Evaluation of Character Recognition Algorithm Based on Feature Extraction

Bishakha Sharma(✉) and Arun Agarwal

ITM Group of Institutions, Gwalior, MP, India
bishakhasharmajmp@gmail.com, arun.agarwal@itmgoi.in

Abstract. At display circumstance there is creating enthusiasm for the item system to see characters in a PC structure when information is investigated paper records. This paper presents point by point review in the field of Optical Character Recognition. Diverse methods are settled that have been proposed to comprehend the point of convergence of character affirmation in an optical character affirmation structure. Decision and feature extraction in light of Optical Character Recognition (OCR). By using the OCR, we can change the information of picture into the information of substance which is definitely not hard to control. In our proposed method, Select the any particular number and crop the selected image and then extract the feature. The text from the OCR process will be compared with the selected number from the loaded image. The overall accuracy of the proposed method is 92%.

Keywords: Image processing · Optical Character Recognition
Feature extraction

1 Introduction

A piece of software through which printed text and images can be converted into digitized form such that it can be manipulated by machine is known as character recognition system. The human brain which has the capability to very easily recognize the text/characters from an image, but machines have not enough capability to perceive image information. Therefore, a large number of research efforts have been put forward that attempts to transform a document image to format understandable for machine.

OCR is a mind boggling issue in light of the assortment of dialects, scholarly styles and styles in which substance can be made, and the versatile tenets of tongues and so on. Thusly, frameworks from various solicitations of programming outlining (i.e. picture dealing with, design depiction and trademark vernacular arranging and so forth are utilized to address grouped difficulties. This paper acclimates the peruser with the issue. It enlightens the reader with the historical perspectives, applications, challenges and techniques of OCR. [1]

Optical Character acknowledgment has been a subject of research. Example acknowledgment has three fundamental advances: perception, design division, and example arrangement. Optical Character Recognition (OCR) frameworks is changing

S. Verma et al. (Eds.): CNC 2018, CCIS 839, pp. 76–88, 2019.
https://doi.org/10.1007/978-981-13-2372-0_7

substantial measure of records, either printed letters in order or transcribed into machine encoded content with no change, tumult, affirmation blends and different segments.

At the point when all is said in done, handwriting affirmation is portrayed into two sorts as offline and On-line character recognition. Offline recognition includes programmed change of content into a picture into letter codes which are usable inside PC and content preparing applications. Offline recognition is more troublesome, as different people have assorted handwritten styles. Be that as it may, in the on-line framework, On-line character acknowledgment manages an information stream which originates from a transducer while the client is composing.

The normal equipment to gather information is a digitizing tablet which is electromagnetic or weight delicate. At the point when the client composes on the tablet, the progressive developments of the pen are changed to a progression of electronic flag which is remembered and investigated by the PC. Optical Character Recognition (OCR) is a field of research in design acknowledgment, manmade brainpower and machine vision, signal processing. It is additionally aforementioned that Optical character recognition (OCR) is reffered to as associate Off-line character recognition system during which system scans and static image of the characters ought to be recognized. It alludes to the mechanical or electronic interpretation of pictures of manually written character or printed content into machine code with no variation [2] (Fig. 1).

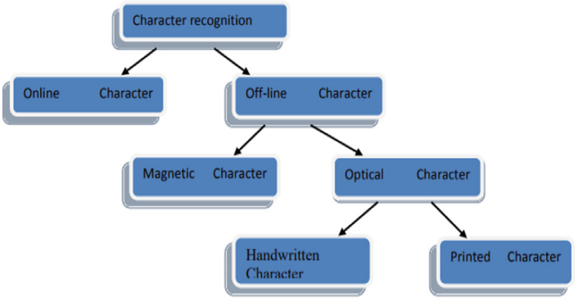

Fig. 1. Character recognition system

OCR comprises of many stages, for example, Pre-handling, Segmentation, Feature Extraction, Classifications and Recognition. The contribution of one stage is the yield of subsequent stage. The undertaking of preprocessing identifies with the evacuation of clamor and variety in written by hand. A few region where OCR utilized including mail arranging, bank preparing, record perusing and postal address acknowledgment require Off-line character recognition systems, design acknowledgement (Fig. 2).

