Noise Detection and Reduction in Printed English Text

Ishita Patel, Dhwanil Patel, Foram Patel, Bhumika Chauhan, Rinal Mistry, Khushbu Patel

Abstract— Printed character recognition has been one of the most interesting and challenging research areas in field of image processing and pattern recognition in the recent years. OCR has an improve efficiency and less computational cost, to use database recognize English character which is managed and very simple. It is digitizing process by handwritten or printed text that can be scanned electronically and used in. machine process. Its applying appropriate methods to scanned image and denoised image to obtain saved for further processing.

Keywords— OCR, Relation, Scanned Images, Type machine.

1. INTRODUCTION

OCR is the contraction for Optical Character Recognition. It automatically recognizing characters through an optical mechanism. In case of human being, our eyes are optical method. The ability to understand these inputs varies in each person according to many factors [1]. There are two categories of printed character recognition: off-line and on-line. In Online character recognition, handwriting is captured using a unique pen in combination with electronic surface. In Offline character recognition input has been scanned from an inside such as sheet of paper and stored digitally. Offline character recognition includes recognition of machine printed, hand-printed and hand-written character [10]. It's provides a full alphanumeric recognition of printed or Handwritten characters at electronic speediness by just scanning the document [3]. Noise is an unsystematic difference of image Intensity and perceptible as grains in the image [8].

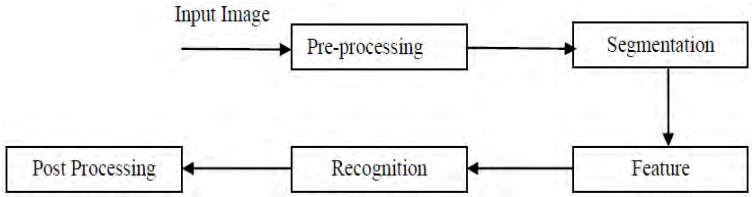


Fig. 1: Proposed Recognition System.

2. LITERATURE REVIEW

How OCR works?

Document is scan using a scanner and is given to the OCR systems which recognize the Character in the scanned document and converts them into ASCII

data [1]. OCR systems used to remove the human interactions for improved performance and efficiency. The area of OCR is proper an important part of document scanner, and is used in several application such as postal process, script recognition, banking, security (i.e. passport authentication) and words identification [3]. It is the mechanism to convert machine printed, hand-printed or hand-written document file into editable text format [2]. Recognized characters which are unpredictability in shape and Irrespective of abnormality are reproducing the actual characters from abnormal documents based on algorithms and methods [4]. OCR is a field of research in pattern recognition, artificial intelligence and machine vision, signal processing [5].

Major Stages of OCR

1. Preprocessing
2. Segmentation
3. Feature Extraction
4. Recognition

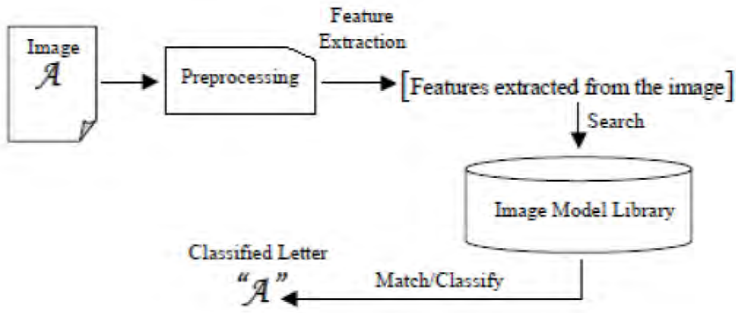


Fig. 2: Block Diagram of character Recognition.

1] Pre-processing:

Pre-processing is an most important step of applying a number of procedure for smoothing, enhancing, filtering etc., for making an image usable by consequent algorithm in order to improve their readability for OCR software [4]. The pre-processing step, there is an order of operations performed on the scan input image. It enhance the image reproduction it suitable for segmentation the grey-level character image is normalized into a window size. After noise reduction, we produced a bitmap image [5]. Pre-Processing can be defined as cleaning the document image and making it appropriate for input to the OCR engine [2]. Major steps under pre-processing are:

A. Noise removal:

Noises degrade the image quality. Noise can occur at different stage like image capturing, transmission and compression. Different filter and morphological operation are available for removing image noise [4].

