Use OCR for Multiple Form Filling System

Dhanawade Komal P[1] Dhamal Pratiksha H[2] Jagtap Siddhesh S[3] Jagtap Suraj S[4]
[1,2,3,4]Department of Computer Science & Engineering
[1,2,3,4]SVPM COE Malegaon (BK) Pune, 413115, India

Abstract— Optical Character Recognition(OCR) is one of the most interesting and challenging research area in the field of image processing. Nowadays, different methodologies are in widespread use for character recognition. Different form filling is newly coming application area of digital document processing. Proposed system for recognition of hand-written as well as printed characters. Various approaches of handwritten and printed character recognition related with forms are discussed along with their performance. OCR to translate images of typewritten or handwritten characters into electronically editable format by preserving font properties. OCR can do this by pattern matching algorithm. Proposed system work as a time consuming system. Due to this system entry done in the database automatically and also reduces manual work.

Key words: Character Recognition, Feature Extraction, Segmentation, Preprocessing, Classification, Post-processing, Training and Recognition, Online Handwriting Recognition

I. INTRODUCTION

Optical character recognition is a process of extracting text from images. We have to noticed that when-ever we scan text document we can't able to edit it with the help of text editor but with the help of OCR we can do it. This technology is very important not only for editing scanned document but for future of formal ling.

Nowadays, in colleges, banks, hospitals there are many forms lled like exam form, admission form, en-try form, bank account forms etc. We have to all these forms and after submit that forms and for this process we have to wait in a long queue. After that form checkers have to check individual form manually. But by using optical character recognition for formlling, this manual work is done automatically. Due to this reason people time will be save and process become faster. So, we are motivated by this idea.

A. Previous Work

1) An Overview of Character Recognition Focused on On Line Handwriting

Character recognition (CR) has been extensively studied in the last half century and progressed to a level sufficient to produce technology driven applications. Now, the rapidly growing computational power enables the implementation of the present CR methodologies and creates an increasing demand on many emerging application domains, which require more advanced methodologies. This material serves as a guide and update for readers working in the CR area. First, the historical evolution of CR systems is presented. Then, the available CR techniques with their superiorities and weaknesses are reviewed. Finally, the cur-rent status of CR is discussed, and directions for future research are suggested. Special attention is given to the online handwriting recognition since this area requires more research to reach the ultimate goal of machine simulation of human reading.

2) Review on Handwritten Character Recognition

Nowadays character recognition has gained lot of attention in the field of pattern recognition due to its application in various fields. Optical Character Recognition (OCR) and Handwritten Character Recognition (HCR) has specific domain to apply. OCR system is most suitable for the applications like multi choice examinations, printed postal address resolution etc. While application of HCR is wider compare to OCR. HCR is useful in cheque processing in banks; almost all kind of form processing systems, handwritten postal address resolution and many more. In coming days, character recognition system might serve as a key factor to create paperless environment by digitizing and processing existing paper documents. In this paper, we have provided the detail study on existing methods for handwritten character recognition.

3) Optical Character Recognition with Fast Training Neural Network

Optical character recognition has been extensively investigated in the past few years. Many existing techniques are able to provide high recognition rate, but at the cost of long training time. In this work, we present a neural network based approach to reduce the training time while maintain the high recognition rate. The main idea is to perform a preprocessing stage to partition the training data prior to the training stage. A multi-stage approach is then used to deal with various types of input source. Our experiments on real image datasets have demonstrated that the balance between the training time and recognition time can be achieved using the proposed method.

4) OCR Accuracy Prediction Method Based on Blur Estimation

In this paper, we propose an OCR accuracy prediction method based on a local blur estimation since blur is one of the important factors that mostly damage OCR accuracy. First, we apply the blur estimation on synthetic blurred images by using Gaussian and motion blur in order to investigate the relation between blur effect and character size regarding OCR accuracy. This relation is considered as a blur-character size feature to de ne a classifier. Finally, the classifier can separate characters of a given document into three classes: readable, intermediate, and non-readable classes. Therefore, the quality score of the document is inferred from the three classes. The proposed method is evaluated on a published database and on an industrial one. The correlation with OCR accuracy is also given to compare with the state-of-the-art methods.

