行政院國家科學委員會專題研究計畫　成果報告

序列比對在古籍活字印刷字體辨識之應用

計畫類別：個別型計畫
計畫編號：NSC94-2213-E-216-021-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：中華大學生物資訊學系

計畫主持人：侯玉松

計畫參與人員：陳華，林婉婷

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 10 月 12 日

行政院國家科學委員會專題研究計畫成果報告

序列比對在古籍活字印刷字體辨識之應用

計畫編號：NSC 94-2213-E-216 -021
執行期限：94 年 8 月 1 日至 95 年 7 月 31 日

主持人：中華大學生物資訊學系助理教授侯玉松
E-mail: yshou@chu.edu.tw

一、中文摘要

古籍數位化是國家圖書館的重要計畫之一，但是現有光學中文辨識軟體的古籍文字辨識準確度太低，只能藉由打字等人工方式，完成文字輸入與校對工作，耗費許多人力、財力與時間，因此增加古籍數位化的難度。

本計畫擬研究古籍的活字印刷字體的辨識問題，我們認為，活字印刷的字形與字體大小的一致性高，但因為古代印刷技術不如現今，可能會有筆劃粗細不均、斷裂、汙點等情形，而產生「大同小異」的現象，這種情形恰可用生物 DNA 序列突變的三種情形：替代、插入、刪除來比擬，所以如果將待辨識的單字圖案，以列為主的方式轉換成 0/1 序列，即可嘗試套用相似 DNA 序列比對方法，如：廣域比對法等，比對兩字的 0/1 序列的相似度，達到辨識的效果。

在本計畫中，將研究生物資訊學的 DNA 序列比對法，應用於字庫搜尋比對的可行性，檢討其辨識準確度與執行速度，期望對於活字印刷的古籍數位化，有所貢獻。

關鍵詞：古籍數位化、光學中文辨識、生物資訊學、序列比對

Abstract

Digitization of ancient books was an important project of National Central Library. Current Chinese optical character recognizers were not enough to apply to digitization of ancient books since their correct rates were low. So the input and correcting works were implemented by hands. It was wasteful in man power, money and time. Therefore digitization of ancient books was harder.

In this project, we will study the recognization problem of movable type Chinese. We think that the font and size of movable type Chinese is consistent. But little difference such as different line width, breaks, blots were exist since ancient printing technique was poor than the modern. This case is like to the mutation of DNA sequences including substitution, insertion, and deletion. After transforming an image of a word to a 0/1 sequence in row-major, we can compare two words by using comparing methods about DNA sequences, such as the global alignment method. Therefore, we can obtain the similarity of two words, and then recognize an unknown word.

In this project, we will study the application of DNA sequence comparing methods in bioinformatics to searching in word bases. The correct rate and execution time will be analyzed. We hope it is useful to digitization of ancient books in future.

Keywords: Digitization of ancient books, optical character recognization, bioinformatic, sequence alignment.

二、緣由與目的

隨著資訊事業蓬勃發展，國內對於資訊中文化的需求也益行迫切。但是，由於中文字數繁多，字體結構也複雜，字型間相似字多，使得中文資訊與電腦的溝通在中文輸入、儲存與輸出，皆造成相當困擾，同時也阻礙中文古籍數位化。

古籍數位化是龐大且繁重的工作，以東吳大學中文系陳郁夫教授製作的「故宮・東吳」數位古今圖書集成系統[1]為例，《古今圖書集成》全套有八百冊、一萬卷、五十多萬頁、一億七千多萬字。

在資料製作流程中，在選取善本之後，分為二系，一系是製作本文，需將古籍資料以打字輸入，再做校對等工作，《古今圖書集成》共一億七千多萬字，製作本文耗費許多時間、人力與財力。另一系是製作圖檔，用意在保存珍貴古籍原貌，《古今圖書集成》共五十多萬頁，掃瞄圖檔也非常耗費人力。

為何不採用中文光學文字辨識比對(Optical Character Recognize，OCR)軟體以節省人力、物力？這是因為現有的中文 OCR 軟體應用在古籍文字辨識的準確度偏低，我們曾從《古今圖書集成》中，任意取出 30 個字，一字一字輸入至「丹青中文辨識系統」測試，只正確辨識 10 字，準確度僅達 33%，無法應用在古籍數位化[2]。

