

Introduction

Documents form an indispensable part of our civilization. From ancient times, we have always encountered documents in our life, be it vedic scripts, palm leaf, newspaper, forms, etc. They are the essential component in every organisation, such as schools, hospitals, government/private offices, etc. With their increasing number, we need to preserve them and the easiest way to archive and make them accessible is to convert them to digital format.

Digitisation enables easier access to resources and preserves the historical data (Likforman-Sulem et al. [2011]). It is the process of converting physical documents into digital formats by either scanning or taking pictures from a camera. However, there are challenges involved in this process. It can result in low resolution images with degradation due to noise resulting from low hardware quality or the inherent poor quality and ageing of the document page itself. Therefore, preprocessing techniques are essential to any kind of automated analysis in order to overcome various degradations that arise due to the imaging and digitisation processes (Lu et al. [2006]; Mitianoudis and Papamarkos [2015]).

1.1 DOCUMENT IMAGE PROCESSING

Document Image Processing is the conversion of images containing textual content into machine-readable formats. It is a broad term that encapsulates various processes involved in enhancing, analyzing, and extracting information from document images. According to (Marinai [2008]), "Preprocessing operations in document image analysis transform the input image into an enhanced image more suitable for further analysis." It aims to facilitate later processing stages by providing inputs that can satisfy the requirements specified for the next stage. Popular preprocessing methods include binarization, noise removal, restoration, resolution enhancement, skew detection and correction, slant removal, layout analysis etc.

Document images primarily consist of handwritten and machine-printed text along with lines, tables, graphics, and figures. Machine-printed text in document images are usually organised as parallel text lines and have the same skew, which make their processing relatively easy as compared to handwritten text which has unconstrained skew, inter-word spacings, shapes of characters, etc. Analysis of handwritten documents faces various challenges due to unconstrained handwritten text with issues like, ink bleed, slant, skew, scaling, (Zagoris et al. [2017]) etc. When these documents are scanned, they get further degraded due to resolution issues, noise, etc (Cao and Govindaraju [2009]). These issues pose challenges for subsequent tasks like, word recognition, spotting, document classification, and retrieval (Chen et al. [2016]). Therefore, handwritten document images need more preprocessing compared to printed ones to facilitate subsequent analysis tasks.

Challenges involved in document image processing common to both handwritten and machine-printed text images include:

1. Noise and degradation: Documents may have noise like speckle noise, blur, or distortions

caused by scanning or image capture. Eliminating or decreasing these artifacts while retaining essential information is a difficulty

2. Low resolution: Low resolution images reduce the accuracy of the OCR process. Enhancing an image without compromising textual information is difficult.
3. Distortions: Skewed documents might result from scanning at an angle. Distortion may also be introduced during printing. It is crucial to rectify distortions without compromising the text's readability.
4. Multi-script: According to the "world writing system" there are 293 scripts present all over the world, with more ancient scripts still being discovered. Supporting documents may contain text in many languages or scripts. This increases the difficulty of OCR systems because of the distinct characters, syntax, and grammar rules of each language.
5. Text Recognition Accuracy: Optical Character Recognition (OCR) systems must precisely identify text from photographs despite differences in fonts, sizes, styles, and languages. Handwritten text recognition is more challenging because of the heterogeneity in handwriting styles.
6. Document Understanding: It is a more complex process which involves more than just extracting text. It also includes comprehending the semantics of the document.
7. Document classification: It involves automatically sorting documents based on their content or purpose, which can be difficult when dealing with several domains.
8. Text Recognition Accuracy: Optical Character Recognition (OCR) systems must precisely identify text from photographs despite differences in fonts, sizes, styles, and languages. Handwritten text recognition is more challenging because of the heterogeneity in handwriting styles.

Presence of these challenges makes the document image processing domain popular among researchers. With the advancement in technology, there is a shift in research methodology from handcrafted features to deep learning. Deep learning techniques however, need large amounts of labelled dataset to train the system. Annotation of document images is a laborious task. Accurate annotations cannot be performed by automated systems. Presence of these challenges motivates us to work in this area. Our aim is to improve the image quality for different tasks and work with limited data by harnessing unpaired learning techniques.

