

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

00007018 PROVISIONAL

JEE - 2000
APPLICATION FOR ADMISSION AND DECLARATION FORM
Indian Institutes of Technology
Bombay, Delhi, Guwahati, Kanpur, Kharagpur and Madras
&
Institute of Technology, BHU and Indian School of Mines, Dhanbad

B08

1. JEE Main Regn. No. 9152444 2. All India Rank 0060


3. JEE Screening Regn. No. 2144358 4. Category SC

5. JEE Application No. 192776

6. Name Pradeep Nagar

7. Father's/Mother's Name Sh. Natthi Singh Nagar

8. Date of Birth 20/07/83 9. Sex Male



Litho Form No. 133454

Date July 5 1941

Deputy Clerk

MARRIAGE LICENSE APPLICATION

In re. Sherman J. Kachela and Evelyn Dickson

The undersigned respectively make application for a Marriage License and upon oath state: That

He is 20 years of age on the 24 day of Jan 1941

His residence is Highland Park

His place of birth Franklin County, Ohio

His father's name is J. S. Dickson Occupation Farmer

His mother's maiden name was Anna Adams

That she was never previously married and is not a widow or divorced woman.

Her residence is Highland Park

Her place of birth Franklin County, Ohio

Her father's name is J. S. Dickson Occupation Farmer

Her mother's maiden name was Anna Adams

That she has no legal husband living.

Said parties are not near of kin than second cousins, and there is no legal impediment to their marriage. Neither of said parties is an habitual drunkard, epileptic, imbecile or insane, and neither is under the influence of any intoxicating liquor or narcotic drug.

It is expected that Mr. Sherman J. Kachela and Evelyn Dickson will solemnize the marriage of said parties.

Seems to be before me and signed in my presence this 5 day of July 1941

Deputy Clerk Probate Court

MARRIAGE CERTIFICATE

I do hereby certify that on the 5 day of July A. D. 1941, I solemnized the marriage of Mr. Sherman J. Kachela with M. Evelyn Dickson

Frank C. O. Conwell, J. W. C. P. D. O.

Semi-Structured Documents

भारतीय स्टेट बैंक Fort Road Kasaragod Kerala 671121

State Bank of India IFSC Code : SBIN0006715

18032016
DDMMYYYY

PAY Manish Thakural को पा उनके आदेश पर OR ORDER

रुपये RUPEES Fourteen Thousand Seven Hundred Only

अंश को ₹ **14700**

VALID FOR PA. 1000000 & UNDER

Prefix: 151500002

ONLINE JOB WORK PVT. LTD.

MULTI-CITY CHEQUE Payable at Par at All Branches of SBI

959911 695004049: 002860 29

UNION POSTALE UNIVERSELLE
Straits Settlements (Établissements du Détroit)
Post Card - Carte Postale
THE ADDRESS ONLY TO BE WRITTEN ON THIS SIDE.

*Monsieur Moura, officier
d. ad. d. art. en retraite au
Ministère des Colonies à
Paris
France*

Structured Documents

3.12 -ing form

The OV's below show two other examples where which has been omitted. Note how the words *area* and *distinction* are repeated. This repetition is not considered bad style in English scientific writing.

Increase the length & width.

Using the method described by Jones et al. (2010), we assessed the state of pollution of three sites in a coastal area (which was characterized by high levels of agricultural, industrial, and urban activity) as well as an occasional wildlife activity (the major criterion was in 1997).

Using the approach described by Smith and Jones (2011), we assessed the state of pollution of three sites in a coastal area. This area is characterized by high levels of agricultural, industrial, and urban activity (the major criterion was in 1997).

The tourism industry does not follow the protocol. Hence,

3.12 -ing form

Another way writers typically link phrases together is to use the -ing form a verb. If using the -ing form will significantly add to the length of a sentence, you can use another form of the verb and begin a new sentence.

Download the software & give illustrations.

