International Journal of Engineering & Technology, 7 (4) (2018) 3148-3153

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.18952
Research paper

An Efficient Character Recognition Technique Using K-Nearest Neighbor Classifier

Nawaf Hazim Barnouti[1]*, Mohammed Abomaali[2], Mohanad Hazim Nsaif Al-Mayyahi[3]

[1]Al-Mansour University College, Baghdad, Iraq
[2]AlSafwa University College, Computer Techniques Engineering Department, Karbala, Iraq
[3]Al-Mansour University College, Baghdad, Iraq
**Corresponding author E-mail: nawaf_hazim87@yahoo.com*

Abstract

Optical Character Recognition (OCR) Systems offers human machine interaction and are commonly used in several important applications. A lot of research has already been accomplished on the character recognition in different languages. This paper presents a technique for recognition of Printed text with noise using Optical Character Recognition (OCR). The main steps of this system are preprocessing of the text including converting the text image to black/white and remove the noise from the text image, segmentation of the text image to each character, Feature extraction using zoning-based technique and classification. The System is implemented using MATLAB 2016a software application program and is still under development. Noise is removed from all the text images. The quality of the input document is very important to achieve high accuracy. The system is able to recognize characters in different 50 images.

Keywords: *Optical Character Recognition; Feature Extraction; Segmentation; Zoning; K-Nearest Neighbor.*

1. Introduction

The field of Optical Character Recognition (OCR) has gained more attention in the recent years because of its importance and applications. Character recognition is a process of detecting and recognizing characters from input image (handwritten, printed, or typewritten) and converts it into American Standard Code for Information Interchange (ASCII) or other equivalent machine editable form [1]. Character recognition is one of the most interesting and fascinating areas of pattern recognition and artificial intelligence [1]. Optical character recognition (OCR) is the branch of technology that deals with the automatic reading of text. The goal is to classify optical patterns (often contained in a digital image) corresponding to alphanumeric or other characters. And to imitate the human ability to read - at a much faster rate - by associating symbolic identities with images of characters. As the emphasis shifts from recognizing individual characters to recognizing whole words and pages, more general terms being used include optical text recognition and document image processing.

The process of OCR involves several steps including segmentation, feature extraction, and classification. Recognizing text in scene images is much more challenging due to the many possible variations in backgrounds, textures, fonts, and lighting conditions that are present in such images. Two major types of character recognition in computer science are in place: (1) Optical Character Recognition (OCR): techniques based solely on image processing techniques which include extracting features in the image, comparing those features with predefined ones and finally character recognition, and (2) Intelligent Character Recognition (ICR): includes machine learning algorithms within the recognition process. Also targets handwritten print script or cursive text one glyph or character at a time [2].

In development of computerized OCR system, few problems can occur. First: there is very little visible difference between some letters and digits for computers to understand. For example, it might be difficult for the computer to differentiate between digit "0" and letter "o". Second: It might be very difficult to extract text, which is embedded in very dark background or printed on other words or graphics [3]. There are many applications of OCR, which includes: License plate recognition, image text extraction from natural scene images, checks process in banking, handwriting recognition, extracting text from scanned documents checks process in banking, handwriting recognition, extracting text from scanned documents, data entry for business documents, assistive technology for blind and visually impaired users, extracting business card information into a contact list etc. [4].

Understanding scene text is more important than ever. One could, for instance, foresee an application to answer questions such as, "What does this sign say?". This is related to the problem of Optical Character Recognition (OCR), which has a long history in the computer vision community. However, the success of OCR systems is largely restricted to text from scanned documents [5]. At the high level, the general problem of end-to-end text recognition consists of two primary components: text localization and word recognition. First, in text localization, the goal is to locate individual words or lines of text. Then, once we know where the regions of text are located in the image, we seek to identify the actual words and lines of text in those regions. Over the years, much time and effort have been invested in solving different components of the text-recognition problem. As a direct result, there now exist algorithms that achieve extremely high performance on specialized tasks such as digit recognition in constrained settings [6].

