行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※　　　　　　資訊隱藏於文字檔之研究　　　　　　　※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：■個別型計畫　　□整合型計畫
計畫編號：NSC　90－2213－E－009－163－
執行期間：90 年 08 月 01 日至 91 年 07 月 31 日

計畫主持人：陳玲慧 教授
共同主持人：
計畫參與人員：

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

執行單位：國立交通大學資訊科學學系

中　華　民　國　91 年　8　月　　　日

行政院國家科學委員會專題研究計畫成果報告

資訊隱藏於文字檔之研究

A Study on Data Hiding in Audio Signals

計畫編號：NSC 90-2213-E-009-163
執行期限：90 年 8 月 1 日至 91 年 7 月 31 日
主持人：陳玲慧　國立交通大學　資訊科學系

一、中文摘要

同一份文件可以利用幾近不可覺的方法如重新定位或修改每一行或每一個字出現的位置而得到許多不同的版本。每一個版本實質內容相同，但卻可以在文件中隱藏著視覺不可察之擁有者相關資訊。一旦後來出現未經許可的傳播，就可以藉此去查出這份文件的原始擁有者，以達到版權保護的目的。

然而將資訊隱藏在一份文件檔中是很困難的。這乃因為文件檔不像影像或是聲音有大量的多餘資訊可供隱藏資料。我們可對一張影像做些微的更改而不會被人查覺，然而假如我們對一份文件檔做修改，即使只是一個字元或是一個句號，都很容易被發覺有異。因此，對文件做資訊隱藏乃是要發現那些可供修改而不被察覺的地方。

目前已有一些方法被提出，然而這些方法在資料擷取時需要有原始文件。因此此類方法實用性很低。在本計劃中，我們將提出一不需原文件的資料隱藏法。此方法主要是先將文件檔轉成 postscript 檔，再將機密資訊透過行距或字距調整達到隱藏資料目的。在擷取資料時，可由 postscript 檔直接抽取，亦可列印出來，看成一張影像，利用影像處理之技巧擷取機密資訊。實驗結果顯示此方法相當有效。

關鍵詞：資訊隱藏、文件檔、版權保護

Abstract

Each copy of a text document can be made to be different in a nearly invisible way by repositioning or modifying the appearance of text such as lines or words. Through this way, the information of the original owner can be embedded in a text document. Thus, if the text document is illicitly disseminated, the original owner information can be extracted to prove the illegal use. This will reach the goal of copyright protection.

Some methods have been provided to hide data in text files. However, most of these methods need the original file to do data extraction, this is impractical. In this proposal, we will provide a method to hide data in text document. And the hidden data can be extracted directly without referring to the original file. The method first converts the text document to a postscript file, then the secret data is embedded in the file by adjusting the spaces of lines or words. The hidden data can be extracted directly from the postscript file or a printed copy. Experimental results are also given to show the effectiveness of the proposed method.

Keywords: data hiding、postscript file copyright protection

二、緣由與目的

由於文字檔在現今資訊社會流通非常的多且普遍，因此我們選擇將我們的資訊隱藏在文字檔當中，同一份文件可以利用

幾近不可覺的方法如重新定位或修改每一行或每一個字出現的位置而得到許多不同的版本。每一個版本實質內容相同，但卻可以在文件中隱藏著視覺不可察之擁有者相關資訊。一旦後來出現未經許可的傳播，就可以藉此去查出這份文件的原始擁有者，以達到版權保護的目的。而在所有文字檔中，postscript file 流通非常普遍，且它允許任意地精確的對 text 作控制，是現在普遍使用的 page description language，故我們選擇其來隱藏我們的資訊。

現今已經有一些方法[1~10]被提出去抵抗不合法的文件拷貝與散播，方法便是將擁有者的資訊隱藏在每一份文件中，因此每一份文件的接收者所擁有的文件中包含一連串唯一的標記來定義擁有者或一些秘密資訊在其中，而每一個標記表示一個 0 或 1 的 bit 並相對於一個文字不可覺的竄改，而這個文字可能是一行 textline 也可能是一個 word，接下來我們將敘述其嵌入程序與萃取程序。