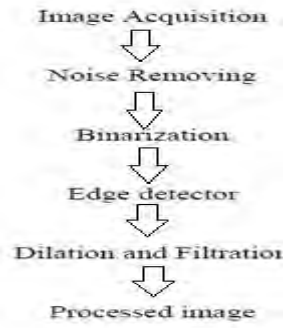


Fig.3: Stages of Process

Types of Noise:

- Gaussian Noise (Amplifier Noise)
- Salt and Pepper Noise (Impulse Noise)
- Shot Noise (Poisson Noise , Photon Noise)

- **Gaussian Noise (Amplifier Noise)**

Gaussian noise is cause by random fluctuation in the signal. It is modeled by random values added to an image. In Gaussian noise, each pixel in the image will be transform from its original value by a small amount. Each pixel in the noisy image is the amount of the true pixel value and a random, Gaussian scattered noise value [8].

- **Salt and Pepper Noise (Impulse Noise):**

Salt and pepper noise is also call fat-tail distribute or impulsive noise or spike noise. An image containing salt-and-pepper noise will have dark pixel in bright region and bright pixel in dark region. It presents itself as sparsely occurring white and black pixel [7]. These noises arise in the image because of sharp and sudden changes of image signal [8].

- **Shot Noise (Poisson Noise , Photon Noise):**

Poisson noise is the noise that can cause, when number of photon sensed by the sensor is not sufficient to provide visible arithmetical information. This noise has root mean square value proportional to square root intensity of the image [7].

B. Image Denoising:

Digital images are prone to a variety of type of noise. There are several ways that noise can be introduce into an image. De-noising filters can be categorized in the following categories:

- Averaging filter
- Order Statistics filter
- Adaptive filter

- **Mean filter:**

Mean filter is an easy to implement method of reducing noise from an image. The idea of mean filtering is easily to restore each pixel value in an image with the mean 'average' value of its neighbors, including itself. This has the effect of eliminate pixel values which are unrepresentative of their surroundings. Mean filtering is usually thought of as a convolution filter. 3*3 square kernel/mask is used. Mean filters show very good performance for the removal of many noise types (e.g. Gaussian noise).

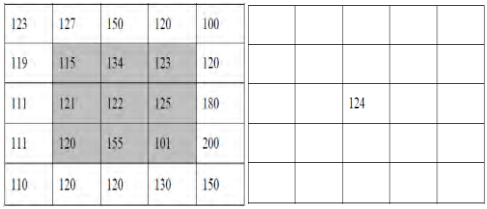


Fig. 4: Mean Filter

- **Median Filter:**

Median filter is a nonlinear method. The median filter is effective for removing salt and pepper noise. The main idea of the median filter is to run during the signal entry by entry, replacing each entry with the median of neighboring entries. The pattern of neighbor is called the "window", which slide, entry by entry over the entire signal.

The median filter takes an area of an image (3x3, 5x5, 7x7, etc.)

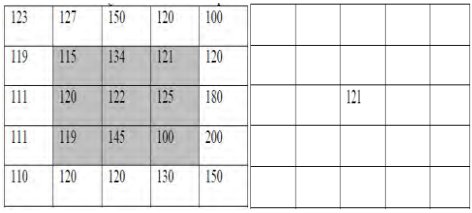


Fig. 5: Median Filter

C. Skew detection/correction:

Skew detection is used to align the paper document with the co-ordinate system of scanner [2].

D. Binarization:

Transformation of a gray scale image into a binary image is called as binarization or thresholding. There are two approached for conversion of gray level image to binary form; i.e. global threshold and local or adaptive threshold [4].

2] Segmentation:

In orders divide text from graphs, images, line, text/graphics segmentation is required. Character segmentation will divide each character from another [2].

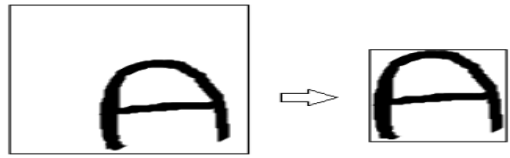


Fig. 6: Segmented image

Classification of recognition-based segmentation.

- **Line segmentation:**

The two text lines between, there are narrow horizontal band with either no pixel or very few pixels [6].

- **Character segmentation:**

Character segmentation is each and every line which is segmented before going through the process of character segmentation [6].

3] Feature Extraction:

Feature extraction is find the set of parameter that define the shape of a character precisely and uniquely [2].

Figure 8: Character extraction image

4] Recognition:

The image from the segment phase is related with all the templates which are preloaded into the system [1].