Using automatic translation software (e.g. Google Translate, Babelfish, and Systran) can considerably ease the work of researchers when they need to translate scientific documents. However, the fee they might have otherwise had to pay to a professional translator and increasing the amount of time they have to spend in the laboratory rather than at the PC.

Submit the draft regarding this.

Whatever written must be checked.

In the OV below, the -ing form is used instead of a relative clause; the author could have written *the -ing form*. In such cases, you can break the sentence immediately before the -ing form and then start a new sentence with *This*.

Handwritten notes and diagrams on a page of text. The notes are written in red ink and include phrases like 'The OV's below show two other examples where which has been omitted', 'Increase the length & width', and 'Using the method described by Jones et al. (2010)'. There are also diagrams and arrows pointing to specific parts of the text.

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.

recognition free → OCR (VUI/008) ICDAR 13

OOV → *examples → a model or pattern to be copied or imitated. A typical eg of instance.*

2013 12th International Conference on Document Analysis and Recognition

Character N-Gram Spotting on Handwritten Documents using Weakly-Supervised Segmentation

visual feature space

Udit Roy, Navdeep Sankaran, Pramod Sankar R. J., C. V. Jawahar
Center for Visual Information Technology, IIT Hyderabad, India
Xerox Research Center India, Bengaluru, INDIA

Abstract—In this paper, we present a solution towards building a retrieval system over handwritten document images that is recognition free. It allows text-querying. It can retrieve at sub-word level. In search for individual characters or word-level patterns that operate at either character or word-level, we use character n-gram images (CNG-Img) as the retrieval primitive. CNG-Img is a character n-gram image that is represented as a matrix of character segments, that are not yet treated as a linguistic unit. The word images are matched in the feature space. This allows for recognition free search query formulation, which can retrieve semantically similar words that have matching sub-words. Further, to enable search flexibility, we build an information scheme as general labeled exemplars for characters and character n-grams, from unannotated handwritten documents. We provide a pipeline of weakly-supervised learning, where character n-gram labeling is obtained automatically from the word labels. The resulting retrieval system can answer queries from an unannotated vocabulary. The dataset is demonstrated on the Google Webfont collection, results show, making improvement in retrieval performance as compared to word-recognition and word-embedding methods.

1. INTRODUCTION

The ability to build text-based retrieval systems over handwritten documents is very much an open problem, despite of much research over printed documents. Handwritten documents possess unique variations such as cursive writing, varying styles across writers, and even from the same writer. As a result, segmentation and recognition of handwritten characters is difficult and unreliable. This is typically addressed in literature by two popular approaches that avoid explicit character segmentation/recognition, i) whole-word recognition and ii) word spotting.

In word recognition, the word is over-segmented into frames of characters, which are recognized individually using their labels, to produce a word-label. The underlying recognition mechanism typically uses a Hidden Markov Model (HMM) [1], or an Artificial Neural Network (ANN) [2] that outputs the most likely lexicon word for the given feature sequence. One of the main reasons HMMs and ANNs are popular for handwriting recognition is that they do not require character level labels to train the models or segmented character during testing. However, such models require large amounts of training data and expert design of the model structure. They also have a high computational cost during recognition.

On the other hand, word-spotting [3], [4], [5] uses a query-by-example (QBE) approach. The query image is matched across the word-images in the document images, similar to a CBIR system, and relevant images are retrieved. This approach circumvents the need for textual transcription of the document images. However, word-spotting approaches are not amenable to query-by-word (QBW), unless one uses expensive manual labeling [3].

Most holistic level based recognition and word-spotting schemes are limited to a constrained vocabulary that can be used during the training phase. This is a serious limitation to build a retrieval system, since labeled data is difficult to obtain. Moreover, neither approaches can efficiently retrieve morphologically similar words, without using computationally intensive Dynamic Time Warping (DTW), sliding window technique [6] for textual language models.