1. Line-Shift Coding[4,8]:

其作法便是垂直的移動 textline 的位置來藏入資訊，以三行 textlines 為單位，上下兩行不動，修改中間那行的位置使其稍微上移或下移來藏入 0 或 1，而遺留不移動的兩行乃用於萃取的步驟。由於在一份文件中，其 textline 與 textline 之間的 spacing 是相同的，因此我們在萃取隱藏資訊的過程中是不需要原始文件的資訊的。

2. Word-Shift Coding[4,8]:

此方法乃是水平的移動 word 的位置來藏入資料，其方式與 line-shift coding 類似，亦是以三個 word 為一組，左右兩個 word 保持不動，修改中間那個 word 的位置使其稍微左移與右移來藏入 0 或 1，遺留不移動的兩個 word 乃用於萃取的步驟。除此之外，為了達到每一行邊界都對齊，因此每一行的第一個 word 和最後一個 word 必須保持不動以達此要求。另外，在一份文件中因為 word 與 word 之間的 spacing 是不相同的，因此在作萃取步驟時是需要原始文件的資訊，藉由原始文件與已進行嵌入的文件作比較，來判斷 word 是往左移或右移以取出藏入的資訊。

3. Character-Shift Coding[4]:

此方法僅可應用在 Bitmap 的文件影像且其只能對 Bitmap 進行萃取資訊的動作，character-shift coding 乃是利用修改單一字元的高度或是其相對於其他字元的位置來嵌入資訊，當然我們亦要留下一些沒有作過修改的字元以便於作萃取資訊之用。對讀者而言，一些字元微幅的修改是幾乎不可覺的，除非是兩個相同的字元正好相鄰，其中一個有被修改過一個沒有，這種情形才有可能被讀者注意到可能有問題。另外，此方法在進行萃取資訊的程序時是需要原始影像的。

4. Open Space Methods[9]:

此方法乃是去竄改在已列印頁中那些未使用的空白 spaces，而那些 spaces 包含有段落間的空白，每一行結尾的空白，以及兩個 word 間的空白。但是其有個嚴重的缺點，那就是在列印後所隱藏的資訊可能會遺失，故此方法適用在檔案內容是 ASCII 格式的情形下。

5. Syntactic Methods[9]:

一些 Syntacic 方法是利用標點來藏入資訊，因為標點是模擬兩可的，甚至是可有可無的，所以其對文章的意義上有較少的影響，因此我們可以把有標點的當作是藏 1，沒有的當作是藏 0。雖然標點的使用是模擬兩可，但是不一致的使用標點還是可能會引起讀者的注意，除此之外，我們把原來有標點的地方改成沒有的話，亦有可能會使文章的意義有所改變，因此使用此方法要特別小心。另有一些方法是去改變措辭及文字的結構而沒有去改變文字的意義，像倒裝句，但由於此方法比用標點來隱藏資訊的方法更明顯，故其應用的機會就更受到限制了。

6. Semantic Methods[9]:

這類方法乃是去修改每一個字本身，因為有些字都會有其同義字存在，我們就去定義其中一個字代表藏 0，另一個同義字

代表藏 1, 如此便可藏入我們所隱藏的資訊了，萃取程序亦同，但此方法問題在於同義字仍有意義上的些微差異，這可能會在我們做嵌入程序時對文章意義有所改變。

上述所提及的乃是一般對一份 text file 進行嵌入與萃取的方法，由於文件亦可直接由印表機印出，再由掃描器掃描成 BMP 檔, 經由偵測此 BMP 檔中 line 與 word 細微的移動來取出之前藏入的資訊，接下來便是提到這一方面的方法[1,4,10]。

1. Feature Detection[4]:

首先，以圖一為例，為了在此圖檔中定義出 line 與 word 的位置，因此先對整個圖檔作水平方向 projection 得到其 histogram，其結果如圖二所示，我們可以發現每一行均有很明顯的兩個 peaks, 左邊那個 peak 是 midline，而右邊那個是 baseline, 現在的方法是以 baseline 所在的位置來定義此 line 的位置，我們在得到每一行的位置後，即可藉此來判斷此行是上移或下移，因此就可以取出藏在 line 中的資訊了。因為 line 與 line 的 spacing 是相同的，因此在對圖檔進行萃取程序時是不需要原圖作比對的。

Data hiding in text is in the discovery of modifications that are not noticed by readers. Some methods have been provided to hide data in text files. Most these methods need the original file to do data extraction, this is impractical.

