Handwritten Form Processing

Bandhan V, K Neetish Bhat, Karthik C Borkar, Mamtha HR

Abstract—Analysis of document images for information extraction has become vital in the modern day. These days so much variety of information is being conventionally stored on paper. For better storage and accurate processing, the paper is being converted into electronic form. This involves a lot of processing of documents using image processing techniques and other computer vision concepts. Pre-Processing techniques like Gaussian Blur, Otsu Thresholding, Median Filter and morphological operations are adopted to increase accuracy of recognition. Based on contours each fields of form are segmented. Character segmentation is done based on bounding box. MNSIT SD-19 database is used for training of characters. SVM and k-NN techniques are used for classification. Our implementation was tried for 10 requisition for certificate forms. Out of 10 forms 8 forms was correctly generated. So the accuracy of result is found to be 80%.

Keywords – Object Character Recognition; Pre-Processing; Segmentation; Classification; Post-Processing;

I. INTRODUCTION

The objective of Optical Character Recognition (OCR) is automatic reading of optically sensed document text materials and to translate human readable characters to machine-readable format. The translation performance is directly dependent upon the quality of the input documents. In a traditional OCR systems input characters by the user are digitized by an optical scanner. Each character is then isolated and segmented. The output of the previous step is the input to a preprocessor for normalization and noise reduction. Certain characteristics from the output are the extracted from the character for classification and analysis. The feature extraction is critical and many different techniques exist, each having its strengths and weaknesses. The current state of OCR has moved from primitive schemes for character set, which is limited to the application of more complicated techniques for Omni font and handprint recognition. The dominant problems in OCR are in the area of segmentation of degraded symbols, which are joined or fragmented. Though they have been significant amount of algorithms that has been processed for character recognition, the problem is not solved completely, especially not in the cases when there are no strict limitations on the handwriting or quality of print. As OCR research has gone to a new level and have made increasing progression, the demand on handwriting analysis has increased since there is a lot of data to mine (such as addresses written on cards, letters, amounts written on checks, names, addresses, identity numbers, and rupee values written on invoices and forms) were written by hand and they had to be manually processed. Initially OCR techniques were based mostly on template matching, simple line and geometric features, stroke detection, and many such features of the image.

Manuscript Received on May 13, 2015.
Bandhan V, ISE department, PESIT, Bangalore-560085, India
K Neetish Bhat, ISE department, PESIT, Bangalore-560085, India
Karthik C Borkar, ISE department, PESIT, Bangalore-560085, India
Dr. Mamtha HR, ISE department, PESIT, Bangalore-560085, India

Those techniques were not sophisticated enough for practical use of handwritten text on various forms and papers. To extract symbolic information from millions of pixels in document images, each component in the character recognition system is designed to reduce the amount of data. As the first important step, image and data preprocessing serve the purpose of extracting regions of interest, enhancing and cleaning up the images, so that they can be directly and efficiently processed by the feature extraction component.
To demonstrate the conversion of form to editable format, sample application form of college requisition for certificate form is selected. This application form is applied to college by students to obtain bonafide study certificate and letter for passport, bank loan, scholarship, hostel admission, visa, expenditure certificate, producing for IT purpose, extension of accommodation and various other purposes.
Given the testing sample, which is a scanned handwritten document of College requisition for certificate form, each character needs to be recognized in the form. Choosing proper and appropriate pre-processing steps and segmentation algorithms plays a vital role in the process as this being the initial step. Efficient feature extraction and classification methods should be used to maintain good performance and accuracy of results.

II. MOTIVATION

Human intervention errors while processing the hand written forms is more in manual process. It may be due to negligence of the person who is analyzing the form or bad handwriting of the person who has filled the form. This may lead unintentional results. In manual process of producing certificate it consumes more time. Instead of that this product can be used in generating letter and save the time of student.

III. DESCRIPTION

The structural complexity and increased character set of English characters, Numerical digits and the different handwritten styles are the inherent problems that exist. There lies the space and time complexity in storing the database should be considered as the complexity of the application. The process of recognizing characters involves different stages. The stages of recognition are depicted in the figure 1.