In this paper, we overcome the limitations of the approaches that operate either at component level or at word-level, by using the character n-gram image (CNG-Img) as a retrieval primitive [7], [8]. CNG-Images are framed as sequences of character segments from a given word-image. This formulation is quite different from text-n-grams which are used to provide a statistical prior on character labels [9]. The CNG-Img, on the other hand, are represented and matched entirely in the image space. With the CNF-Img-Query-Response, we can build a QBE retrieval system over handwritten documents. Due to the representative capability of CNG-Img, the system allows for retrieving morphologically similar words also.

Further, we extend the work towards building a text-based retrieval system (or QBR) to avoid explicit word-recognition, we instead convert the QBR into an exemplar image, which can query the QBE system. While it is straightforward to identify exemplar images for in-vocabulary queries, word-level exemplars are unavailable for out-of-vocabulary (OOV) queries. However, the advantage of the CNG-Img-spotting scheme, is that it allows to obtain labels/exemplars for the CNGs in the query. In order to obtain labels/exemplars for the CNGs, the word-images need to be accurately segmented/annotated, which could be challenging over cursive-written documents. We address this challenge by proposing a weakly-supervised scheme for CNG-Img segmentation/annotation using labels given at the word-level. In [10], [11], there are a few recent works that explore automatic character level annotation given the word-label.

While much of previous transcript-alignment work operated at the word-level [10], [11], there are a few recent works that explore automatic character level annotation given the word-label.

ICDAR 13

Examples → a model or pattern to be copied or imitated. A typical eg of instance.

2013 12th International Conference on Document Analysis and Recognition

Character N-Gram Spotting on Handwritten Documents using Weakly-Supervised Segmentation

visual feature space

Udit Roy, Navdeep Sankaran, Pramod Sankar R. J., C. V. Jawahar

Center for Visual Information Technology, IIT Hyderabad, India

Xerox Research Center India, Bengaluru, INDIA

Abstract—In this paper, we present a solution towards building a retrieval system over handwritten document images that is recognition free. It allows text-querying. It can retrieve at sub-word level. In search for individual characters or word-level patterns that operate at either character or word-level, we use character n-gram images (CNG-Img) as the retrieval primitive. CNG-Img is a character n-gram image that is represented as a matrix of character segments, that are not yet treated as a linguistic unit. The word images are matched in the feature space. This allows for recognition free search query formulation, which can retrieve semantically similar words that have matching sub-words. Further, to enable search flexibility, we build an information scheme as general labeled exemplars for characters and character n-grams, from unannotated handwritten documents. We provide a pipeline of weakly-supervised learning, where character n-gram labeling is obtained automatically from the word labels. The resulting retrieval system can answer queries from an unannotated vocabulary. The dataset is demonstrated on the Google Webfont collection, results show, making improvement in retrieval performance as compared to word-recognition and word-embedding methods.

1. INTRODUCTION

The ability to build text-based retrieval systems over handwritten documents is very much an open problem, despite of much research over printed documents. Handwritten documents possess unique variations such as cursive writing, varying styles across writers, and even from the same writer. As a result, segmentation and recognition of handwritten characters is difficult and unreliable. This is typically addressed in literature by two popular approaches that avoid explicit character segmentation/recognition, i) whole-word recognition and ii) word spotting.

In word recognition, the word is over-segmented into frames of characters, which are recognized individually using their labels, to produce a word-label. The underlying recognition mechanism typically uses a Hidden Markov Model (HMM) [1], or an Artificial Neural Network (ANN) [2] that outputs the most likely lexicon word for the given feature sequence. One of the main reasons HMMs and ANNs are popular for handwriting recognition is that they do not require character level labels to train the models or segmented character during testing. However, such models require large amounts of training data and expert design of the model structure. They also have a high computational cost during recognition.

On the other hand, word-spotting [3], [4], [5] uses a query-by-example (QBE) approach. The query image is matched across the word-images in the document images, similar to a CBIR system, and relevant images are retrieved. This approach circumvents the need for textual transcription of the document images. However, word-spotting approaches are not amenable to query-by-word (QBW), unless one uses expensive manual labeling [3].

