

Autonomous Deblurring Images and Information Extraction from Documents Using CycleGAN and Mask RCNN

Oishee Bintey Hoque*, Maisha Binte Rashid[†], K M Tawsik Jawad[‡]
Department of Computer Science and Engineering,
Ahsanullah University of Science and Technology Dhaka, Bangladesh
{*bintu3003, [†]rashid.maisha05, [‡]tawsikzawad}@gmail.com,

Abstract—In this era of automation technology, there has been a resurgence of interest for extracting data from documents so that information can be used more efficiently. The age of storing sensitive information on paper is on the brink of extinction. Almost every organization in the world is shifting towards an efficient cloud storage system. With the global impact of Covid-19 and the norm of “work from home”, the importance of proper data extraction from digital documents has increased manifold. So, in this paper, we proposed a method to extract data from scanned document images by identifying handwritten texts and respective label fields. While scanning images various issues can appear such as - background noises, blurred images due to camera motion, out of focus images, watermarks, stains, or anything that can cause hindrance to readability of users. So, for our method to work better our first approach was to filter the scanned images by removing noises. We trained our dataset on Cycle Generative Adversarial Networks (Cycle-GAN) to generate clean scanned images from noisy images. Later on, for detecting labels and text fields we trained Mask R-CNN on our cleaned dataset. Finally, we extract the information using Tesseract from the detected fields and assign the labels filed with corresponding information on a text file.

Index Terms—document deblurring,unpaired data,cycle gan,information extraction,mask r-cnn

I. INTRODUCTION

Data is considered to be gold for large companies and organizations these days. That’s why storing information by extracting data from printed documents is of utmost importance. Critical analysis of data can help marketing analysts get a good grasp of mass preferences which will help them target their products to the proper audience. Researchers around the world would love a robust and effective data extraction tool to lessen their workload. Data is the most vital aspect for statistical information like fatality rate, birth rate, health statistics, etc. So, all of these explain the necessity of extracting vast data in a way that can be used for various

purposes. Till today, many government offices and institutions are storing sensitive information in handwritten forms. Storing and retrieving data from those documents for further use is both time-consuming and challenging. Many researchers have conducted experiments to extract handwritten information from printed documents containing basic forms. They used image processing for scanning the layout of the documents in order to detect text fields for information extraction. In our paper, we propose a method to detect labels and text fields from scanned images by training our dataset using a deep neural network. The performance of any model depends highly on how the dataset is pre-processed. So, the first step of our research is filtering out different kinds of noises from document images. The printed documents will be scanned and stored to get rid of the long process of manual data entry. But while scanning document images, numerous challenges like salt and pepper noise, stains on old documents, image blurriness due to camera motion or improper focusing, etc. can occur easily. As these noises can be detrimental to our system’s accuracy, denoising the dataset is the first step for our system to work accurately. We use Cycle-GAN to train our dataset using an image to image translation for deblurring, as it is known to give remarkable performance.

The oldest form of data entry is to fill up a printed form by handwriting, where a human can easily detect where to write the required information. But while retrieving data from document images, various challenges occur as – to detect labels and text fields, recognizing text fields with their corresponding labels. Text field recognition and data extraction are done using several methods, but the most general approach is template matching using image processing and machine learning. In template matching, the process is to define a familiar layout and train dataset using the template. A more convenient approach is to make a model that can be used in any

printed form, regardless of whether it depends on the layout of the form.

Section II contains research and analysis of related projects used for background study, Section III contains dataset information and purposes, dataset pre-processing and labeling, document deblurring and information extraction techniques, and an overview of Cycle-GAN and Mask R-CNN methods. Section IV contains details of the training process and analysis of the results obtained from Cycle-GAN and Mask R-CNN methods. Finally, Section V discusses the research overview and potential ways to improve the results of this research in the future.