1.2 PROBLEM SPACE

Document image processing is a well explored research domain. However, the crux in this field is the presence of noise and degradation. Working with clean images always adds up to the benefits of good results in a down stream tasks. But when it comes to dealing with old documents, low resolution document images, it becomes a real challenge to complete the desired task.

In literature, researchers have proposed work addressing several steps intended to produce a clean image. These include denoising, restoration, enhancement, etc. The majority of work in this thesis revolves around dealing with noisy, degraded images.

Our work addresses the task of document image quality enhancement in the application domain of newspaper and form images. In the application of form processing, we focus on form field extraction along with form field association. Word spotting in a cluttered environment also

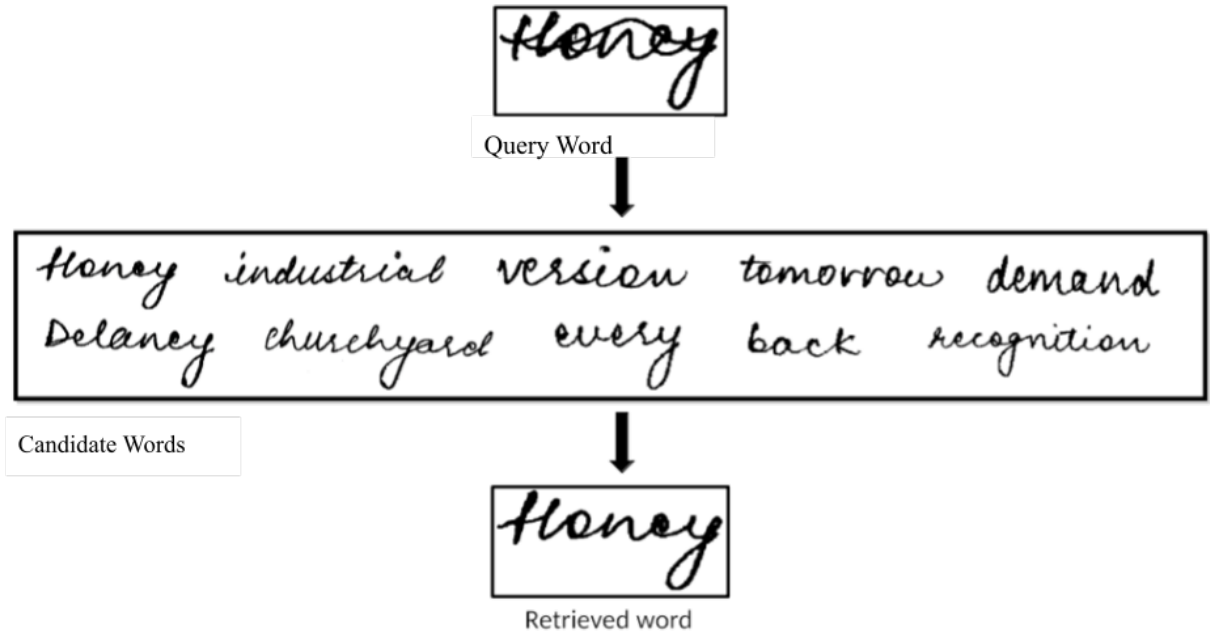


Figure 1.1 : Word spotting after distortion by strike-through

has its application on form processing, where some words can be striked out. We address the tasks of segmentation on degraded newspaper images, simultaneous denoising and resolution enhancement and selective denoising.

1.3 OUR CONTRIBUTIONS

1.3.1 Word Spotting in a cluttered medium

Handwritten words can be altered by strike-throughs, either unintentionally or purposely. While filling the forms, handwritten words can overlap with cell borders in structured documents. In this work, we present a novel problem of handwritten word spotting in a scenario where the word is cluttered by a strike-through with a line stroke. This is illustrated in Figure 1.1 These line strokes can be straight, slant, broken, continuous or wavy in nature. We have used the Vertical Projection Profile (VPP) feature and formulated its modified version, which we call as the combinatorics Vertical Projection Profile (cVPP) feature for this task. The matching strategies identifies the stroke combination in the cluttered word that produces a projection value closest to the corresponding column in the cluttered word image. A modified DTW algorithm is proposed to address the computationally expensive Brute Force matching approach for combination and selection issues. Here, the DTW cost is adjusted twice, first normalizing the text by the path length to adjust for word length, and then normalizing it by the cluttered word length after applying a penalty to ensure invariance to the presence of clutter-induced blobs. A dataset for the proposed problem is unavailable so we prepared our dataset with cluttered and uncluttered words. This work and dataset description is explained in chapter 3.