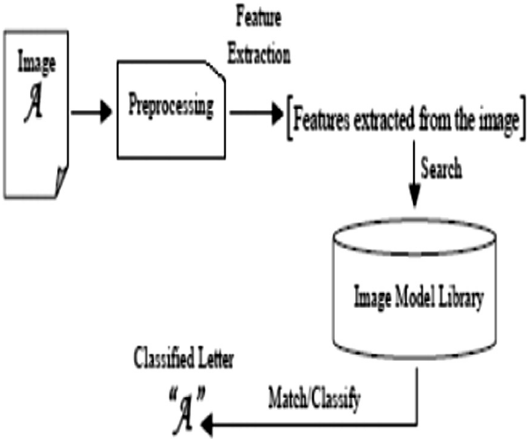

Fig. 2. Periods of general character recognition system

1.1 Using Techniques

The process of OCR is a composite activity comprises different phases. These stages are as per the following: Image securing: To catch the picture from an outer source like scanner or a camera and so forth.

Preprocessing: Once the photo has been gotten, unmistakable preprocessing steps can be performed to improve the idea of picture. Among the diverse preprocessing strategies are clamor evacuation, thresholding and extraction picture benchmark and so on. Along these lines, Pre-Processing helps in expelling the above challenges. The outcome after Pre-Processing is the paired picture containing content as it were. Along these lines, to accomplish this, few stages are required, to start with, some picture upgrade strategies to expel clamor or right the differentiation in the picture, second, thresholding (described beneath) to evacuate the foundation containing any scenes, watermarks as well as commotion, third, page division to isolate illustrations from content, fourth, character division to isolate characters from each other and, at long last, morphological preparing to improve the characters in situations where thresholding or potentially other pre-handling methods disintegrated parts of the characters or added pixels to them. This technique is utilized generally in different character acknowledgment usage.

Thresholding: Thresholding is a methodology of changing over a grayscale input picture to a bi-level picture by using a perfect farthest point. The motivation behind thresholding is to extricate those pixels from some picture which speak to a protest (either message or other line picture information, for example, diagrams, maps). In spite of the fact that the data is paired the pixels speak to a scope of forces. In this way the goal of binarization is to check pixels that have a place with genuine frontal area districts with a solitary force and foundation areas with various powers (Fig. 3).

Archaelogical Image
Thresholding
Binarized Image

Fig. 3. Thresholding process

For a thresholding calculation to be extremely successful, it should protect legitimate and semantic substance. Different types of thresholding algorithms are as follows.

1. Global thresholding calculations
2. Neighborhood or flexible thresholding counts

In overall thresholding, a lone farthest point for all the photo pixels is used. At the point when the pixel estimations of the segments and that of foundation are genuinely predictable in their individual esteems over the whole picture, global thresholding could be used. [3]

Character segmentation: In this progression, the characters in the picture are isolated to such an extent that they can be passed to acknowledgment motor. Among the least difficult systems are associated segment examination and projection profiles can be utilized. However in complex circumstances, where the characters are covering/broken or some clamor is available in the picture. In these circumstances, propel character division strategies are utilized. In this progression, the picture is sectioned into characters before being passed to characterization stage.

The division can be performed expressly or verifiably as a side-effect of grouping stage [2]. Also, alternate periods of OCR can help in giving logical data valuable to division of picture.

Feature extraction: The fragmented characters are then procedures to separate diverse highlights. In light of these highlights, the characters are perceived. Different types of features that can be used extracted from images are moments etc. The removed features should be capably measurable, restrain intra-class assortments and lifts between class assortments.

Character classification: This step maps the features of segmented image to different categories or classes. There are distinctive kinds of character order systems. Basic characterization procedures depend on highlights removed from the structure of picture and uses diverse choice guidelines to group characters. Statistical pattern classification methods are based on probabilistic models and other statistical methods to classify the characters.

Post processing: After classification, the results are not 100% correct, especially for complex languages. Post planning methodologies can be performed to upgrade the precision of OCR systems. These systems uses normal dialect handling, geometric and semantic setting to redress blunders in OCR comes about. For instance, post processor can utilize a spell checker and lexicon, probabilistic models like Markov chains and n-grams to enhance the exactness. The time and space multifaceted nature of a post processor ought not be high and the use of a post-processor ought not cause new blunders [4].

