A Study on Preprocessing Techniques for the Character Recognition

Poovizhi P
Assistant Professor
Dept of Computer Science and Engineering
SNS College of Engineering
Coimbatore
TamilNadu
Email Id: poovizhiponnusamy27@gmail.com

Abstract—**Image processing encompasses a large range of techniques which are used in a very wide range of applications. Handwritten Recognition is an emerging field in the pattern recognition. Documents which are scanned may contain unnecessary information or some translation of the document or de-skew should be performed in order to process the document further. The main objective of the paper is to explain about the importance of the preprocessing steps for the Optical Character Recognition. For the development of OCR for any language, preprocessing step is necessary.**

Keywords—***Preprocessing, OCR, Noise, Binarization, Normalization***

I. Introduction

Optical Character Recognition (OCR) is the electronic conversion of scanned images of the handwritten or printed text into machine encoded text.

There are two types of character recognition. They are Offline and Online Handwriting Recognition. Off-line Handwriting Recognition is the technique which involves the automatic conversion of text in image to machine code which can be used within computer and text processing applications. On-line Handwriting recognition is the automatic conversion of the text which is written on a PDA and the sensor is used to sense the pen-tip movements.

As compared with printed character recognition, handwritten character recognition (HCR) is still a challenging task due to the following factors:

- Individual styles of writing,
- Speed of writing,
- Size of letters,
- Physical and mental condition of the writer,
- Overlap of letters etc.

Again the physical devices those are used during recognition also affect the recognition rate such as the acquisition devices, pen width, pen ink color, etc [2].

The following are the steps of the Optical Character Recognition (OCR): Preprocessing, Feature Extraction, Classification.

A. *Off-line Handwriting Recognition*

Handwriting is different for each persons and it may differ according to his/her situation. Hence, recognition of the handwritten character is very difficult when compared to on-line characters. There is no OCR that supports handwriting recognition [3].

B. *On-line Handwriting Recognition*

Pen or Stylus is used for writing the character on the PDA screen. These characters are then converted into machine readable form. This is not much difficult to recognize when compared to off-line character recognition [4].

Both the on-line and the off-line character recognition contain the following steps:

- Preprocessing
- Feature Extraction
- Classification

II. PREPROCESSING

The process of enhancing the image, which should be used for further processing, is called preprocessing. Preprocessing is the major step in handwriting recognition system.

Noise in a document image is due to poorly photocopied pages. Median Filtering [7], Wiener Filtering method [8] and morphological operations can be performed to remove noise [5]. Median filters are used to replace the intensity of the character image [6], Where as Gaussian filters can be used to smoothing the image [9].

The initial images which we tend to use contain some information which is not necessary for the next step which is feature extraction. The images which are scanned may also contain noises.

The scanned images not only have noises which are inbuilt within it, but also the noise may be during the

P.Poovizhi is an Assistant Professor in the department of Computer Science and Engineering in SNS College of Engineering, Coimbatore, TamilNadu, India. (e-mail: poovizhiponnusamy27@gmail.com) scanning of that image. So the noises and the unwanted information should be removed from the image. Preprocessing is not the single step rather it contains sequences of steps. They are

- Binarization
- Normalization
- Sampling
- Denoising
- Thinning

A. *Binarization*

The conversion of the grayscale image to black and white is called binarization. Binary images are also called as Bi-level or two-level. First, the original RGB image should be converted to grayscale and then the image is converted to black and white image. Most of the OCR packages work on the binarized images.

The conversion is possible because of the threshold values and the values which are higher than the threshold are white and the values which are lower than this threshold are black. Otsu's method is used to perform threshold based on cluster i.e. from gray level image to binary image.

The threshold value 0.5 yields better for all type of images. The laser printer, fax machines can handle the binarized images.

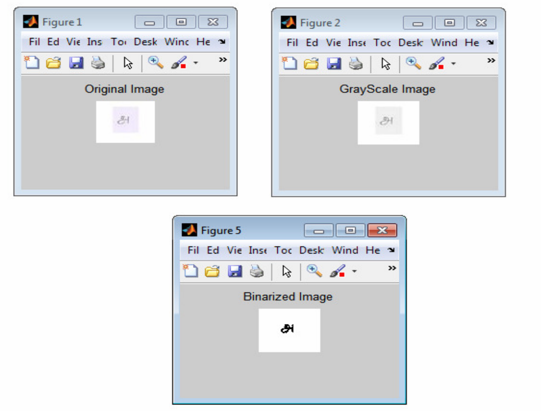

Figure.1. Conversion of the Original Image to Binary Image

B. *Normalization*

The process of changing the intensity value of the pixel to the range of [0,1] is called normalization in image processing. The conversion of various dimension images into fixed dimensions is also called as normalization. Normalization is used in digital signal processing [10]. The matrix values of the image can be normalized along the column and row using the normc and normr commands in Matlab. Further, complications during feature extraction are removed if normalization is done in the earlier stages.

