1
Introduction

For hundreds of years, papers have been the key instrument to gain perpetual mode of communication for humankind. Though there is widespread use of computers for document editing (e.g. word processors), most information is still recorded, stored and distributed in paper format. Several studies demonstrate the increase in the use of paper as a media for information exchange. In this day and age of current technological trends, we are marching towards a paperless world, yet some studies demonstrated a strong cumulative use of paper as a media for information [Lyman and Varian, 2003]. Likewise, there are still application domains where the hardcopy persists to be the preferred media [Sellen and Harper, 2003].

The objective of automatic document processing is to recognize text, graphics, and pictures in digital images and extract the intended information as a human would. In recent years, OCR technology has revolutionized the scanned document image processing and has been applied throughout the entire spectrum of industries [Lin, 2005]. OCR has enabled scanned documents to become more than just image files, turning into fully searchable documents with text content that is recognized by computers. With the help of OCR, people no longer need to manually retype important documents while entering them into electronic databases. The characters are automatically recognized and this results in accurate information processing in less time.

The work mentioned in this thesis particularly aims at

- Localizing and segmenting the handwritten annotations from printed documents.
- Identifying the author who penned the annotations on the printed documents.

Section 1.1 focuses on the research gaps that motivated us to look upon the annotations extraction problem as a contribution to research. Section 1.2 enlists our contributions and new methods for annotation extraction and writer identification. It also entails the key insights while experimenting with various datasets and literature. Finally, Section 1.3 presents the organization of chapters in the thesis.

1.1 CHALLENGES/PROBLEMS ADDRESSED IN THIS THESIS

Many documents have both printed and handwritten content. Such mixed mode documents are termed as *annotated documents* and the handwritten content is called *annotations*. Examples of such documents are postal letters, forms, annotated study materials, examination objective sheets, bank cheques, official documents etc. Due to incongruity in the properties of handwritten and printed text, OCR's process both of them differently [Pal and Chaudhuri, 1999]. This indicates the requirement for the separation of handwriting from the printed text. Hence for accurate recognition, the separation task is formulated as a simple two-class problem where one is the handwritten content while the other is printed content.

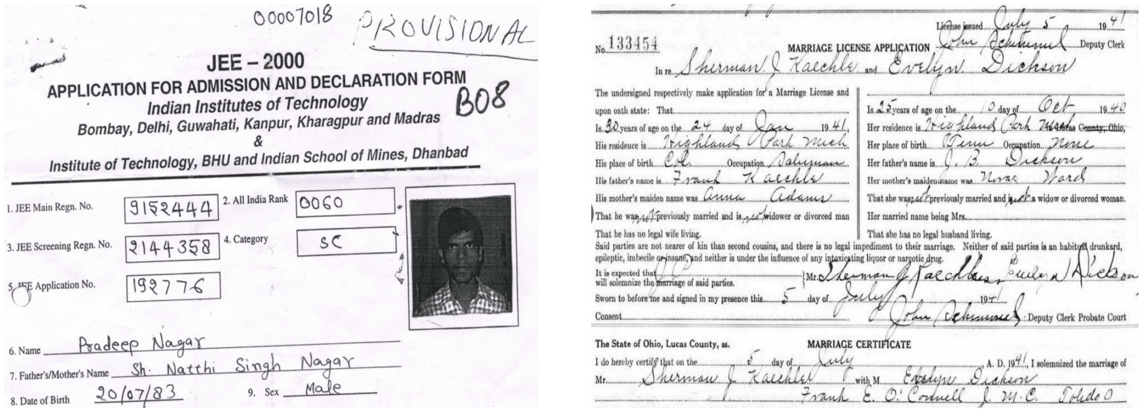
The problem of segmentation of annotations from printed text has been addressed by several research works and good results have been obtained. The recognition of free-form handwritten annotations still remains a considerable challenge. With the progression of technology, improved OCR's are being developed that can recognize textual annotations in addition to standard typewritten text. Such OCR's are termed ICR (Intelligent Character Recognition). Their development has added flexibility to the realm of OCR, as recognition process no longer depends on the conformity of text to strict standards for the shape of the characters. For such systems, segmentation is typically performed at the character [Imade *et al.*, 1993; Kuhnke *et al.*, 1995; Koyama *et al.*, 2008; Song *et al.*, 2011a], text line [Fan *et al.*, 1998; Pal and Chaudhuri, 1999, 2001; Kavallieratou and Stamatatos, 2004], and connected component levels [Franke and Oberlander, 1993; Srihari *et al.*, 1996; Santos *et al.*, 2002; Kavallieratou *et al.*, 2002; Jang *et al.*, 2004; Likforman *et al.*, 2006; Shetty *et al.*, 2007; Kandan *et al.*, 2007; Chanda *et al.*, 2010; Awal and Belaïd, 2017].