5) OCR Error Correction Using Character Correction and Feature-Based Word Classification

This paper explores the use of a learned classifier for post-OCR text correction. Experiments with the Arabic language show that this approach, which integrates a weighted confusion matrix and a shallow language model, improves

the vast majority of segmentation and recognition errors, the most frequent types of error on our dataset.

6) Handwritten Character Recognition A Review

In the eld of pattern recognition, HCR is one of the most intricate and tricky area. Plenty of works were proposed for foreign languages but a few works exists for south Indian languages due to the complex shape and varying writing styles of individuals. This paper introduces a review of online and online recognition of different natural languages. HCR is an optical character recognition, which convert the textual document in to machine readable format. To attain 99.9% accuracy in the field of HCR is very difficult. The efficiency of HCR depend the features extracted and the classifier used.

7) Review on Text Detection Methodology from Images

Texts in an image directly carry high-level semantic information about a scene, which can be used to assist a wide variety of applications, such as image understanding, image search and indexing, navigation, and human computer interaction. However, lot of existing text detection and recognition systems are considered for horizontal or near horizontal texts. With the increasingly popular computing on the go devices, detecting texts of random orientations from images taken by such devices under less controlled conditions has become an increasingly important and yet challenging task. Di erent techniques have been proposed to address this problem, and to classify and review these related algorithms. This paper gives detail ex-planation of work done for automatically detection of text from images, localize and extract text in images having complex backgrounds.

8) A Survey of OCR Applications

Optical Character Recognition or OCR is the electronic translation of handwritten, typewritten or printed text into machine translated images. It is widely used to recognize and search text from electronic documents or to publish the text on a website. The paper presents a survey of applications of OCR in different fields and further presents the experimentation for three important applications such as Captcha, Institutional Repository and Optical Music Character Recognition. We make use of an enhanced image segmentation algorithm based on histogram equalization using genetic algorithms for optical character recognition. The paper will act as a good literature survey for researchers starting to work in the eld of optical character recognition.

9) A Literature Survey on Handwritten Character Recognition

Handwriting recognition has gained a lot of attention in the field of pattern recognition and machine learning due to its application in various fields. Optical Character Recognition (OCR) and Handwritten Character Recognition (HCR) has specific domain to apply. Various techniques have been proposed to for character recognition in handwriting recognition sys-tem. Even though, sufficient studies and papers describes the techniques for converting textual content from a paper document into machine readable form. In coming days, character recognition system might serve as a key factor to create a paperless environment by digitizing and processing existing paper documents. This paper presents a detailed review in the field of Handwritten Character Recognition.

II. PROPOSED SYSTEM

Our Proposed System is for different form filling using OCR which is a character recognition system that supports recognition of the characters of multiple forms. This feature what we call form filling which eliminates the problem of heterogeneous character recognition and support multiple functionalities to be performed on the document. The supports only editing of the document. In this context, form filling means the infrastructure that supports group of specific set of forms. These OCR for form filling is for multiple forms.

III. MATHEMATICAL MODEL

1) Capture Image
S = {I,O,F,success,failure}
I = Input.
O = Output.
F = Function.
I = {d0,d1,d2dn}
d0 - college form d1-banking form d2-hospital form dn = hostel form O={o1}
o1 - scanned image F={f1} f1=capture_image();
success={Sucessfully scanned image}; Output get successfully.
failure={ UnSucessfully scanned image.};
2) Image Preprocessing
S = {I,O,F,success,failure} I1={i1}
i1= captured image O={o1,o2 ,o3}
01= noise removed from image. 02=blur removed from image. O3=clear image
F = {f1} f1=image_preprocessing ();
success={Sucessfully image are preprocessed};
failure={Unsuccessful image are preprocessed.};
3) Segmentation of Image
S={I,O,F,success,failure} I1={i1}
i1=clear image O={o1}
01=separated character. F={f1}
f1= segmentation_image.();
success={Sucessfully character are separated};
failure={Unsuccessful character are separated};
4) Feature Extraction
S={I,O,F,success,failure} I1={i1}
i1=separated character O={o1}
01=extracted individual character. F={f1}
f1= Feature_extraction.();
success={Sucessfully extracted individual character};
failure={Unsuccessful extracted individual character};
5) Classification
S={I,O,F,success,failure}
I1={i1}
i1=separated individual character O={o1}
01=build classes of extracted characters. F={f1}
f1=. Classification ();
success={Successfully extracted Classified character};
failure={Unsuccessful extracted Classified character };
6) Post Processing
S={I,O,F,success,failure} I1={i1}
i1=classes of extracted characters. O={o1}
01=electronic data. F={f1}
f1=. Post_processing ();

