簡易影片字幕文字辨識法及其詢答應用

林川傑　劉哲嘉　陳信希

國立臺灣大學資訊工程學研究所

{cjlin, jjliu}@nlg2.csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

摘要

影片字幕通常反應影片部份內容，可以輔助影片內容檢索。雖然在新的影片格式如 MPEG2 以上，字幕內容可以輕易取得，但是仍有大量早期的影片，需要進行影片文字辨識，才能擷取字幕內容。本文提出一套簡易的中文字幕辨識法，包括影像擷取、字幕尋找、背景去除、字元切割、光學文字辨識、及後處理。我們以 Discovery Channel 影片作為訓練和測試的資料，以兩部影片作集外測試，其辨識率分別為 82.3%和 85.9%，而集內測試可以達 94.2%的正確率。在 Pentium-4 1.7G，256M RAM，40G 7200 轉速的 IBM 硬碟等配備下，處理平均 495MB 大小的影片，需要 29 分 11 秒。這套影片文字辨識法，對於影片數位圖書館的建立，以及後續的影片內容檢索有很大的助益。本文以影片檢索和詢答系統為例，說明影片文字辨識的應用。

1. 緒論

在多媒體的資訊時代，影片數量相當龐大，也隱藏豐富的知識，如何有效的檢索與擷取影片內容，就成為重要的考量要素之一。在著名的數位圖書館計畫 Infomedia (Wactlar, 2000)，影片資料庫的管理使用就是個顯明的例子。由於每一幕影片片段透過聲音和影像傳遞重要信息，因此以語音檢索或影像相似性比對，找出相關的影片片段，是最直接的方式。但是在相關的影像相似性比對技術，還未完全成熟之前，影片字幕仍是重要的檢索依據。在新的影片格式如 MPEG2 以

上，字幕內容可以輕易取得，但是仍有大量早期的影片，需要進行影片文字辨識，才能擷取字幕內容。本論文擬提出一套簡易的中文字幕辨識法，擷取影片中的文字，供後續的檢索使用。

光學文字辨識的研究歷史很早，且已經有很好的成果。紙本的文字資料透過掃描器輸入成影像檔，透過文字辨識系統的處理，將影像檔辨識成文字檔。另外，手寫文字辨識也有突破性的發展。相對的，影片文字辨識比傳統的文字辨識挑戰性高，主要的原因是後者所辨識的格式，大多是白底黑字，而影片上的文字大多出現在複雜的背景上，並且字通常不大，前者會遇到解析度較差及複雜背景的問題。

過去已有些論文與影片文字辨識相關，Wu 等人(1997, 1998)嘗試以 connected component 的方式尋找圖片中的文字，其方法在圖片中的結果不錯。但應用在影片中的文字尋找時，會因為影片中的文字大多有著複雜的背景，造成字會與其它圖形物件相連在一起，因而產生不好的結果。Lienhart 等人(1998, 2000)則利用 color segmentation、contrast segmentation、geometry analysis、texture analysis 等方法尋找影片中的文字。Li、Doermann 和 Kia (2000)採用類神經網路的方式，來找尋影片中的字串。Li 和 Doermann (1999) 也利用多張影像的整合，來加強文字影像的解析度。在影片的整合部份，Smith 和 Kande (1997)利用字幕、影像的移動及臉部辨識等方法，來簡化影片的大小。Sato 等人(1998)利用文字的修補及多張影像的文字擷取，來提昇影片文字辨識的正確率。

本文以影片字幕的文字辨識為主，研究的對象是中文字。論文共分十節，第二節將介紹影片文字辨識可能的問題，以及系統架構。第三節至第八節分別描述系統每個模組採用的策略和方法，並以 Discovery Channel 影片為訓練和測試材料，實驗各個模組的效能。第九節討論影片字幕文字辨識結果在詢答系統上的應用。第十節做總結，並探討未來的方向。

2. 影片文字辨識系統架構

影片中會出現的文字有兩種：一種是字幕文字(subtitles or captions)，一種是畫面文字(texts in image)。字幕文字常出現在畫面上特定的地方，例如字幕和標題常出現在影片的下方橫書，而中文的人名頭銜等文字通常出現右方或左方直書。畫面文字指的則是出現在字幕文字外畫面中的文字，例如商店的招牌，車子的車牌號碼等。它本身本來就是畫面的一部份，因此不僅會因鏡頭的移動而改變位置，而且在與字幕等重合時，會被字幕所遮蓋。由於字幕正是影片中旁白或是對話的文字呈現，字幕往往提供了影片內容的重要資訊，因此本文的重點是“字幕文字”的處理。

影片字幕文字辨識首先面對的問題就是如何擷取畫面，以及如何去除背景。一般影片中的文字，大多出現在很複雜的畫面上，因此系統第一步工作就是要去除複雜的背景，並將影像轉成白底黑字，再交給文字辨識軟體辨識。最後文字辨識後的結果,再運用自然語言處理技術來提昇正確率。圖一顯示整體系統架構圖。

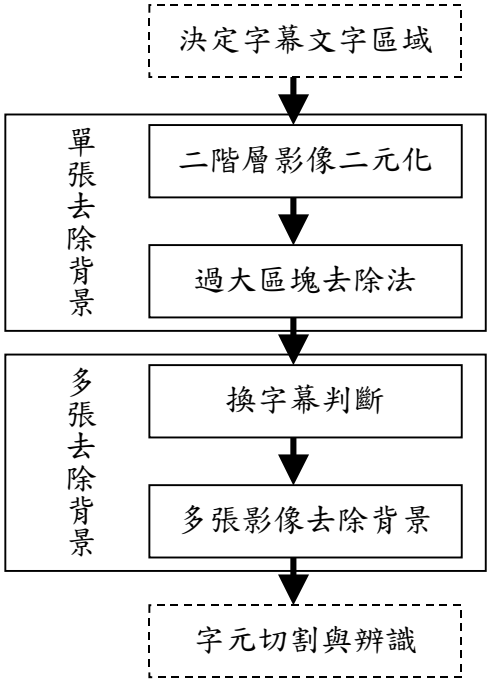
為了訓練及測試這個系統的效能，實驗的素材取自於 Discovery Channel 影片，內容包羅萬象，有自然生態、歷史文化、科技新知、軍事飛行、旅遊冒險、生活集錦等知識性題材。在第九節將結合詢答系統，對此影片集提出問題，透過字幕辨識結果，快速尋找相關的影片片段並做評估。

