

Information Retrieval from Handwritten Documents (Duration: 6 Hrs.)

Speakers

Achint Oommen Thomas and Anurag Bhardwaj

CUBS, University at Buffalo, USA

Email: aothomas@buffalo.edu, ab94@cedar.buffalo.edu

Achint Oommen Thomas completed his MS in Computer Science and Engineering from University at Buffalo (2007) and is expected to complete his PhD under the supervision of Dr. Venu Govindaraju by Sep., 2010 at UB. His areas of interests are computer vision, information retrieval, Human Interactive Proofs, handwriting generation and recognition. His work has been published in Pattern Recognition, IJDAR, ICFHR, ICPR and has also been recognized by MIT Technology Review. He is joining Yahoo! Research starting July, 2010 as a Research Scientist. Additional information can be found at: www.achintoommenthomas.net

Anurag Bhardwaj completed his MS in Computer Science and Engineering from University at Buffalo (2008) and is expected to complete his PhD under the supervision of Dr. Venu Govindaraju by Sep., 2010 at UB. His areas of interests are handwriting analysis, information retrieval and machine learning. His work has been published at IJDAR, PR, ICDAR, ICFHR as well as NIPS and SIGIR workshops. He is joining CUBS starting Sep., 2010 as a Post-Doctoral Fellow. Additional information can be found at: www.cse.buffalo.edu/~ab94

Abstract

This tutorial aims to provide a basic understanding of existing techniques for information retrieval from handwritten documents. The scope of the tutorial is to familiarize researchers from various backgrounds with this key topic and help them apply such techniques to related research problems as well as real world applications. The first part of the tutorial would focus on introductory concepts of information retrieval, handwriting recognition and image processing. Text based retrieval models involving noisy text retrieval as well as adaptation of traditional retrieval models will also be described. The second part of the tutorial would survey the existing IR toolkits available to researchers and discuss important concepts from keyword spotting. Topics from recognition based as well as recognition free keyword spotting would be covered in detail. Content based retrieval of handwriting in form of signature based document retrieval, writer style retrieval and temporal document retrieval would also be discussed.

Topics to be Covered

Introduction

- **Image Processing Basics (image representation, Binarization, histogram analysis, recap of probability)**
- **Handwriting Representation (Chain code, run length encoding, ligatures and strokes, projection profiles)**
- **Information Retrieval Basics (Bag-of-word representation, Vector Space Model - VSM, Evaluation Metrics)**

Handwriting Recognition

- **Challenges**
- **Applications (i.e. postal, bank, medical)**
- **Paradigms (lexicon-free, lexicon-driven, interactive models)**
- **Motivation for Information Retrieval**

Text Based Retrieval

- **OCR Correction Techniques**
- **LSA and VSM based Retrieval**
- **Probabilistic IR, Kernel Methods for IR**

Document Retrieval Models

- **Modified VSM**
- **MMSE Estimation Based Model**
- **Cross Lingual IR**

Demo of Existing IR Toolkits

- **Intro to Lemur, Lucene**
- **Adaptation to image based IR**

Keyword Spotting - I

- **Feature Based Approaches (i.e. Bio-features, Gabor, GLOH, GSC, Moments)**
- **Recognition-free approaches**

Keyword Spotting - II

- **Recognition Based approaches**
- **Partial OCR Indexing**
- **Probabilistic Keyword Spotting**

Content Based Retrieval

- **Signature Based Retrieval**
- **Writer Style Retrieval**
- **Temporal Retrieval**