II. RELATED WORKS

While scanning images from documents, several obstacles can occur such as blurred images, stains, or spots in the background, a watermark of the company, and anything that can create difficulties for extracting data. To clean images, Monika Sharma, Abhishek Verma, and Lovekesh Vig used GAN and Cycle GAN to denoise images. The image to image translation method was used by making pairs of clean and noisy images. Kaggle Document Denoising Dataset, Document Deblurring Dataset, Watermark Removal Dataset, and Document Defading Dataset were used to clean images with stains, blurriness, or watermarks. CycleGAN obtains a Peak Signal-To-Noise Ratio (PSNR) value of 31:774 dB for Kaggle Document Denoising Dataset and a PSNR value of 19:195 dB for deblurring dataset [5].

Previous works on feature extraction from documents or scanned images used algorithms like SVM, GROBID, Naive Bayes, etc. Some researchers used template matching and image processing approaches for this purpose. Template matching is one of the most popular techniques for object detection from images. Ying Yi Sun, et al. used template matching and image processing techniques to retrieve data from invoices [1]. The first step was to pre-process the models by separating the foreground and background in scenarios where the background didn't contain relevant information. Contour extraction performs well for position information using edge detection, erosion, and dilatability. Template matching was used to extract the required position. For image to text conversion, optical character recognition was helpful to obtain the expected data from invoice images. A segmentation method was applied by Yefeng Zheng, Huiping Li, and David Doermann in their paper to detect handwritten regions from document images. The first approach was to identify handwritten or printed text. It was done by finding an aspect ratio of a region depending on font size using the histogram technique and black pixel density. For classifying whether the region is handwritten or printed, Fisher classifier was used and achieved a classification accuracy of 97.3% [2]. Ahmad B. Hassanat et al. introduced an invoice classification approach using deep features and machine learning. Deep convolutional neural network AlexNet was used for feature extraction. Several machine learning algorithms such as – Random Forests, K-Nearest Neighbors, Naive Bayes were applied for classifying whether the invoice is

handwritten, printed, or receipts. Among all of the algorithms, the K-Nearest Neighbor performed most precisely with an accuracy rate of 98.4% [3].

Ozair Saleem and Seemab Latif used a hybrid approach to extract header information from research papers like title, keywords, conference name, authors, etc. To get better accuracy, three different algorithms were used to extract various fields. For extracting title field header parser tool Grobid, ParCit, and Mendeley were used. For extracting author names, affiliations, and email address fields, Grobid and ParCit were used, leading to an overall accuracy of 95.97% [4]. Rasmus Berg Palm, Ole Winther, and Florian Laws presented an invoice analysis method - CloudScan which is generalized to analyze invoices of any layouts. They used a recurrent neural network model - Long Short Term Memory(LSTM). LSTM can model context directly which helps to increase performance. Their work has achieved an accuracy of 0.84 on unseen invoice templates [9]. Jerzy Sas and Jerzy PEJCZ discussed a method of document type detection and recognition in handwritten medical text documents in their paper. For document detection, they focused on the graphical features of the documents such as horizontal and vertical line segments, and used them for template matching. They were able to find an accuracy close to 99% for document detection by using Template Matching. For document type recognition in handwritten medical text documents, they used a handwritten text recognizer to apply word classifiers based on the probabilistic lexicon prepared for each recognized text invoice templates [10].

III. ANALYSIS PLAN

In this section, we first describe the dataset that we use for our analysis, followed by our approach for cleaning the noisy documents and information extraction from documents.

A. Datasets

We employ two types of document datasets: the first one for background noise removal and the other one for information extraction with bounding boxes.

a) *Document Deblurring Dataset:* We formed an artificial deblurring dataset for our CycleGAN [6] network to deblur. We utilize the internet as the primary source to collect several types of documents admission form, NID form, etc. to keep variation in our dataset. Initially, we have downloaded 500 images (see Fig. 1), and from those, we separate 100 images for further processing. Later, each image goes through three types of blurring, i.e., stack blur, gaussian blur, and motion blur. These procedures are randomly applied to images based on some random probability, radius, kernel size, and standard deviation. From each image, we generated additional ten images and manually filtered the dataset. Finally, our dataset consisted of 1000 images in total, split into 8:2 ratio as a training and validation set. For further evaluation, we collected additional 20 images to check our model's accuracy on the real noisy dataset.