3. EXPERIMENTS AND ANALYSIS

Table 1: Comparative Study of types of noise

Parameters	Gaussian Noise (Amplifier Noise)	Salt and Pepper Noise	Shot Noise
Characteristics	Gaussian noise is caused by random fluctuations in the signal. It is modelled by random values added to an image.	An image containing salt-and-pepper noise will have dark pixels in bright regions and bright pixels in dark regions.	Shot noise is the noise that can cause, when number of photons sensed by the sensor is not sufficient to provide detectable statistical information.
Danification	In this noise, each pixel in the image will be changed from its original value by a small amount.	In this noise, it presents itself as sparsely occurring white and black pixels.	In this noise, root-mean-square value proportional to the square root of the image intensity, and the noises at different pixels are independent of one another.
Use of method	Mean Filter	Median Filter	-

4. CONCLUSION

Pre-processing technique used in document image as an initial step in character recognition system was presented. The feature extraction stage of OCR is the most required. Noise arises at the time of pre-processing stage creates problem during the recognition of the character. We can eliminate noise from the image by applying Mean Filter and Median Filter.

5. REFERENCES

[1] Ravina Mithe, Supriya Indalkar and Nilam Divekar, " Optical Character Recognition", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-1, March 2013.

[2] Nisha Sharma, Tushar Patnaik, and Bhupendra Kumar, " Recognition for Handwritten English Letters: A Review", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013.

[3] Shalin A.Chopra , Amit A. Ghadge , Onkar A. Padwal , Karan S. Punjabi and Prof. Gandhali S. Gurjar , "Optical Character Recognition" , International Journal of Advanced Research in Computer and Communication Engineering.

[4] S.K.Thilagavathy and Dr.R.Indra Gandhi, "Recognition Of Distorted Character Using Edge Detection Algorithm", International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 4, June 2013, ISSN (Print): 2320 – 9798, ISSN (Online): 2320 – 9801.

[5] Er. Neetu Bhatia Kurukshetra institute of Technology & Management Kurukshetra, India , "Optical Character Recognition Techniques: A Review" , International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 5, May 2014.

[6] Jagruti Chandarana , Mayank Kapadia Department of Electronics and Communication Engineering ,Uka Tarsadia University , "Optical Character Recognition" , International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014).

[7] Mr. Rohit Verma School Of Information Technology APJIMTC , Jalander , India And Dr. Jahid Ali SSCIMT , Badhani , Pathankot , India ," A Comparative Study Of Image Noice and Efficient Noise Removal Techniques " , International Journal Of Advanced Research in Computer Science and software Engineering , Volume 3, Issue 10, October 2013 , ISSN: 2277 128X.

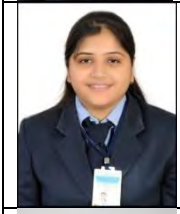
[8] Purna Vithlani , Department of Computer Science Saurahtra University , Rajkot , Gujarati , India , "Pre-processing Techniques in character Recognitin " , International Journal Of Advanced Research in Computer Science and software Engineering , Volume 4, Issue 11, November 2014 , ISSN: 2277 128X.

[9] Poovizhi P Assistant Professor Dept of Computer Science and Engineering SNS College of Engineering Coimbatore TamilNadu, "A Study on Preprocessing Techniques for the Character Recognition", International Journal of Open Information Technologies ISSN: 2307-8162 vol. 2, no. 12, 2014.

[10] Pardeep Kaur and Pooja Choudhary, "Review On: English Scanned Documents ", International Journal of Engineering Research-Online A Peer Reviewed International Journal Vol.3, Issue.2, 2015.

[11] Gaurav Kumar Department of Information Technology, Panipat Institute of Engineering & Technology Samalkha, Haryana, India , "Analytical Review of Preprocessing Techniques for Offline Handwritten Character Recognition " , Special Issue: Proceedings Of 2nd International Conference on Emerging Trends in Engineering and Management, ICETEM 2013.

6. AUTHORS' PROFILE

	**Ishita Patel** is studying in third year MCA programme of Shrimad Rajchandra Institute of Management and Computer Application affiliated to Uka Tarsadia University-Bardoli.
	**Dhwanil Patel** is studying in third year MCA programme of Shrimad Rajchandra Institute of Management and Computer Application affiliated to Uka Tarsadia University-Bardoli.
	**Foram Patel** is studying in third year MCA programme of Shrimad Rajchandra Institute of Management and Computer Application affiliated to Uka Tarsadia University-Bardoli.
	**Bhumika Chauhan** is studying in third year MCA programme of Shrimad Rajchandra Institute of Management and Computer Application affiliated to Uka Tarsadia University-Bardoli.
	**Rinal Mistry** is studying in third year MCA programme of Shrimad Rajchandra Institute of Management and Computer Application affiliated to Uka Tarsadia University-Bardoli.
	**Khushbu Patel** is working as a Teaching Assistant at SRIMCA. Her area of interest is Data Mining and Image Processing.