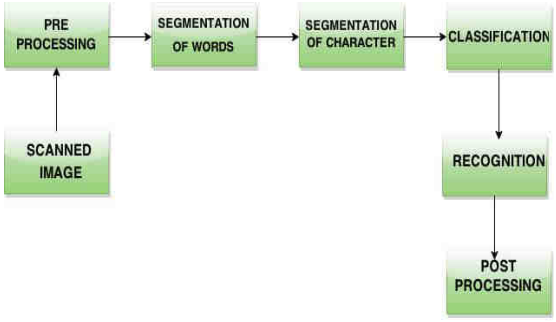
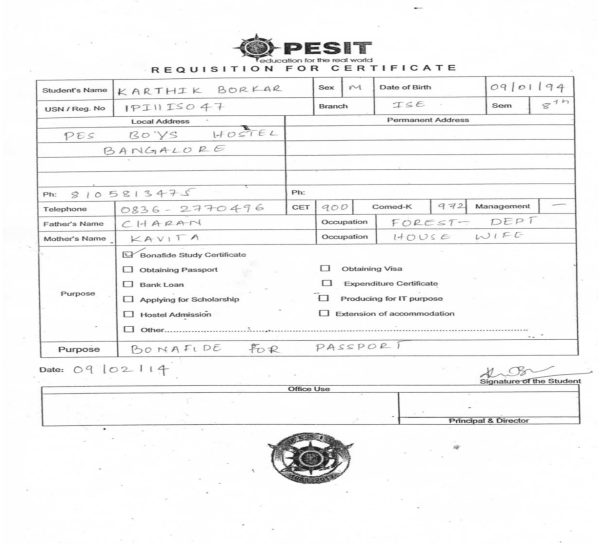


Figure-1: Flow of Recognition System

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

A. Pre-processing

Preprocessing is done to enhance the document by removing noise and other distortions in the written material. Preprocessing steps are applied for the image shown in figure 2. The preprocessing steps used are described below.


PESIT
education for the real world
REQUISITION FOR CERTIFICATE

Student's Name	KARTHIK BORKAR	Sex	M	Date of Birth	09/01/94
USN / Reg. No	1PI11IS047	Branch	ISE	Sem	8th

Local Address	Permanent Address
PES BOYS HOSTEL BANGALORE	
Ph: 9105813475	Ph:
Telephone: 0836 - 2770496	CET: 900 Comed-K: 992 Management: —
Father's Name: CHARAN	Occupation: FOREST- DEPT
Mother's Name: KAVITA	Occupation: HOUSE WIFE

Purpose:
- ☑ Bonafide Study Certificate
- ☐ Obtaining Passport
- ☐ Bank Loan
- ☐ Applying for Scholarship
- ☐ Hostel Admission
- ☐ Other
- ☐ Obtaining Visa
- ☐ Expenditure Certificate
- ☐ Producing for IT purpose
- ☐ Extension of accommodation

Purpose: BONAFIDE FOR PASSPORT

Date: 09/02/14 Signature of the Student

Office Use

Principal & Director

Figure-2: Scanned Image

To get an image with reduced noise, Gaussian blur (Gaussian smoothing) is result of smoothing or blurring an image by a Gaussian function. Gaussian blur is used in many applications of image processing, basically to reduce image noise. The effect of this Gaussian blurring is a smooth blur showing that of viewing the image through a screen, it is very different from that of bokeh effect.

After noise reduction in image Otsu method is used to separate foreground region from background region. Otsu method performs histogram based thresholding. In which image is assumed to be bi-modal histogram i.e. image contains two classes of pixels black pixels and white pixels.

Median filtering technique is also used to reduce noise. The main idea of Median filtering is to traverse through each entry by entry and replacing each entry with median of neighbouring entries. The pattern of neighbours is called window. Window for 1D signals are few preceding and following entries. For higher dimensional signal windows can be defined as box or cross patterns.