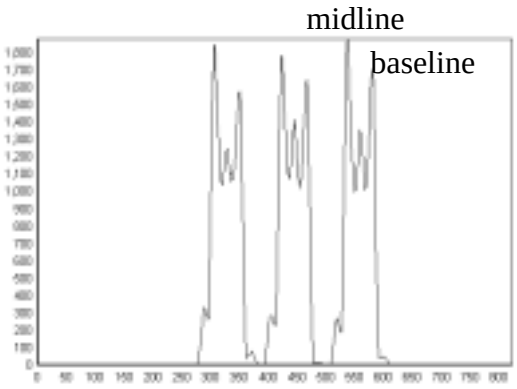
圖一、 包含三行文字的黑白圖檔

當 textline 被分割出來後，對每一 textline 作垂直的 projection, 如圖三所示。由圖三我們可以發現其 histogram 並沒有很明顯的 peak 去定義出每一個 word 的位置，因此此方法並不適用去偵測出 word 的移動。

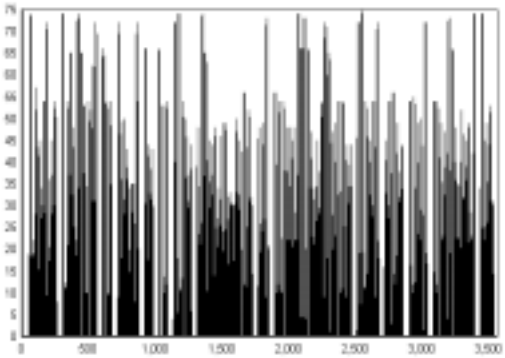
2. Centroid Detection[1,4,10]:

此一方法適用於偵測 line 與 word 的移動, 其方法便是取出每一個 line 或 word 的質心，以此質心來當作此 line 或 word 的位置，藉由判斷其移動的方向後，即可得到藏在 line 和 word 的資訊，其中需要注意的是若是要去偵測藏在 word 中的資訊是需要原圖去做比對的, 因為 word 與 word 的 spacing 是不相同的。

本計劃的目的是提出一個方法使我們在從一份文件去做資料擷取時是不需要原始文件的，而且除了可以從文件檔案中擷取出我們所藏的資訊外，亦可由列印後的文件再掃描成 BMP 檔來擷取資料。



圖二、 水平方向 projection 所得之結果



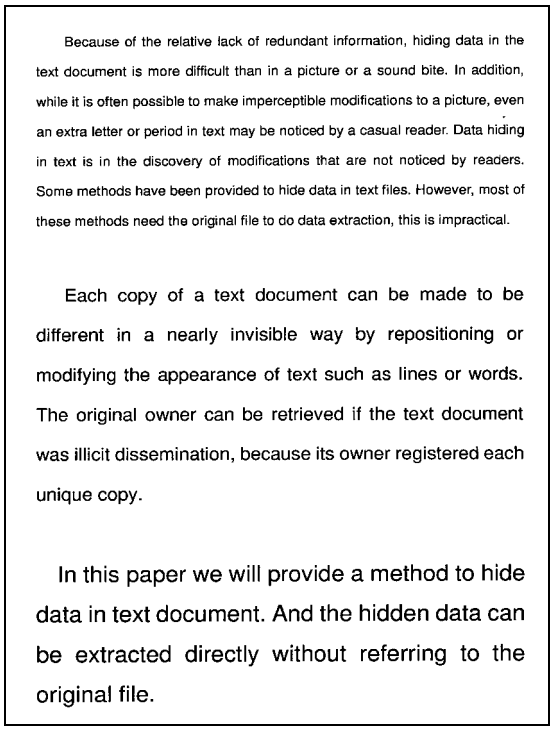
圖三、 對圖一中的第一行作垂直方向 projection 所得之結果

三、結論與討論

在這個段落中，我們將介紹利用我們所發展出來的演算法所做的實驗結果。在我們提出的方法中，我們將資料嵌入在一 postscript 檔案, 會選擇 postscript 檔案乃是因為其被廣泛的使用。

