中文文件影像中之特殊字體偵測

林裕淵、曾逸鴻

E-mail: 9422469@mail.dyu.edu.tw

摘 要

光學文字辨識(optical character recognition, OCR)是近幾十年來一個熱門且被廣泛研究的題目，利用文字辨識技術將文件數位化，既可減少紙張的存放空間，又可以自動將文件分類以方便日後搜尋，有利於後續的知識分享。 目前的文字辨識相關產品說明中，都聲稱文字辨識率可高達90%以上。然而這些數據大都是根據正常字體印刷文字影像的辨識結果統計。對印刷文件中常見的特殊字體(如粗體、中空、底線與斜體)，辨識效果會與正常字體有明顯的差異。若使用多個辨識核心辨識，對於數量龐大的中文字集，將會導致辨識速度下降。本論文將提出方法來偵測印刷文件影像中的斜體、底線、中空、粗體等特殊字體文字出現的位置。首先分析文件中文字區塊影像的投影輪廓，擷取出文字行與文字元件，再統計各元件大小、元件間距離、元件筆劃寬度與元件黑點群長度等特性，以判斷文件影像中各字元的字體歸屬。後續進行文字辨識時，再用其相對的辨識核心去比對，如此便可在維持辨識速度下提升多字體文件的辨識效果。

關鍵詞 : 特殊字體 ; 投影輪廓 ; 文字辨識

目錄

參考文獻

[1] Y. Yu, A. Samal and S. C. Seth,“ A system for recognizing a large class of engineering drawings, ”IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 8, pp. 868-890, 1997.
[2] K. C. Fan, C. H. Liu, Y. K. Wang,“ Segmentation and classification of mixed text/graphics/image documents, ”Pattern Recognition Letters, vol. 15, pp. 1201-1209, 1994.
[3] C. L. Tan, and P. O. NG,“ Text extraction using pyramid, ”Pattern recognition, vol. 31, no. 1, pp. 63-72, 1998.
[4] B. F. Wu, C. C. Chiu, and Y. L. Chen,“ Algorithm for compressing compound document images with large text/background overlap, ”IEE Proceedings-Vision, Images, and Signal Processing, vol. 151, no. 6, pp. 453 – 459, 2004.
[5] G. Nagy,“ Twenty years of document image analysis in PAMI, ”IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 38-62, 2000.
[6] Y. H. Tseng and H. J. Lee,“ Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm, ” Pattern Recognition Letters, vol. 20, pp. 791-806, 1999.
[7] K. C. Fan, L. S. Wang and Y. T. Tu,“ Classification of machine-printed and handwritten texts using character block layout variance, ” Pattern Recognition, vol. 31, no. 9, pp. 1275-1284, 1998.
[8] Y. H. Tseng, C. C. Kuo and H. J. Lee,“ Speeding up Chinese character recognition in an automatic document reading system, ” Pattern Recognition, vol. 31, no. 11, pp. 1601-1612, 1998.
[9] X. Ye, M. Cheriet, and C. Y. Suen,“ Stroke-model-based character extraction from gray-level document images, ”IEEE Transactions on Image Processing, vol. 10, no. 8, pp. 1152-1161, 2001.
[10] L. Y. Tseng, R. C. Chen,“ Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming, ”Pattern Recognition Letters, vol. 19, no. 10, pp. 963-973, 1998.
[11] H. M. Suen and J. F. Wang,“ Text string extraction from images of colour-printed documents, ”IEE Proceedings-Vision, Images, and Signal Processing, vol. 143, no. 4, pp. 210-216, 1996.
[12] Y. H. Tseng, C. C. Kuo and H. J. Lee,“ Typeface identification for printed Chinese character, ”International Journal of Pattern Recognition and Artificial Intelligence, vol. 12, no. 2, pp. 173-190, 1998.
[13] Y. Zhu, T. Tan and Y. H. Wang,“ Font recognition based on global texture analysis, ”IEEE Transactions on pattern analysis and machine intelligence, vol. 23, no. 10, pp. 1192-1200, 2001.

[14] A. Zramdini and R. Ingold, ” Optical font recognition using typographical features, ” IEEE Transactions on pattern analysis and machine intelligence, vol. 20, no. 8, pp. 877-882, 1998.
[15] Z. Li, Y. Lu and C. L. Tan, “ Italic font recognition using stroke pattern analysis on wavelet decomposed word images, ” Proceedings of the 17th International Conference on Pattern Recognition, vol. 4, pp. 23-26, 2004.
[16] Y. Li, S. Naoi, M. Cheriet and C.Y. Suen, ” A segmentation method for touching italic characters ” , Proceedings of the 17th International Conference on Pattern Recognition, Vol. 2, pp. 23-26, 2004.
[17] B. Gatos, N. Papamarkos, and C. Chamzas, “ Skew detection and text line position determination in digitized documents, ” Pattern Recognition, vol. 30, no. 9, pp. 1505-1519. 1997.