Bank Form Classification using Document Layout Analysis and Image Processing Techniques

Shashikala H K
Department of Computer Science and Engineering
School of Engineering and Technology, Jain University
Bangalore Rural, India
shashi.hk85@gmail.com

***Abstract*- Every day thousands of forms are filled out and submitted across the world, in banks, post offices, government organizations, educational institutions etc. These include electronic forms as well as physical forms. All of these forms irrespective of their origin are at some stage made digital and stored electronically to address issues of physical storage, form degradation and data accessibility. Document layout analysis is a basic step in converting document images into electronic form. This conversion is laborious and can be made more efficient (in terms of throughput and human resource) by automating most of the conversion process using document layout analysis techniques. Document classification is an important step in Office Automation, Digital Libraries, and other document image analysis applications. Physical forms require human supervision for any operations done on the form. Digitization of these forms reduces human resources, also reduces any human redundancy involved with the operation on the physical forms. This paper addresses the initial stage of this automation, namely, bank form classification and decipherment of fields. The former recognizes the type of the bank form and the latter extracts regions of useful data from the classified bank form. The proposed work aims to provide accurate bank form classification along with noise removal, skew detection and correction, finally layout analysis is carried out to extract fields like name, address, signature from the classified forms.**

Keywords—Document processing, Document classification, Classification algorithms. Document layout Analysis, Skew detection and correction, Noise removal, Bank forms, optical character recognition.

I. INTRODUCTION

Over decades, paper was used as a principal medium for storing the information. Most of information is still stored on paper. Due to the introduction of computers and word processors, the amount of information stored on paper is increased instead of reducing. .However, there are many applications where the paper is the preferred media [1].Every day thousands of forms are filled out and submitted across the world. Storage and maintenance of information on paper is expensive and conversion to electronic document is laborious and can be made more efficient by automating most of the conversion process using document layout analysis techniques. There are many advantages of electronic documents over paper documents, i.e. lossless storage, easy to maintenance and fast transmission.

Currently, most information is still recorded and transmitted through paper, since people are most comfortable with paper documents. Storage and maintenance of all this paper is expensive. Apart from these expenses, it also presents the problem of paper degradation especially in less than ideal conditions. Paper degradation can occur due to the following reasons:

- Natural calamities such as destruction by fire, floods, etc.
- Temperature and Relative Humidity
- Biological Infestation

Physical forms require human supervision for any operations done on the form. Digitization of these forms reduces the human resources required by limiting their participation to build a system which imitates the human operations. This also reduces any human redundancy involved with the operation on the physical forms.

An advantage of converting paper documents to electronic form is, better archiving. Document processing methods usually consists of: scanning, layout analysis, and optical character recognition [3]. Generally skew will be introduced into the image when a paper document is scanned due to various factors. Many character recognition techniques and document layout analysing methods are more sensitive to skew and it may affect further analysis. Moreover, while scanning the document it is very difficult to prevent skew. Therefore, during pre-processing stage only skew has to be detected and corrected to avoid problem in further processing.

Another problem in pre-processing stage is, noise may be introduced into scanned images. Noise may be introduced during scanning document or image transmission. Noise reduction is a very important step in most of the image processing applications [4]. Many linear and non linear filtering techniques are available to reduce noise in the image. Non linear filters are more effective for removing impulse noise.

Document classification assigns a one-page document image to one of the predefined document classes sets. Classification can be done based on variety of features, like structural features or textual features. There is variety of document classifiers. Each Classifier solves a various document classification problems, and they differ in terms of how they are used to train the data, the features used for classification and the choice of recognition algorithm [5].

IJARE

This is an open access article under the CC BY-SA 4.0 license (https://creativecommons.org/licenses/by-sa/4.0/)

II. LITERATURE SURVEY

Over the decades, there are many document layout analysis techniques [6, 7] that are used for different document layouts and are more effective to against noise and skew in the document. Fundamental step in layout analysis is to get a digitized image of the paper document using a scanning system followed by page layout analysis. The skew has to be detected and corrected in the pre-processing stage. There are two type of skew in document images: Single Skew and Multiple Skew. Most of the optical character recognition algorithms are very sensitive to the skew in the scanned image. So, skew detection and correction is necessary in the scanned image automatically.

While capturing the image, various noises can be introduced into the scanned document. Noise removal [9] is used for improving the quality of the image and it is also used in pre-processing step for most of the applications. Most common type of noise in scanned image is salt and pepper noise [10, 11]. Noise in a scanned image can affect the information contained in it. Many noise removal methods can be used to remove the noise from the image from the degraded copy of an image. Most popular non-linear filter used to remove Salt and Pepper noise is median filter.

Document Layout Analysis is carried out for dividing the scanned document into regions with a homogeneous content, and then a meaning will be assigned to all the divided regions. Document layout analysis is divided into two phases: segmentation step also referred as physical layout analysis (identifies the geometric page structure) and logical layout analysis also referred as functional labeling (assigns logical meaning to each region).