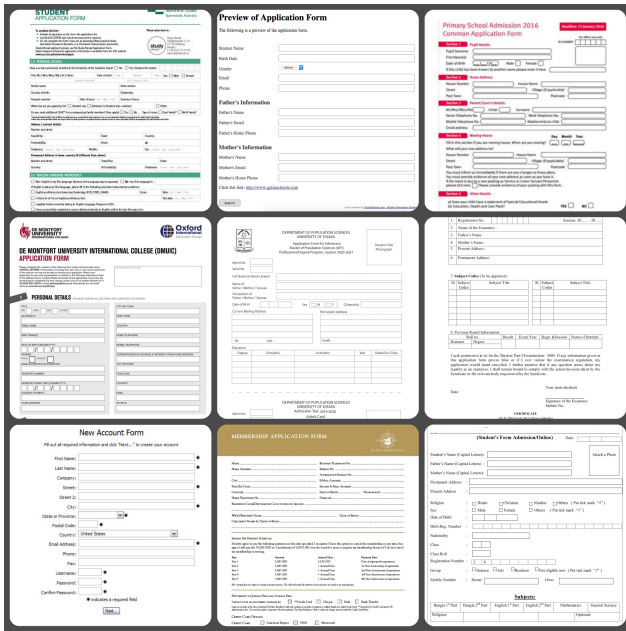


Fig. 1: Sample document images in our collected dataset.

b) Information Extraction Dataset with Bounding Boxes:

The impetus for region proposal-based networks is bounding box labeled images. But, this sort of dataset is unavailable. We generated our dataset to train the deep learning model and have made this dataset open-source for further research as well. This dataset consists of the same photos as the previous dataset. We have hand labeled each document into two types of bounding boxes - one for the label field name and another for corresponding field information. One hundred images consist of over 10,000 bounding boxes for each field. We have used `labelImg` software to generate the bounding boxes as shown in Fig. 3, and these files include class id for each box and the position (xmin,ymin,xmax,ymax). After labeling the images, we again perform several manual image transformations to make the dataset robust. Some of the conversion, such as rotating, translation, etc. - change the bounding boxes' position. We also applied changes to the bounding boxes so that the class information does not get distorted. Finally, our dataset consisted of 2000 copies in total, split into 8:2 ratio as a training and validation set.

B. Document Deblurring

We treat document deblurring as an image to image translation task - where blurred images of documents are converted to real document images. Documents can of be several types with several patterns. So, a document deblurring classifier should be accurate enough, not limited to our dataset. For this reason, we chose an unpaired image-to-image translation network - Cycle-Consistent Adversarial Networks, known as CycleGAN.

a) CycleGAN Overview: Generative Adversarial Networks (GANs) [8] have been successfully used for Image-to-Image translation tasks. Transforming images from one domain to another domain, changing art styles, going from

sketch photos to real photos, etc. have gained popularity in recent years. Since most of the time, it's not possible to get paired data in most domains, not even possible in some, the unsupervised training capabilities of CycleGAN are quite useful. CycleGAN is a GAN that uses two generators (G) and two discriminators(D), where each G has a corresponding D, which attempts to distinguish the real images from the synthesized ones. To bypass the issue of learning meaningful transformations from an unpaired dataset, CycleGAN uses cycle consistency loss. After converting an image to the other domain and back again to the source domain, we should get back something similar to the source by successively feeding it through both generators.

The first generator will generate photos of blurry documents given pictures of real ones, and the second one will do the exact opposite. The discriminators of corresponding generators will predict how accurate the generate images are. It is evident in Figure 1 that each discriminator takes two inputs - the original image in the source domain and the generated image via a generator. The task of the discriminator is to predict and defeat the generator by rejecting images generated by it. While competing against discriminator, the generator learns to produce images very close to the original input images. (see fig 4). We use the same network of CycleGAN, as proposed in paper [7].