3. 字幕文字區域之決定

“字幕文字” 的特徵如下： (1) 呈垂直或水平排列；(2) 字的本身和影片會有強烈的對比色，一般會有明顯的邊框；(3) 一定是在影片前方，不會被影片畫面所遮蓋；(4) 會連續出現兩字以上；(5) 不會太大，一般不會高於影片高度的1/3；通常也不會太小，因為太小，人類也無法識別；(6) 固定的高度 (或寬度)與字體大小；(7) 固定的顏色。我們根據這些特徵來尋找字幕的位置。

3.1 影像二元化

在進一步處理影像文字之前，我們會先將影片轉成二元化影像 (Binary



圖一、影片文字辨識系統架構圖

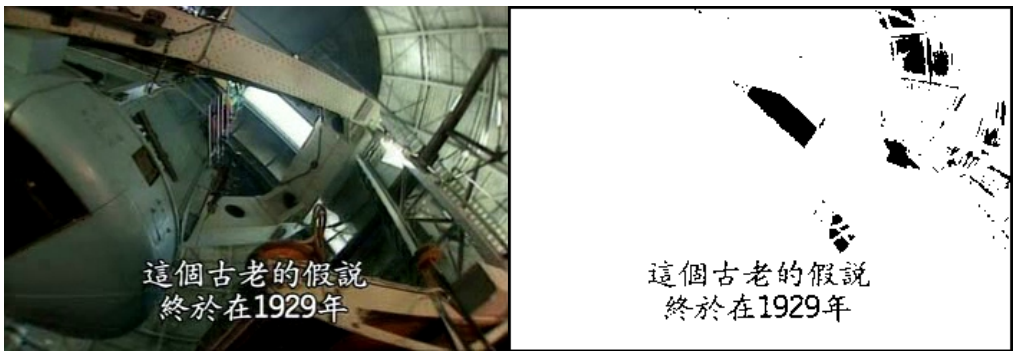
Image)。這個步驟是做影片文字處理過程中常用的方法，它可以幫助將複雜的背景單純化，並讓字幕文字更易於顯現出來。

我們在撥放影片的過程中擷取出影像畫面來，一秒取 2 幕，並將之存為 BMP 檔。在 BMP 檔中，圖片上每一點的色彩均是以其 RGB 值來記錄。所謂 RGB 值，就是該色彩由多少亮度的紅(Red)、綠(Green)、藍(Blue)色光所合成，記成 RGB(色值,色值,色值)，其中色值的範圍由 0~255，0 表最暗(無此色彩)，255 表最亮。

利用影像中各點的 RGB 值，我們就能將原影像轉換成二元化的影像。演算法如下：

設定二元化門檻值 SegColorScore
對影像中的每一點而言：
若 該點色彩 R 值、G 值、B 值均 > SegColorScore
則 更改該點色彩為黑色(RGB(0,0,0))；
否則 更改該點色彩為白色(RGB(255,255,255))。

其中 SegColorScore 值在實驗中設定為 190，這是多次實驗所得結果最好的經驗值。圖二為影像二元化結果的範例。我們可以很清楚地看到，字幕文字在二元化影像中更容易與背景分離開來。此次實驗中字幕文字在原影片中為白色字體、黑色邊框，轉換為二元化影像後成為黑色字體。



圖二、影像二元化範例

3.2 決定字幕文字區域

將影像二元化之後，接著要決定畫面中字幕文字的區域在哪裡。這裡我們利用了字幕文字的另一項特性：在橫書的字幕中，若在字幕上劃上一條水平線，則此線會通過不少的直豎筆劃。由於在文字書寫的習慣上，直豎筆劃的寬度大致一定。而且影片畫面的其他地方，也不容易出現此種多個連續相同寬度的黑色區塊，我們便可利用這個資訊，計算得知字幕所在位置。

考慮在同一水平高度位置 $height_i$ 的各點，將連續相鄰的黑色點視為同一區段(segment)，則可得在 $height_i$ 的水平位置上的黑點區段集合為 $SEGMENT_i$= ($segment_{i1}$, $segment_{i2}$,…)。考慮各 $segment_{ij}$ 中所含的黑點數，若相鄰區段所含黑點數相差不超過一定值 δ 時(本實驗中 δ 值設為 3)，則將這些區段視為同一組(group)，如此可得在 $height_i$ 的水平位置上的黑點組集合為 $GROUP_i$= ($group_{i1}$, $group_{i2}$,…)。令 $Seg(group_{ij})$為構成此 $group_{ij}$ 的區段個數，我們定義 $height_i$ 為字幕區域的可能分數 $SASA_i$ (Score as Subtitle Area)為：

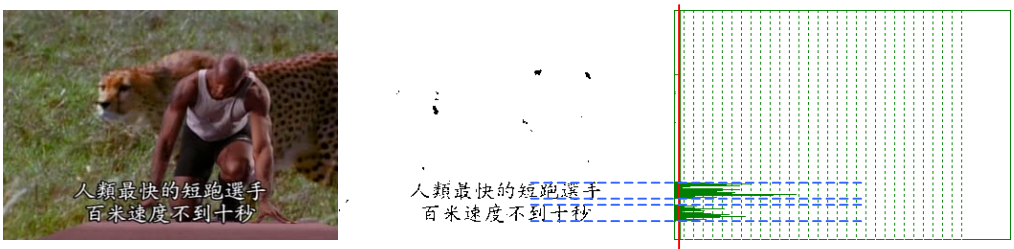
$$SASA_i = \sum_{j=1}^{|GROUP_i|} Seg\left(group_{ij}\right) \times \log_2 Seg\left(group_{ij}\right) \tag{1}$$