2. Related Work

Existing work on text recognition has focused primarily on optical character recognition in printed and hand-written documents since there exists a great demand in and market for document readers for office automation systems. These systems have attained a high degree of maturity. Further text recognition work can be found in industrial applications. Most published methods for text localization and recognition are based on sequential pipeline processing consisting of three steps - text localization, text segmentation and processing by an OCR for printed documents.

In [2], OCR system has been presented. The presented system is based on character extraction, characters recognition, and text correction to recognize text in poor quality images. The character extraction preprocessing step is performed to ease the recognition process using chain-code. The performance of the proposed system is evaluated using ICDAR 2011 dataset. The proposed system achieves 74.02% correctly recognized word rate.

In [4], proposed an effective method to recognize scene text. Our model combines bottom-up cues from character detections and top-down cues from lexica. We infer the location of true characters and the word they represent as a whole jointly. Our results show that scene text can be recognized with a reasonably high accuracy in natural, unconstrained images.

In [7], produced a text detection and recognition system based on a scalable feature learning algorithm and applied it to images of text in natural scenes. Our results point out that it may be possible to achieve high performance using a more automated and scalable solution.

In [8], have presented the solution for optical character recognition (OCR) in printed document image with considerably improved accuracy in various noisy environments and less memory consuming. Our proposed approach uses minimal character set. However, it is not specified for different writing styles and font size issues. The following key challenges can be further covered by adding those in training data.

3. Text Recognition Dataset

In this work CHARS74K and ICDAR2003 datasets are used. CHARS74K dataset that is shown in Figure 1 contains a total of over 74K images which explains the name of the dataset. ICDAR2003 dataset that is shown in Figure 2 contains 249 images with 5370 characters and 1106 words [9]. Each dataset split into a number of training and testing portions [10].

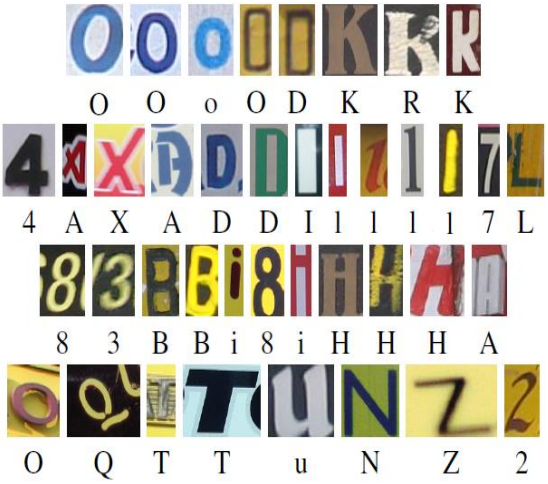


Fig. 1: Chars74K dataset Samples

Fig. 2: ICDAR2003 dataset Samples

4. The Proposed System

In this work, automatic character recognition system is implemented on noisy text images. Figure 3 shows the block diagram of the proposed character recognition system. The system involves different stages including: Image acquisition, Preprocessing, Segmentation, Feature Extraction and Classification.