為提高辨識準確度，我們認為應針對每一本古籍的字體差異性，設計個別的字庫及辨識演算法。而不像目前市面販售的中文 OCR 軟體，一套字庫及辨識演算法，就想應用在各式中文字體的辨識。

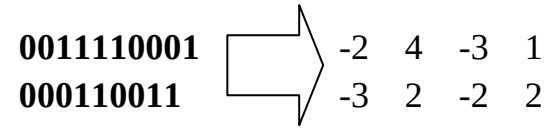
在本計畫中，針對字形一致性較高的活字印刷體，研究其辨識問題，並利用生物資訊中廣為採用的序列比對法(sequence alignment)，設計簡易而快速的辨識演算法，以及以整數序列代表字形的字庫建構方法，將詳述於下節。

三、研究方法

(一) 活字印刷字體的 0/1 序列化

活字是經由掃瞄及數位化後的數位影像，以二元化處理後即可將灰階值範圍改為白或黑(即 0 或 1)。一個字若預設為 256*256 圖素大小，若以列為主(row major)，從上到下連續讀出各圖素的 0/1 序列值。如此整個中文文字影像將轉換成長度 65536 的 0/1 序列。

若將 0/1 序列的連續數個 0 或連續數的 1，以行程長度編碼法(Run Length Encoding ，簡稱 RLE)方式壓縮。如:連續 5 個 1(11111)以+5 表示，而連續 10 個 0(0000000000)以-10 表示，可將數列做一個壓縮。如圖一。中文字 0/1 序列的特性是常會出現連續的 0/1，所以 RLE 的數列大小，將遠小於 0/1 序列的大小，可用做建構字庫時，中文字形的儲存方式。

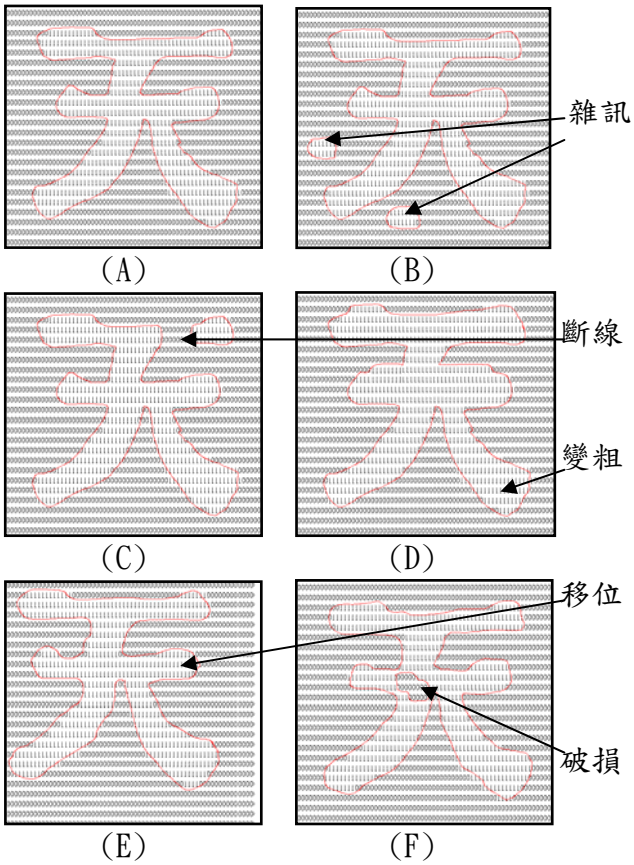


圖一：行程長度編碼法

(二) 活字印刷字體辨識與 DNA 序列比對

兩個相同字的活字印刷，受限於古代印刷技術，其字跡不是完全相同，轉換成 0/1 序列後，序列上會有些許差異。以圖二為例，以 0/1 序列之差異處做比較。

以圖二為例，(A)為正常情形。(B)表示有雜訊發生，在轉換為 0/1 序列後，會出現多餘 1 的序列。(C)表示字跡有斷線，在字跡部分某 1 區塊變成 0，使筆劃分離。(D)表示筆畫粗細發生變化，變粗表示 1 的序列增加，變細表示 1 的序列減少。(E)切字位置不同，導致 0/1 序列有位移發生。(F)表示字跡破損，在字跡部分某區塊 1 變成 0。



