Development of an OCR System

Nathan Harmata

TJHSST Computer Systems Lab 2007 - 2008

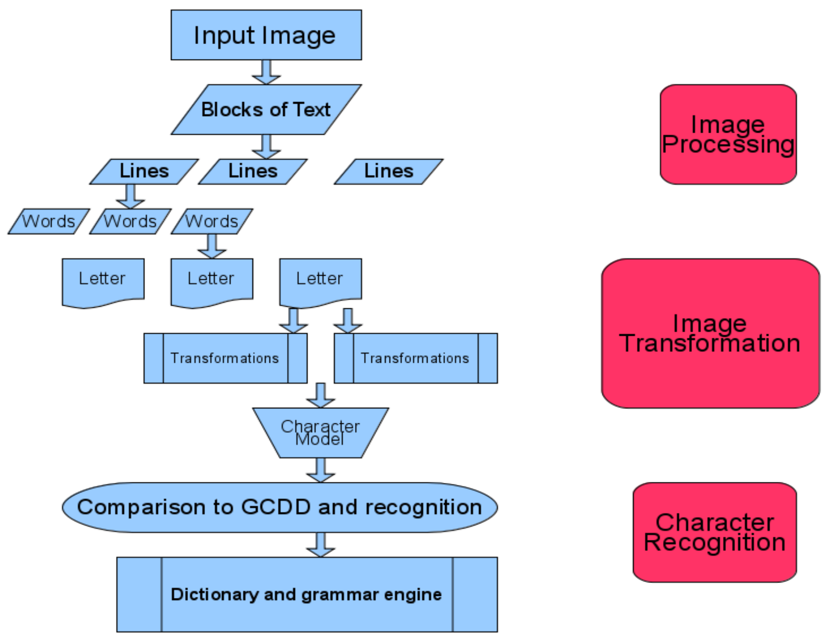
Abstract

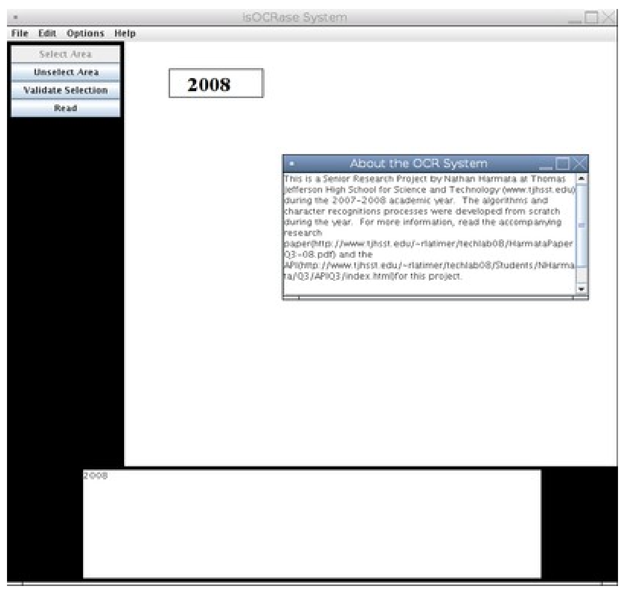
OCR (Optical Character Recognition) is a very practical field of Computer Science. Since the late 1980's, researchers have been developing system to identify text from non electronic sources, such as pictures or newspapers. The use of OCR systems has spanned from making books in Braille to sorting mail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

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 from scratch that is simple to use and can handle most formatting and fonts.

Overview of OCR System





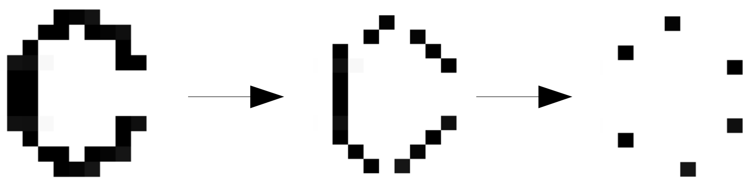
This is a Senior Research Project by Nathan Harmata at Thomas Jefferson High School for Science and Technology (www.tjhsst.edu) during the 2007-2008 academic year. The algorithms and character recognitions processes were developed from scratch during the year. For more information, read the accompanying research paper(http://www.tjhsst.edu/~rlatimer/techlab08/HarmataPaperQ3-08.pdf) and the API(http://www.tjhsst.edu/~rlatimer/techlab08/Students/NHarmata/Q3/APIQ3/index.html)for this project.

Procedures

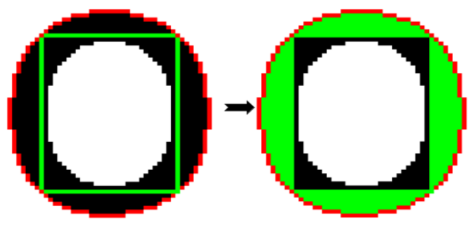
The user opens a supported image file and selects a portion of it to be read. Using the Java BufferedImage class, locations and colors of the pixels in the image can be determined. The program uses these to find the positions of horizontal straight lines of whitespace in the image. It pairs together lines of whitespace and ignores those so that only lines of text remain. Each line is parsed into words using a similar method involving vertical lines. After making spacing analysis, each word is parsed into letters.

Then, each letter is converted into a form called a "CharacterModel." Each model is a collection of "Attributes," which are the results of a series of complex transformations. These "Attributes" are the:

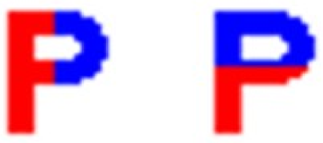
Sector Vector



Gap Vector



Pixel Concentration Vector



The result is that a letter is simplified into a few pieces of generic information. This procedure is applied to each letter of several different fonts, and information from the results is averaged. Using this, a cache is created to which results from OCR analysis can be dynamically compared, and the best matching character can be determined. Combinations of the top matches for each character are run through a dictionary and grammar engine, and the best word is chosen.

Results

The system was tested to be very accurate at recognizing single characters, with an overall success rate of 93.7%, which is comparable to commercial OCR programs.

Table 1: System Accuracy Results

Font	Size			
	18	20	24	28
Dialog	68.9	91.5	95	90.4
Serif	93.2	100	89.6	100
Times New Roman	97.2	100	97.3	95
Courier	100	95	96.4	89

The system is much weaker, however, at recognizing words accurately. This could be improved by implementing a more powerful grammar engine.