考慮如下的範例：point 列中以 0 表示白點，1 表示黑點。

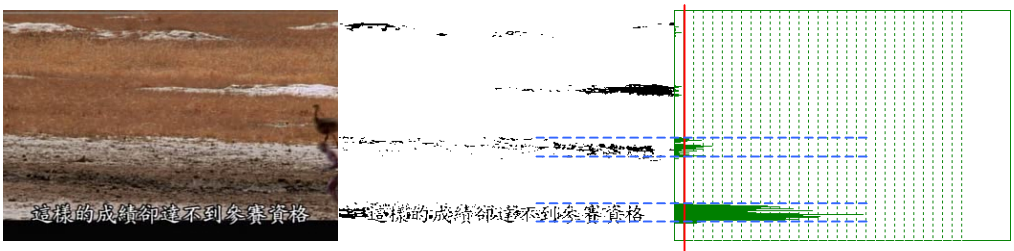
```
point:    00111011111001100011101111111011111111100111101110110111111
segment: --111-22222--33---444-5555555-666666666--7777-888-99-AAAAAA
group:    --1          -2              --3        -4
Seg(group):   =4             =2              =3         =1
```

這個例子的 $SASA$ 分數為 $4\log_2 4 + 2\log_2 2 + 3\log_2 3 + 1\log_2 1 = 14.75$。

若影像畫面總高度有 m 列，算出各列的 $SASA$ 值，求其平均值 $\overline{SASA}$。我們



圖三、決定字幕文字區域結果範例一



圖四、決定字幕文字區域結果範例二

將各列中 *SASA* 值高於平均值者視為字幕區域，如此便可決定出字幕出現在畫面中的位置了。圖三和圖四為決定字幕文字區域的範例，其中左圖為原影像圖片，中圖為其二元化圖片，右圖則為各列之 *SASA* 值。右圖中縱軸表畫面高度，橫軸即各列 *SASA* 值之大小，垂直紅線為平均值 $\overline{SASA}$ 的所在，而水平的藍色虛線即為判斷所得的字幕文字區域。

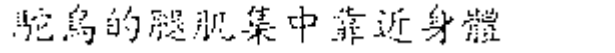
3.3 實驗評估

字幕文字區域的實驗資料來自三部 Discovery Channel 的影片：「閃電」、「動物之最」和「鯨魚探奇」，其中各出現 69、66 和 71 行字幕。實驗結果如表一所示，正確率各為 76.7%、39.8%和 82.0%，召回率幾乎可以達到 100.0%。如同圖四所示，在畫面中央的碎白石子路，因為出現多個連續相同寬度的色塊，因此被誤判為字幕文字區域。對於不正確的字幕文字區域，尚可在 OCR 處理的過程中，因為與標準字庫過低的相似度而被過濾。因此這裡高召回率就比正確率要來的重要，如此才能確保擷取出所有的字幕。

表一、字幕尋找結果評估

字幕個數	實際	系統判斷	正確	正確率	召回率
閃電	69	90	69	76.7%	100.0%
動物之最	66	161	64	39.8%	97.0%
鯨魚探奇	41	50	41	82.0%	100.0%

圖五、SegColorScore 設為 140 的影像二元化結果



圖六、SegColorScore 設為 180 的影像二元化結果

4. 單張影像去除背景法

調整 SegColorScore 的值將影像二元化時，我們發現一個有趣的現象。若是將 SegColorScore 值調得太低，會有字幕文字和殘留背景交雜重疊的情形出現；但若是 SegColorScore 值調得太高，得到的字幕文字又會過於模糊破碎，不利於 OCR 的進行。在第 3 節所用的值為 190，就只能留下較模糊的字幕文字影像。

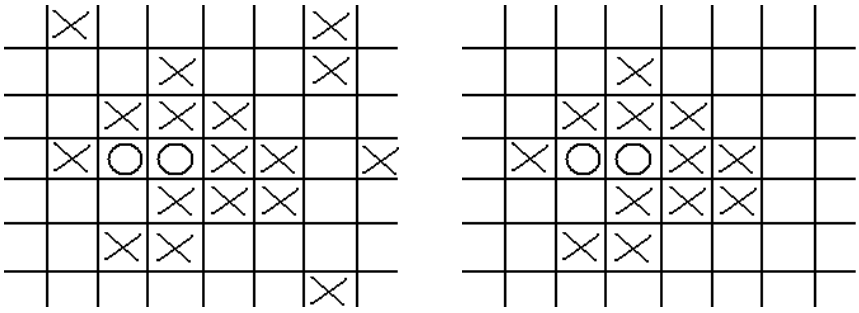
為了讓 OCR 結果能更準確，就需試著留下清晰的文字，並做去除背景的動作。本節先介紹在單張圖片中，利用一些文字與背景畫面不同的特性，將背景去除的方法。下一小節再討論利用多張同字幕文字的畫面去除背景的方法。

4.1 二階層影像二元化

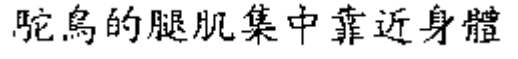
在第 3 節中，我們設定了 SegColorScore 值。色彩 RGB 三值均高於 SegColorScore 的點會被轉為黑色，而其中一值低於 SegColorScore 者則被轉為白色。然而 SegColorScore 的值大小，影響著二元化後的字幕清晰度。例如圖五和圖六分別為 SegColorScore 設為 140 和 180 的結果，可見 SegColorScore 設為 140 時，字幕文字清晰，但是有過多背景殘存下來。而 SegColorScore 設為 180 時，背景雖多已去除，但所留下來的字幕文字較為模糊。

這裡我們提出了一種新的方法，稱為二階層的影像二元化方法。利用二個不同的 SegColorScore 值所得的二元化影像，我們可以把字幕文字清晰的留下，且背景去除。方法如下：

將同一張圖片利用二個高低不同的 SegColorScore 值(分別為 HiSegColorScore 和 LowSegColorScore)，所得的二元化影像重疊在一起，參考圖七左圖。其中○為 HiSegColorScore 值所得到的黑點處，而×則表示



圖七、二階層影像二元化示意圖



圖八、由圖五及圖六二階層影像二元化的結果

LowSegColorScore 值所得的黑點處，值得注意的是○也會是 LowSegColorScore 值所得的黑點處。接著我們只保留與○相連接的×黑點區塊，其餘未與任何○相連的×區塊均改為白色點，結果如圖七右圖。圖八即為圖五和圖六利用二階層影像二元化所得的結果，為更清晰的字幕文字圖。

