Doctoral Dissertation Defense:

Statistical Techniques for Efficient Indexing and Retrieval of Document Images

Anurag Bhardwaj
Department of Computer Science and Engineering
University at Buffalo - SUNY, Buffalo, NY

Committee
Dr. Venu Govindaraju (Chair)
Dr. Aidong Zhang
Dr. Bharat Jayaraman

Venue: CEDAR Conference Room
Date: Aug. 10, 2010 (12:30 PM)

Abstract. We have developed statistical techniques to improve the performance of document image search systems where the intermediate step of OCR based transcription is not used. Previous research in this area has largely focused on challenges pertaining to generation of small lexicons for processing handwritten documents and enhancement of poor quality document images. However, in practice one must deal with several additional challenges such as processing multilingual documents which are predominantly in non-Latin scripts. In this dissertation we have developed script-independent and content-based retrieval techniques to access document images from multilingual digital libraries containing both printed and unconstrained handwritten documents.

Our work advances the state-of-art in retrieval of Indic documents. Our two-fold solution involves keyword spotting for scripts with existing OCR solutions and a semi-supervised recognition-free approach when an OCR option is unavailable.

We have also designed a novel framework for content based retrieval of handwritten documents that captures the stylistic properties of handwriting. This framework is adapted from the Latent Dirichlet Allocation (LDA) model for handwriting to learn the latent handwriting styles (i.e. cursive, loopy) present in a given corpora without any manual annotations or grammar. We have successfully applied this (style) modeling technique to forensic document analysis tasks of writer identification.

Finally, we have extended the idea of content based retrieval of historical documents by formulating for the first time, the problem of temporal indexing and retrieval of such manuscripts. We use a novel subspace learning technique for estimating the age of a scanned document image and apply it to retrieve other documents in the collection of similar age. The proposed subspace learning technique (hGLRAM) is based on a globally as well as locally optimized hierarchical generalized low rank approximation of matrices (GLRAM) that learns a tree based low-dimensional representation of documents images for robust modeling of aging patterns.

The methods developed in this dissertation have been validated on publicly available datasets: handwritten documents from the IAM database, George Washington's letters dataset, and printed documents datasets available from the Google Book Project and the Million Book Project. The accuracy of our methods is significantly superior to results reported in the literature.