Development of an OCR System

Nathan Harmata

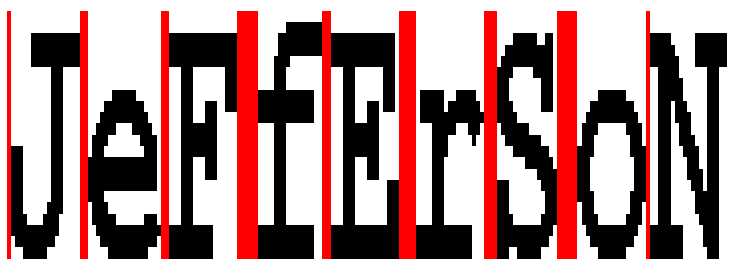
TJHSST Computer Systems Lab 2007 - 2008

Abstract

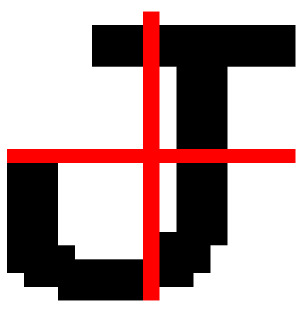
OCR (Optical Character Recognition) is a very practical field of Computer Science. Since the late 1980's, researchers have been developing system to identify text from non electronic sources, such as pictures or newspapers. The use of OCR systems has spanned from making books in Braille to sorting mail by zip code.

Procedures

The input for the current prototype is a PNG picture file that contains text in the Courier font. Using the Java BufferedImage class, locations and colors of the pixels in the image can be determined. The program uses these to find the positions of horizontal straight lines of whitespace in the image. It uses a simple dynamic programming technique to match the lines together to break apart the image into separate lines. Each line is parsed into words using a similar method involving vertical lines. After making spacing analysis, each word is parsed into letters, as seen below.

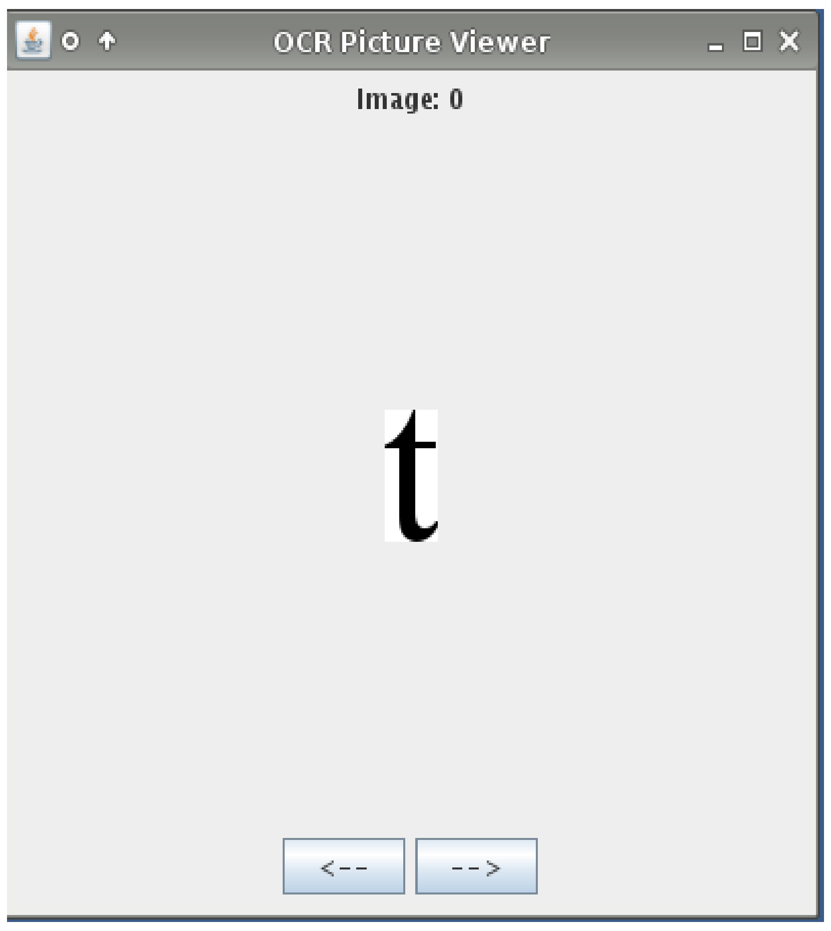


Then, each letter is considered as a sum of its four quadrants. The number of non-white pixels in the quadrants is used to find to best match from a cached list of pixel data for every character in the Courier font.



Background

Although there are a few options currently available to the public, like Microsoft Document Imaging, most of them are either unused or not accurate enough. The goal of this project is to create an OCR system that is simple to use and can handle most formatting and fonts.



Expected Results

The first prototype is fairly simple since all comparisons should be direct matches with the cache. In order to deal with different fonts, however, more complex and generic techniques will need to be developed.

A GUI for viewing the process of the prototype in parsing and analyzing input has already been made. In order to make the final product as simple as possible, a GUI for inputting and perhaps even cropping pictures will need to be devloped as well.