NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

PH.D DEFENCE - PUBLIC SEMINAR

Title: **Handwritten Document Image Retrieval**

Speaker: Ms Zhang Xi

Date/Time: 10 November 2014, Monday, 10:00 AM to 11:30 AM

Venue: Executive Classroom, COM2-04-02

Supervisor : Dr Tan Chew Lim, Professor, School of Computing

Abstract:

A vast amount of information is stored as text format in large databases or digital libraries. Users can easily access them by traditional text retrieval methods which many researchers have worked on for decades. However, paper-less life is impossible nowadays and many important and valuable documents are available only as imaged format. Therefore, it is now an important and urgent issue to let users access these imaged documents effectively and efficiently, similar to retrieving text format documents produced by computer software. Information retrieval for handwritten document images is more challenging due to the difficulties in complex layout analysis, large variations of writing styles, and degradation or low quality of historical manuscripts. Optical Character Recognition (OCR) can convert word or text line images directly to their transcriptions and traditional text retrieval methods can be used to retrieve user specified information. However, OCR needs large segmented and labelled training data, and recognizing the entire documents is a waste of time if the objective is to just to retrieve an imaged document without having the process the recognized text. Furthermore, OCR may provide poor recognition results due to unconstrained writing styles. In order to overcome the limitations of OCR, keyword spotting becomes an alternative way to retrieve handwritten documents. It only needs the features extracted from the imaged documents, and has no use of the ASCII content. In view of large variations in handwriting styles, this thesis proposal will present a method for extracting text lines from multilingual handwritten documents. Then, a combination two well-trained networks is used to increase the recognition performance for word image recognition. Finally, Heat Kernel Signature (HKS), which can better tolerate non-rigid deformations than gradient information, is used to represent the key points detected on the documents, and to achieve word image matching and segmentation-free keyword spotting.