success={Sucessfully electronic document is generated};
failure={Unsuccessful electronic document is generated};
7) Insert Data
S={I,O,F,success,failure} I1={i1}
i1=electronic documents O={o1}
01=entry is done successfully. F={f1}
f1=. Post_processing (); success={Sucessfully entry is done }; failure={Unsuccessful entry is done };

IV. ALGORITHMS

A. *Template Matching*

- Input: Image
- Output: Recognize character template

1) Steps

- Firstly, the character image from the detected string is selected.
- After that, the image of the size of the rst tem-plate is rescaled.
- After the rescale the image to the size of the rst template (original) image, the matching matrix is completed.
- Then the highest match found is stored. If the image is not matched repeat again the third step.
- The index of the best match is stored as recognized character.

B. *K-Algorithm*

- Input: Scanned Image with impurities.
- Output: Produce cleaned up version of image.

1) Steps

- Firstly, Filtering technique is used for removal of noise.
- After that, Binarization technique is used to convert alter image to binary image.

C. *Support Vector Machine*

- Input: Cleaned Image.
- Output: Separate out characters from image.

1) Steps

- Firstly, scan the image from left to right and from bottom to top.
- After that, for each black pixel encountered which is not part of an area already found do.
- Tag the up, left and right directions as possible expansions.
- If there is a direction of which frontlier contains no black pixels, mark this direction as not possible for expansion.

D. *Context Based Error Correction*

- Input: OCR text.
- Output: Best words sequence for the strings in the sentence.

1) Steps

- Firstly, Read the sentence from the input OCR text.
- After that, Retrieve up to M candidates from the lexicon for each possible error.
- Rewrite the N candidates by their conditional probabilities to the error. Keep only the Top N candidates for the next processing steps (In the current system M is 10,000 and N is 10).
- Use the viterbi algorithm to get the best words sequence for string in the sequence.

V. CONCLUSION

The purpose of this system is to develop a new OCR technology for data entry of multiple forms. OCR is powerful tool for data entry from image text of different forms. To obtain this objective first step is to capture the high resolution image. After that perform the different operations on image like normalization, Binarization, blur estimation and noise removal. In the second step, separate out the individual character of the image and extract the different features of that individual character. In the third step, according to that features characters are classified for decision making. In the recognition stage, if some unrecognized characters or misspelled characters found those characters mistakes correct in this stage and the nal step, the successful entry of form is done in the database without seriously damaging the data quality.

REFERENCES

[1] Na z Arica and Fatos T. Yarman-Vural, \An Overview of Character Recognition Focused on O -Line Handwriting," IEEE Transactions On Systems, Man, and Cybernetic-Spart C: Applications And Reviews, VOL. 31, NO. 2, MAY 2001.
[2] Surya Nath R S, Afseena S, "Handwritten Charac-ter Recognition A Review," International Journal of Scienti c and Research Publications, Volume 5, Issue 3, March 2015 1 ISSN 2250-3153.
[3] Ayush Purohit, Shardul Singh Chauhan Ayush Purohit, \A Literature Survey on Handwritten Character Recognition," (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (1), 2016, 1-5.
[4] Suruchi G. Dedgaonkar, Anjali A. Chandavale, Ashok M. Sapkal, "Survey of Methods for Char-acter Recognition," International Journal of Engi-neering and Innovative Technology (IJEIT), Vol-ume 1, Issue 5, May 2012.
[5] Professor Latika R. Desai, Miss. Poonam B. Kadam, Professor Swati Shinde, "Review on Text Detection Methodology from Images," Interna-tional Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 2, February 2014.