Most holistic level based recognition and word-spotting schemes are limited to a constrained vocabulary that can be used during the training phase. This is a serious limitation to build a retrieval system, since labeled data is difficult to obtain. Moreover, neither approaches can efficiently retrieve morphologically similar words, without using computationally intensive Dynamic Time Warping (DTW), sliding window technique [6] for textual language models.

In this paper, we overcome the limitations of the approaches that operate either at component level or at word-level, by using the character n-gram image (CNG-Img) as a retrieval primitive [7], [8]. CNG-Images are framed as sequences of character segments from a given word-image. This formulation is quite different from text-n-grams which are used to provide a statistical prior on character labels [9]. The CNG-Img, on the other hand, are represented and matched entirely in the image space. With the CNF-Img-Query-Response, we can build a QBE retrieval system over handwritten documents. Due to the representative capability of CNG-Img, the system allows for retrieving morphologically similar words also.

Further, we extend the work towards building a text-based retrieval system (or QBR) to avoid explicit word-recognition, we instead convert the QBR into an exemplar image, which can query the QBE system. While it is straightforward to identify exemplar images for in-vocabulary queries, word-level exemplars are unavailable for out-of-vocabulary (OOV) queries. However, the advantage of the CNG-Img-spotting scheme, is that it allows to obtain labels/exemplars for the CNGs in the query. In order to obtain labels/exemplars for the CNGs, the word-images need to be accurately segmented/annotated, which could be challenging over cursive-written documents. We address this challenge by proposing a weakly-supervised scheme for CNG-Img segmentation/annotation using labels given at the word-level. In [10], [11], there are a few recent works that explore automatic character level annotation given the word-label.

While much of previous transcript-alignment work operated at the word-level [10], [11], there are a few recent works that explore automatic character level annotation given the word-label.

1520-5295/13/2600-0000 © 2013 IEEE
DOI: 10.1109/ICDAR.2013.128

277

Explain the advantage of CPS

Applied → made personal implicit → automatic.

Recognition free search → query by eq. → query by keyword

26 BRAIN DEVELOPMENT
EXCHANGE NOVEMBER/DECEMBER 2010

Reprinted with permission from Exchange magazine.
Visit us at www.ChildExchangeMag.com or call (800) 221-8264.
Multiple use copy agreement available for educators by request.

Early brain development research review and update

by Pam Schiller

Touch, movement and learning are critical

Breaking up one long sentence in to four short sentences

Thanks to imaging technology used in neurobiology, we have access to useful and critical information regarding the development of the brain. This information ~~is~~ *is* more and more effective in helping children in their early development. In fact, when we base our practices on the findings from medical science research, we optimize learning for all children. This article will review five research findings and new areas under investigation.

Review

The findings from the advancement of technology in neuroscience field made their way into the early childhood profession in *Reframing the Brain: New Insights into Early Development* published by the Families and Work Institute (1996). This publication expounds five major findings and their relevance to the development of young children and to those who work with young children.

Finding 1: The brain of a three year old is two-and-a-half times more active than an adult's.

Infants are born with a limited amount of neurological wiring. Their vision is intrinsically wired as are their hearing and other senses. Nothing is wired in the higher region of the brain, known as the cerebellum. Hardware is in place and ready to wire but requires 'earthy' experiences and human interactions for the cells to forge the neurological net, works that will become the foundation for thinking and reasoning, language, physical movement and social and emotional behavior. During the first three years of life, a child builds an estimated 1,000 trillion synapses through the experiences she encounters.

Finding 2: Brain development is contingent on a complex interplay between genes and the environment.

One of the most dramatic findings from medical research is that genes do not dictate the environment plays in the structure and capacity of the brain. Daniel Goleman (2006) says, seventy percent of what is given to us genetically is of what is given to us environmentally brought to fruition by our environmental experiences. The richer the environment and the more intentional and purposeful the interactions and experiences, the greater the number of neurological connections children are able to forge.