Handwriting styles vary greatly from person to person. Similarly there are a number of ways by which a person can annotate a document. A single document can have varied annotations. They can be irregular and include marks, cuts, underlined text, characters; single and multiple words; overlay text and special symbols, along with the regular text. Apart from varied annotations, there exists a wide diversity in document layouts, content, quality and structure. Documents can be structured (eg. tables), semi-structured (eg. forms) and structure-free [Belaïd *et al.*, 2013]. Figure 1.1 represents the prevailing structural diversities in documents.

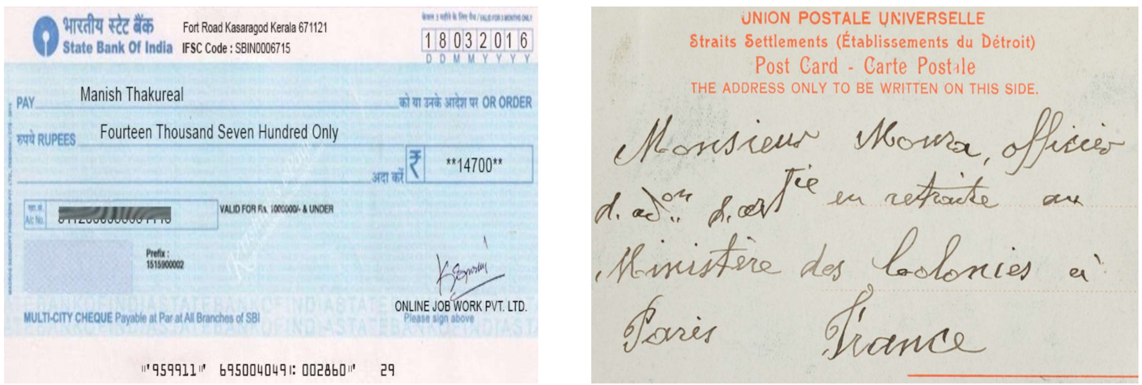
Arbitrary orientation of content in a document drastically affects the extraction of hand-made content from the printed content. There is a need to devise an approach that could encounter such adverse effects, and still maintain the reliability and robustness of ICR's.

Most of the methods are proposed for textual annotation extraction. Although, few works [Guo and Ma, 2001; Peng *et al.*, 2013; Seuret *et al.*, 2014] are available for separating overlapping or overlay handwritten annotation on the printed text. These encompass documents with homogeneous layouts, especially structured or semi-structured documents. Such environment is termed as controlled environment where we restrict the location of the annotation. A more complicated environment exists which concerns with non-predictable layouts having structure-free documents. Writing within the margins, between the paragraphs, multi-oriented text-lines, overlapping with the printed text, and presence of symbolic annotations like arrows, underlines, cuts, and encirclement are the examples of the unconstrained annotations. Annotations on these documents are marked in an unstructured way which results in an unconstrained non-controlled environment. Figure 1.2 shows the specimens of the unconstrained annotations for non-controlled environment. Consequently, extracting multi-oriented handwritten annotations in varied layouts remains a difficult task. As futuristic ICR technology improves, there arises this need to automate more and more processes and to separate variety of annotations from unconstrained document environment. Developing techniques for robust extraction of handwritten content from heterogeneous layouts is the need of the hour.