圖二：各種活字印刷問題對應序列差異處

綜合上述各狀況，(B)、(C)、(D)及(F)四種狀況，發生雜訊、斷線、粗細改變及破損，可視為部分 0/1 序列發生 0、1 相互替代(substitution)的情形。(E)移位的狀況，可視為插入(insert)或刪除(delete)連續的 0、1 序列，使得字體發生移位。這些情形與生物資訊中 DNA 序列發生替代、插入、刪除之情形非常相似，引發我們將生物資訊的 DNA 序列比對方法，應用在活字印刷字體比對的研究動機。

(三) RLE 序列比對方法

序列比對技術在生物資訊中是最基本，也是非常重要的工具。過去學者曾針對序列比對發展出以下演算法：廣域比對(global alignment)法、區域比對法(local alignment)、BLAST、FASTA 等方法，其演算法請參閱文獻[3]。為加快比對速度，本計畫改良 FASTA 應用在 RLE 序列比對。

FASTP 的演算法主要的三個步驟[3]：

1. 利用 k-tup 的方式，先計算出一字串的位置表，再計算對另一字串的位移雜湊表(hash table)及位移向量(offset vector)。
2. 取出位移量出現最多次的前 5 段區域。
3. 再利用區域序列比對演算法，將 5 段最佳的比對結果算出。

應用 FASTA 於 RLE 序列比對，需先設定下列參數：

1. 比對到黑色部分的分數：

即 RLE 數列為＋的部分互相比對，亦即黑色筆劃部分比對的得分。

2. 比對到白色部分的分數：

即 RLE 數列為－的部分互相比對，亦即空白部分比對的得分。

3. 沒有比對到的部份分數：

亦即筆劃跟空白比對的扣分。

4. 容許誤差：

因為銅活字的特性，同樣的字，但是 RLE 數值會有少許的不同,所以當 RLE 數值相差在容許誤差內可當作是相同數值。

RLE 序列比對的流程步驟：

1. 先把兩個序列存入結構陣列中。
2. 把兩個序列的結構陣列中的 RLE 數值由小到大排序。
3. 將兩個結構陣列使用合併排序法合併成一個結構陣列。
4. 使用前述方法，把位移向量算出來。
5. 經過位移後，再使用簡化的廣域比對法，將比對成績算出。

演算法細節請參閱[4]。

四、結果與討論

(一) 測試資料取得

本計畫實驗之字體，以東吳大學中文系陳郁夫教授提東的《古今圖書集成山川典》的銅活字字體為主。所採用的影像解析度為 600dpi。取字方面是利用自行撰寫的程式人工取字，不論原本字的大小，抓取後都縮放成 100 × 100 像素的大小，沒有去除雜訊直接比對。

辨識率比對的實驗方式，先從《古今圖書集成山川典》中選取一段文章，作為欲辨識比對之資料，分別以丹青中文辨識系統與本計畫所提出之方法，比較其準確率及比對速度。

實驗方法是由《古今圖書集成山川典》其他書頁中，選取 1145 個字作為銅活字字庫，以欲辨識之文章與銅活字字庫做比對，找出銅活字字庫中比對後最佳相似度之字體做為預測結果。

(二) 實驗結果

測試的版本以《古今圖書集成山川典》，字庫總共有 1145 個字，測試比對的字有 152 個字，都是在不同頁數中取出的字，分數參數設定為：比對到黑色的部份分數為 7 分，比對到白色部分的分數為 2 分，沒有比對到的分數為-2 分，容許誤差率為 0.03。

測試結果：比對出第一名的字有 140 個字，準確率達到 92.1%。比對出前五名的字有 145 個字，準確率達到 95.3%。

平均比對出一個字的時間為16.89秒。

(三) 討論

0/1 序列廣域比對法,比較二字需耗時 3 分鐘，利用本計畫開發的 RLE 序列的 FASTA 比對方式，只需 16.89 秒，約加速

10.66 倍。但就實務應用來看，16.89 秒仍嫌慢，還有改善空間。

本計畫主要針對中文辨識與匹配比對部分做探討，比對整個字庫 1145 字需 16.89 秒，若加入其他影像處理技巧，並將字庫字體妥善分類，預先過濾比較字體，使得只要比對 100 字以內，即可預測正確字體，就可以將比對時間降至 1.48 秒左右，約每分鐘 40 字。