2 Literature Review

Jain et al. [5] As of late the distinguishing proof and stopping of vehicle has turned into a troublesome errand in light of the expansion in the quantity of cars. In the existing surveillance system the maintenance of incoming and outgoing vehicles is difficult. To determine this issue various strategies can be utilized out of which Optical Character Recognition (OCR) is most the appropriate innovation. OCR has been the subject of research for more than decades. OCR is characterized as the change of examined pictures into machine encoded content. The proposed system is implementing the OCR technology to park the vehicles in smart way and keep the track of the vehicles which are entering and leaving. The framework will catch the picture of number plate of the vehicle utilizing the OCR procedure and will in a flash refresh the database.

Badwaik et al. [6] as more and more learners are opting for online learning, e-learning industry is working on improving learning experience of online user by providing relevant substance and part of extra references. Since online students generally incline toward video instructional exercises, recognizing real subjects and subtopics canvassed in video instructional exercise is a major test. As of late, for productive information sharing and interoperability over web parcel of consideration is given to semantic web. In this paper, we propose a semantic electronic structure for programmed subject ID from video instructional exercises so as to recognize the ideas and their related semantically significant resources. Our system distinguishes pertinent theme utilizing disambiguation in e-learning asset which helps students in more engaged examination.

Chiron et al. [7] In this paper, we plan to assess the effect of OCR mistakes on the utilization of a noteworthy online stage: i.e. Gallica digital library from the National Library of France. It accounts for more than 100M OCRed documents and receives 80M search queries every year. In this uncommon situation, we show two basic obligations. Initial, an exceptional corpus of OCRed records made out of 12M characters near to the differentiating most bewildering quality level is shown and given, with an equivalent offer of English- and French-written documents. Next, statistics on OCR errors have been computed thanks to a novel alignment method introduced in this paper. Making utilization of all the client inquiries submitted to the Gallica entrance more than 4 months, we exploit our blunder model to propose a marker for anticipating the relative hazard that questioned terms confound focused on assets because of OCR mistakes, underlining the basic degree to which OCR quality effects on computerized library get to.

Xiaoxiao et al. [8] Another strategy for computerized number acknowledgment for mechanical advanced meters in substation is clarified in this paper, which acknowledge straight SVM unending supply of Oriented Gradients (HOG) highlights. The grids of Histograms of Oriented Gradient descriptors considerably exceed for feature detection of the gray image which has more information than binary image. A unique approach with division of locale of character picture is proposed in this paper, which is critical to the further HOG include location. SVM classifier is utilized as a part of the recognition parade and result demonstrates that HOG has better execution on digit arrangement in the substation examination robot instrument recognition.

Lusa et al. [9] Programmed activity sign acknowledgment by PCs is winding up broadly attractive actually. Techniques for programmed movement sign discovery are utilized as a part of the car business, in models of car autos, as well as in mass-created models and cell phones. In this paper, a two-stage calculation in view of key focuses include locators to identify and perceive street signs will be exhibited. The principal phase of the calculation finds objects show in the scene and decides their shape in light of geometric properties. In order to reduce the number of found objects first phase includes two additional steps to remove too large and too small objects, and to merge objects of the same shape found in a similar area of the scene into one object. The second stage includes appropriate examination of recognized question with street signs from the information database in light of distinguished keypoints.

Cho [10] This paper gives a novel scene content location calculation, Canny Text Detector, which takes advantage of the contrast between picture edge and content for viable content limitation with enhanced review rate. As closely associated edge pixels construct the structural information of an object, we observe that consistent characters compose a meaningful word/sentence which can shared a parallel properties such as spatial location, size, color, and stroke width in spite of language. In any case, regular scene content discovery approaches have not completely used such likeness, but rather generally depend on the characters characterized with high certainty, can lead to a low review rate. With a specific end goal to rapidly and heartily confine an assortment of writings we can misuse a correlation. By the utilization of unique Canny edge indicator, our calculation makes utilization of twofold limit and hysteresis following to recognize writings of low certainty. As indicated by exploratory outcomes on open datasets we can show that our calculation beats the state-of the-art scene content identification techniques in wording of detection rate.

Hengel et al. [11] This paper communicated the detail study and examination of different character acknowledgment techniques and methodologies: in subtle elements like as stream and kind of moved toward procedure was utilized, sort of calculation has worked with help of innovation has actualized foundation of the proposed system and development best outcomes stream for the every system. This paper and furthermore communicated the primary destinations and philosophy of different OCR calculations, as neural systems calculation, auxiliary calculation, bolster vector calculation, factual calculation, format coordinating calculation alongside how they classified, recognized, govern shaped, surmised for acknowledgment of characters and pictures.