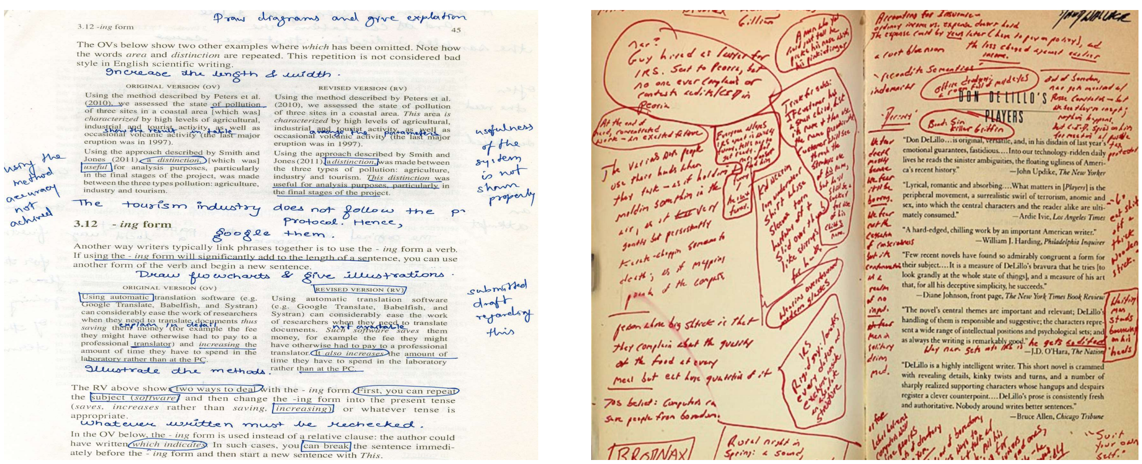
Industrial grade systems have the added requirement of faster recognition speed. Modern OCR technology is near-perfect, however, most of the trainable systems cannot be adequately trained with labeled training data. It is desirable to have trainable systems for document automation which can give accurate performance with less training data. ICR systems should be capable of interpreting handwritten text and characters, extracting various quotations and automatically labelling them according to their type. Sometimes, it may be of interest to know who edited a document or wrote a specific word on the document. Handwriting as a personal biometric is considered to be unique to a person [Pervouchine and Leedham, 2007]. Every individual has a consistent handwriting which is distinct from the handwriting of another individual [Srihari *et al.*, 2001]. Hence, the individuality of handwriting makes the identification of the writer of a handwritten sample possible. Trainable systems have not yet succeeded in giving satisfactory results because of (a) variability and variation of handwritten patterns, (b) the limited amount of



Semi-Structured Documents



Structured Documents



Structure-free Documents

Figure 1.1 : Structural Diversities Among Documents

image data, and (c) the presence of noise patterns.

Much of writer identification is linked with forensic and criminology applications. This area still requires a great deal of research, due to the difficulty of interpreting handwriting styles that vary greatly from person to person. Intra-writer variability adds difficulty while learning the variability among different writers. To overcome this, efficient feature extraction schemes are needed, along with methods that are robust towards intra-writer variability.