Morphological operations are applied to alter the thickness of the characters. There are two types of morphological operations. Dilation is used to reduce the thickness of characters. Window slides through the image. Pixel in the image (either 0 or 1) will be considered as 0 only if all the pixels under the kernel are 0. Otherwise pixel is set to 1. Erosion is used to increase the thickness of characters. The window slides through the image. The Pixel element is '0' if at least there is one pixel under the kernel that is '0'. So it increases the black region in the image or thickness of foreground object increases

Subtraction of unwanted inputs is necessary since An input form has all the unwanted things like boxes, horizontal lines, vertical lines, printed text which are on same position, to get only the handwritten data, we maintain an original copy of the form without filling it, we use this form to retain only handwritten characters. This is done by subtracting the filled form with unfilled original form. Converting image into matrix does this operation and subtraction of matrix is performed to get the required result

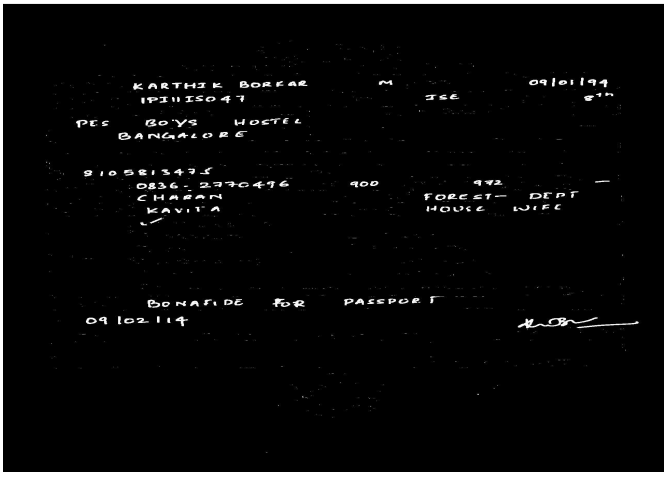


Figure- 3: After Subtraction

B. Segmentation

For an optical character recognition (OCR) system, segmentation phase is an important phase and accuracy of any OCR heavily depends upon segmentation phase. Incorrect segmentation leads to incorrect recognition. Segmentation phase includes line, word segmentation. Before word segmentation, line segmentation is performed to find the number of lines and boundaries of each line in any input document image. Incorrect line segmentation may result in decrease in recognition accuracy. The simplest and most widely used method to segment the lines is to use the inter-line gap in horizontal projection as line boundaries. In word segmentation method, a text line is taken as an input. After a text line is segmented, it is scanned vertically and till a space is encountered which forms a word. For character segmentation static method like vertical projection profile is been used.

Figure-4: An Example of Segmented Fields

Line Segmentation is used to segment lines for the address column if needed, as address can contain more than one line .The simplest and most widely used method to segment the lines is to use the inter-line gap in horizontal projection as line boundaries.

Word Segmentation is used to segment words. In word segmentation method, a text line has taken as an input. After a text line is segmented, it is scanned vertically.

Figure-5: An Example of Character Segmentation

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

C. Feature Extraction & Classification

Bounding box is merely the coordinates of the rectangular border that fully encloses a digital image when it is placed over a page, a canvas, a screen or other similar bi dimensional background.

Our method uses contours to draw bounding box around each character in a field and store them in respective folder, which can be used by classifier to recognize the character. Contours are often obtained from edges, but they are aimed at being object contours. Thus, they need to be closed curves. You can think of them as boundaries. When Contours are obtained from edges, then connect the edges in order to obtain a closed contour.

Once you find contours in image then you sort them by finding moment, moment can be used to find centroid of image depending upon that you can sort contours from left to right, i.e. you can get characters from left to right which can be sent for detection.

In machine learning, classifying the images to which of a set of categories a new observation belongs, on the ground of a set of data, which is subject to training also which contains observations (or instances) whose category membership is known. For an example would be like assigning a given blood group into classes of the blood groups (A+, A- etc.).