Normalization along the row:

```
nm = [1 4; 5 4];
normr(nm)
```

```
ans =

    0.2425    0.9701
    0.7809    0.6247
```

Normalization along the column:

```
nm = [1 4; 5 4];
normc(nm)
```

```
ans =

    0.1961    0.7071
    0.9806    0.7071
```

C. *Sampling*

Discretization of analog signal is called sampling. The smallest element which is the result of discretization of the space is called pixel.

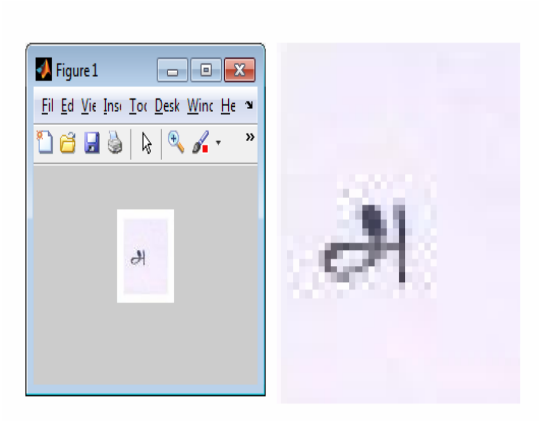

Figure.2. Sampling applied to the character

The process of selecting the subset of individuals from the large sample of population and examining those samples, can generalize the results to the whole population.

Examples

- Conducting a poll to predict the winner of an upcoming election.
- Inspecting a sample of parts to determine if the entire lot meets requirements.

One main advantage of the sampling technique is that it saves much time and it is very reliable.

D. *Denoising*

Digital images are prone to a variety of types of noise. There are several ways that noise can be introduced into an image. For example:

- If there is scanned image then there will be the noise.
- During the electronic transmission of image, the noise can be introduced.

a. *Salt and Pepper Noise*

This type of noises is present in many of the images. This noise can be removed by using Medfilt2 in Matlab. In this paper, Salt and Pepper noise is manually added to the image and medfilt2 removes those noises and the noise which is actually present in the image. But there is no noise in the image. Figure.3(a) and (b) shows the noise removal using ordfilt2 and medfilt2.