4.2 過大區塊去除法

當字幕文字區域出現高亮度的背景時，第 4.1 小節的做法便不足以去除之。如果此背景為一大片的高亮度區塊(二元化後則大片黑色區塊)，則我們提出一個方法在單張圖片中清除此一過大區塊。對於零碎的小區塊，下一節會試著利用多張同字幕的圖片資訊來清除背景。

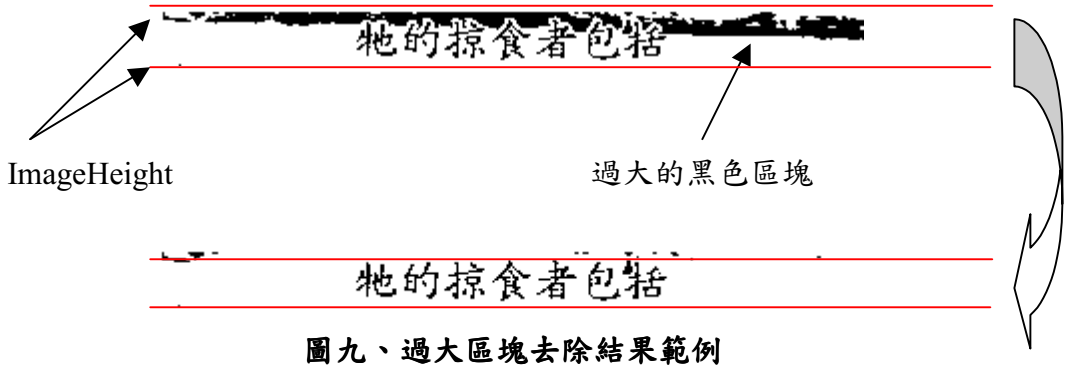
圖九上圖為字幕文字區域中有過大黑色區塊的範例，下面為提出的演算法:

Range = (字幕文字區域高度) ÷ 4;
Total = Range×Range×0.9;
對字幕文字區域中每一黑點**均做如下檢查**：
　檢視以該黑點為左上角、邊長為 Range 的正方形區域，
　若 正方形區域中黑點數 >= Total (即此區域中有九成以上均為黑點)
　　則 把該黑點及與其相連接的黑色區塊都改成白色
結束

圖九下圖為其去除過大區塊後的結果影像。

5. 多張影像去除背景法

雖然利用單張去除背景的方法已可很有效地去除大部份的背景，但小區塊高



圖九、過大區塊去除結果範例

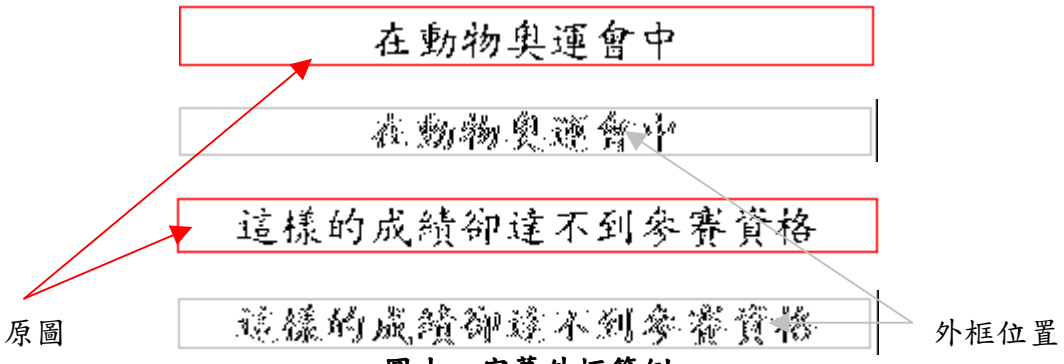
亮度的背景卻不易被清除。如圖九下圖在字幕文字周圍仍有一些背景圖案存在。

字幕文字的另一項特性是：它不會隨著畫面鏡頭的移動而改變位置，然而背景圖案卻會隨之更動。利用這項特性，把同一字幕文字的二元化影像重疊在一起，留下出現頻率高的黑點，即為字幕文字部份。Sato 等人(1998)就是利用這樣的想法，去除移動中的背景。

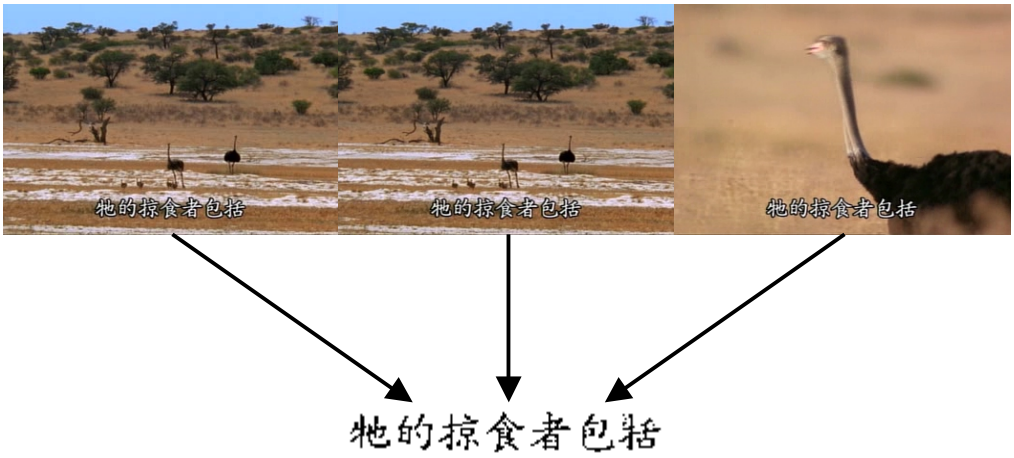
要用多張影像去除背景，就得先判斷那幾張影像為同一字幕。此後兩小節分別介紹判斷更換字幕，以及多張影像去除背景的方法。