We have built two classifiers SVM and KNN nearest both them have almost same accuracy we use hog features for both classifiers.

SVMs (Support vector machines) are supervised learning models with algorithms which analyze data and recognize patterns. SVM's perform linear classification and also SVMs solves non-linear classification, which is called the kernel trick, mapping some of their inputs into high-dimensional feature spaces.

K-NN is a lazy learning or instance-based learning classifier, where the function is calculated locally and all computation is delay until everything is classified. The k-NN algorithm is one of the easiest to implement and less time consuming.

For classification and regression, it can be very useful to weight the influence of the neighbours, so that the neighbour who is nearest contributes more to the average than the more distant ones. The neighbours are taken from the object set for which the k-NN classification class or property object value for k-NN. This is the training set for k-NN algorithm, and there is no explicit training step is required. One of the disadvantages of the k-NN algorithm is that KNN is sensitive to the structure of the data.

MODEL	TOTAL CHARACTERS	RECOGNIZED CHARACTERS	CORRECTLY CLASSIFIED
SVM	296	285	149
K-NN	296	285	155

Table-1: Comparison of SVM and k-NN

D. Training

Once the classifier is built we need to train it suitable input data, input data is carefully select input training set these determine the how test set is classified by the classifier. We choose various databases MNIST SD-19. The MNIST database of handwritten digits has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST, which is known for it's diverse dataset spread across multiple languages. The digits have been size-normalized and centred in a fixed-size image. We calculate HOG feature for all the training images based on that we make our classifier undergo training. Similarly for test data we calculate the HOG features then the classifier predicts the characters. Both SVM and KNN use HOG feature. HOG (Histogram of Oriented Gradients) is feature descriptors used in image processing for the purpose of object detection. The method counts the occurrences of gradient orientation in localized portions of an image.

E. Post Processing

The expected result of this project is to generate appropriate letter based upon the tick box ticked by the user. So in segmentation phase each tick box will be segmented and examined for the presence of tick mark using extreme detection method. If min extrema and max extrema is 0 then tick box doesn't contains tick mark. Else the tick box contains tick mark. If the tick mark is present appropriate subject and body will be generated. For generation of body, it takes recognized fields like name, usn, semester and branch. Next this generated body and subject will be sent to the later stages to generate letter. The format for the letter generation is taken from the existing word document called 'Letter.docx'. The letter format is shown in figure 6. This document will be copied and in the copied document subject and body will be replaced dynamically. Python libraries like python-docx was used to read and replace the words in Microsoft document file. RE(Regular Expressions) were used to replace subject and body in the document. The generated document is shown in the figure 7.