C. Information Extraction

Information extraction from scanned forms is a significant problem domain. In this paper, we initially identify the position of the labels i.e. name, phone number, date of birth, and information field (see Fig. 3). Later, from each prediction box, we extract the information with the help of OCR technology and systematically assign them in a text file (see Fig. 6). We interpret the prediction of the label and information field problem as an object detection problem where we treat the label and information field as two different objects. For this purpose, we employ several deep learning architectures on our datasets and get the best result with region-based Mask RCNN.

a) Mask RCNN Overview: Mask RCNN is a deep neural network aimed to solve the instance segmentation problem by separating different objects in an image or a video. With a given image as input, it provides different object bounding boxes, classes, and masks present in that image. Mask RCNN consists of two stages. First, it generates proposals about the regions where there might be an object based on the input image. Second, it predicts the class of the object, refines the bounding box, and generates a mask in the object's pixel-level based on the first stage proposal. Both stages are connected to the backbone architecture - Resnet. Initially, the image passes through the backbone network and converts from 1024x1024px x 3 (RGB) to a feature map of shape 32x32x2048. This feature map becomes the input for the next stages - Feature Pyramid Network(FPN), Region Proposal Network(RPN), ROI Classifier & Bounding Box Regressor.

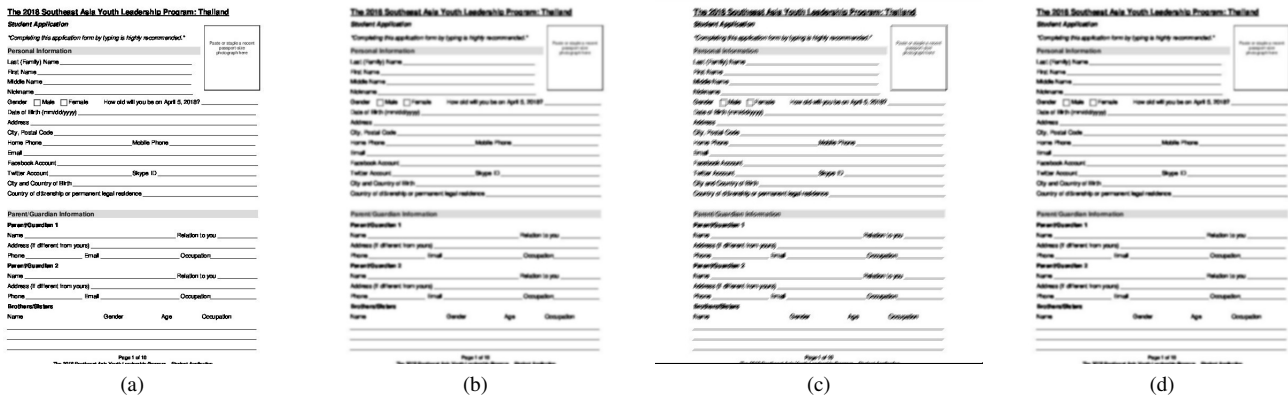


Fig. 2: Some results of randomly applied blur effect on image (a). In (b) gaussian blur has been applied with std. dev = 3 and kernel size = 13, (c) motion blur with 45 degree and (d) stack blur with radius 3.

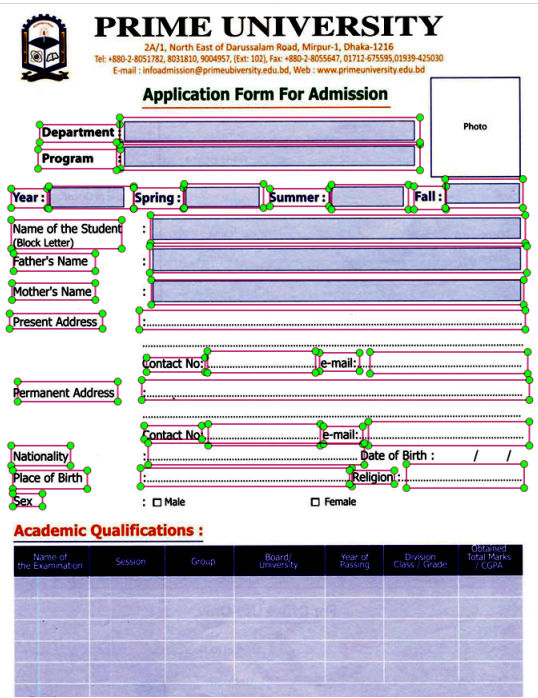


Fig. 3: Sample bounding box labeled image in our dataset. Here in this image the blank boxes are labeled as information field and other's are label as label field.

