Handwriting recognition for Chinese Characters

Motivation

Chinese is the mostly spoken language in the world with over a billion speakers. Given that multiple variants of the Chinese language such as Mandarin and Cantonese use the same character set, there is a significant reach of the language throughout the world. Our project will develop an algorithm that can convert an input image to a digital character, also known as OCR (optical character recognition) or offline character encoding. Alternatively we will explore algorithms for online character encoding. Online characters are handwritten characters or text written on a digital surface such as a tablet. In this case, there is more information available than just a bare image. One challenge will be the handling the encoding of the Chinese characters on computer systems. Through this work, we hope to democratize and accelerate research in Chinese language technologies.

Source of Dataset

HIT-OR3C Database

HIT_OR3C is a dataset of handwritten Chinese characters. Both online and offline information is available. The characters have been collected using a handwriting pad and are recorded and labelled automatically via the handwriting document collection software: OR3C Toolkit. The software used to collect the characters is also made available (supplied version is in Chinese).

http://www.iapr-tc11.org/mediawiki/index.php?title=Datasets_List