四、計畫成果自評

本計畫原先預期目標為：

1. 研究如何加快序列的廣域比對的執行速度。
2. 建立大規模的字庫資料
3. 藉由大量測試，分析辨識準確度，以評估廣域序列比對應用在活字印刷字體辨識的可行性。
4. 針對測試結果，反覆研究提高準確度與執行速度的方法。
5. 將最後研究成果，製作研究報告與發表論文，以供未來相關研究參考。

我們主要耗費大量研究時間於設計較快速的 RLE 序列比對演算法，利用 FASTA 的運作觀念，將 FASTA 原本只能針對 DNA 或蛋白質序列比對，改良為正負 RLE 數列比對，達到 95%的預測準確度，對比於丹青中文辨識系統的 33%準確度，大幅提升，已完成上述 1、3、4 點目標。

至於第 2 點，我們受限於掃瞄圖檔與研究人力不足，無法建立《古今圖書集成》的 25 萬個銅活字庫[5]。

關於第 5 點，我們即將整理本計畫成果，撰寫投稿論文。

五、參考文獻

(一) 成果報告相關參考文獻

1. 陳郁夫，「故宮‧東吳」數位古今圖書集成技術報告，2004 漢學研究國際學術研討會，雲林科技大學主辦，2004。
2. 黃偉峰、侯玉松，廣域比對在銅活字中文辨識之應用，中華大學資訊工程系碩士論文，2004。
3. J. Setubal and J. Meidanis, Introduction to Computational Molecular Biology, PWS PUBLISHING COMPANY, 1997.
4. 林世勇、侯玉松，銅活字中文辨識之序列比對演算法研究，中華大學資訊工程系碩士論文，2005。
5. 錢存訓，中國古代書籍紙墨及印刷術，北京圖書館出版社，2002。

(二) 其他計畫相關參考文獻

6. B. S. Jeng, et al., “A Study on Optical Chinese Character Recognition”, Technical Reports of the Telecommumication Lab., Vol. 20, Special Issue, 1990, pp. 1-27.
7. S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequences of two proteins.” J. Mol. Biol. 147, pp.195-197, 1970.
8. Y. Tsuji. and K. Asai, “ Character Image Segmentation, Based Upon minimum Variance Criterion”, NEC Res. & Develop., July, 1985, pp. 23-30.
9. F. Ali and T. Pavlidis, “Syntactic recognition of handwritten numerals”, IEEE Trans. Systems Man Cybernet., Vol. 7, 1977, pp. 537-541.
10. S. F. Altschul and B. W. Erickson, “Optimal sequence alignment using affine gap costs.” Bull. Math. Biol., 48, 1986, pp.603-616.
11. R. G. Casey and G. Nagy, “Recursive segmentation and classification of composite character patterns”, Proc. IEEE, 1982, pp. 1023-1026.
12. J. W. Fickett, “Fast Optimal alignment.” Nucleic Acids Res., 12, 1984 ,pp.175-180.
13. W. M. Fitch and T. F. Smith, “Optimal sequence alignments.” Proc. Natl. Acad. Sci. US, 80, 1983, pp.1382-1386.
14. T. Fujita, M. Nakanishi, M. Miwa, “Kanji Character Recognition System”, Fujisu Scientific and technical Journal, Sept. 1977.
15. H. A. Glucksman, “Multicategory classification of pattern represented by high-order vectors of multilevel measurements”, IEEE Trans. Comput.,

