Abstract

It is a very important problem, characters be overlapped with lines in forms, to recognition character in known forms. The interfered-characters can't be extracted from the text lines exactly and the traditional OCR engines will fail to recognize characters with interference. The first method is to remove form lines and reconstruct characters. Characters are broken with line removal and strokes are separated into two sets of stroke-ends. The colinearity and position of the stroke-ends are used to find out correct connecting correspondence. Gaps of the broken strokes are filled to reconstruct the original characters. The second method is to modify the OCR model to fit interfered-characters. Printed characters with form lines are uniformly segmented according to projection profiles. The locations of form lines in the interfered-characters are extracted and both CDFs (contour-direction features) and CCFs (crossing-count features) of form lines are calculated. Trained features of the OCR engine are modified by the features of form lines to match interfered-characters.
In the first experiment, 938 handwritten characters with form lines are tested, and the recognition rate is 23.7%. After using the first method, the accuracy is raised to 78.3%. In the second experiment, 695 printed characters with form lines are tested, and the recognition rate is 64.3%. After using the second method, the accuracy is increased to 77.3%.

Keywords: form database, form line removal, character recognition

三、結果與討論

我們測試的環境是 Pentium II 233 PC, 128MB RAM 的機器上，作業平台是 Microsoft Windows NT 4.0，使用程式語言和模組是 Microsoft Visual C++ 5.0 的 MFC (Microsoft Foundation Class)。

我們拿了兩種測試資料來測試，一種是手寫字的表格資料，一種是印刷字的表格資料，我們將手寫字中有將線段除去再補字的資料和沒有去線的資料作比較，發現無去線的辨識率只有 23.7%，有去線補字的辨識率高達 78.3%，由這個結果可以得知，我們所採用去線補字的方法是有辦法明顯改善辨識率的。印刷字的測試資料方面，我們和手寫字做同樣的處理，另外加上測試調整過辨識核心後的辨識系統，發現沒有去線的辨識率為 64.3%，去線補字的辨識率為 79.6%，修改辨識核心的辨識率為 77.3%，由這個結果看來，修改辨識核心似乎沒有多大的作用，但是由於去線補字必須對影像作處理，而修改辨識核心，不需要動到影像，相對來講，修改辨識核心在速度上有明顯的改善，而辨識率只下降了一些而已，因此就某些用途來講，也算是一個值得考慮的用法。

四、計畫結果自評

在本年度的計劃中，我們針對字和表格線的重疊現象加以改善，這個現象之前的計劃雖然有考慮進去，但是由於並沒有詳細的去研究解決之道，形成辨識率的瓶頸，因此這份研究有助於使我們辨識系統更加完善。

本年度的計劃雖然沒有大幅度變動整個系統，不過卻可以讓辨識率大幅度的提昇，雖然離希望達到的目標還有一段距離，但是只要再加上後續的研發，整個系統的完成將指日可待。

五、參考文獻