III. PROPOSED SYSTEM

Most of the existing document analysis systems are restricted to relatively small application domains [13,14] like Journal pages, technical reports, newspaper pages, e-mail filtering, biomedical journals, web documents, patent applications. Physical forms require human supervision for any operations done on the form. Digitization of these forms reduces human resources, also reduces any human redundancy involved with the operation on the physical forms. The proposed work aims to provide accurate bank form classification along with noise removal, skew detection and correction, finally layout analysis is carried out to extract fields like name, address, signature from the classified forms. The main classification objective is to simply access to and processing of information in the document. Classification eases retrieval, organization and visualization. Document classifier has two aspects The Proposed method has been divided into two major stages called Training Stage and Classification stages. As the name suggest the training stage concerns with the training the classifiers for all possible document forms layout. Once the classifier is trained and the knowledge base is created, the classifier will take the input image and performs the classification.

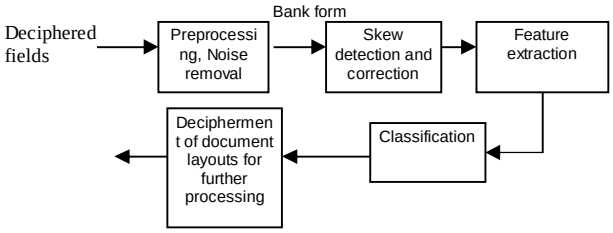

Fig 1: System Architecture

Proposed method provides an automated way for classifying various banking forms, based on training data. The class model is defined by the document image, their features and one or more logos.

IV. CLASSIFICATION PROCESS

In the classification process, the test document will be processed and its class shall be identified. The major stages of classification process are almost same as that or the Training process except the knowledge base creation. In the classification process the knowledge base created during the Training process is used to perform the classification. The procedure for the proposed classification process is mentioned below.

Given image, is preprocessed and converted to the YUV color space and further processing is performed in the luminance channel. The document is skew corrected using the skew detection and correction method. The skew corrected image is resized and logo region is cropped and correlation features are extracted.

Extracted correlation features are matched against the correlation features stored in the knowledge base which was created during the training process. The Euclidian distance measure has been used as a measure to estimate the similarity between the test feature vectors against the stored feature vectors. The document is classified into the class with which the computed Euclidian distance is minimum.

Further, the raw feature vector of the image has been transformed into the kernel sub-space and the reduced feature vector in the transformed space is projected into the Principal Components (PC) computed during the training phase. Again the Euclidian distance has been considered for measuring the similarities between the feature vectors. The test form will be categorized into a class with minimum Euclidian distance.

Subsequently, the two scores from two different classifiers are fused to arrive at global score to decide on the final class for which the document has belonged to. This is done using the score level fusion using the weighted fusion approach. Once the class of the test form is identified the Configuration file has been retrieved and the corresponding positions of the form are cropped and presented to the user for further processing.

V. EXPERIMENTAL RESULTS

Proposed work takes scanned bank form as an input and performs noise removal and skew correction, then classifier classifies the given bank form and layout analysis extracts the

IJARE

This is an open access article under the CC BY-SA 4.0 license (https://creativecommons.org/licenses/by-sa/4.0/)

fields like name, address, signature, account number from the classified bank form.

