Doctoral dissertation defense

Probabilistic Random Field Based Approach to Text Identification in Annotated Machine Printed Documents

Xujun Peng
Department of Computer Science and Engineering
State University of New York at Buffalo, Amherst NY

Committee: Dr. Venu Govindaraju (chair)
Dr. Peter D. Scott
Dr. Jason J. Corso

Date: Sept 29, 2010 (Wednesday)
Time: 9:00 - 10:00 am
Place: CEDAR conference Room

Abstract.

Today, the convenience of search, both on the personal computer hard disk and on the web, is essentially limited to machine-printed text documents and images because the accuracy of handwriting recognizers continues to be inadequate. This thesis advances the state-of-the-art in search of hand-annotated documents (possibly) by multiple writers in an office/collaborative environment. The annotations are assumed to be action instructions, such as, “make 4 copies”, “remove Figure X” etc.

One of the main research tasks we undertake to solve the overarching problem of search in the context of the above described documents is that of “ink separation”. This is essentially the task of segmenting handwritten annotations from machine printed matter, while allowing for noisy and overlapped blobs. Prior techniques have primarily used histogram thresholding and analysis of the connectivity of strokes to tackle the problem. These previous methods, although effective, rely on heuristic rules of spatial constraints, and are not scalable across applications. We have developed a Machine Learning based principled approach which has two parts: (i) binarization of document images (with special emphasis on documents captured by hand-held devices) using a non-linear transformation scheme and a novel edge potential feature in the framework of MRF, and (ii) classification of handwritten and machine-printed matter using an innovative adaptation of the decision tree classifier followed by a MRF based approach that exploits the statistical dependencies within a neighborhood in both the spatial and feature spaces. It also avoids the overfitting problem commonly encountered with implementation of decision trees by an approach inspired by the AdaBoost algorithm. The isolated handwritten textual blocks are indexed (unsupervised) based on writing instrument, style, and ink color, which are all indicators of different writers. Our thesis offers a unique way of selectively removing annotations belonging to a particular writer and allowing the end system user to view an unmarked document even though the original document image is marked up.

We have performed experiments on large public and private (HP) datasets which include both scanned and mobile camera captured document images under various illumination conditions and obtained results superior to previously reported methods in the literature.