我們的實驗包含兩個部分:嵌入程序與萃取程序，在嵌入程序中，我們將資料藏在一 postscript 檔案中經由 line-shifter 與 word-shifter, 而在萃取程序中, 則是將資料從一 postscript 檔案

中或者一文件影像中擷取出來。其中嵌入程序實驗是將一份文件藏入機密資料，其中這個機密資料是由七個位元組字元‘AIPNCTU’所組成。圖四是原始未藏入資料的文件，而圖五則是將機密資料藏入後的文件。

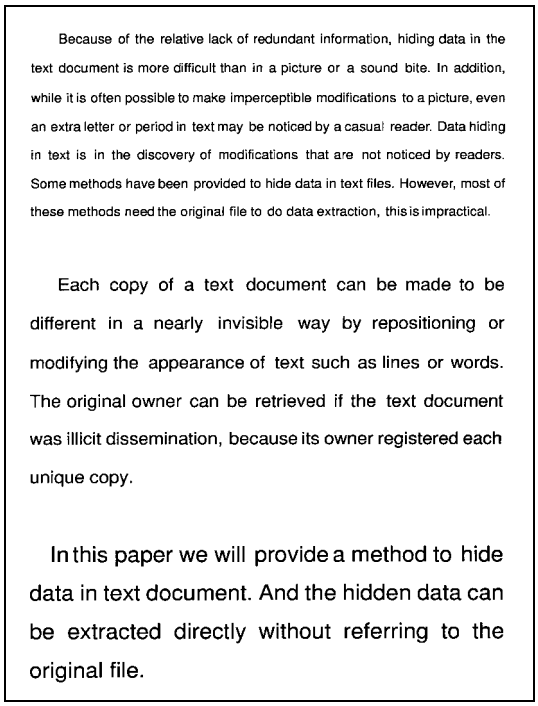


Because of the relative lack of redundant information, hiding data in the text document is more difficult than in a picture or a sound bite. In addition, while it is often possible to make imperceptible modifications to a picture, even an extra letter or period in text may be noticed by a casual reader. Data hiding in text is in the discovery of modifications that are not noticed by readers. Some methods have been provided to hide data in text files. However, most of these methods need the original file to do data extraction, this is impractical.

Each copy of a text document can be made to be different in a nearly invisible way by repositioning or modifying the appearance of text such as lines or words. The original owner can be retrieved if the text document was illicit dissemination, because its owner registered each unique copy.

In this paper we will provide a method to hide data in text document. And the hidden data can be extracted directly without referring to the original file.

圖四、 未嵌入資料的文件



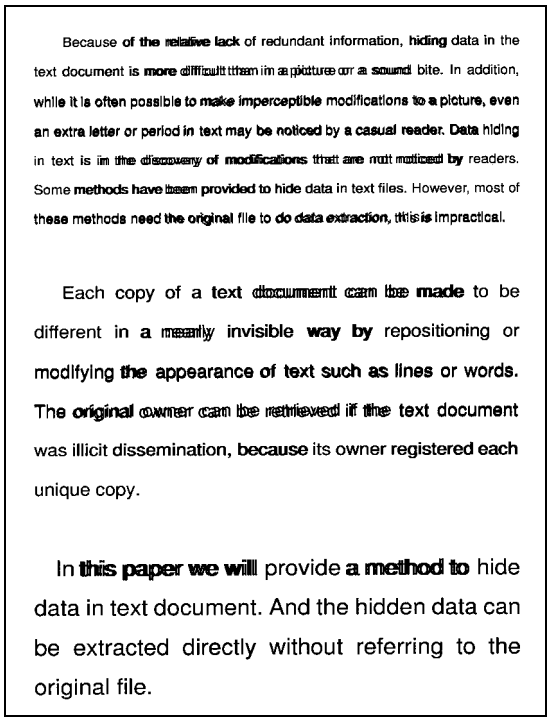
Because of the relative lack of redundant information, hiding data in the text document is more difficult than in a picture or a sound bite. In addition, while it is often possible to make imperceptible modifications to a picture, even an extra letter or period in text may be noticed by a casual reader. Data hiding in text is in the discovery of modifications that are not noticed by readers. Some methods have been provided to hide data in text files. However, most of these methods need the original file to do data extraction, this is impractical.

