Reviewing Technique of Text Extraction and Document Image

[1]Navdeep Kaur, [2]Er. Sumit Chopra

[1]Department of CSE [2]HOD of Department
[1]Computer Science,
[1]KC College of Engineering and Technology, Nawanshehar, Punjab, India.

Abstract: Content Extraction through Picture is the mechanical or electronic transformation of pictures of type written or printed content into machine-encoded content. This is comprehensively utilized as a kind of data area from printed version records, paying little mind to whether global ID chronicles, sales, bank declarations, electronic receipts, business cards, mail, printouts of static-data, or any sensible documentation. It is an ordinary strategy for digitizing printed messages with the objective that it might be electronically adjusted, looked for, set away more moderately, appeared on-line, and used in machine methods, for instance, machine elucidation, substance to-talk, key data and substance mining. Content EXTRACTION is a field of research in model acknowledgment, electronic thinking and PC vision.

Early structures ought to have been set up with photos of each character, and managed one content style on the double. Impelled structures prepared for conveying an abnormal state of acknowledgment precision for most content styles are right now typical. A couple of structures are prepared for recreating sorted out yield that eagerly approximates the main page including pictures, areas, and other non-printed parts.

***IndexTerms* – OCR, Text Extraction, Image processing.**

I. INTRODUCTION

Penmanship recognition is a wide territory of research in the field of picture preparing and design recognition. With the developing computational power character recognition techniques have been enhanced and expanding its interest in different applications. It is a troublesome errand to build up a handy arrangement of written by hand character recognition with high precision of recognition. In the current frameworks the exactness of perceiving the content depends gigantically on the nature of the information record. Optical character recognition (OCR) is generally alluded to as a disconnected character recognition procedure to imply that the framework examines and perceives static pictures of the character's Different classifiers are utilized for upper and lower case English letter sets so as to build the exactness. Penmanship recognition can be separated into two of its sort, on the web and disconnected. As in Online technique depends on the pen direction information while disconnected strategy depends on the pixel information as it were. Presently online technique gives leeway that the spatially covering characters does not make any issue in division. Then again it represents a trouble in disconnected strategy. First the manually written or printed content is changed over into the machine meaningful frame with the assistance of Optical Character recognition framework (OCR).

II. **Literature Review**

Prior optical character recognition could be utilized for exercises like growing telecommunication and making perusing gadgets for all the visually impaired individuals. Amid 1914 a researcher named Emanuel Goldberg had built up a gadget that extract characters and changes over them into broadcast code. Amid that time, Edmund Fournier was building up an otophone, a scanner which moved over on printed papers, that helped in perceiving explicit characters. Be that as it may, it neglected to peruse non optical characters for which diverse looks into occurred. The improvement occurred and ICR (Shrewd Character Recognition) was presented by M. Sheppard in the year 1951. Clever character recognition is a progressed optical character recognition (OCR) or rather progressively explicit penmanship recognition framework that permits text styles and distinctive styles of penmanship to be learned by a PC amid handling to enhance exactness and recognition levels.

Most ICR programming has a self-learning framework alluded to as a neural system, whose activity is to naturally refresh the recognition database for the crisp penmanship designs, along these lines broadening the value of examining gadgets with the end goal of record preparing, from the printed character recognition (an element of OCR) to written by hand matter recognition, as this procedure is associated with the recognition of hand writing[5], once in a while the precision levels may not be great but rather can accomplish 97%+ exactness rates in perusing the transcribed substance in organized structures. For the most part to accomplish these high recognition rates a few read motors are utilized inside the product and each is given elective casting a ballot rights to decide the genuine perusing of characters. In the numeric fields, motors that are intended to peruse numbers take inclination, though in alpha fields, motors are intended to peruse transcribed letters which have higher elective rights. At the point when these are utilized related to a bespoke interface center, the hand composing can be consequently being populated into a back office framework maintaining a strategic distance from relentless manual keying and can be more exact than conventional human information section Clever word recognition (IWR) can perceive and separate printed-written by hand data, as well as cursive penmanship also.

III. BASIC PROCESS OF A CHARACTER RECOGNITION SYSTEM

A. Pre A. Pre A. Pre Processing

It is the first and the real advance of optical character recognition programming. At that particular stage activities are performed on the filtered picture, changing over a picture from shading to highly contrasting, tidies up boxes and lines, recognizes sections, passages, subtitles as various squares and standardization.

B. Segmentation

Segmentation is the widely used technique in image processing concepts, basically pixels in the image show the different value to the other pixel. Image division is used to find edges of the items with in the image. This technique of division is known as edge detection technique.

C. Feature extraction

The point of highlight extraction is to catch the basic qualities of the images, and it has been acknowledged this is one of the most serious issues of example recognition. In this the methodology is to separate certain highlights that describe the images, yet forgets the immaterial trait. The Choice of the suitable element removing strategy is presumably a standout amongst the most vital factors in accomplishing high recognition execution.