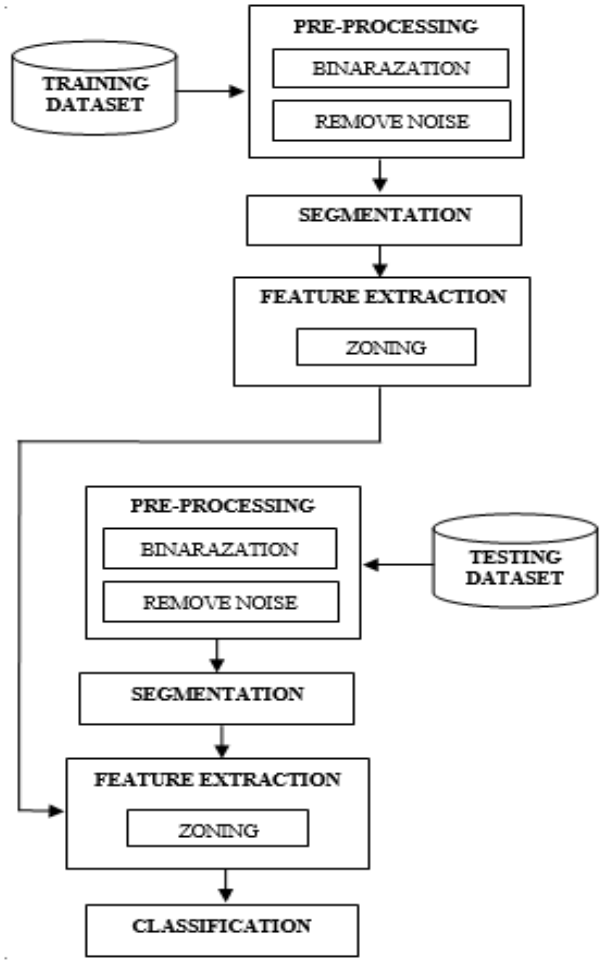


Fig. 3. The Proposed Text Recognition System.

The recognition of the text and characters begins with obtaining a digitized image of the text using a suitable scanning system. In the second stage the Preprocessing of the image goes on (Binarization and noise removal). In the third stage the segmentation of the text to individual characters goes on. Segmentation of individual text to characters is an essential and difficult stage in text recognition system. The fourth stage is the Feature Extraction stage. The main

benefits of Feature Extraction is in fact to removes redundancy from the data and indicate the character image by some numerical features. The last stage is the Classification Process. This will compare feature vectors to the different models and find the closest match.

4.1 Image Pre-Processing

The digital image may contain a certain amount of noise depending on the resolution of the scanner. The recognition rates could be poor since the character may be smeared or broken. This can usually eliminate by using a pre-processing technique to smooth the digital image [11].

4.1.1 Binarization

Binarization is the process of converting a pixel image to a binary image as shown in Figure 4. Binary image also called bi-level or two-level has only two possible values for each pixel which represent the color black or white.

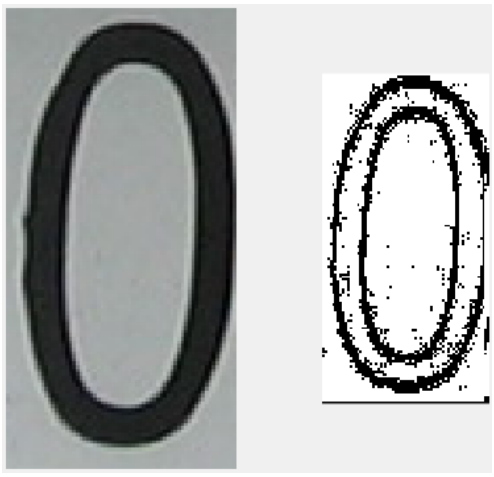


Fig. 4: Convert PNG Image to Binary Image

4.1.2 Remove Noise

Noise can significantly impact the quality of digital images. Various techniques (Mean Filter, Median Filter, Local Pixel Grouping, Adaptive Filter, Wiener Filter, etc.) can be used to connect unconnected pixels, to remove isolated pixels, to smooth pixels boundary [12]. Median filter is used to remove the noise that is added on character 'a' [13] for example as shown in Figure 5.

