International Journal of Advanced Research in IT and Engineering

ISSN: 2278-6244

AN INTRODUCTION TO TYPES OF PRE-PROCESSING ON OPTICAL CHARACTERS

Navid Samimi Behbahan*

Milad Samimi Behbahan*

Abstract: *In this paper an effective method for pre-processing of images on handwritten data with high variations in order to extract features and finally reaching the more desirable than has been studied with recognition rates. So far, the process is generally what is a used Persian handwritten digit prepared including thinning of binary data, and the size is. This study has shown that the pre-processing steps proposed adding significantly improve recognition rates.*

*Department of Computer Engineering, Behbahan Branch, Islamic Azad University, Behbahan, Iran

1- INTRODUCTION

All the processing which is done on the raw image signals to facilitate or increase the accuracy of the subsequent phases includes:

- ✓ Noise reduction
- ✓ Normalizing data
- ✓ Compression of the amount of data which must be protected

2- NOISE REDUCTION

Noises generated by the optical scanning machines or writing instruments, cause fraction, connection between lines, filling the pores in the picture of some letters and etc.

Before letter recognition, it is essential to correct these defects. Different techniques of noise reductions can be classified into three main groups:

2-1- filtering

This method helps to remove noise, rough body of letters usually written by a rough surface (in the case of handwritten text), or reduces weak sampling rate of data acquisition systems. The main idea in this method is convolutioning a pre-defined mask with the picture in order to allocate the new amount to the pixel according to the function (mathematic function) of the adjunct pixels. Filters can be designed for different purposes like smoothing, sharpening, applying to threshold level, removing the background texture and contrast setting [1].

2-2- Morphologic operators

The main idea underlying these operators is filtering the text files by using logical operators instead of convolutions. Many morphologic operators can be designed to smooth cantors, reduce points of errors and thinning the letters. Therefore morphologic operators can successfully eliminate noises on the images which are created due to the low quality documents or irregular motion of author's hand [2].

2-3- Noise modeling

In case there is a model for the existing noise, noise can be removed by using some calibration techniques. However, noise modeling is not possible for most applications. Little research has been done on noise modeling caused by optical defects such as some spots appearing on scanned pictures, blurred pictures or on skewed pictures [3].

3- NORMALIZING DATA

Normalization methods help to remove changes in writing and therefore it results in standardized data [4]. Basic methods of normalizing are:

3-1- Normalizing Skewedness of words

Due to lack of attention in scanning stage or carelessness of author while creating the handwritten text, characters might be slightly rotated or deviated. This issue can affect the operation of algorithms used in next stages of OCR system.

Developed algorithms used to detect deviation of the screen are roughly the same number of algorithms for binary images. All of these methods operate accurately on text pages with uniform alignment.

3-2- Normalizing the Skewedness

This phenomenon is known as “skewedness”. Skewedness is defined as the slope angle among the longest stroke in a word and vertical direction. Normalization of the skewed in order to normalize all the characters is used in a standard form. The most common ways to determine the skewed, is calculating the mean angel of the components which are near the vertical line. Extracting vertical lines of characters is done by following chain code elements by a pair of one-dimensional filtering. Coordination of the start and end of each line, gives us the angel of the skewed.

A B

Figure 1: skewed estimation with consideration A-components near the vertical line B-average slope angle

3-3- Normalizing the size

In OCR systems most of the images which are very small or very big, will be normalized to a standard size. This is done by re-sampling the image. Methods like Bilinear or Bicubic operate properly on the images with gray surface. (Alusefi and Ayda 1992)

ISSN: 2278-6244

Normalizing could be done as a part of training stage and estimate the measure parameters separately for each individual training data. In Figure 2, two characters gradually become smaller to the optimal size therefore recognition rate in training data will be maximum.

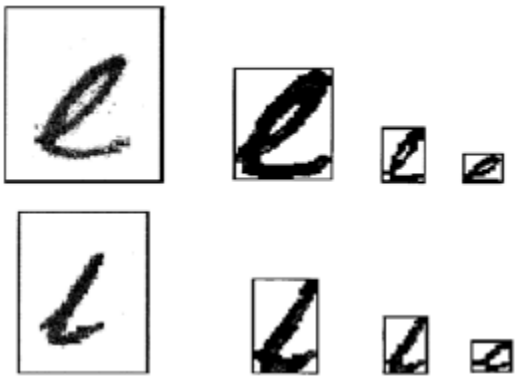

Figure 2: normalizing λ and 3

3-4- Contour smoothing

In handwritten texts, due to the unwanted vibration or movement of the author's hand while writing the text, shape of the letters' contour may be uneven or lumpy. (Figure 3) according to Arika and Yamin(2013) the amount of each pixel of the text is replaced with the average of its neighboring pixel. By repeating this procedure twice, smoother image may be obtained from handwritten text; therefore, the effect of hand-shake will reduce. (Figure 3)

