



Master Thesis Defense

Speaker: Ning Wang

Supervisor: Dr. Suen

Examining Committee: Drs. Bergler, Lam and Dr. Yan (Chair)

Title: Noise Tolerant Oriental and English Document Language Identification Methods using Downgraded Pixel Density Feature

Date: Monday December 7, 2009

Time: 10:00 am.

Place: EV3.309

ABSTRACT

Document processing is one of the major applications of computers. One well-known branch of document processing, Optical Character Recognition (OCR) has been researched heavily and has achieved promising performance; on the contrary, another branch, Document Language Identification (DLI) has achieved much less progress so far. Despite not being studied well enough, DLI is a very important process for other document processing applications.

A new noise tolerant feature for DLI, Downgraded Pixel Density feature, is introduced in the thesis. Compared to other features widely used in existing Document Language Identification solutions, the new feature is much less sensitive to noises like slant, font and style. It is also very straightforward and more similar to how human beings recognize characters and words.

Two Document Language Identification solutions using the new feature are introduced afterwards: Component Matching solution and Two Step SVM solution, and the objective of the two solutions is to identify three oriental and English languages. Both solutions identify language by searching specific templates in document images. The two solutions focus on different points of the language identification process. The Component Matching solution, it is easy to tell from the name, focuses on a new template matching algorithm. The other one, the Two Step SVM solution, focuses on a new method of picking and processing templates instead.

Several experiments were conducted on an oriental language document database and an English document database at CENPARMI. Five hundred and ninety scanned documents in different languages are included in the two databases and the images are in really different styles to simulate the variety of documents in real life. The solutions and the results are described, discussed and compared in the thesis, with references to a few other existing solutions.