5.1 換字幕判斷

在換字幕的判斷上，我們保留了字形的外框，藉以偵測字幕文字是否已經更換。以圖十為例，我們先將每一張圖中所有黑色區塊的外框位置記錄下來。當讀入下一張圖後，將其外框位置與前一張的外框位置做比較。若位置不同的比率超過某一門檻值 SceneChangeScore 時，就將其判斷為字幕轉換點。實驗得 SceneChangeScore=0.6 時，有最佳結果。



圖十、字幕外框範例



圖十一、多張去背結果範例

表二、字幕轉換點判斷結果

	換字幕次數	判斷錯誤次數	正確率
閃電	69	0	100.0%
動物之最	66	3	95.5.0%
鯨魚探奇	41	0	100.0%

我們也做了一個小小的評估。同樣以第 3 節實驗所用的三部影片，來看看各行字幕的轉換點是否判斷正確。由表二結果可知判斷正確的成功率相當高。

5.2 多張背景去除法

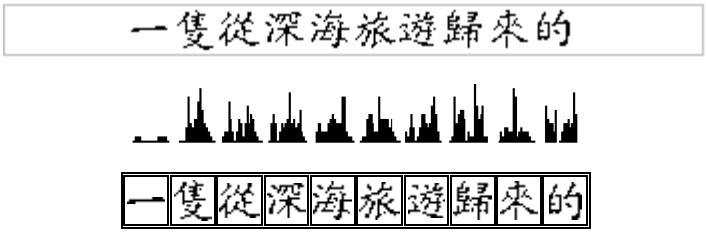
對於同一字幕文字的連續圖片，我們利用多張影像去除背景的方法來得到一個字幕文字區域大小、內容僅含字幕文字的圖片。令 NumFrames 為此連續圖片的張數，考慮字幕文字區域中的任一點位置。若在連續圖片中該點位置有九成(NumFrames×0.9)以上的圖片出現黑點，則在結果字幕文字圖片上該點位置亦設為黑色，否則設定為白色。

圖十一為多張影像去除背景的結果範例，可以看到在背景部份比圖十清除地更乾淨。

6. 字元切割

經由前幾節對影像處理的步驟，現在每一個字幕文字都已有對應的白底黑字結果字幕文字圖片，接下來就可以用傳統 OCR 的方法辨識出字幕中的文字。

OCR 的第一步是決定每個字元的邊界。由於我們先前決定字幕文字區域



圖十二、字元切割結果範例

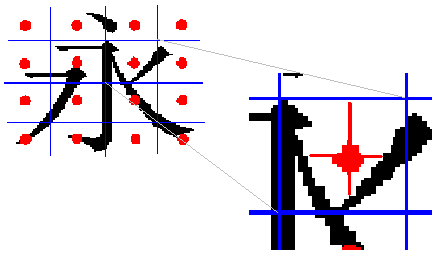
時，等於已經決定了各字元的上下邊界，因此現在所要判斷的，是各字元的左右邊界。

字元切割的方法，大多都用垂直投射的方法(Lu, 1995)：如圖十二所示，對於結果字幕文字圖片的每一個水平位置，將不同高度上的所有黑色點投影到水平線上。在中文字與字之間因為有空間，因此會出現投影量為 0 的間隔 (gap)。由於中文字元大多在正方形區域內，真正是字元間隔的兩間隔之間寬度(即字的寬度)也會大約等於結果字幕文字圖片的高度 ImageHeight (即字的高度)。我們的方法如下：若兩間隔間的距離為 ImageHeight×0.7~ ImageHeight×1.4 之間，則此兩間隔切出一個字元。圖十二的最下圖即是字元切割的結果範例。

7. 文字辨識

OCR 的研究主要分 on-line 及 off-line 兩種。而早在 1970 年代，就有許多的研究針對 on-line 手寫或是簡單的印刷字體辨識，到了 1980 年代 off-line 的研究才慢慢變多，而 off-line 的辨識系統又主要分統計式模型及結構分析兩種。論文中所採用的是統計式模型，為 Oka 在 1982 年所提出的方法。

以圖十三為例，首先將讀入的影像檔等分為 16 區塊，由每一區塊中點開始，觀察其上下左右四個方向。如果在這區塊中該方向上有黑點存在，則記錄特徵值為 1，否則為 0。如此一來共可以得到 64 個(16 區塊×4 個方向)特徵值。



圖十三、記錄影像特徵值(Oka, 1988)

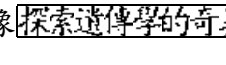
圖十四、OCR 辨識後所得前十名的候選字集

我們蒐集了一些文字圖片做為標準字庫以做比對之用，將這些圖片也都以特徵值的方式記錄下來。當一個新的文字影像要做辨識時，首先找出其特徵值，再與標準字庫中的特徵值做比對。我們以計算相同特徵值個數來做為計分依據，當對應的特徵值相同時，相似度加一分，因此相似度的分數會界於 0 到 64 之間，分數越高代表兩個影像越相似。

下面舉一個簡單的例子：設有一文字影像為  ，分別和標準字庫中的「傳」與「博」字做比對，其特徵值分別如下：

 1010100010001101110100100000101111101010010001011111111111001111
傳: 1010100000001101010101100000100111101010010001011111111111001111
博: 1110100000011111010001000100100111101100010001011111111111101101

比對結果與標準字庫的「傳」字相似度為 60 分，與「博」字相似度為 50 分，因此「傳」成為此文字影像的第一名候選字。

圖十四是影像  經過 OCR 辨識後所得前十名的候選字集。可以看到第一名的候選字即為正確答案的比例非常高，而且正確答案也都出現在前十名的候選字集中。下一節我們會利用 OCR 後處理的方法，選出不在第一名的正確答案。

僅選取第一名做為辨識結果的實驗評估數據記錄在表三中，其中影片「萬象雜誌：基因的秘密」為集內測試，而「金字塔之王」、「埃及豔后」影片則為集外測試。實驗數據顯示，以 OCR 辨識第一名做為答案的正確率，集內測試可達 91.5%，集外測試也可達到 78.5%和 81.5%，已有不錯的成績。