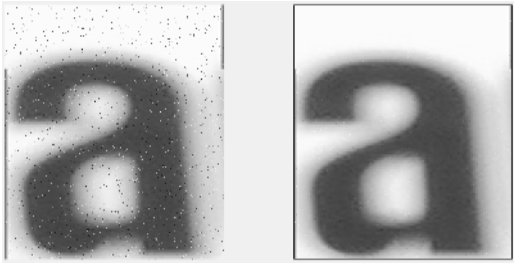


Fig. 5: Remove Noise Using Median Filter

4.2 Segmentation

Segmentation is among the most crucial and is an essential step in an OCR. The most of optical character recognition techniques will segment the words into individual characters which can be recognized individually [14]. You must find the regions of the document in which data are printed and separate them from figures and graphics. A poor segmentation process gives misrecognition or rejection.

4.2.1 Line Segmentation Process

The lines of a text block are discovered by checking or scanning the input image horizontally. Frequency of black pixels in each row is counted in order to build the row histogram [15]. When black pixels frequency in a row is zero it indicates a boundary between two white pixels consecutive lines as shown in Figure 6.

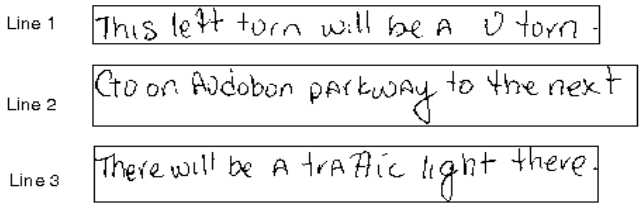


Fig. 6: Line Segmentation Process

4.2.2 Word Segmentation Process

When a line has been discovered, after that each line is scanned vertically for word segmentation. Number of black pixels in each column is determined to build column histogram. When no black pixel is found in vertical scan which is proved to be the space between two words [16]. Therefore, we can separate the words as shown in Figure 7.

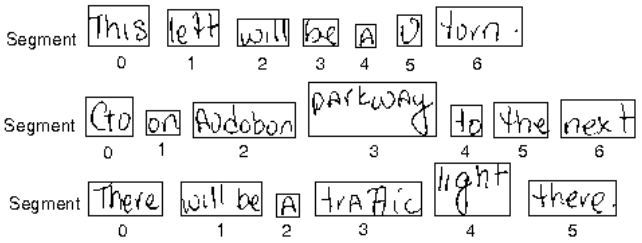


Fig. 7: Word Segmentation Process

4.3 Feature Extraction

The purpose of feature extraction would be to capture the important characteristics of the symbols, which is usually accepted that this is among the most difficult problems of pattern recognition [16]. Two kinds of features are found, statistical features and structural features. The majority of researchers accept that statistical features could be found quickly using easy methods and might carry out high recognition results especially in closed testing data. Structural features are more conformed to the intuitive thinking of human mind, its more robust for the deformation of symbols. However, they usually depend on human summarized rules for the recognition algorithm [17].

4.3.1 Zoning

Zoning is a well-known technique used in character recognition. In this technique, the rectangle character images are divided into a number of overlapping or non-overlapping regions (zones) of predefined sizes [15]. These predefined sizes are usually of the order 2×2, 3×3, 4×4 etc. Then features are computed for each zone. The average pixel density was found by dividing the number of foreground pixels by the total number of pixels in each zone [16][18].

4.4 The Classification Process

The purpose of character classification would be to assign certain part of text to one or more predefined classes or categories. The part of text could be a document, news article, search query, email, tweet, support tickets, customer feedback, user product review etc. Applications of classification include categorizing newspaper articles and news wire contents into topics, organizing web pages into hierarchical categories, filtering spam email, sen-

timent analysis, predicting user intent from search queries, routing support tickets, and analyzing customer feedback [16]. There are two steps in creating a classifier: training and testing [19]

4.4.1 K-Nearest Neighbor (K-NN)

This algorithm is non-parametric machine learning which used for classification. The viewpoint behind k-Nearest Neighbor algorithm is very straightforward [20]. To classify a new character, the system discovers the k nearest neighbors among the training datasets, and uses the categories of the k nearest neighbors to weight the category candidates. Several researchers have discovered that the k-NN algorithm comes up with very good performance for character recognition in their research projects on different data sets [21].