Chopra et al. [12] This paper shows a straightforward, proficient, and ease way to deal with develop OCR for perusing any record that has settle text dimension and penmanship style. Optical Character Recognition in this paper utilizes database to perceive English characters which makes this OCR extremely easy to oversee which accomplishes proficiency and less computational cost. The component extraction advance of optical character acknowledgment is the most imperative. It can be utilized with other existing OCR techniques with the end goal of English content acknowledgment. This system offers an upper edge by having an advantage i.e. its scalability, i.e. in spite of the fact that it is arranged to peruse a predefined set of report designs, as proposed in this paper for English records, it can be arranged to perceive new composes.

3 Propose Work

In this paper propose work define that how to extract feature from the image using various steps and technique, we will define propose work using flow chart that will define in below (Fig. 4).

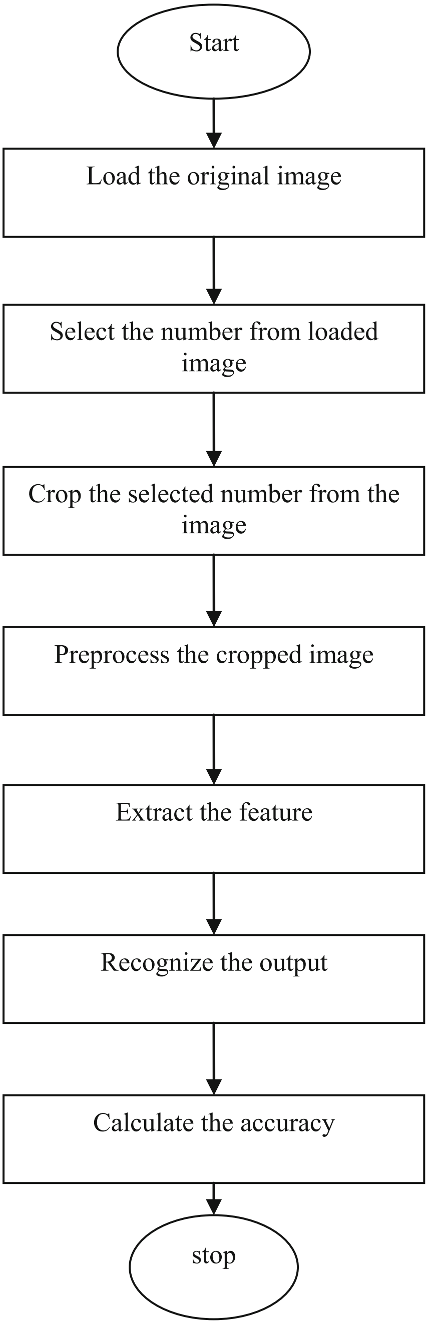
Start
Load the original image
Select the number from loaded image
Crop the selected number from the image
Preprocess the cropped image
Extract the feature
Recognize the output
Calculate the accuracy
stop

Fig. 4. Flow chart of propose work

```
Propose Algorithm-
Step 1- first we takes an original image.
Step 2-After choosing an image now select the number from the image.
Step 3-after selecting the number crop the selected number.
If (RGB)
Cropping=rgb2gray
Else
Gray=gray
Step 4-after crooping the number prepocess the cropped gray image.
Step 5- Extract the feature from the preprocessed image.
Step 6- after feature extraction we can recognize the output.
Step 7-the last step calculate the accuracy.
```

4 Result Analysis

See Figs. 5, 6, 7, 8, 9, 10, 11, 12 and Table 1.