表三、OCR 辨識實驗結果

影片	TOTAL	CORRECT	ERROR	MISS
基因的秘密	809	739(91.5%)	69 (8.5%)	0
金字塔之王	684	537(78.5%)	110(16.1%)	37(5.4%)
埃及豔后	750	611(81.5%)	86(11.5%)	53(7.1%)

8. OCR 後處理

在前面的小節中我們知道 OCR 第一名的集內測試正確率約為 90%，而前十名的正確率約為 95%。所以我們的目標就是希望我們能將 OCR 出來的字正確率能逼近 95%，要讓第二名以後的正確答案能被選中，而原本就辨識第一名即為正確答案的則保持不變。所以如何提高辨識率，是本節主要要介紹的課題。

8.1 基本後處理方法

每個文字影像辨識出來的結果都取其前十名做候選字，並有相對應的相似度分數，圖十四中標示為 **(分數)候選字**。首先我們將分數與第一名差在 4 分以上(包含 4 分)的候選字剔除，以減少後處理比對時所帶來的雜訊(在圖十四中以灰色表示被剔除的候選字)。接下來從第一個字開始，每次連續看三個文字影像(令其為 ABC)，查詢其候選字組 A_iB_j 或 B_jC_k 是否在字典中為二字詞或是多字詞的一部份。若是，則分數為兩候選字相似度相乘，否則分數為零。比較所有 A_iB_j 和 B_jC_k 所得分數的高低，若最高分的候選字組為 A_iB_j，就選定 A_iB_j 為 AB 的辨識結果，然後由第三個文字影像 C 開始，重覆前面的步驟(看 CDE)。若是最高分的候選字組為 B_jC_k，則選擇第一個文字影像 A 的第一名候選字 A_1 為其辨識結果，然後由第二個文字影像 B 開始，重覆前面的步驟(看 BCD)。

8.2 後處理實驗策略

為了了解在做後處理時，是否有必要考慮所有相似字的任意組合，或是第一名的候選字有其重要性，甚至候選字組是否出現在字典中的資訊有何幫助，我們提出了三種不同的後處理策略，並且和僅取第一名候選字、以及僅以長詞優先法則所得的結果做比較。

[策略一]在選取文字影像的候選字時，考慮所有的候選字組合。

[策略二]在兩兩一組查詢字典的時候，其中一個候選字一定是第一名的候選字。例如在辨識文字影像 AB 時，只查詢 A_1B_1、A_1B_2、…、A_2B_1、A_3B_1、…。這樣的做法是為了增加對第一名候選字的信任。

[策略三]連續看四個文字影像，若其所有的候選字組合中出現一組在字典中為四字詞，則以這四字詞為其辨識結果。否無則再連續看三個文字影像，是否有三字詞的候選字組合。如果有則選為辨識結果，無則接著以策略二的方式選擇結果。

8.3 實驗評估

標準字庫的蒐集來自六部 Discovery Channel 的影片(「動物之最」、「蛇類奇觀」、「萬象雜誌：基因的秘密」、「天然景觀—落磯山脈」、「神戶大地震」、「達爾文之島」)，總共有 7,818 個文字影像檔，得到 2,256 不同字的特徵值。

表四到表六分別是針對三部影片做字幕文字辨識所得的實驗結果，其中「萬象雜誌：基因的秘密」影片為集內測試，而「金字塔之王」、「埃及豔后」影片則為集外測試。表格中各欄位資訊解釋為：

TOTAL: 影片中字幕總字數
CORRECT: 辨識正確的字數
ERROR: 文字出現在標準字庫中但辨識錯誤的字數
MISS: 文字未收錄在標準字庫中的字數
Improve: OCR 後處理所得的改進值
最佳優先: 選取第一名候選字為辨識結果
長詞優先: 選取候選字組中出現在字典最長詞者為辨識結果

表四、OCR 後處理結果(影片「萬象雜誌：基因的秘密」)

	TOTAL	CORRECT	ERROR	MISS	Improve
最佳優先	809	739(91.5%)	69(8.5%)	0	------
長詞優先	809	753(93.1%)	56(6.9%)	0	1.6%
策略一	809	751(92.8%)	58(7.2%)	0	1.3%
策略二	809	759(93.8%)	50(6.2%)	0	2.3%
策略三	809	762(94.2%)	47(5.8%)	0	2.7%

表五、OCR 後處理結果(影片「金字塔之王」)

	TOTAL	CORRECT	ERROR	MISS	Improve
最佳優先	684	537(78.5%)	110(16.1%)	37(5.4%)	-------
長詞優先	684	544(79.5%)	103(15.1%)	37(5.4%)	1.0%
策略一	684	546(79.8%)	101(14.8%)	37(5.4%)	1.3%
策略二	684	559(81.7%)	88(12.9%)	37(5.4%)	3.2%
策略三	684	563(82.3%)	84(12.3%)	37(5.4%)	3.8%

表六、OCR 後處理結果(影片「埃及豔后」)

	TOTAL	CORRECT	ERROR	MISS	Improve
最佳優先	750	611(81.5%)	86(11.5%)	53(7.1%)	------
長詞優先	750	614(81.9%)	83(11.1%)	53(7.1%)	0.4%
策略一	750	635(84.5%)	62(8.3%)	53(7.1%)	3.0%
策略二	750	640(85.3%)	57(7.6%)	53(7.1%)	3.8%
策略三	750	644(85.9%)	53(7.1%)	53(7.1%)	4.4%