We use the Mask RCNN, pre-trained on the COCO dataset in our experiment, with the same parameters suggested in the paper [6].

IV. EXPERIMENTS AND RESULTS

This section is divided into the following subsections: Training Details and Experiment Results. First, we give an elaborate overview of the training details of our experiments. Subsequently, we discuss the results of our systems.

A. Training Details

With the same training procedure in paper [7], we train the CycleGAN network on our dataset. The discriminators are fed a history of generated images, rather than the ones produced by the latest versions of the generators to prevent the model from changing drastically from iteration to iteration. To do this, we keep a pool to store the 50 most recently generated images. The training approach was fairly typical for an image-to-image translation task. We use Adam optimizer and the learning rate is 0.0002 for the first half of training. Then the learning rate linearly reduces to zero over the remaining iterations. The batch size was 1, which is referred to as instance normalization, rather than batch normalization.

For the second phase of our training, we use a similar base configuration in paper [6] of Mask-RCNN to train our dataset with a batch size of 1. The initial learning rate is 0.001, weight decay of 0.0001 with an RPN anchor size of 1.

Our technique is implemented in Tensorflow-GPU V2 and cuda V10.0. The experiment has been conducted on a machine having CPU from Intel @. Core TM i7-7500U of 2.7GHz, GPU Nvidia 1050GTX with 4.00GB and with 16.00GB memory on a Windows10 operating system

B. Experimental Result

First, we present the results obtained by CycleGAN for the document deblurring process. We evaluate the performance of CycleGAN using Peak Signal-to-Noise Ratio(PSNR). We observed that CycleGAN obtains 31.334dB. We show some examples of deblurring images that have been produced with our model (See Fig. 5).

Secondly, we evaluate the system trained by the Mask-RCNN network. This network detects objects by sliding the window on a feature map to scan the potential objects and then classifies them and regressing the corresponding box coordinates. Here, the objects refer to **id** and **info** where id is the label of the text fields and info is the information written in the text fields. Average precision is a measure that combines recall and precision for ranked retrieval results. We measure the

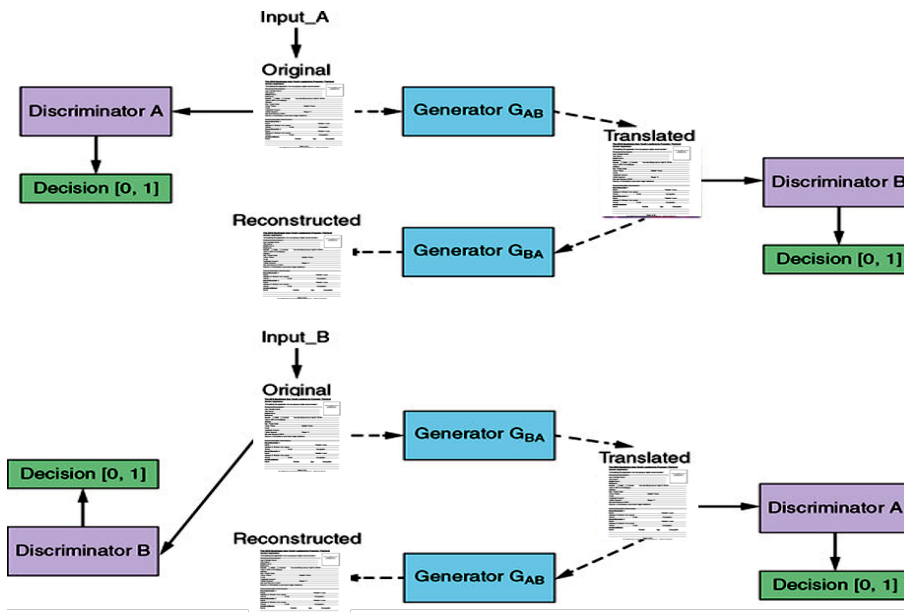


Fig. 4: Overview of CycleGAN - It consists of two generators, G_{AB} and G_{BA} which map blur images to clean images and clean to blur images, respectively using cycle-consistency loss. It also contains two discriminators A and B which acts as adversary and rejects images generated by generators.