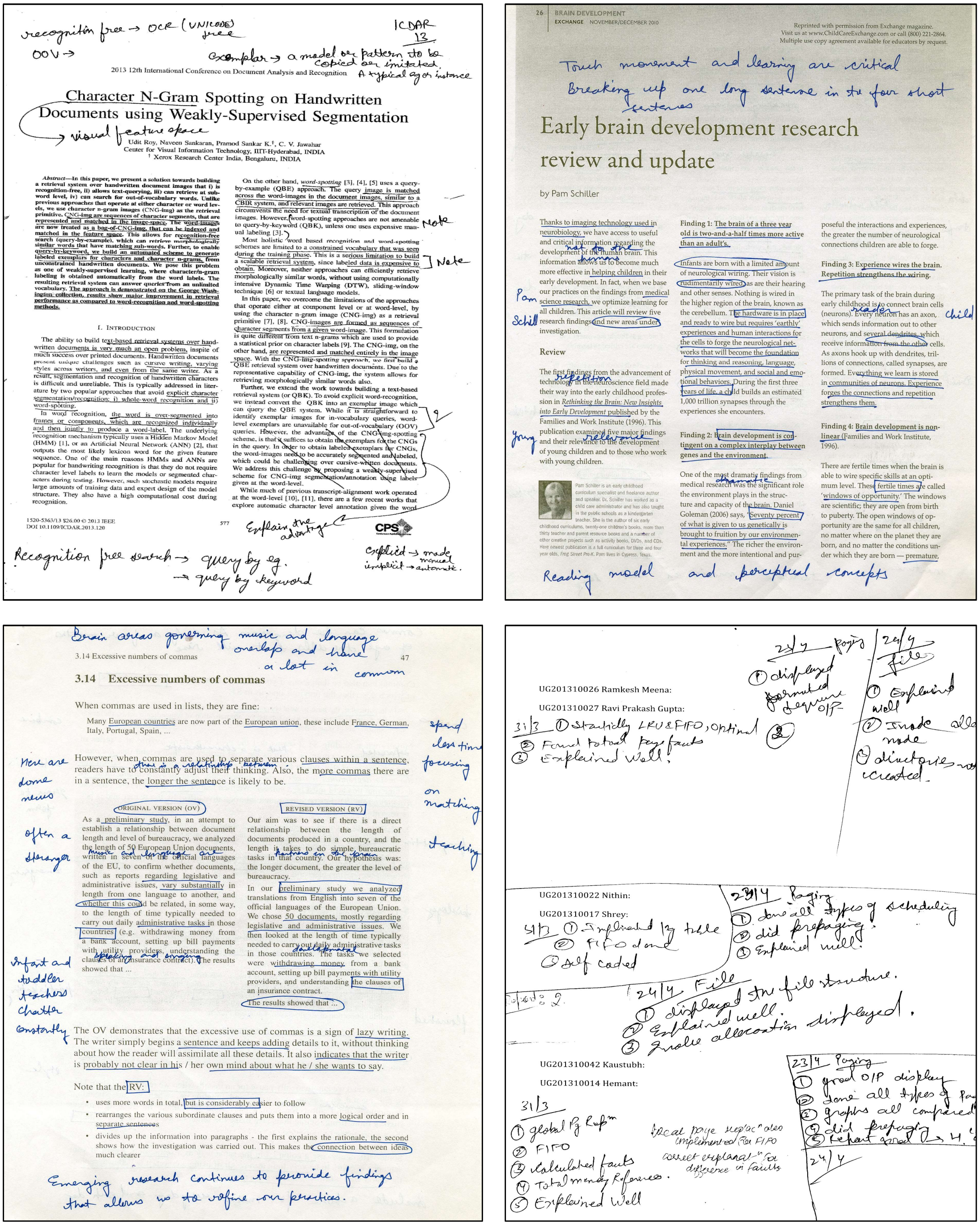


Figure 1.2 : Unconstrained Annotations for Non-controlled Environment. Writing within the margins, between the paragraphs, multi-oriented text-lines, overlapping with the printed text, and presence of symbolic annotations like arrows, underlines, cuts, and encirclement are the examples of the unconstrained annotations.

In many application domains, a document passes through a hierarchy of officials.

Consequently, it gets annotated by multiple writers in multiple ways. Under such circumstances, a robust application is demanded that can segment the required annotation and identify the author for each segment.