由表四到表六可以發現，策略三是所有方法中效果最好的。它在集外測試可有 82.3%和 85.9%的正確率，在集內測試更可達到 94.2%。

另外因為標準字庫僅蒐集 2,265 個字，在集外測試中就分別有 5.4%和 7.1%的字幕文字無法做比對。擴充標準字庫絕對是未來的重要工作之一。

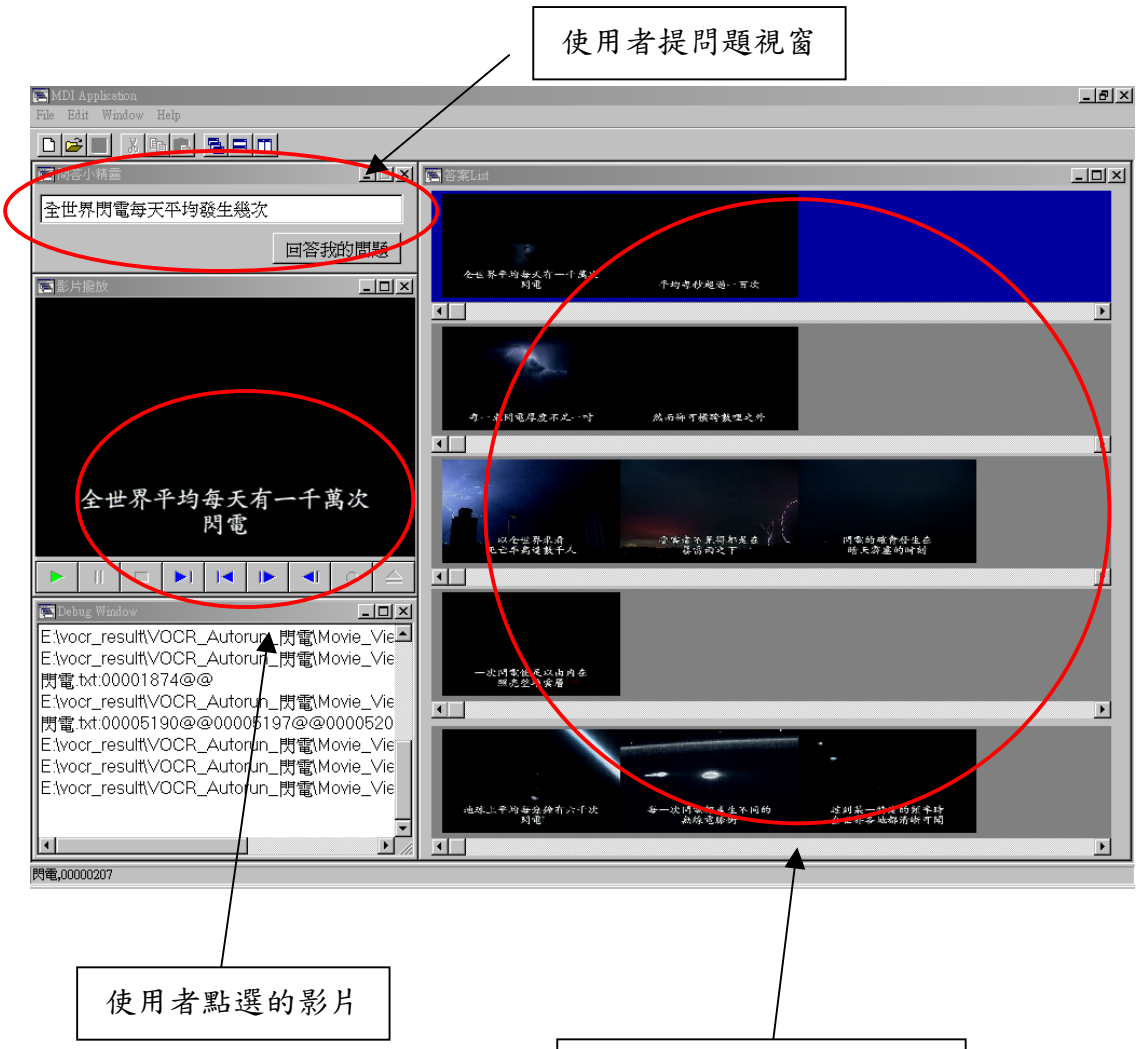
將 OCR 後處理策略三的實驗結果與未做 OCR 後處理(最佳優先)的實驗結果做分析，比較結果記錄在表七之中。其中 “True to False” 表示原本依最佳優先判斷正確、策略三卻判斷錯誤的情形，而 “False to True” 表示原本依最佳優先判斷錯誤、策略三可判斷正確的情形。由表七可見 OCR 後處理僅會造成 0.7%左右多餘的錯誤，卻能多判斷正確 3.0%到 5.2%的文字，由此可知後處理的幫助。

表七、策略三與最佳優先之結果比較分析

	Total	Result	True to True	True to False	False to True	False to False
基因的秘密	809	94.2%	738(91.2%)	6(0.7%)	24(3.0%)	41 (5.1%)
金字塔之王	647	87.0%	532(82.2%)	5(0.8%)	31(4.8%)	79(11.2%)
埃及豔后	697	92.4%	608(87.2%)	3(0.4%)	36(5.2%)	50 (7.2%)

9. 影片檢索與詢答系統

藉由先前各節所描述的影片 OCR 方法，我們已經能將影片中的字幕擷取出來。接著就可結合資訊檢索技術，或是問題詢答技術，發展查詢影片的檢索與詢答系統。



圖十五、影片詢答系統界面

9.1 影片詢答系統

圖十五是本實驗室所發展的影片詢答系統的界面。使用者可以輸入感興趣的問題，系統所找到的答案會陳列在右方視窗中，每一名答案均列出它的代表字幕畫面讓使用者參考。若使用者想要觀看其中一個答案的原始影片，可點選其畫面，系統就會自動將影片撥放出來。

詢答系統部份的技術，採用的是 Lin 等人(2001)對各類異質的資料進行詢答的技術中，針對 Video OCR 所提的方法。其中特別需要注意的是，因為搜尋的文本是文字辨識以後的結果，而由表四和表五得知，文字辨識的效能是 82.3%以上，尚未能完全正確。因此傳統詢答技術所用到的字串比對(不論是關鍵詞、同義詞，甚至是語意關係樹等)，都必須考慮到 OCR 產生的錯誤。因此含答文句的

計分方式，就要引入 OCR 相似度的分數：

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

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

其中 qw_i 和 pw_j 是做字串比對的兩個詞，$|qw_i|$ 表示詞 qw_i 中的字元個數，qc_k 是詞 qw_i 的第 k 個字元 (pw_j 的表示法同 qw_i)。$Ocr(qc_k, pc_k)$ 是 qc_k 和 pc_k 的 OCR 相似度分數，為第 7 節特徵值比對所得分數除以 64，以使值的範圍落在 0~1 之間。$weight(qw_i)$則為原先 qw_i 和 pw_j 相同時所得的分數。

