The high-accuracy extraction and recognition of characters in engineering drawings

Yalin Wang

P O Box 84-187, Tsinghua University Postoffice
Beijing 100084, P.R.China. Email: wyl@sun5.thit.tsinghua.edu.cn

Abstract

With the project of Automated Callout Capture-Rrecognition Engine, Bell&Howell/PSC wants to get a character recognition engine and associated pattern editing/creation functionality to automate the callout capture process which will reduce labor costs associated with this function. They set the expected accuracy as no less than 98% and had tracked the progress in the OCR/ICR world for three years before they found our group.

For the recognition module, two main functions are realized in our software: Extraction and Recognition. In the Extraction stage, we use a new method named scanline algorithm(get the topology of the subarea by checking the consecutive pixel lines) to find all the subareas after only one pass of the image. We then extracts all the character candidates by using the area criteria(characters are connected areas.) and the size criteria(the most frequently appeared image parts should be the characters in the drawings.) and group them into strings by using group criteria(the characters in one string have almost the same distance to each other and their centroid points are in one line.) During the Recognition stage, we use features getting from eroding and tracking the pixel images of the character candidates and a precisely matching template method to recognize the characters. The advantages of this method include precise and efficiency by matching one character with several template, high adaptability by learning new characters. Other efforts are also made to achieve such a high target of 98% accuracy, e.g. we use a specific pattern file for a class of characters, the maximum size of character candidates is restricted and erasure of some simple touching problems is also done.

For the training module, additional operation are added to the recognition functions. The additional operation make it possible to let operator to decide whether the character candidate is a real character and if it is what character it is. Then the results of the operation with the ten features of the character are written in a pattern field, which will be loaded in the memory when to recognize the characters.

We did a lot of tests to our software. At last we reached the goal and contracted the project.

Key Words: Text Segmentation, OCR, Engineering Drawings Recognition and Understanding, Document Analysis