100% yet (82.3% or better is shown in Tables 5 and 6), so pattern matching in traditional techniques (i.e., matching keywords or synonyms, or searching in other semantic trees) has to take OCR similarity into account. The score for extracting answers can be calculated as follows:

$$score(qw_i, pw_j) = 0 \quad \text{if } |qw_i| \neq |pw_j|$$

$$\text{else} = \left(\sum_{k=1}^{|qw_i|} Ocr(qc_k, pc_k) \Big/ |qw_i| \right) \times weight(qw_i), \qquad (2)$$



```
E:\vocr_result\VOCR_Autorun_閃電\Movie_Vie
E:\vocr_result\VOCR_Autorun_閃電\Movie_Vie
閃電.txt:00001874@@
E:\vocr_result\VOCR_Autorun_閃電\Movie_Vie
閃電.txt:00005190@@00005197@@0000520
E:\vocr_result\VOCR_Autorun_閃電\Movie_Vie
E:\vocr_result\VOCR_Autorun_閃電\Movie_Vie
E:\vocr_result\VOCR_Autorun_閃電\Movie_Vie
```

__Figure 15__ The Interface of Video QA System.

where $|qw_i|$ denotes the number of characters in qw_i, and qc_k is the k^{th} character in qw_i (the same convention is used for pw_j). Ocr(qc_k, pc_k) is the OCR similarity of characters qc_k and pc_k.

9.2 Evaluation

9.2.1 Questions

Testing questions were collected from "Assignment Discovery" at the web site of Discovery, traditional Chinese version (http://chinese.discovery.com/sch/index.html). "Assignment Discovery" is a project that provides many learning lessons from Discovery programs. This project provides lesson plans, activities, and comprehension questions and answers for teachers to use in designing study programs for students.

We selected the comprehension questions for six films as our testing questions to do the evaluation. We collected questions from this website in order to avoid bias. The films were "Elephants," "On Jupiter," "Hubble: Secrets from Space," "Eye of the Serpent," "Whales," and "Lightning."

9.2.2 Performance

The performance of the QA system was measured in MRR (Mean Reciprocal Rank), which was used in the QA evaluation of TREC QA-Track [Voorhees, 2000].

There were 43 questions in total for these six films. The experiment results are listed in Table 10. The MRR result was 0.1848 (=(4+5/2+3/3+1/4+1/5)/43). 32.6% (14/43) of the questions were answered correctly.

Table 10. *Evaluation of the Video QA System.*

Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Not Answered
4	5	3	1	1	29

From our investigation, the main sources of errors were as follows:

(1) Characters in keywords were not collected in the standard character corpus,

for example, “**雹**” in the question “**冰雹如何形成？**”

(2) Paraphrase problem.

For the question “**木星繞行太陽一週需時多久？**”, the answer text is “**木星環繞太陽一周，須地球時間十二年。**” The two phrases “**繞行一週**” and “**環繞一周**” are paraphrases.

(3) More precise rules for deciding question focus are required.

Consider the question “**閃電可以到達多熱的程度？**” It is classified as “QUANTITY,” so all quantity expressions become possible candidates. But we should only look for temperature expressions as answers.

(4) World knowledge is needed.

Consider the question “**歷史上第一位做閃電實驗的人是誰？**” The correct answer mentions that Franklin did an experiment in 1752, but “the first” is not mentioned. Therefore, it is hard to decide whether he was the first experimenter.

We only employ information consisting of question foci, question keywords, and Named Entities in our Chinese QA system. From the above observations, world knowledge and semantic analysis are needed to answer these questions, especially “How” and “Why” questions. This is a challenging problem.

10. Conclusion

This paper has introduced a Chinese video OCR system, including image capturing, caption regions deciding, background removal, character segmentation, OCR, and NLP post-processing. The correctness achieved is above 90% for the inside test, and above 80% for the outside test. Its application to video retrieval and a QA system have also been discussed.

There are mainly four kinds of OCR errors: (1) the standard character corpus is not complete; (2) the background is not clear enough; (3) character segmentation errors; and (4) errors in OCR post-processing. In our standard character corpus, there are only 2,256 characters. But there are 5,401 frequently used Chinese characters, not to mention 7,659 less frequently used characters. This is why many characters could not be recognized. In our experiments, most of the backgrounds could be cleared successfully. But if the objects do not move, or if small fragments appear behind the captions, it is not easy to remove them using our method. This will affect the performance of character segmentation and OCR. The OCR errors may also propagate to the post-processing module. For example, a character image that is not in the standard character corpus will not have a correct answer among its candidates, and these ten candidates will affect the choice of other characters.

References

Discovery Channel, http://chinese.discovery.com/.

Lee, Yue-Shi and Hsin-Hsi Chen, "Analysis of Error Count Distribution for Improving the Postprocessing Performance of OCCR," *Communication of Chinese and Oriental Languages Information Processing Society*, 1996, pp. 81-86.

Li, Huiping and David Doermann, "Text Enhancement in Digital Video Using Multiple Frame Integration," *Proceedings of SPIE, Document Recognition IV*, 1999, pp. 1-8.

Li, Huiping; David Doermann and Omid Kia, "Automatic Text Detection and Tracking in Digital Video," *IEEE Transactions on Image Processing*, 9(1) 2000, pp. 147-156.

Lienhart, Rainer and Axel Wernicke, "On the Segmentation of Text in Videos," *Proceedings of IEEE International Conference on Multimedia and Expo (ICME2000)*, 3 2000, pp. 1511-1514.

Lienhart, Rainer and Effelsberg Wolfgang, "Automatic Text Segmentation and Text Recognition for Video Indexing," *Technical Report TR-98-009, Praktische Informatik IV*, 1998.

Lin, Chuan-Jie, Hsin-Hsi Chen, Che-Chia Liu, Jin-He Tsai and Hong-Jia Wong, "Open-Domain Question Answering on Heterogeneous Data," *Proceedings of Workshop on Human Language Technology and Knowledge Management, ACL*, 2001, pp. 79-85.

Lu, Y., "Machine Printed Character Segmentation – An Overview," *Pattern Recognition*, 28, 1995, pp. 67-80.

Oka, R. I., "Handwritten Chinese-Japanese Characters Recognition by Using Cellular Feature." *Proceedings 6th International Joint Conference on Pattern Recognition*, 1982, pp. 783-785.

Sato, Toshio, Takeo Kanade, Ellen K. Hughes, Michael A. Smith and Shin' ichi Satoh, "Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Caption," *ACM Multimedia Systems*, 7(5) 1999, pp. 385-395.

Smith, Michael A. and Takeo Kande, "Video Skimming and Characterization Through the Combination of Image and Language Understanding Technique," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 775-781.

Voorhees, Ellen, "Overview of the TREC-9 Question Answering Track," *Proceedings of the Ninth Text Retrieval Conference (TREC-9),* 2000, pp. 71-80.

Wactlar, Howard, "Informedia - Search and Summarization in the Video Medium," *Proceedings of Imagina 2000 Conference*, 2000.

Wu, Victor and Edward M. Riseman, "TextFinder: An Automatic System to Detect and Recognize Text in Images," *IEEE Transactions on pattern analysis and machine intelligence*, 21(11) 1998, pp. 1224-1229.

Wu, Victor; Manmatha, R. and Riseman, Edward. M., "Finding Text in Images," *Proceedings of the 2nd intl. conf. on Digital Libraries*, 1997, pp. 1-10.