1.3.2 Cell extraction and horizontal scale correction in a form image

In structured documents like forms, cheques, etc., there is a predefined space called a frame field/cell for the user to fill the entry. Extracting these cell boundaries from form images is essential to obtain the text written inside it. In the case of a structured document, the cell boundaries are



Figure 1.2 : Horizontal Scale Correction applied on left word, output shown in right word

the line segments. Existing algorithms return all the line segments present in the form image. These line segments also include possible straight strokes in characters and shirorekha in case of Devanagari, which adds to noisy detection. Also, during digitization the document image gets degraded and the cell boundary may get distorted. To address the challenges, we propose a modified region growing algorithm to extract the cells present in the form.

When the user is writing inside a cell, the non-uniformity of inter-character spacing becomes an issue. Many times the starting characters of the word are written with normal spacing between the characters but gradually the writer starts writing with a more compact spacing so as to accommodate the word within the frame field. To deal with this variation in intra word spacing, we develop a horizontal scale correction for the text in the extracted form fields. An example result of our normalization approach is shown in figure 1.2. We have explained this work in chapter 4.

1.3.3 Field association in form images

In a typical form image, the handwritten data entries are made corresponding to the printed fields in designed places. Association of the handwritten entries to their respective printed fields can be done using a rule-based approach utilizing their spatial locations. However, the heterogeneity in the form templates, which varies from one form structure to another, renders the use of rule-based systems infeasible for this application. Hence, we focus on using machine learning techniques to perform the intended data association for a given form image.

We propose a graph-based deep network for predicting the associations pertaining to field labels and field values in heterogeneous handwritten form images. We consider forms in which the field label comprises printed text and the field value is the handwritten text. Figure 1.3 depicts our problem formulation, where the form is shown with the desired associations between the fields. Inspired by the relationship predicting capability of the graphical models, we use a Graph Autoencoder to perform the intended field label to field value association in a given form image. At a time of the work, it was the first attempt to perform label-value association in a handwritten form image using a machine learning approach. We have prepared our handwritten form image dataset comprising 300 images from 30 different templates having 10 images per template. These form images are carefully prepared keeping in view the real form images which exist in various organisations such as colleges, hospitals, government offices, recruitment agencies, insurance policies, etc. Details of methodology and dataset are mentioned in chapter 5.

1.3.4 Simultaneous denoising and super resolution

Certain documents have low resolution, which can adversely affect the performance of downstream document reading tasks. In most real cases the noise affects the document unevenly. Moreover, to train any learning-based denoising model the high resolution clean ground truth is not generally available. In this work, we address the task of enhancing the resolution of a

APPLICATION FORM
for
Student Character Certificate

Roll No. 41151135

Paste Photograph here

Student Name	<u>Shubham Goe</u>	Section	<u>TECHNICAL</u>
Department/Branch	<u>Computer Science</u>	Degree	<u>B.TECH</u>
Address	<u>H-115, E-BLOCK, Sector-E, Housing Board JODHPUR - RAJASTHAN</u>	Course	<u>IT Tech M.Tech</u>
Pin Code	<u>342001</u>	E mail ID	<u>Shubham@ gmail.com</u>
Hostel/Day School	<u>Hostel</u>	Hostel No.	<u>5</u>
		Room No.	<u>119</u>

9876543210
Mobile No.

For the use of Academic Section

Certificate No.	<u>51151</u>
Issue Date	<u>8/1/18</u>
Issued By	<u>Rajesh</u>
Signed By	<u>Rajesh</u>