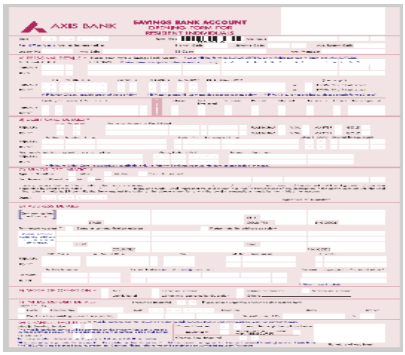

Fig 2. Classified bank form

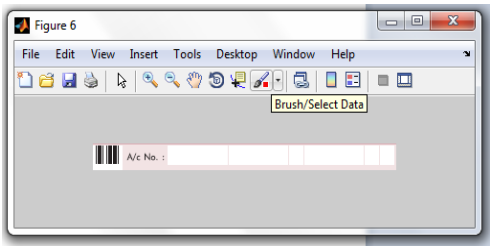

Fig 3. Account field extracted from classified bank form

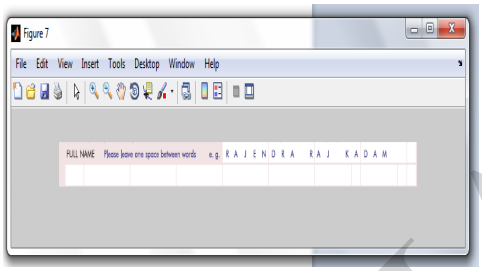

Fig 4. Name field

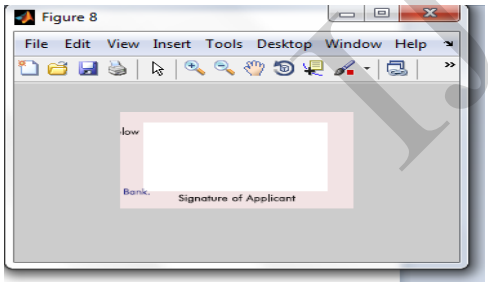

Fig 5. Signature field

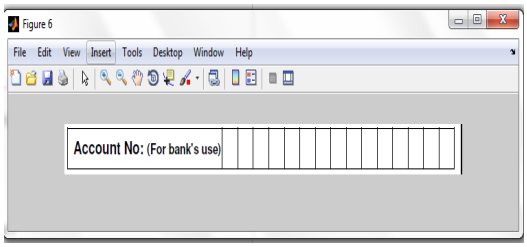

Fig 6. Account number

VI. CONCLUSION

This paper describes a method for automatic classification of document images into appropriate classes, with a special treatment towards classification of banking application forms image. The proposed method has two phases, the training phase and classification phase. The training phase involves training the system to learn about the predefined set of document images. For the purpose of training the system two set of classifiers are used; the first one is based on the Normalized Cross Correlation method based on the features of correlation coefficients. The other one is based on the sub-space analysis of the image intensities; for the sub-space transformation concept of Principal Component analysis is used. The trained classifier information is stored in the knowledge base and is used in the classification process. In the classification process the test document is subjected to classification using both classifier and finally the scores from both classifiers are fused using the score level fusion approach. Once, the document has been classified, the relevant desired fields are segmented and stored for further process.

The proposed method has been tested on various banking application forms and from the experiment it is concluded that the proposed method performs extremely well in classifying the banking application forms.

REFERENCES

[1] Sellen,A.J, Harper,R:The myth of the paperless office.MIT press(2001)

[2] Lymann, P,Varian,H.R: How Much Information, Technical report: http://www.sims.berkely.edu/how-much- info-2--3 on 6-1-2007 (2003)

[3] Yang Cao, Heng Li, "Skew Detection and Correction in document Images Based on Straight- Line Fitting", Polytechnic university, Hong Kong, Email: caoyang@yaulee.com

[4] Mrs. C Mythili, Dr. V Kavitha, 'efficient technique for color image noise reduction', University College of Engineering Nagercoil.

[5] Nawei Chen, Dorothea Blostein, "A survey of document image classification: problem statement classifier architecture and performance evaluation" , Springer-Verlag 2006

[6] Cattoni R, Coianiz T, Messelodi S, Modena C '.M: 'Geometric Layout Analysis Techniques for Document Image Understanding':a review'. Technical report 9703-09, IRST,Trento,Italy (1998)

[7] Nagy G. : 'Twenty years of document image analysis' in PAMI, IEEE Transaction, Pattern Anal. Mach. Intell. 22 (1), 38–62 (2000)

[8] Rafael Dueire Lins and Bruno Tenório Ávila, 'A New Algorithm for Skew Detection in Images of Documents', ICIAR 2004, LNCS 3212, pp. 234–240, 2004. Springer-verlag, Berlin Heidelberg, 2004

[9] Chi Chang-yanab, Zhang Ji-xiana, Liu Zheng- juna" study on methods of noise reduction in a stripped image", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B6b. Beijing(2008)

This is an open access article under the CC BY-SA 4.0 license (https://creativecommons.org/licenses/by-sa/4.0/)

[10] R H.Chan ,W H.Chung, M.Nikolova, 'Salt- and-Pepper Noise Removal by Median-Type Noise Detectors and Details-Preserving Regularization', Image processing,IEEE,Vol.14,No.10,2005, PP.1479-1485.

[11] Mr. Salem Saleh Al-amri,, Dr. N V Kalyankar,, Dr. Khamitkar S D, 'A Comparative Study of Removal Noise from Remote Sensing Image ' ,IJCSI International Journal, Vol. 7, Issue., No. 1, January 2010, 32 ISSN (Online): 1694-0784 ISSN (Print) : 1694-0814, 2011.

[12] Youngjoong Koa , Jungyun Seo, "Text classification from unlabeled documents with bootstrapping and feature projection techniques", Information Processing and Management 45 (2009) 70–83

[13] Burget, R: Layout Based Information Extraction from HTML Documents, Brno Univ. of Technol., Brno, (2007)624 - 628

[14] Feng, J: A learning approach to discovering Web page semantic structures, AT&T LABS RESEARCH (2005)1055 - 1059 Vol. 2

IJARE

This is an open access article under the CC BY-SA 4.0 license (https://creativecommons.org/licenses/by-sa/4.0/)