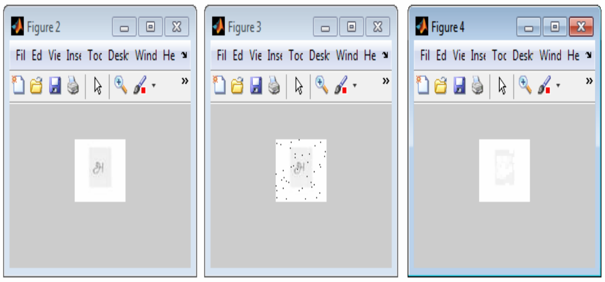

Figure.3 (a) Removal of noise from the Tamil character using Ordfilt2

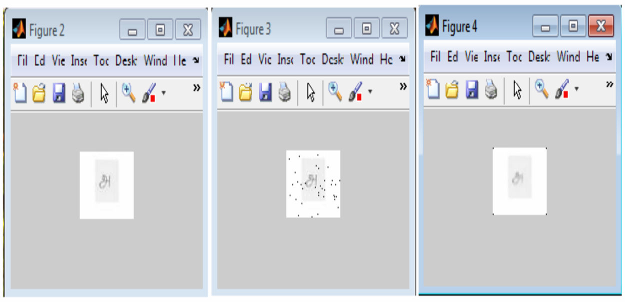

Figure.3 (b). Removal of noise from the Tamil character using Medfilt2

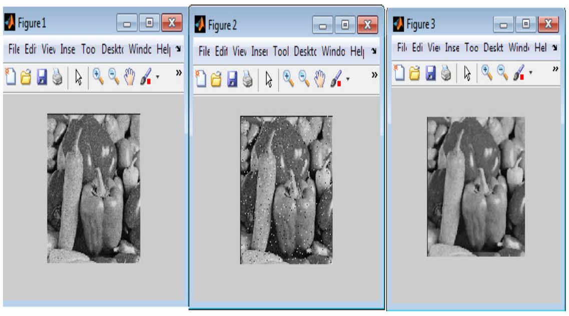

Figure.4. Removal of noise from the sample image using Ordfilt2

The Figure.4 shows the output using Ordfilt2. The original image is converted into grayscale. The salt and pepper noise is added manually to the gray scale image and the noise which is actually present in the image is removed. But for the input image here used does not have any salt and pepper noise.

The noises filtered using median filter gives the best output when compared to the ordfilt2. The Figure.3(a) does not show any image; actually it erases the entire image. Here, Figure.5 shows one sample image with salt and pepper noise is taken to show the noise removal strategy.

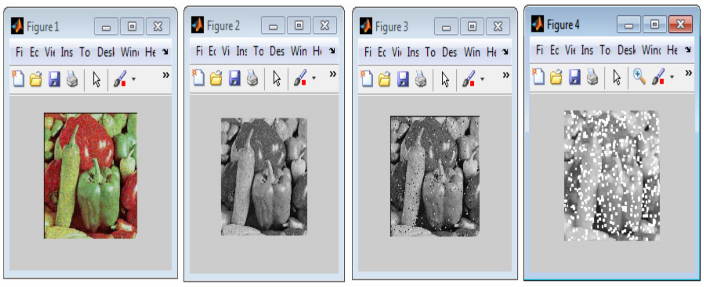

Figure.5. Noise Removal for the character using Ordfilt2

Medfilt2 gives the exact result when compared to wiener filter. The Figure.6 shows the output for wiener filter.

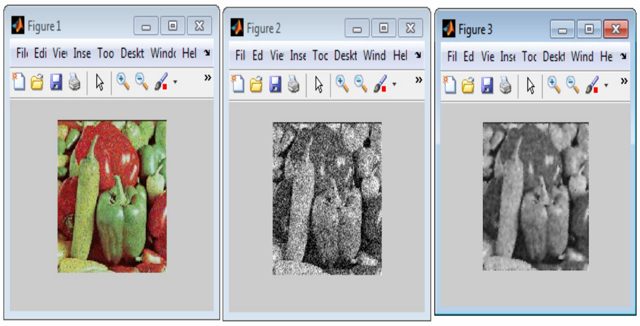

Figure.6. Removal of noise using Medfilt2 to sample image

The result shows that the noise is added using Gaussian and it removes some noise which are inbuilt in the image. The Figure.7 represents the noise removal using the wiener filter.

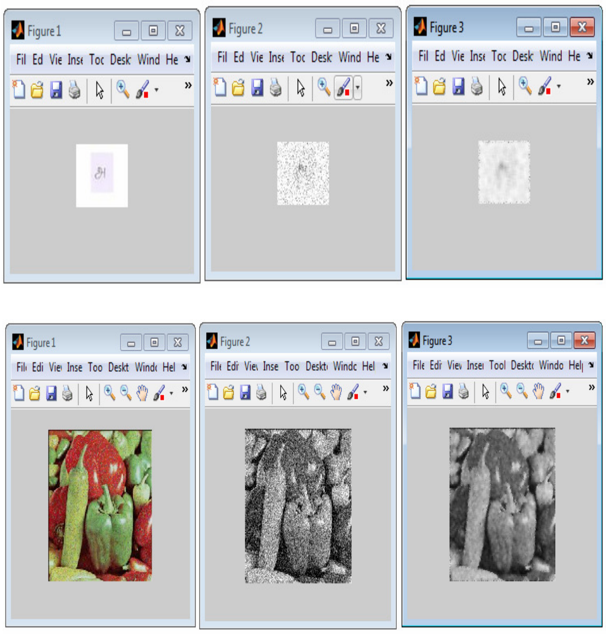

Figure.7. Gaussian noise added and the noise removal using Wiener Filter

The Figure.8 shows the noise removed by imfilter, the noise is added manually using gaussian noise. Imfilter performs well when compared to wiener filter.