Finding 3: Experience wires the brain. Repetition strengthens the wiring.

The primary task of the brain during early childhood is to connect brain cells (neurons). Neurons has an axon, which sends information out to other neurons and several dendrites which receive information from the axons. As axons hook up with dendrites, trillions of connections, called synapses, are formed. Everything we learn is stored in connections of neurons. Experience forges the connections and repetition strengthens them.

Finding 4: Brain development is nonlinear (Families and Work Institute, 1996).

There are fertile times when the brain is able to wire specific skills at an optimum level. These fertile times are called windows of opportunity. The windows are scientific; they are open from birth to puberty. The open windows of opportunity are the same for all children, no matter where on the planet they are born, and no matter the conditions under which they are born — premature.

Reading model and perceptual concepts

Brain areas governing music and language

3.14 Excessive numbers of commas

13.14 Excessive numbers of commas

When commas are used in lists, they are fine:

Many European countries are now part of the European union, these include France, German, Italy, Portugal, Spain, ...

However, when commas are used to separate various clauses within a sentence, readers have to constantly adjust their thinking. Also, the more commas there are in a sentence, the longer the sentence is likely to be.

ORIGINAL VERSION (OV)

As a preliminary study, in an attempt to establish a relationship between document length and level of bureaucracy, we analyzed the length of 50 European Union documents, written in Section 40 of Official Languages of the EU, to confirm whether documents, such as reports regarding legislative and administrative issues, vary substantially in length from one language to another, and length from one language to another, to the length of time typically needed to carry out daily administrative tasks in those countries (e.g. withdrawing money from a bank account, setting up bill payments with utility providers, understanding the clauses of an insurance contract). The results showed that ...

REVISED VERSION (RV)

Our aim was to see if there is a direct relationship between the length of documents produced in a country, and the length it takes to do simple bureaucratic tasks in that country. Our hypothesis was: the longer document, the greater the level of bureaucracy.

In our preliminary study we analyzed translations from English into seven of the official languages of the European Union. We chose 50 documents, mostly regarding legislative and administrative issues. We then looked at the length of time typically needed to carry out daily administrative tasks in those countries. The tasks we selected were withdrawing money from a bank account, setting up bill payments with utility providers, and understanding the clauses of an insurance contract.

The results showed that ...

The OV demonstrates that the excessive use of commas is a sign of lazy writing. The writer simply begins a sentence and keeps adding details to it, without thinking about how the reader will assimilate all these details. It also indicates that the writer is probably not clear in his / her own mind about what he / she wants to say.

Note that the RV:

- uses more words in total, but is considerably shorter to follow
- rearranges the various subordinate clauses and puts them into a more logical order and in separate sentences
- divides up the information into paragraphs - the first explains the rationale, the second shows how the investigation was carried out. This makes the connection between ideas much clearer

Emerging research continues to provide findings that allows us to refine our practices.

Explain the advantage of CPS

Applied → made personal implicit → automatic.

Recognition free search → query by eq. → query by keyword

UG201310026 Ramkesh Meena:

UG201310027 Ravi Prakash Gupta:

31/3

- 1) Datability LK & FFO optional
- 2) Final total page faults
- 3) Explained well.

24/4

- 1) displayed formatted & sorted
- 2) Explained well
- 3) Inside all made
- 4) Directories not created.

UG201310022 Nithin:

UG201310017 Shrey:

31/3

- 1) Displayed by table
- 2) FIFO done
- 3) Self coded

24/4

- 1) Displayed the file structure.
- 2) Explained well.
- 3) Inode allocation displayed.

UG201310042 Kautstuh:

UG201310014 Hemant:

31/3

- 1) global of luf*
- 2) FIFO
- 3) Calculated faults
- 4) Total memory differences.
- 5) Explained Well

24/4

- 1) good O/P display
- 2) gave all types of fa
- 3) graphs all compared
- 4) did program
- 5) neat and clear

24/4

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.