A distance function is needed to compare points similarity. Euclidean Distance can be used between the test point and all the reference points in order to find K nearest neighbors, and then arrange the distances in ascending order and take the reference points corresponding to the k smallest Euclidean Distances. A test sample is then attributed the same class label as the label of the majority of its K nearest neighbors [22]. Euclidean Distance can be calculated using the equation below:

$$EuclideanDistance(X,Y) = \sqrt{\sum_{n=1}^{No.of\ Images} (X_n - Y_n)^2} \qquad (1)$$

The overall performance of this algorithm very much depends on two conditions or factors, that is, a suitable similarity function and an appropriate value for the parameter k.

5 Result and Discussion

Experiments have been performed to test the proposed system. The developed character recognition system has been tested using randomly selected scanned text. CHARS74K and ICDAR2003 datasets are used and divided into training and testing portions. The system is implemented using MATLAB R2016a software application program.

5.1 Create Our Own Dataset

You can also create your own characters dataset additional to the two datasets that has been used in this work. Basic step is to create character images including numbers from 0 to 9 and characters from A to Z as shown in Figure 8. These images are cropped and has the same image size.

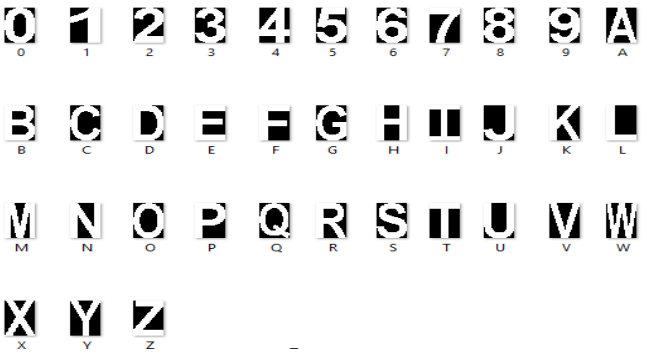


Fig. 8: Dataset for English Numbers and Characters

In this work it's not necessary to create this dataset because two datasets with much more different character shapes and style are used.

5.2 Loading Input Image

Images can be imported into the GUI by clicking on the Select Text Button as shown in Figure 9. In this application software PNG, TIF and JPG file formats are supported. The selected image will be used for the recognition process.

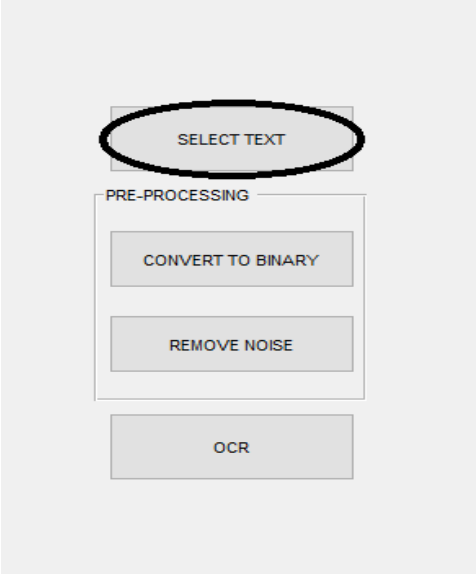


Fig. 9: How to Select the Text Image

5.3 Image Pre-Processing

Image are selected from a specific file and then converted to gray scale images. Median Filter is used to remove the noise from the selected image by clicking on the Remove Noise button. The last step of pre-processing is converting the image to black and white just by clicking on the Convert to Binary button. The pre-processing steps is shown in Figure 10. In order to maintain size uniformity of all the character images, they are resized into a standard dimension.