1.2 OUR CONTRIBUTIONS

Localization and segmentation of annotations from varied document images is a challenging problem. This thesis presents approaches to localize generic annotations and to identify the writer for handwritten text. The developed methods are applicable to a diverse collection of printed document images such as books, magazines, newspapers, scientific research papers and official documents. Moreover, these documents may be scanned or camera-captured, binary or grayscale, noisy or bleed through.

This thesis contributes as follows:

1. For documents comprising mixed content unconstrained annotations are segmented in an unsupervised manner. Not only textual annotations are extracted but also overlapping, encircled, underlines, arrows, and special symbols are also extracted.

2. We designed a novel feature called Envelope Straightness that separated printed text from complex annotations.

3. We categorized annotations into multiple types and developed a method to identify a specified type of annotated region among other types of annotations. If a document has been annotated by multiple writers, the method can identify the writer for every handwritten word. The method makes use of graphemes in the word to recognize a particular word. We developed a new method for detecting core regions of a handwritten word. Accurate detection of core-region makes the extracted features robust to handle the diversity in annotation.

4. We investigate the use of two top-down visual saliency models for categorizing annotations. The first model makes use of supervised learning in the form of conditional random fields with a sparse encoding of feature vectors. The second model makes use of a weakly supervised learning formulation for discriminant saliency.

5. We created dataset and ground truth for annotation extraction and core region detection problem.

1.3 ORGANIZATION OF THIS THESIS

Figure 1.3 illustrates the organization of this thesis.

1.3.1 State of the Art: Printed and Handwritten Content Separation (Chapter 2)

This chapter provides a review of the methods that have been used to separate printed text from handwritten text. The distinguishing properties of handwritten and printed text are presented in detail. Past work addressing segmentation of handwritten and printed text has been reviewed at 6 levels: pixel level, word level, line level, block level, character level and connected component level. Several feature extraction techniques are reviewed and their limitations are analyzed. Information about public and unpublished datasets is given and the challenges relating to handwriting segmentation are discussed. In the end, the chapter gives an overview of the miscellaneous applications of annotation extraction pertaining to ease the real-life challenges.