9.2 實驗評估

9.2.1 問題的來源

問題的來源是 Discovery 繁體中文網站(http://chinese.discovery.com)，其“教育工程”內所擺放的影片與相關問題。這個網站提供一個免費而龐大的影像記錄片庫給教師使用，每個節目鎖定一個主題，讓老師在課堂上透過影像與特別編製的教師手冊、特別設計的活動、以及相關的網路資源，輔助學生在課程內或課程外的學習。

經由該網站上的資料，挑選了幾個影片的相關問題，來對我們的詢答系統做評估。至於為何要用該網站上的問題，主要原因是因為其問題較具公平性、一般性。其中，影片的片名有「大象」、「木星」、「哈伯望遠鏡：太空的奧秘」、「蛇之眼」、「鯨」及「地球科學面面觀：閃電」等。

9.2.2 詢答系統準確率

我們在此使用 MRR(Mean Reciprocal Rank)評估詢答系統的準確率，這是在詢答系統評比(TREC QA-Track)中所用的評量方法(Voorhees, 2000)。

在六部影片中共有 43 個問題，其結果如表八所示。MRR 分數為 0.1848 (0.1848=(4+5/2+3/3+1/4+1/5)/43)，答題率為 32.6% (14/43)。

表八、影片詢答系統評估結果

第一名	第二名	第三名	第四名	第五名	沒答出來
4	5	3	1	1	29

為何此處 MRR 只有 0.1848，觀察問題後，主要為下列幾點原因：

(1) 與問題有關的關鍵字文字未收錄在標準字庫中。

例如問題「冰雹如何形成？」中的「雹」字。

(2) 問題的用詞與影片中的用詞不一樣。

例如問題「木星繞行太陽一週需時多久？」，在影片中出現的是「木星環繞太陽一周，須地球時間十二年。」

(3) 需要更精準的問句規則處理。

以問題「閃電可以到達多熱的程度？」為例，目前系統在處理以"多"字來詢問程度的問題時，僅試著找尋屬於數量的答案，而不是更精確地找尋溫度描述詞"華氏五萬度"來做為答案。

(4) 需引入常識或理解文字。

以問題「歷史上第一位做閃電實驗的人是誰？」為例。影片中有提到 1752 年，富蘭克林做了閃電實驗，但是並沒有提到"第一位"這樣的字眼，系統並無法得知他就是第一人。

扣除掉第一點影片文字辨識系統的錯誤，目前的詢答系統大多只做到關鍵字比對與依問題類型尋找專有名詞答案的程度。而本文中所用的問題大多需要很多其它的相關知識或是語意上的分析，才能夠回答的出來。而問題的類型也偏重於 Why 及 How，這類型的問題本來就比較難解。因此，未來這部分的研究是個很大的挑戰。

10. 結論

本文介紹 Video OCR 的所有步驟，從影片的畫面擷取、尋找畫面中字幕文字區域的位置、去除背景、字元的切割、OCR 及透過自然語言的技術提高 OCR 的辨識率，並介紹了 Video OCR 的一些應用。Video OCR 的辨識率在集內測試，約有九成以上的正確率，而集外測試也有八成以上的正確率。

從得到的結果回去看錯誤的地方大概分下列幾種：一、收集的字元不夠多；二、背景去的不夠乾淨；三、字元切割錯誤；四、OCR 後處理錯誤。以下我們一點一點來討論。

第一個問題是收集的字元不夠多。目前我們所收集的字元共有 7,818 個文字影像檔，其中共有 2,256 個不同的字，而一般的常用字共有 5,401 個，所以很多字辨識不出來。

第二個問題是背景去除地不夠乾淨。在大部分的情況下，背景都可以經由單張影像去除背景的方法，加上多張連續同字幕畫面去除背景的方法來清除。但是遇到不會移動的背景，加上字幕後的背景又是破碎的黑白色素混雜時，往往就無法順利清除乾淨。而未去除的背景會影響後面的字元切割及 OCR 辨識。

第三個問題是字元切割錯誤，這常常是因為背景沒去除乾淨所造成。

第四個問題是 OCR 後處理錯誤。這個問題主要也是來自第一和第二個問題，因為原本所收集的字元中就沒有出現，也就不會出現在候選字中，所以常會將原本辨識為第一名正確的字，因為前後字的關係反而被辨識錯了。

此外，本次實驗資料均取材自 Discovery Channel 的節目，字幕的字型、顏色或是大小都比較一致。未來在處理其他來源的影片文字時，就必須再更進一步地探討不同字型、不同格式所帶來的影響。

在未來的工作中，也將試著以現有字型(例如標楷體等)建立標準字庫，以更完整的字元集來作實驗。找出更好的去除背景方法以及套用更好的 OCR Model，並且和以現有的語言模型做後處理結果來比較，以期能更進一步利用這樣的工具去挖掘出影片所帶有的資訊。

參考文獻

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

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

Li, Huiping; Doermann, David and Kia, Omid (2000). “Automatic Text Detection and

Tracking in Digital Video." *IEEE Transactions on Image Processing,* Vol. 9, No. 1 , pp. 147-156.

Lienhart, Rainer and Wernicke, Axel (2000). "On the Segmentation of Text in Videos." *IEEE Int. Conference on Multimedia and Expo (ICME2000),* Vol. 3, pp. 1511-1514, also as *Technical Report MRL-VIG00005.*

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

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

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

Oka, R. I. (1982). "Handwritten Chinese-Japanese Characters Recognition by Using Cellular Feature." *Proc. 6th Int. Joint Conf. on Pattern Recognition*, pp. 783-785.

Sato, Toshio; Kanage, Takeo; Ellen K.Hughes; Smith, Michael A. and Satoh, Shin'ichi (1998). "Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Caption." *ACM Multimedia Systems Special Issue on Video Libraries.*

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

Voorhees, (2000) "QA Track Overview (TREC) 9." [on-line]
Available: http://trec.nist.gov/presentations/TREC9/qa/index.htm

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

Wu, Victor and Riseman, Edward M. (1998). "TextFinder: An Automatic System to Detect and Recognize Text in Images." *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 21, No. 11.

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