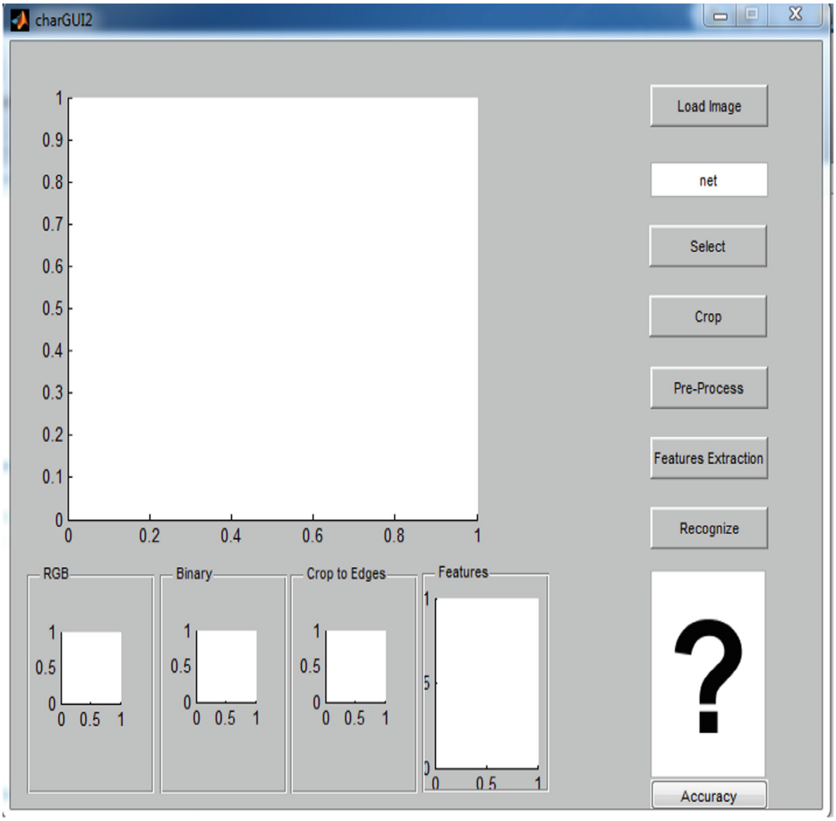

Fig. 5. First run the our code than we obtain this type of figure.

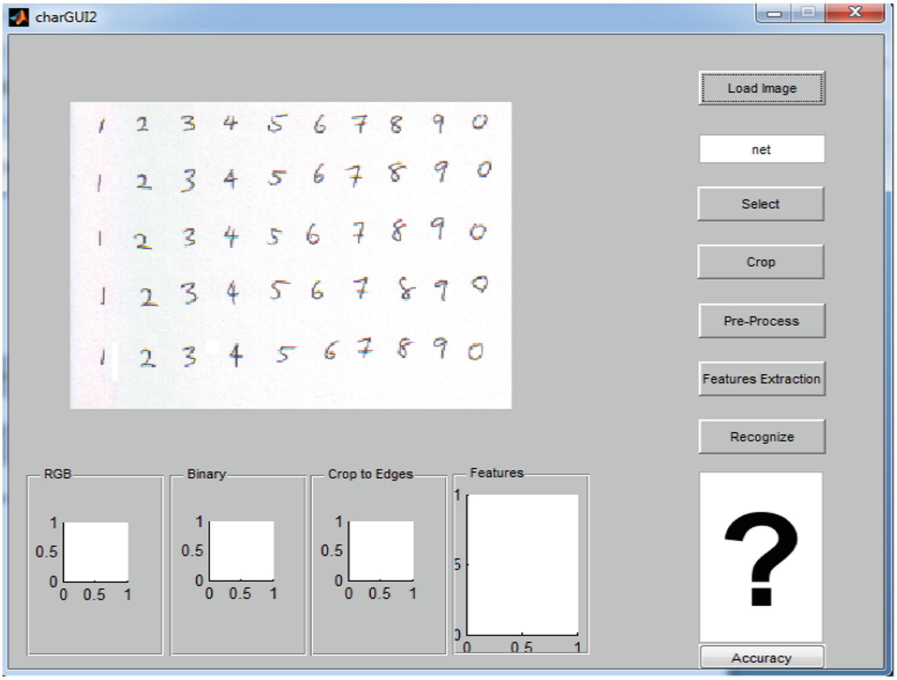

Fig. 6. Now browses the original image.