Vol. 20, 1971, pp. 1593-1598.
16. O. Gotoh, “An improved algorithm for matching biological sequences.” J. Mol. Biol., 162, 1982, pp.705-708.
17. G. H. Grandlund,”Fourier preprocessing for hand print character”, IEEE Trans. Comput., Vol. 21, 1972, pp. 195-201.
18. R. M. Gray, “Vector Quantization”, IEEE ASSP Magazine, April 1984, pp. 4-29.
19. G. F. Groner, “Real-Time Recognition of Handprinted Text”, Proc. Fall Joint Comput. Cont., AFIPS, Vol. 29, Nov., 1966, pp. 591-601.
20. W. H. Highleyman, “Linear decision functions with application to pattern recognition”, Proc. IRE, Vol. 50, 1962, pp. 1501-1514.
21. R. L. Hoffman and J. W. McCullough, “Segmentation Method for Recognition of Machine-Printed Character”, IBM J. Res. & Develop., 15, March 1971, pp. 153-161.
22. C. E. Kim, “ On cellular straight line segments”, Comput. Graphics and Image Processing 18, 1982, pp. 369-381.
23. C. E. Kim and A. Rosenfeld,” Digital straight lines and convexity of digital regions”, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-4, 1982, pp. 149-153.
24. T. Kohonen, “ Self-Organization and Associative Memory”, Mar. 1988, pp. 36-41.
25. P. Krause, W. Schwerdtman and D. Paul,” Two modifications of a recognition system with pattern series expansion and Bayes classifier”, Proc. 2nd Int. Jt. Conf. Pattern Recgnition, 1974, pp. 215-219.
26. J. K. Lin, B. S. Jeng, et al., “Recognition of printed Chinese character utilizing two-stage classification with mesh and peripheral features”, Proceedings of Telecommunications Symposium, 7B-1, 1987, pp. 533-537.
27. Umeda Michio, “Recognition of Multi-Font printed Chinese Character”, ICPR. 1982, pp. 793-799.
28. K. Mori and I. Masuda,” Advances in Recognition of Chinese Character”, Proc. 5th ICPR., 1980, pp. 692-702.
29. S. Naito, I. Masuda, “Handprinted Kanji Recognition by Feature Matching Methods and Its Application to Personal OCR”, J. IECE Japan, April, 1984, pp. 480-487.
30. W. C. Naylor, “ Some studies in the interactive design of character recognition system”,IEEE Trans. Comput., Vol. 20, 1971, pp. 1075-1086.
31. R. Oka, “Handpronted Chinese Character Recognition by using Cellular Feature”, Proc. 6th Int. Joint Conf. Pattern Recognition, 1982, pp. 783-785.
32. W. R. Pearson, “Rapid and sensitive comparison with FASTP and FASTA.” Methods Enzymol., 183, 1990, pp.63-98.
33. W. R. Pearson, “Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms.” Genomics, 11, 1991, pp.635-650.
34. W. R. Pearson and D. J. Lipman, “Improved tools for biological sequence comparison.” Proc. Natl. Acad. Sci. USA, 85, 1988, pp.2444-2448.
35. K. Sakai, T. Kawada, S. Amano and K. Mori, ”An Optical Chinese Character Reader”, Proc. 3rd Int. Joint Conf. Pattern Recofnition, Nov. 1976, pp. 122-126.
36. T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences.” J. Mol. Biol., 147, pp.195-197, 1981.
37. W. W. Stallings, “Recognition of Printed Chinese Characters by Automatic Pattern Analysis”, Comput. Graphics Image Processing, Vol. 1, 1972, pp. 44-65.
38. C. Y. Suen, “Distinctive Features in Automatic Recognition of Handprinted Character”, Signal Processing, 1984,pp. 193-207.
39. C. Y. Suen and R. J. Shillman, “Low error rate optical character recognition of unconstrained handpronted letters based on a model of human perception”, IEEE Trans. System Man Cybernet.,

Vol. 7, 1977, pp. 491-495.
40. H. Tamura, “A comparison of line thinning algorithms from digital geometry viewpoint”, Proc. 4th Int. Jt. Conf. Pattern Recognition, 1978, pp. 715-719.
41. Y. Tsuji and K. Asai, “Character Image segmentation”,SPIE 28th Technical Symp., 504, Aug. 1985, pp. 2-9.
42. K. Yamamoto and A. Rosenfeld, “Recognition of handprinted Kanji characters by a relaxation method”, Proc. 6th Int. Conf. Pattern Recognition, 1982, pp. 395-398.