Form Image with associations (red line) between field components

Figure 1.3 : Form image with association shown between fields

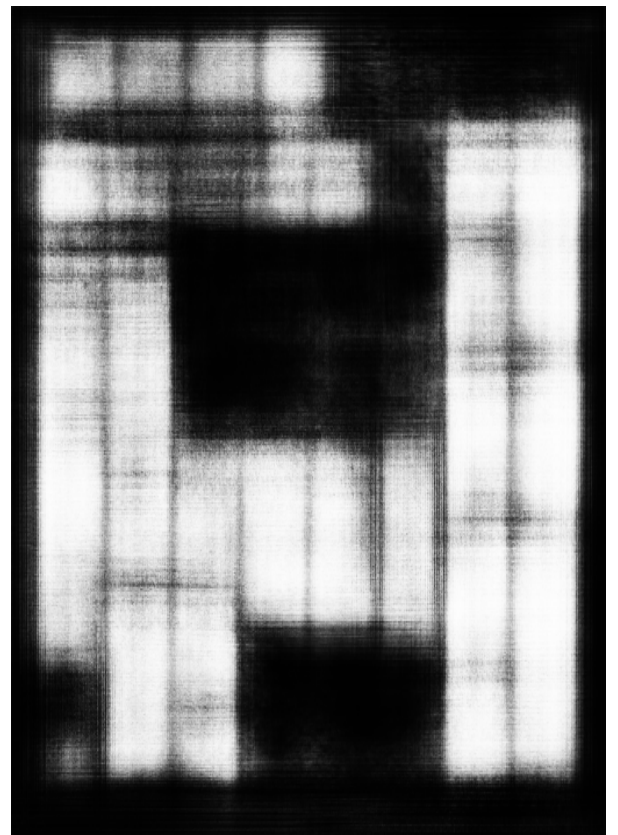


Figure 1.4 : Article Segmentation results on Bangla degraded newspaper image

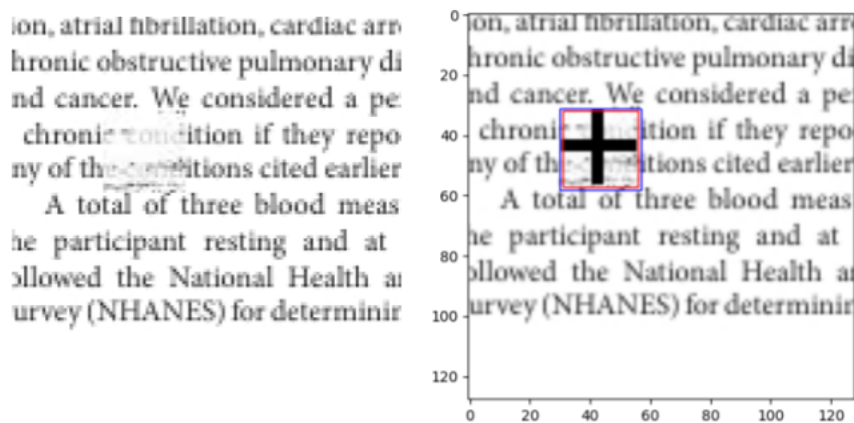


Figure 1.5 : Left: Image with noise patch. Right: Image with noise patch localized

document image while simultaneously reducing the noise content. The approach is a one shot unpaired technique where a single unpaired example is used as reference for training a SinGAN model. The training is carried out in two steps. In the first step, we use a clean reference image to train a SinGAN to learn the characteristics of the clean image. The second step then uses the given input image and performs super resolution while trying to achieve the characteristics of the reference clean image. Hence, we perform unsupervised super resolution of document images along with denoising. Our unique formulation of the loss function helps in this task by prompting the generated images to have characteristics similar to the reference clean image. We conduct experiments on publicly available datasets (Kaggle Dirty Documents Images and DIBCO) and obtain promising results. Sample results are shown in figure 1.6. We also evaluate the performance of our model for OCR and obtain a higher recognition rate compared to competing methods. Details of this work can be found in Chapter 6.