Each copy of a text document can be made to be different in a nearly invisible way by repositioning or modifying the appearance of text such as lines or words. The original owner can be retrieved if the text document was illicit dissemination, because its owner registered each unique copy.

In this paper we will provide a method to hide data in text document. And the hidden data can be extracted directly without referring to the original file.

圖五、 嵌入資料的文件

圖六則是將圖四與圖五重疊之後的結果，由圖六我們可以清楚地發現文件中 line 與 word 均已經改變。根據我們的實驗，我們所提出的方法對於一頁二十五行、字體大小為十二的文件，我們所能嵌入的字元數大約是七個字元。



Because of the relative lack of redundant information, hiding data in the text document is more difficult than in a picture or a sound bite. In addition, while it is often possible to make imperceptible modifications to a picture, even an extra letter or period in text may be noticed by a casual reader. Data hiding in text is in the discovery of modifications that are not noticed by readers. Some methods have been provided to hide data in text files. However, most of these methods need the original file to do data extraction, this is impractical.

Each copy of a text document can be made to be different in a nearly invisible way by repositioning or modifying the appearance of text such as lines or words. The original owner can be retrieved if the text document was illicit dissemination, because its owner registered each unique copy.

In this paper we will provide a method to hide data in text document. And the hidden data can be extracted directly without referring to the original file.

圖六、 將圖四與圖五重疊之後的結果

四、計畫成果自評

這一個計畫於執行期間的進度與工作目標與當初所提的計畫內容大致吻合。在本計畫中，實驗結果證明了我們所提出的方法之可行性，也驗證了本計劃提出的資訊隱藏在文件檔的機制，可以克服目前已有的方法之缺點。換言之，我們所提出的方法在作資料擷取時不需要有原始文件，並且能正確的取出所隱藏的資料，因此實用性很高。

除此之外，我們在本計畫中也建立了一套，簡單且富有彈性的全盤性整合人機介面。

五、參考文獻

[1] S. H. Low, N. F. Maxemchuk, A. M. Lapone, "Document Identification for

Copyright Protection Using Centroid Detection," *IEEE Trans on Commun.*, Vol. 46, No. 3, pp. 372-381, Mar. 1998.

[2] J. Brassil, S. H. Low, N. F. Maxemchuk, L. O'Gorman, "Electronic Marking and Identification Techniques to Discourage Document Copying," *IEEE Journal on Sel. Areas in Commun.*, Vol. 13, No. 8, pp. 1495-1504, Oct. 1995.

[3] N. F. Maxemchuk, "Electronic Document Distribution," *ATT Technical Journal*, pp. 73-80, Sept. 1994.

[4] J. Brassil, S. H. Low, N. F. Maxemchuk, "Copyright Protection for the Electronic Distribution of Text Documents," *Proceedings of the IEEE*, Vol. 87, No. 7, July 1999.

[5] J. Brassil, S. H. Low, N. F. Maxemchuk, L. O'Gorman, "Marking Text Features of Document Images to Deter Illicit Dissemination," *Int. Conf. on Pattern Recognition Israel*, Oct., 1994(in press)

[6] A. K. Choudhury, N. F. Maxemchuk, S. Paul, H. Schulzrinne, "Copyright Protection for Electronic Publishing over Computer Networks," *IEEE Network Mag.*, Vol. 9, No. 3, pp. 12-21, May/June 1995.

[7] S. H. Low, A. M. Lapone, N. F. Maxemchuk, "Document Identification to Discourage Illicit Copying," *IEEE GlobeCom 95*, Nov. 13-17 1995, Singapore

[8] W. Bender, D. Gruhl and N. Morimoto, "Techniques for data hiding," in *Proc. SPIE*, pp. 2420-2440, Feb. 1991.

[9] N. F. Maxemchuk, S. H. Low, J. Brassil, L. O'Gorman, "Document Marking and Identification using Both Line and Word Shifting," *Infocom '95*, pp. 853-860, Boston, Mass. April 4-6, 1995.

[10] S. H. Low, N. F. Maxemchuk, "Performance Comparison of Two text Marking Methods, " *IEEE JSAC*, Vol. 16, No. 4, pp. 561-572, May 1998.