PES

PES INSTITUTE OF TECHNOLOGY

100 Feet Ring Road, B S K 3rd Stage,
Bengaluru-560085

From:
PESIT,
Bangalore-560085
Date: date

To:
Whoever it concerns

Respected Sir/Madam,

Subject: subject

body

Thanking you,

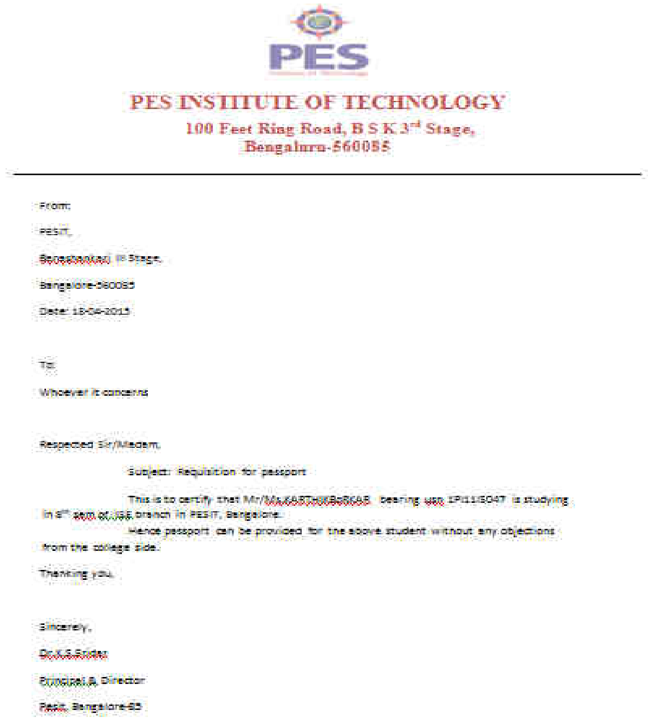
Sincerely,

Director

Bangalore-85

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Figure-6: Letter Template


PES

PES INSTITUTE OF TECHNOLOGY

100 Feet Ring Road, B S K 3rd Stage,
Bengaluru-560085

From:

PESIT,

Banashankari III Stage,

Bangalore-560085

Date: 18-04-2015

To:

Whoever it concerns

Respected Sir/Madam,

Subject: Requisition for passport

This is to certify that Mr/Ms.[illegible] bearing usn 1PI11IS047 is studying in 8th sem of ISE branch in PESIT, Bangalore.

Hence passport can be provided for the above student without any objections from the college side.

Thanking you,

Sincerely,

Dr.K.S.Sridar

Principal & Director

Pesit, Bangalore-85

Figure-7: Generated Letter

IV. EXPERIMENTAL RESULTS

The proposed character recognition method is tried with 10 different forms of college requisition for certificate form. By using SVM classification out of 296 characters 155 characters are correctly classified. So the accuracy of our method using SVM is 52.36%. By using k-NN classification out of 296 characters 149 characters are correctly classified. So the accuracy of our method using k-NN is 50.33%

Since the taken form is used to generate different certificates, we have tried our implementation with 10 different forms to generate appropriate certificate. Out of 10 forms 8 certificates was correctly generated. So the accuracy of generating letter is 80%.

V. CONCLUSION

Developing an OCR for a handwritten certificate requisition form is quite challenging and prone to errors due to structural complexity of hand written characters. An attempt is made in this direction and the recognition of characters is done. Better skew detection and noise removal techniques can be used to enhance the Pre-processing and Segmentation phases. Efficient feature extraction and classification methods are used to get good performance and accuracy of results. This can be further extended for other form-based applications like forms used in handling withdrawals and deposits in banks, educational institutions, applications in government offices etc. The results are found satisfactory for the algorithms used for the current system. The system may require to be modified slightly if used for other form-based applications processing.

REFERENCES

1. Vamsi Krishna Madasu, Brian Charles Lovell, "Automatic Segmentation and Recognition of Bank Cheque Fields" NICTA& School of ITEE, University of Queensland.
2. Ameur Bensfia, "Extraction of Arabic Handwriting Fields by Forms Matching", Journal of Signal and Information Processing, 2015.
3. Vamsi Krishna Madasu, Mohd. Hafizuddin Mohd.Yusof, M. Hanmandlu, Kurt Kubik, "Automatic Extraction of Signatures from Bank Cheques and other Documents", Intelligent Real-Time Imaging and Sensing group, School of Information Technology and Electrical Engineering, University of Queensland, 2003.
4. Roongroj Nopsuwanchai, "Discriminative training methods and their applications to handwriting recognition", Technical Report, \UCAM-CL-TR-652, ISSN 1476-2986, 2005.
5. M.K.Jindal, R.K. Sharma, G.S. Lehal, "Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts", International Journal of Computational Intelligence Research. ISSN 0973-1873 Vol.3, No.4 (2007), pp. 277–286.
6. Nallapareddy Priyanka, Srikanta Pal, Ranju Mandal, "Line and Word Segmentation Approach for Printed Documents", IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition" RTIPPR, 2010.
7. K. Srikanta Murthy, G. Hemantha Kumar, P. Shivakumar,P.R. Ranganath, "Nearest Neighbour Clustering approach for line and character segmentation in epigraphical scripts".
8. Opencv Official Documentation: http://docs.opencv.org/

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.