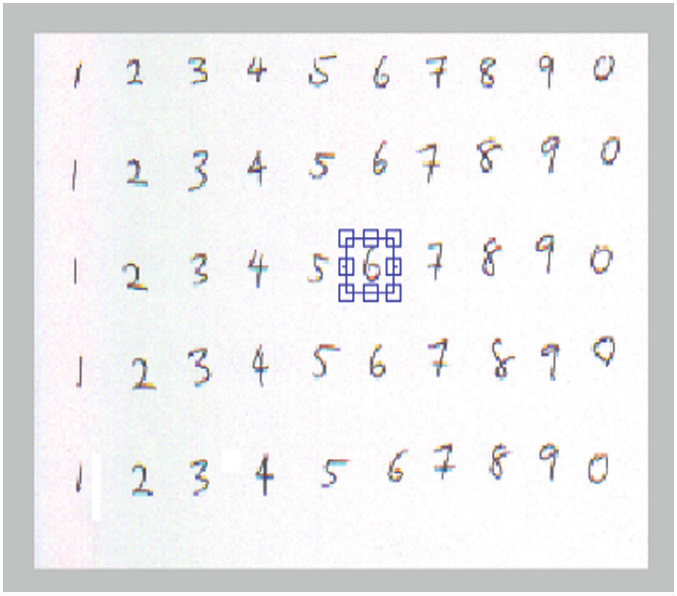

Fig. 7. Now select the number.

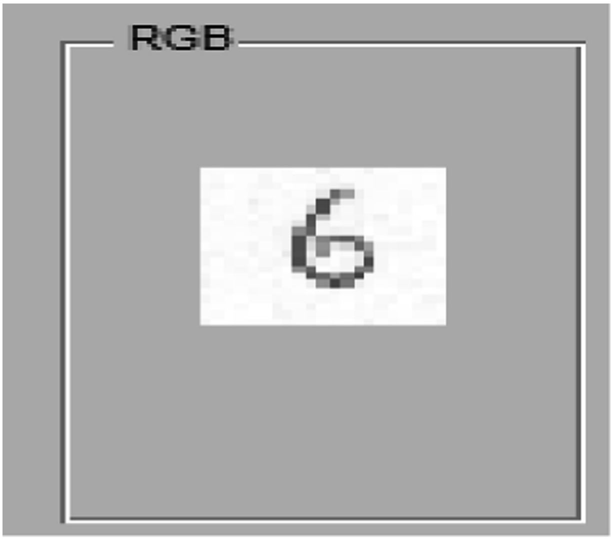

Fig. 8. Now crop the selected number.

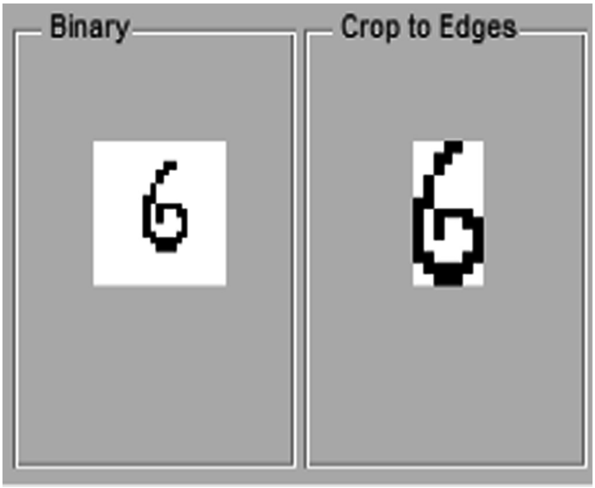

Fig. 9. now preprocesses the cropped image.

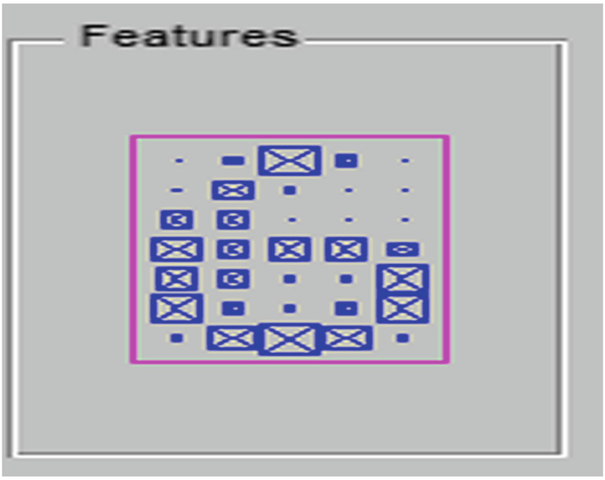

Fig. 10. Now extract feature from the preprocessed image.

Fig. 11. Finally analysis the image.