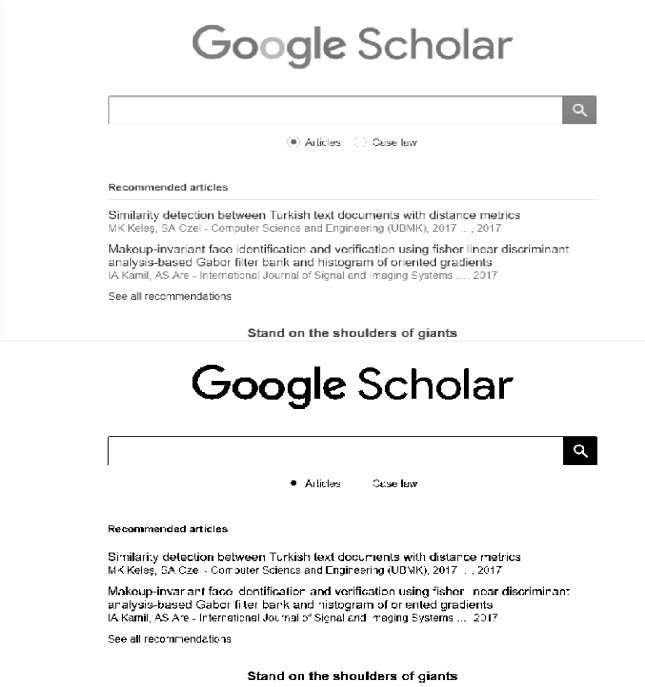


Fig. 10: Remove Noise and Convert Input Image to Binary Image.

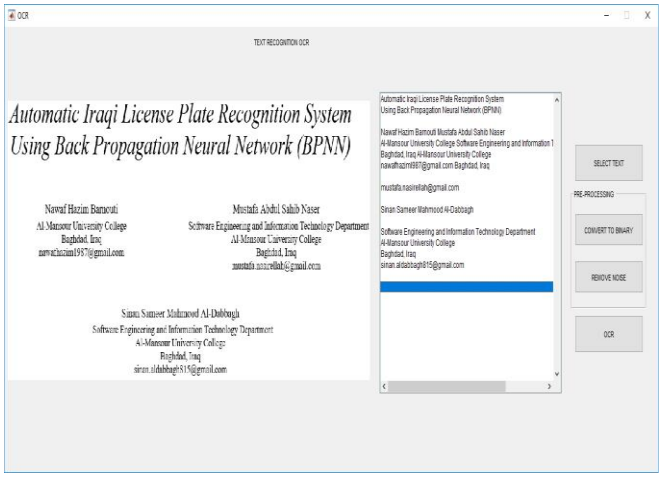
5.4 Segmentation and Feature Extraction

Segmentation and feature extraction are applied after clicking on the OCR button. Then after that the recognition process will continue to recognize all the characters inside the image and display the result. Segmentation and feature extraction process was powerful on images having no noise and acceptable on images with noise. In the proposed system GUI selected image displayed on the left side and the result will be displayed on the right side.

5.5 Recognition Process

The proposed system performance was evaluated on large number of images using two different datasets (CHARS74K and ICDAR2003) with the same parameter settings. An English independent dataset also was created for numbers and characters. Testing is performed in two stages (on documents having no noise and on the documents with noise).

The K-NN classifier is used on both (CHARS74K and ICDAR2003) datasets for classification process. Three options can be occurred (text is localized and recognized meaning that its matched, text is localized correctly but not recognized meaning that it's not matched, or text is not localized at all meaning that the text is not found). Figure 11 shows an example of some input images that is used for recognition.



