Special Typeface Identification in Chinese Document Images

林裕淵、曾逸鴻

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

ABSTRACT

Optical character recognition (OCR) is a famous research subject in recent twenty years. To digitize paper documents by applying OCR techniques can decrease the document storage space. These digitized document images can be classified and retrieved conveniently. At present, commercial OCR products purported to provide a satisfactory recognition results whose recognition accuracy is over 90%. The accuracy is generally measured by recognizing those printed characters whose typefaces are normal. However, several special typefaces such as italic, underline, hollow, and boldface, poor recognition accuracy is obtained by commercial OCR systems. Since the amount of Chinese characters is large, the recognition speed is slow using a multi-engine OCR system. This paper proposes an approach to detect all characters in special typefaces. In the proposed typeface identification system, text lines and character components are extracted by analyzing the projection profiles of text block images. Then, several characteristics such as component sizes, gaps between two components, stroke widths, and black run lengths, are computed and analyzed to identify the typeface of each character. Afterward, a specific recognition engine is applied to recognize each unknown character according to the corresponding typeface identification result.

Keywords : special typefaces ; projection profiles ; character recognition

Table of Contents

REFERENCES

[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.