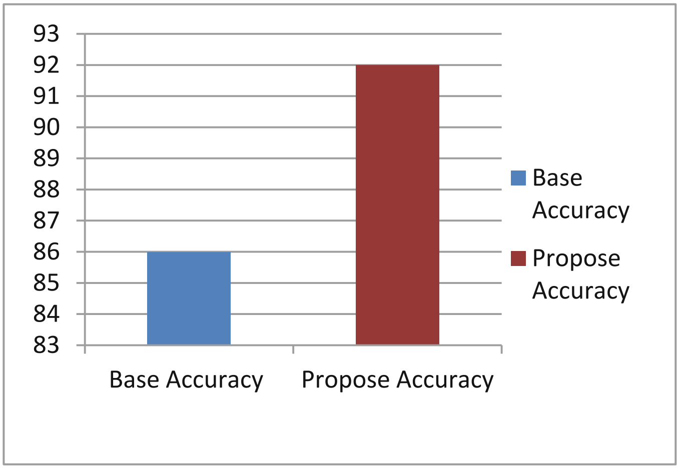

Fig. 12. Comparison on base accuracy and propose accuracy

Table 1. Comparison on base accuracy and propose accuracy

Base accuracy	Propose accuracy
86	92

5 Conclusion

In this paper, we have described the methodology of the Selection and feature extraction based on OCR. The experimental setting is done by capturing 1 image from several website. The outcomes demonstrated that our proposed technique can altogether transpose the choice and highlight extraction. The general precision of the proposed strategy is 92%. Beside, this method can use directly in the other captured image types such as scanned image and photography images etc.

Future work includes comparisons by using other distance measures with the best meta-heuristic -the considered computationally cheaper-, found in this work, in order to determine if the convergence of the algorithm is improved according the cost function used.

References

1. Islam, N., Islam, Z., Noor, N.: A survey on optical character recognition system. J. Inf. Commun. Technol.-JICT **10**(2) (2016). ISSN 2409-6520
2. Bhatia, E.N.: Optical character recognition techniques: a review. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **4**(5) (2014). ISSN 2277 128X

3. Sharma, S., Sharma, R.: Character recognition using image processing. Int. J. Adv. Eng. Technol. Manag. Appl. Sci. (IJAETMAS) **03**(09) (2016). ISSN 2349-3224. www.ijaetmas.com
4. Ciresan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J.: Convolutional neural network committees for handwritten character classification. In: International Conference on Document Analysis and Recognition, Beijing, China, 2011. IEEE, Washington, D.C. (2011)
5. Jain, K., Choudhury, T., Kashyap, N.: Smart vehicle identification system using OCR. In: 3rd IEEE International Conference on Computational Intelligence and Communication Technology (IEEE-CICT 2017) (2017). 978-1-5090-6218-8/17/$31.00 ©2017 IEEE
6. Badwaik, K., Mahmood, K., Raza, A.: Towards applying OCR and semantic web to achieve optimal learning experience. In: 2017 IEEE 13th International Symposium on Autonomous Decentralized Systems (2017). 978-1-5090-4042-1/17 $31.00 © 2017 IEEE. https://doi.org/10.1109/isads.2017.40
7. Chiron, G., Doucet, A., Coustaty, M., Visani, M., Moreux, J.-P.: Impact of OCR errors on the use of digital libraries. 978-1-5386-3861-3/17/$31.00 ©2017 IEEE
8. Xiaoxiao, C., Hua, F., Guoqing, Y., Hao, Z.: A new method of digital number recognition for substation inspection robot" 978-1-5090-3228-0/16/$31.00 ©2016 IEEE
9. Lusa, M.: Recognition of multiple traffic signs using keypoints feature detectors. In: 2016 international Conference and Exposition on Electrical and Power Engineering (EPE 2016), 20–22 October, Iasi, Romania (2016)
10. Cho, H., Sung, M., Jun, B.: Canny text detector: fast and robust scene text localization algorithm. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (2016)
11. Hengel, S.K., Rama, B.: Comprative study with analysis of OCR algorithms and invention analysis of character recognition approached methodologies. 1st IEEE International Conference on Power Electronics. Intelligent Control and Energy Systems (ICPEICES-2016), 978-1-4673-8587-9/16/$31.00 ©2016 IEEE
12. Chopra, S.A., Ghadge, A.A. Padwal, O.A. Punjabi, K.S., Gurjar, G.S.: Optical character recognition. Int. J. Adv. Res. Comput. Commun. Eng. **3**(1) (2014)