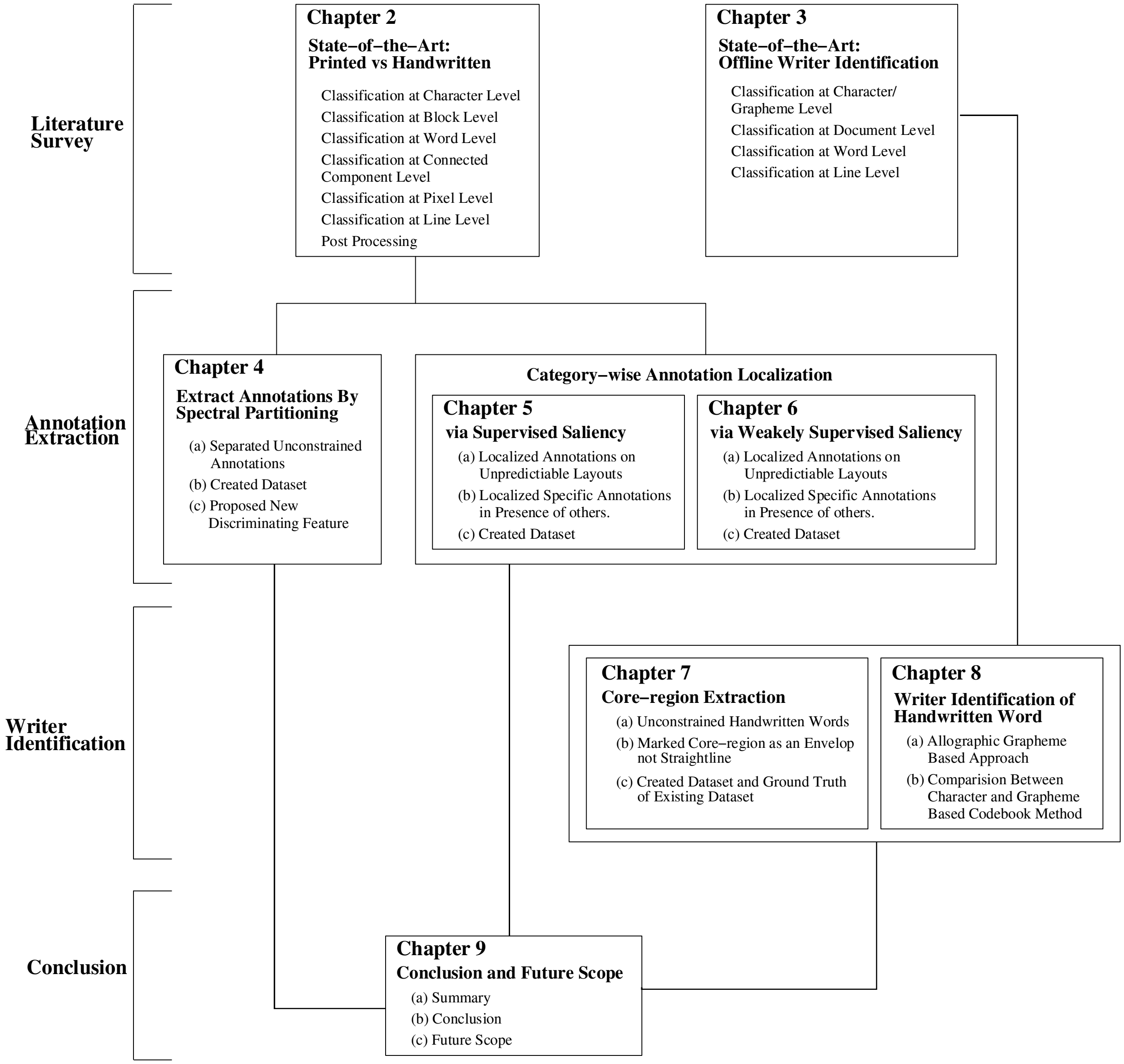


Figure 1.3 : Organization of the Thesis

1.3.2 State of the Art: Writer Identification (Chapter 3)

This chapter discusses the variability inherent in different writing styles. It reviews feature extraction techniques that can qualify intra-class variability and enhance inter-class variability. It reviews different methods for writer identification of words, line, block and complete documents. Performance evaluation methods on several public and unpublished datasets are presented. In the end, the chapter illustrates the miscellaneous applications of writer identification and highlights the challenges that remain.

1.3.3 Printed and Handwritten Annotation Segmentation using Spectral Partitioning (Chapter 4)

This chapter addresses the problem of segmenting handwritten annotations on scientific research papers. It deals with documents that have multi-oriented handwritten annotations rather than annotations in controlled scenario in which the annotations share the same orientation as that of printed text. Most of the previous methods use datasets which have well-separated hand-written text in a predefined layout. This chapter proposes a new approach to identify all possible types of annotations that normally readers make on a document while reading or editing. It presents an unsupervised method using spectral partitioning to segment several types of complex annotations. For experimentation, we created a dataset comprising research proceeding papers annotated by a single writer, including blank regions, lines, printed and handwritten words and digits, noise, underlined text and special annotation symbols.

The segmentation scheme makes use of spectral partitioning to geometrically segment the complex cases of handwritten annotations, including marks, cuts and special symbols along with the regular text. The performance of the proposed method of spectral partitioning is compared with hierarchical clustering and partitional clustering. A new feature Envelope Straightness is developed and included in the feature set. This feature is shown to provide significant discriminating capability among the two classes. This feature gives improved accuracy when used along with conventional features.

1.3.4 Annotation Localization using a Supervised Model for Top-down Visual Saliency (Chapter 5)

This chapter deals with spotting annotations of a specific type on a printed document. It presents the motivation for the task of extracting specific annotated regions in a document. The chapter illustrates the categorization of annotations into two broad categories: Textual and Symbolic. The textual annotations are further divided into Marginal and Inline annotations. The symbolic annotations are subcategorized as Arrows, Encirclements, and Underlines. Visual attention models are adopted to mimic the human behavior of spotting specific annotated regions on a document. The chapter presents a conceptual background of Visual Saliency and gives an overview of the related models.