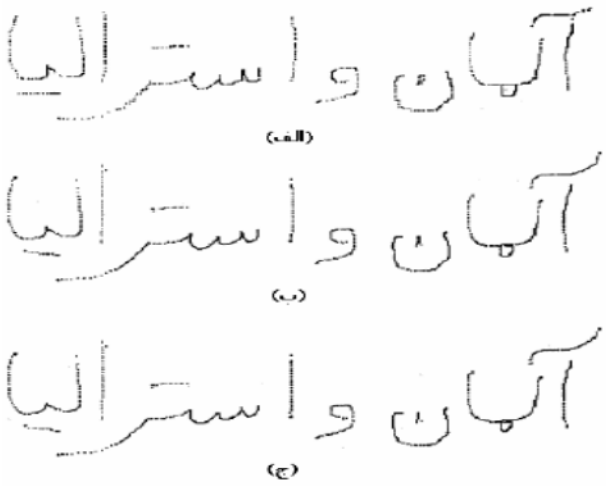

Figure 3: Applying smoothing algorithm on written text A-original image B-one stage of smoothing C-two stage of smoothing

4- COMPRESSION

An accepted issue is that classic techniques of image compression which transfer the image from the local domain to different domains, is not suitable for word recognition. In word recognition, compression requires local domain techniques which preserve the shape information. Two conventional compression techniques are techniques applied to threshold (in order to binary the gray surface of texts) and the other one is thinning [5].

4-1- Binary (bi-leveling) the image of the text

In order to reduce the required storage size and speed up processing, it is often desirable to convert grayscale images or color images by choosing a threshold level. Two methods of applying threshold levels are global and local. In local threshold level application, a threshold amount for the entire image is chosen. This amount is often measured by the estimation of the background surface which is calculated by the histogram of the brightness level.

4-2- Thinning

While it causes a marked decrease in the size of data, it extracts data which are about the shape of characters. Two basic approaches to thinning are: pixel wise thinning or non-pixel thinning. In pixel wise thinning, pixel processes the data locally and repeatedly until only its structure in a pixel size remains from the character. Figure 4 shows a text accompanied by its thinned picture.

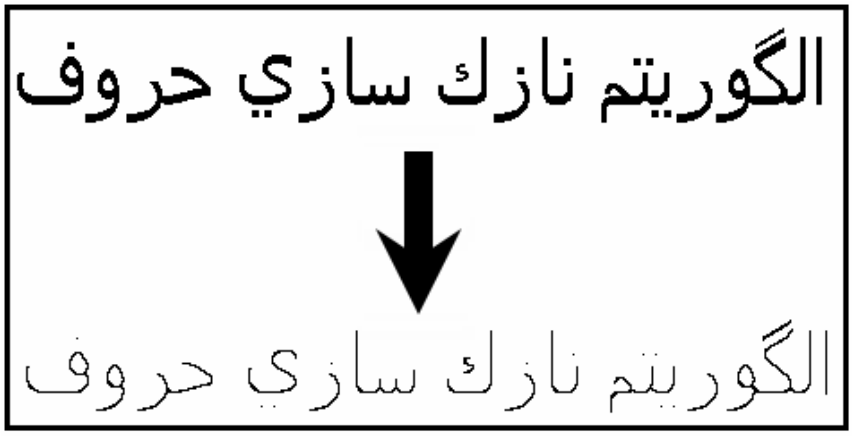

Figure 4: applying thinning on a sample text image

5- CONCLUSION

In addition to word recognition techniques, pre-processing techniques have been examined and studied in many image processing fields. Note that these techniques change data; therefore, it is possible that they direct unexpected distortions to the image document.

Thus, these techniques may result in loss of some important data. Consequently, enough attention must be paid while using them.

REFERENCES

[1] J.T. Tou and R.C. Gonzalez, “Pattern Recognition Principles”, Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1974

[2] F. Chang, “Retrieving information from document images: problems and solutions”, Int. J. Doc. Anal. Recognition 4 (2001) 46–55

[3] C. Wolf, D. Doermann, “Binarization of low quality text using a Markov random field model”, in: Proceedings of the 16th International Conference on Pattern Recognition, vol. 3, 2002, pp. 160–163

[4] D. Lopresti. “Optical character recognition errors and their effects on natural language processing”. In Proceedings of the Workshop on Analytics for Noisy Unstructured Text Data, pages 9–16, Singapore, July 2008.

[5] J. Sauvola, M. Pietikainen, “Adaptive document image binarization”, Pattern Recognition 33 (2) (2000) 225–236.