Automatic Iraqi License Plate Recognition System
Using Back Propagation Neural Network (BPNN)

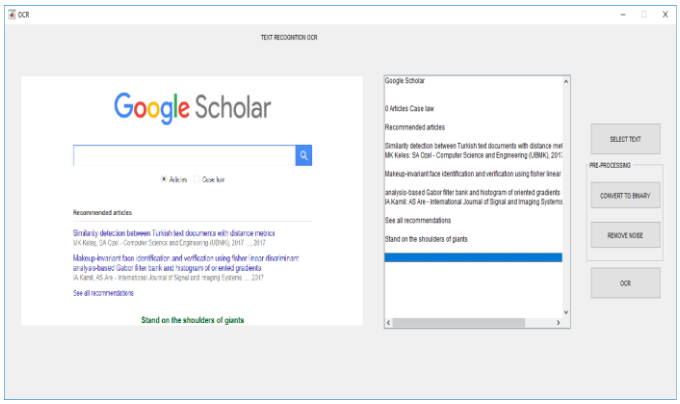
Nawaf Hazim Barnouti Mustafa Abdul Sahib Naser
Al-Mansour University College Software Engineering and Information
Baghdad, Iraq Al-Mansour University College
nawafhazim1987@gmail.corn Baghdad, Iraq

mustafa.nasirellah@gmail.com

Sinan Sameer Mahmood Al-Dabbagh

Software Engineering and Information Technology Department
Al-Mansour University College
Baghdad, Iraq
sinan.aldabbagh815@gmail.com

(A)



Google Scholar

0 Articles Case law

Recommended articles

Similarity detection between Turkish text documents with distance met
MK Keles: SA Ozel - Computer Science and Engineering (UBMK), 2017

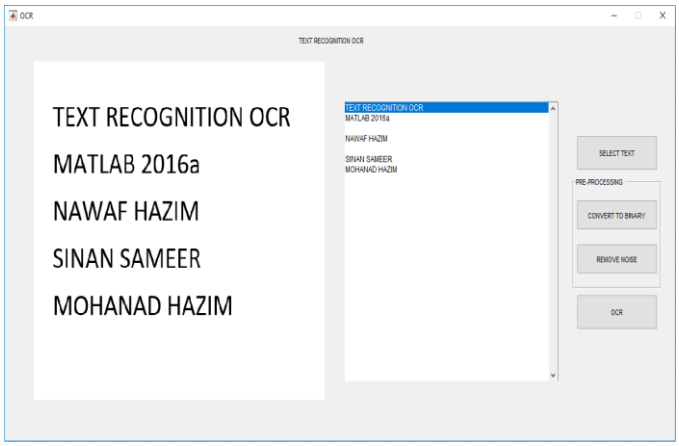
Makeup-invariant face identification and verification using fisher linear

analysis-based Gabor filter bank and histogram of oriented gradients
IA Kamil: AS Are - International Journal of Signal and Imaging Systems

See all recommendations

Stand on the shoulders of giants

(B)



TEXT RECOGNITION OCR
MATLAB 2016a

NAWAF HAZIM

SINAN SAMEER
MOHANAD HAZIM

(C)

Fig. 11: Recognition Results: (A) PDF Document (B) Google Scholar Web Page (C) Microsoft Word Document

6 Conclusion

In this paper, an effective character recognition system using OCR has been proposed. The proposed system is based on image pre-processing, characters segmentation, feature extraction, and classification process. English numbers and characters dataset were created in addition to CHARS74K and ICDAR2003 datasets. Two types of images are used (images having no noise and images with noise). k-Nearest Neighbor algorithm is used for classification process. The last step, Euclidean Distance is used because a distance function is needed to compare similarity. The proposed character recognition system performance was evaluated and high recognition rate was achieved.

References

[1] N. Venkata Rao, A.S.C.S.Sastry, A.S.N.Chakravarthy, and K. P, "Optical Character Recognition Technique Algorithms," Journal of Theoretical and Applied Information Technology, vol. 83, no. 2, pp. 275-282, 2016.

[2] A. H. Ahmed, M. Afifi, M. Korashy, E. K.William, M. A. El-sattar, and Z. Hafez, "OCR System for Poor Quality Images Using Chain-Code Representation," The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015). Beni Suef, Egypt. Springer, Cham, pp. 151-161, 2016.

[3] C. Patel, A. Patel, and D. Patel, "Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study," International Journal of Computer Applications, vol. 55, no. 10, pp. 50-56, 2012.