1.3.5 Selective Denoising of document images

Image denoising deals with removal of unwanted noise from images. Figure 1.5 shows the localized patch after applying the proposed reinforcement learning framework. While there are many techniques that can be applied to denoise a given input noisy image, the methods process an image in its entirety, assuming that the noise uniformly affects the entire image. For inputs where the noise affects a localised part of the image, applying methods that attempt to denoise the entire image can adversely affect the clean portions. To address this problem, we propose a deep reinforcement learning-based framework aiming to overcome this limitation and achieve better results for images with non-uniformly distributed noise. We propose a two-step procedure that first identifies the noisy patch and then denoises the extracted patch. We use a reinforcement learning-based approach for noise localization and use PixelRL(Furuta et al. [2019]) for noise removal. We have prepared a comprehensive dataset specifically for the noise localization problem in which noise patches are induced in clean document images using various noise patterns, such as Gaussian noise, coffee stains, and ink bleeds. This work is presented in Chapter 7.

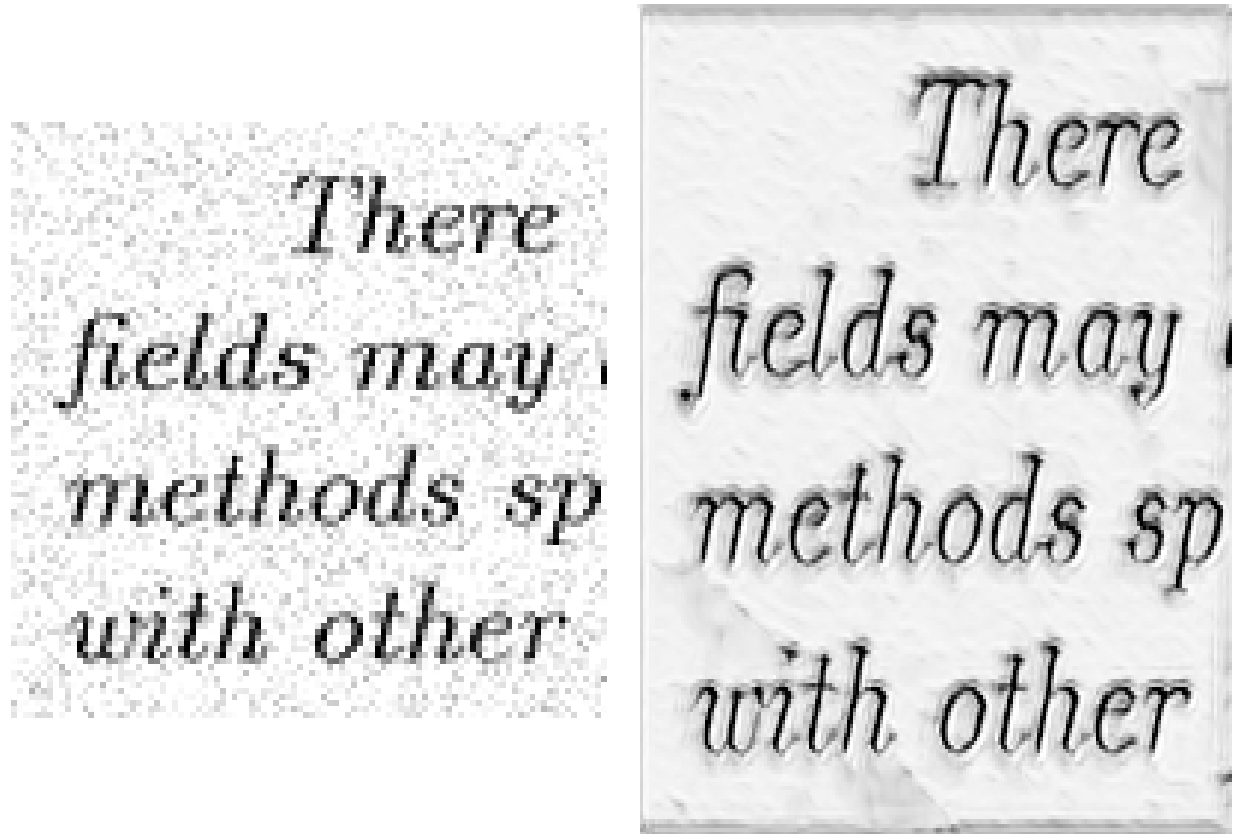


Figure 1.6 : Sample result for resolution enhancement of noisy low resolution image