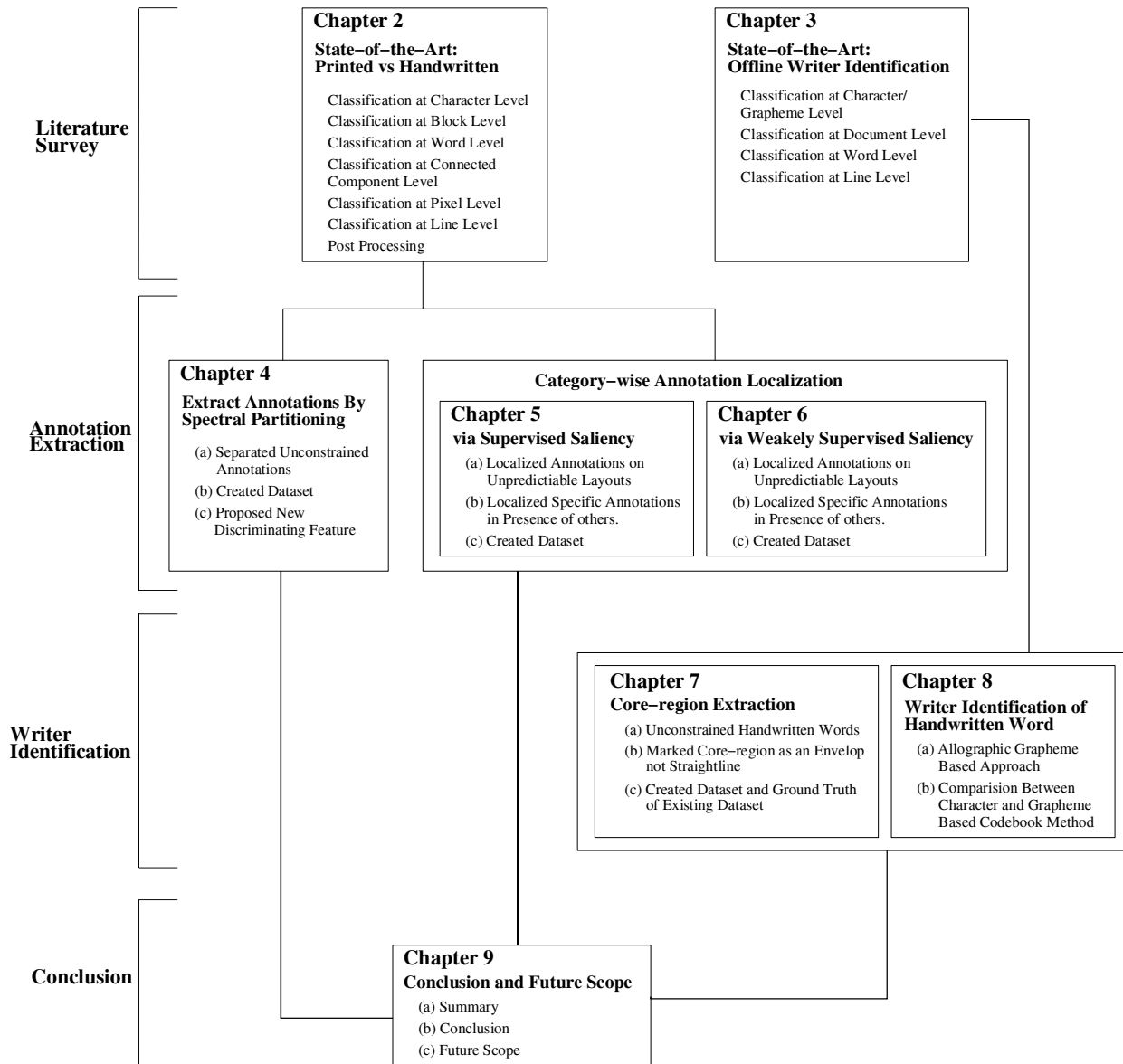


Figure 1.3 : Organization of the Thesis

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.

...