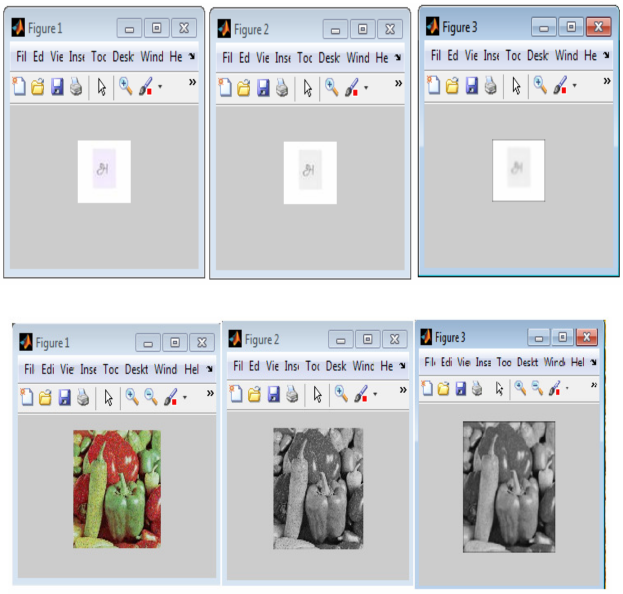

Figure.8. Gaussian noise is added and the noise is removed by imfilter

I. *Thinning*

Thinning is a pre-process which results in single pixel width image to recognize the handwritten character easily. It is applied repeatedly leaving only pixel-wide linear representations of the image characters [1]. Figure.9 shows the thinning process.

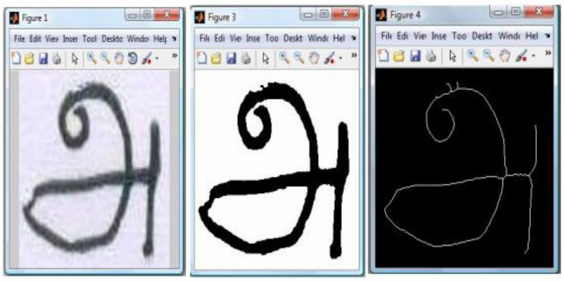

Figure.9. Thinned Image

REFERENCES

[1] Dineshkumar.R, Dr.Suganthi.J "A Research Survey of Sanskrit Offline Handwritten Character Recognition" International Journal of Scientific and Research Publications, Volume 3, Issue 1, January 2013.

[2] Swapnil A. Vaidya, Balaji R. Bombade, "A Novel Approach of Handwritten Character Recognition using Positional Feature Extraction", IJCSMC, Vol. 2, Issue. 6, June 2013, pg.179 – 186.

[3] Jomy John, Pramod K. V, Kannan Balakrishnanm, "Handwritten Character Recognition of South Indian Scripts: A Review", National Conference on Indian Language Computing, Kochi, Feb 19-20, 2011

[4] S. Manke, U. Bodenhausen, .A connectionist recognizer for online cursive handwriting recognition., Proceedings of ICASSP 94, Vol. 2, 1994, pp 633-636.

[5] Shanthi N and Duraiswami K, "A Novel SVM -based Handwritten Tamil character recognition system", springer, Pattern Analysis & Applications,Vol-13, No. 2, 173-180,2010.

[6] Sutha J and RamaRaj N, "Neural network based offline Tamil handwritten character recognition System", International Conference on Conference on Computational Intelligence and Multimedia Vol : 2, page(s): 446 – 450, 2007.

[7] Jagadeesh Kumar R, Prabhakar R and Suresh R.M, "Off-line Cursive Handwritten Tamil Characters Recognition", International Conference on Security Technology, page(s): 159 – 164, 2008.

[8] Anil.K.Jain and Torfinn Taxt, "Feature extraction methods for character recognition-A Survey," Pattern Recognition, vol. 29, no. 4, pp. 641 - 662, 1996.

[9] L. R. B. Schomaker, H. L. Teulings, .A handwriting recognition system based on the properties and architectures of the human motor system., *Proceedings of the IWFHR, CENPARMI*, Concordia, Montreal, 1990, pp 195-211.

[10] Normalization http://en.wikipedia.org/wiki/Normalization_(image_processing)

P.Poovizhi received M.E degree in Software Engineering from Bannari Amman Institute of Technology, 2014. She has completed her Bachelor's degree in Information Technology from Bannari Amman Institute of Technology, 2012. Her email_id is poovizhiponnusamy27@gmail.com