1.3.6 Article Segmentation in degraded Newspaper images

The work delves into the problem of article segmentation for noisy degraded old newspaper images. The problem is formulated as a semantic segmentation task and our method uses deep fully convolutional neural networks to learn segmentation masks. The model is a deep neural network making use of convolution, dilated convolution and skip connections. The network has two branches, one for image reconstruction and the other for the main task of binary segmentation. the reconstruction branch facilitates better feature embedding in our degraded images. To handle the noise element in these images we train our model on clean and noisy newspaper images. Further, the model is fine-tuned on a noisy Bangla newspaper collection. With this experimental setting, we conducted study on article segmentation on noisy, denoised and clean form images. The proposed method is qualitatively evaluated on a collection of Bangla newspaper images published in the era from 1937-1980 and collected as part of NDLI project. Segmentation result on a sample Bangla newspaper image is shown in figure 1.4. This work is described in chapter 8.

1.4 THESIS OUTLINE

The thesis is divided into 9 chapters, as mentioned below:

- In Chapter 1, we outline the purpose and the scope of the thesis, providing an overview of our problem space and areas of interest.
- In Chapter 2, we give an overview of the related work for the problems addressed in this

thesis. The past work is organized according to the level(pixel,patch,page) of their spatial effect.

- In Chapter 3, we perform handwritten word spotting in cluttered scenarios, particularly where words are obscured by strike-through. We have developed a dataset specifically for this problem and demonstrated improved performance over existing methods in handwriting recognition in the presence of strike-through.
- In Chapter 4, we present horizontal scale correction of handwritten words in structured documents, like forms, in order to reduce the variations in intra-word spacing. Our methodology involves cell extraction and horizontal scale correction, tested on a specially created dataset comprising bilingual (English and Devanagari) forms.
- In Chapter 5, we present our work that associates different field components in heterogeneous handwritten form images. We convert form images into an undirected graph where nodes represent field components and edges signify associations. The approach utilizes a graph autoencoder technique to perform association of form fields with handwritten and printed components. We developed our dataset of handwritten forms inspired from forms present in real scenarios.
- In Chapter 6, we present a unified approach for denoising and resolution enhancement of document images. Our method uses a SinGAN model trained in a one shot mode with a single unpaired instance.
- In Chapter 7, we propose a novel problem of selective denoising targeting non-uniformly distributed noise patches in document images. Our methodology uses reinforcement learning to localize the noise patch and then denoises this patch using PixelRL approach.
- In Chapter 8, we perform article segmentation in degraded old newspaper images, specifically Bangla newspapers from 1937 to 1980. We develop a deep learning model that constitutes convolution, dilated convolution, and skip connections, in order to handle the inherent noise in these images. Training is carried out on both clean and noisy newspaper images, and fine-tuning is done on noisy Bangla images.
- In Chapter 9, we present the conclusions and future scope.

1.5 PUBLICATIONS

The work done in this thesis has been presented in the following publications.

1.5.1 Journal publications:

- Divya Srivastava and Gaurav Harit. "Simultaneous denoising and super resolution of document images." *Sādhana* 49.1 (2024): 35.
- Divya Srivastava and Gaurav Harit. "Selective Denoising in Document Images using Reinforcement Learning" Under Review.

1.5.2 Conference publications:

- Divya Srivastava and Gaurav Harit, "Word Spotting in Cluttered Environment" In Proceedings of 3rd International Conference on Computer Vision and Image Processing (pp. 161-172) Springer, Singapore.

- Divya Srivastava and Gaurav Harit, "Cell Extraction and Horizontal Scale Correction in Structured Documents", In Proceedings of 3rd International Conference on Computer Vision and Image Processing (pp. 53-64) Springer, Singapore.

1.5.3 International Workshop:

- Divya Srivastava and Gaurav Harit. "Associating Field Components in Heterogeneous Handwritten form Images using Graph Autoencoder." In 2019 International Conference on Document Analysis and Recognition Workshop on Machine Learning (ICDAR-WML), vol. 5, pp. 41-46. IEEE, 2019. [Australia Oral Paper]