Fig. 5: Examples of blurry images(column 1) deblurred(column 2) by CycleGAN and their corresponding cleaned images.(column 3) from our test set.

classifier by utilizing Average Precision with IoU = [.5,.75]. We get 75.8 and 60.1 respectively.

For the last part of our experiment, we use Tesseract - an

optical character recognition engine, to convert the text and save it to the database with corresponding field labels and information. (See Fig 6).

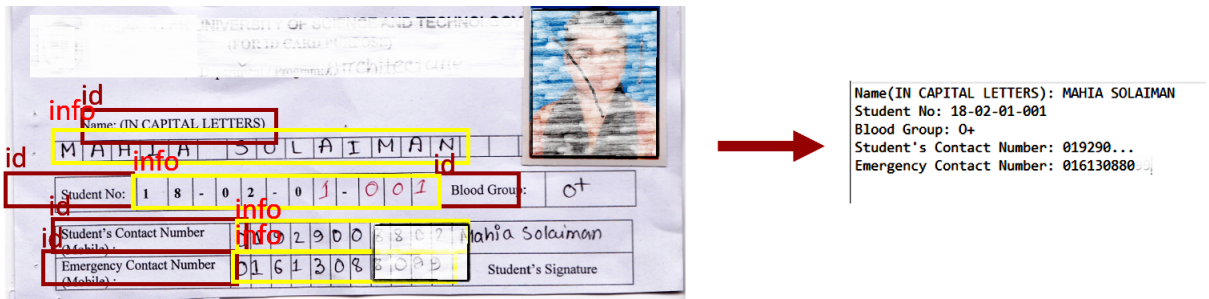


Fig. 6: Information extracting from a raw image with our model. Some portion of the image is blurred due to privacy concern.

V. CONCLUSION

The purpose of our work is to make a system that can retrieve data from scanned images efficiently. We trained our data in two steps - first to remove noises from images to make the quality better by training our dataset to Cycle GAN. Then detecting labels and text fields using Mask RCNN. For removing background noise, deblurring Cycle GAN performed remarkably well in our dataset. Mask RCNN worked well enough to detect fields and labels from scanned document images. But with a larger dataset, our method will work more accurately and it will be more dynamic to detect labels and fields from various types of forms and documents. For handwritten information extraction, we used OCR but, in the future, if we can develop our own algorithm, we can achieve better accuracy to detect and extract handwritten information that is suitable for any handwritten documents. So, our overall outcome was satisfactory as by cleaning images we can detect fields and retrieve data from old documents containing forms. In the future, we will focus on extracting information from any type of documents.

REFERENCES

- [1] Sun, Yingyi and Mao, Xianfeng and Hong, Sheng and Xu, Wenhua and Gui, Guan: Template matching-based method for intelligent invoice information identification. (2019).
- [2] Zheng, Yefeng and Li, Huiping and Doermann, David: The segmentation and identification of handwriting in noisy document images. Springer(2002)
- [3] Tarawneh, Ahmad S and Hassanat, Ahmad B and Chetverikov, Dmitry and Lendak, Imre and Verma, Chaman: Invoice classification using deep features and machine learning techniques.(2019).
- [4] Saleem, Ozair and Latif, Seemab: Information extraction from research papers by data integration and data validation from multiple header extraction sources.(2012)
- [5] Sharma, Monika and Verma, Abhishek and Vig, Lovekesh: Learning to clean: A GAN perspective. Springer(2018)
- [6] He, Kaiming and Gkioxari, Georgia and Dollár, Piotr and Girshick, Ross: Mask r-cnn.(2017)
- [7] Zhu, Jun-Yan and Park, Taesung and Isola, Phillip and Efros, Alexei A: Unpaired image-to-image translation using cycle-consistent adversarial networks.(2017)
- [8] Radford, Alec and Metz, Luke and Chintala, Soumith: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
- [9] Rasmus Berg Palm, Ole Winther, Florian Laws: Cloudscan-a configuration-free invoice analysis system using recurrent neural networks.(2017)
- [10] Jerzy Sas, Jerzy PEJCZ: APPLICATION OF DOCUMENT TYPE IDENTIFICATION IN MEDICAL HANDWRITTEN TEXTS RECOGNITION