A supervised approach is adopted to decide on the presence or absence of annotations in local patches extracted from a given image. In general, printed and handwritten text tends to possess large semantic and geometric ambiguities. Therefore to select the most distinctive parts we need selectivity among the feature set which can be incorporated by sparse coding. Thus to implement sparse codes as latent variables, we train a dictionary modulated by CRF. The joint learning of dictionary and CRF parameters is inspired by the max-margin learning approaches. Once the dictionary and optimal CRF parameters are learned the saliency map of a test image is constructed. For each test image, the presence or absence of an annotation in an image patch is inferred by message passing algorithms. Each test image patch gets its saliency value as the posterior probability of annotation. The saliency map for a given document is constructed by normalizing the posterior probabilities of patches within their context.

1.3.5 Annotation Localization using a Weakly Supervised Model for Top-down Visual Saliency (Chapter 6)

This chapter develops a weakly supervised model to localize category-wise annotations. The task of localizing the annotation defined in this chapter is a one-versus-all classification problem. This defines two groups of stimuli. The first group comprises images containing an object class of interest forming target hypothesis, while the other contains the rest of the classes forming a null hypothesis.

A top-down saliency model is deployed to localize annotations in a weakly supervised manner. To implement top-down saliency, a formulation based on Discriminant Saliency is applied to spot specific handmade annotations in a document. According to discriminant saliency, the salient features of a target class are those, which most efficiently distinguish target class from all other visual classes of recognition interest. To achieve this goal, discriminant saliency is defined with respect to two classes of stimuli: a target class corresponding to stimuli that contributes to top-down saliency and a null hypothesis comprising all the stimuli that are not salient. A formulation is presented for the two pertinent aspects of discriminant saliency: feature selection and saliency detection. This means that in order to best distinguish a target class from all other visual classes, the salient features of the target class must be identified. The implementation of these two fundamental operations is based on Barlow's principle. Barlow proposed an organizational principle for unsupervised learning based on information theory. He pointed to redundancy reduction at several levels of the visual system. According to Barlow's theory, a system detects new statistical regularities in the sensory input that differ from the environment to which the system has been adapted. This chapter explains the combination of Barlow's principle with Information Theoretic decision theory to select the features that deliver the most distinguishable information about the salient object. Finally, the selected features generate a Generalized Gaussian distribution whose parameters are used to define the discriminant saliency to find salient objects in a test document.

1.3.6 Core Region Extraction for Off-line Unconstrained Handwritten Words (Chapter 7)

This chapter describes a novel approach to find the ascender and descender regions from an unconstrained handwritten word. The method estimates correct core-region for complexities like long horizontal strokes, skewed words, first letter capital, hill and dale writing, jumping baselines and words with long descender curves, cursive handwriting, calligraphic words, title case words, and very short words. This method provides a better result in comparison with the *state-of-the-art* core region extraction methods.

1.3.7 Writer Identification for Handwritten Words (Chapter 8)

This chapter addresses the task of identifying the writers for handwritten words using features drawn from allographs. This chapter develops an approach based on construction of a codebook of graphemes using a sliding window. The grapheme based codebook is build with overlapping and non-overlapping windows. K-means clustering is used for codebook generation and one vs. rest SVM is used for further classification. In the end, majority voting decides the author of the given handwritten word. This chapter presents a performance analysis for two codebooks built using graphemes and characters.

1.3.8 Conclusion and Future Scope (Chapter 9)

This chapter summarizes the research work done in this thesis and provides a conclusion with useful insights about annotation localization, categories of annotations, performance analysis of the supervised and weakly supervised saliency-based methods, writer identification, and intra-class variability reduction. It also mentions the open challenges and the problems related to the above-mentioned topics for future extension.

…