[4] A. M. A. M. Asif, S. A. Hannan, Y. Perwej, and M. A. Vithalrao, "An Overview And Applications Of Optical Character Recognition," International Journal of Advance Research In Science And Engineering, vol. 3, no. 7, pp. 261-274, 2014.

[5] AnandMishra, K. Alahari, and C. V. Jawahar, "Top-Down and Bottom-up Cues for Scene Text Recognition," Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 2687-2694, 2012.

[6] L. Neumann, and J. Matas, "A method for text localization and recognition in real-world images," Asian Conference on Computer Vision. Springer, Berlin, Heidelberg, pp. 770-783, 2010.

[7] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng, "Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning," International Conference on Document Analysis and Recognition (ICDAR). IEEE, pp. 440-445, 2011.

[8] S. Babu, Z. A. Masood, S. Munir, S. Adnan, and I. Bari, "Android Based Optical Character Recognition for Noisy Document Images," International Journal of Computer Science and Information Security (IJCSIS), vol. 14, no. 1, pp. 34-37, 2016.

[9] T. E. d. Campos, B. R. Babu, and M. Varma, "Character Recognition In Natural Images," 2009.

[10] L. Neumann, and J. Matas, "Real-Time Scene Text Localization and Recognition," Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 3538-3545, 2012.

[11] M. A. Mohamad, D. Nasien, H. Hassan, and H. Haron, "A Review on Feature Extraction and Feature Selection for Handwritten Character Recognition," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 6, no. 2, pp. 204-212, 2015.

[12] A. Fabijańska, and D. Sankowski, "Image Noise Removal – The New Approach," 9th International Conference on the Experience of Designing and Applications of CAD Systems in Microelectronics (CADSM'07). IEEE, pp. 457-459, 2007.

[13] S. Kaur, "Noise Types and Various Removal Techniques," International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), vol. 4, no. 2, pp. 226-230, 2015.

[14] K. C. Nguyen, and N. Masaki, "Text-Line and Character Segmentation for Off-line Recognition of Handwritten Japanese Text," IEICE technical report 115.517 pp. 53-58, 2016.

[15] M. Sarfraz, S. N. Nawaz, and A. Al-Khuraidly, "Offline Arabic Text Recognition system," Proceedings International Conference on Geometric Modeling and Graphics. IEEE, pp. 30-35, 2003.

[16] P. Singh, and S. Budhiraja, "Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey," International Journal of Engineering Research and Applications (IJERA), vol. 1, no. 4, pp. 1736-1739, 2011.

[17] M. Z. Hossain, M. A. Amin, and H. Yan, "Rapid Feature Extraction for Optical Character Recognition," arXiv preprint arXiv:1206.0238 2012.

[18] P. Vithlani, and C.K.Kumbharana, "Structural and Statistical Feature Extraction Methods for Character and Digit Recognition," International Journal of Computer Applications, vol. 120, no. 24, pp. 43-47, 2015.

[19] Y. Elglaly, and F. Quek, "Isolated Handwritten Arabic Characters Recognition using Multilayer Perceptrons and K Nearest Neighbor Classifiers," pp. 1-6, 2011.

[20] M. Rajalingam, P. Sumari, and V. Raman, "Text Detection and Extraction from Document Images using K-Nearest Neighbor Rule," International Journal of Computer and Information Technology, vol. 3, no. 4, pp. 731-736, 2014.

[21] YingquanWu, K. Ianakiev, and V. Govindaraju, "Improved k-nearest neighbor classi□cation," Pattern Recognition, vol. 35, no. 10, pp. 2311-2318, 2002.

[22] D. K. Patel, T. Som, S. K. Yadav, and M. K. Singh, "Handwritten Character Recognition Using Multiresolution Technique and Euclidean Distance Metric," Journal of Signal and Information Processing, vol. 3, no. 2, pp. 208-214, 2012.