Abstract

Digital documents are now omnipresent. Techniques and algorithms to process and understand these documents are still evolving. This thesis focuses on the non-textual documents of textual content. Example of this category are online handwritten documents and scanned printed books. Algorithms for accessing such documents at the content-level are still missing, specially for Indian Languages. This thesis addresses two fundamental problems in this area – Annotation and Retrieval. Annotated datasets of handwriting are a prerequisite for the design and training of handwriting recognition algorithms. Retrieval from annotated data sets is relatively straightforward. However retrieval from unannotated datasets is still an open problem. We explore algorithms which make these two tasks possible.

Annotation of large datasets is a tedious and expensive process. The problem becomes compounded for handwritten documents, where the characters correspond to one or more strokes. We have developed a versatile, robust annotation tool for online handwriting data. This tool is aimed at supporting the emerging UPX/hwDataset schema, a promising successor of the UNIPEN. We provide easy-to-use interface for the annotation tool. However, still the annotation is highly manual. We then propose a novel, automated method for annotation of online handwriting data at the character level, given a parallel corpus of online handwritten data and typed text. The method employs a model-based handwriting synthesis unit to map the two corpora to the same space. Annotation is then propagated to the word level and finally to the individual characters using elastic matching. The initial results of annotation are used to improve the handwriting synthesis model for the user under consideration, which in turn refine the annotation. The method takes care of errors in the handwriting such as spurious and missing strokes and characters. The output is stored in the UPX format.

Search and retrieval of online handwriting is gaining importance due to the increase in availability of such data. However, the problem is challenging due to variations in handwriting between various writers, digitizers and writing conditions. We propose a retrieval mechanism for online handwriting, which can handle different writing styles, specifically for Indian languages. The proposed approach provides a keyboard-based search interface, enabling the search of handwritten data from any computer, in addition to pen-based and example-based queries. Textual queries are supported for handwritten data sets with the help of a handwriting synthesis module. Synthesis of handwriting has a variety of applications including generation of personalized documents, study of writing styles, automatic generation of data for training recognizers, and matching of handwritten data for retrieval etc. Most of the existing algorithms for handwriting synthesis deal

with English, where the spatial layout of the components are relatively simple, while the cursiveness of the script introduces many challenges. We present a synthesis model for generating handwritten data for Indian languages, where the layout of characters is complex while the script is fundamentally non-cursive. The retrieval framework, which employs handwriting synthesis and holistic matching of online words, also allows for cross-lingual document retrieval across Indian languages.

We also demonstrate the retrieval scheme on a set of offline printed documents. The system for retrieval of relevant documents from large document image collections is developed by adapting existing search engines. We achieve effective search and retrieval from a large collection of printed document images by matching image features at word-level. For representations of the words, profile-based and shape-based features are employed. Our scheme groups together similar words during the indexing process. The system supports cross-lingual search using OM-Trans transliteration and a dictionary-based approach. System-level issues for retrieval (eg. scalability, effective delivery etc.) are the focus.

Digitized books and manuscripts in digital libraries are often stored as images or graphics. They are not searchable at the content level due to the lack of OCRs or poor quality of the scanned images. Portable Document Format (PDF) has emerged as the most popular document representation schema for wider access across platforms. When there is no textual (eg. UNICODE, ASCII) representation available, scanned images are stored in the graphics stream of a PDF file. We propose a novel solution to search the textual data in the graphics stream of the PDF files at content level. The proposed solution is demonstrated by enhancing an opensource PDF viewer (Xpdf). Indian language support is also provided. Users can type a word in Roman (ITRANS), view it in a font, and search in textual and graphics stream of PDF documents simultaneously.

In short, the contributions of this thesis are:

1. Development of a versatile annotation tool for online handwriting data and to store the annotation in UPX/hwDataset format, which is considered as the successor to the popular UNIPEN standard.
2. Development of an algorithm to annotate the online handwriting data with the help of an unaligned parallel text.
3. A synthesis scheme is proposed for online handwriting in Indian languages by addressing specialties of the Indian scripts. Given a text, corresponding realistic handwriting is synthesized.
4. A document retrieval system is presented for online handwriting, which does not require a recognizer. It accepts textual queries,

synthesizes the handwriting and then does a word level elastic matching for finding the most similar words.

5. Scalability of recognition-free retrieval systems is verified by adapting an existing opensource search engine for large collection of document images.

6. Portable document format (PDF) representation of document images are not content-level accessible if the data is stored in the graphics stream of the PDF. We find the application of the recognition-free retrieval scheme in the graphics stream of PDF and verify the performance by integrating it into an opensource PDF reader.