D. Classification and recognition

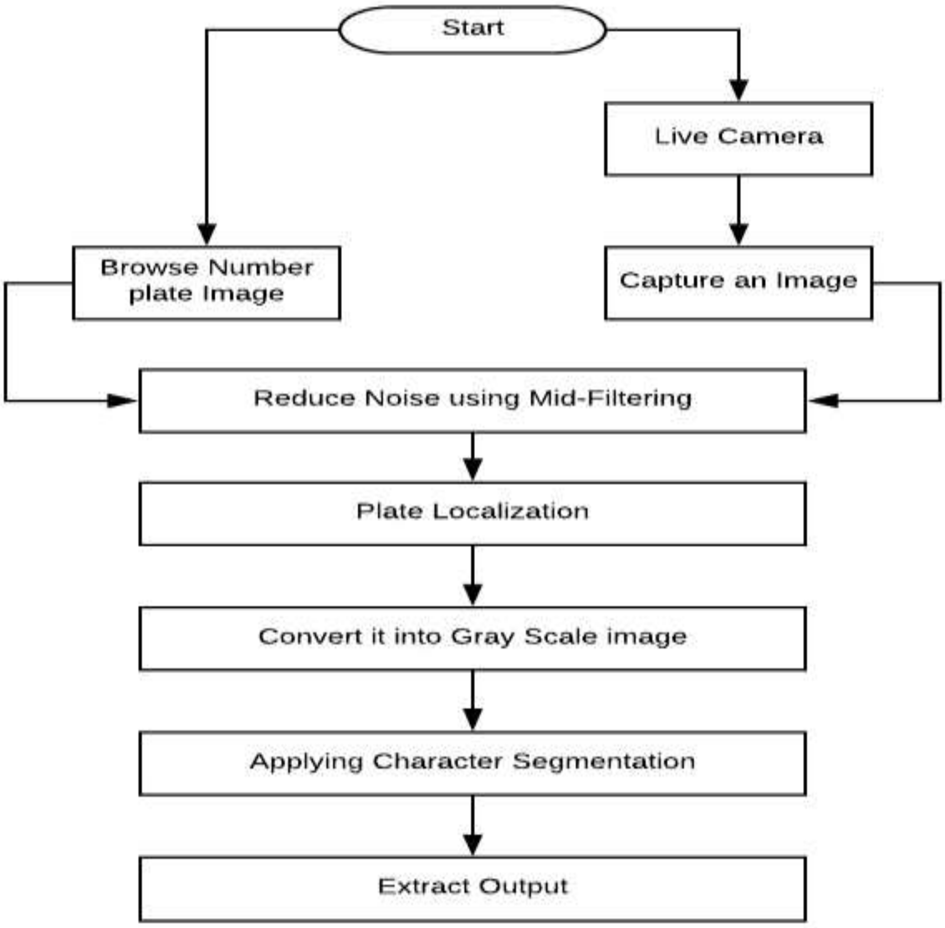
The arranging and distinguishing of each character and doling out to it the right character class is called order. In this stage the basic leadership of an recognition framework utilizes every one of the highlights removed in the before stage.

E. Post processing

It is the last advance of recognition framework being talked about. It prints the relating characters which were perceived in the organized content shape which is finished by the figuring of identical ASCII esteem utilizing recognition file of the test tests.

IV. Proposed Algorithm

To extract the text from number plate is widely used technique now these days. Most systems and programs are shown only offline mode to extract the text, we also use online method (live cameras) to capture the image and detect text from that image. The main and major issue is noise removal from the image. So, we will use mid-filtering method to overcome that problem. The flow chart of proposed work as shown in figure 1.


Start
Live Camera
Browse Number plate Image
Capture an Image
Reduce Noise using Mid-Filtering
Plate Localization
Convert it into Gray Scale image
Applying Character Segmentation
Extract Output

V. CONCLUSION

We have chipped away at the characterization and Recognition Procedures that are utilized for transcribed archive Pictures. This point by point talk will be advantageous knowledge into different ideas included, and support further advances in the region. The precise recognition is specifically relying upon the idea of the material to be perused and by its quality. Ebb and flow explore isn't specifically worried about the characters, yet additionally with words and expresses, and even the total reports. Here, we have utilized the word recognition separations for enhancing the word coordinating precision. From different examinations we have seen that determination of pertinent element extraction and characterization strategy assumes a vital job in execution of character recognition rate. Counterfeit neural systems helped us in performing character recognition which was very useful because of its high commotion resistances. These frameworks can give great outcomes. The component extraction venture of optical character recognition is the most essential. We additionally discovered that an inadequately picked set of highlights will yield poor order rates by any neural system. This strategy gives a gauge for the probabilities of word limit division utilizing the separations between associated segments and consequently consolidating the division and recognition separations to make a probabilistic word coordinating comparability. A great deal of Exploration is as yet required for abusing new highlights to enhance the present execution. We likewise seen that utilization of some explicit highlights that helped in expanding the recognition rate. To perceive strings as words or sentences division stage assumes a noteworthy job for division at character level and modifier level. Thus, there is as yet a need to do the examination in this field of character recognition.

REFERENCES

[1] Vijay Laxmi Sahu, Babita Kubde, ” Offline Handwritten Character Recognition Techniques using Neural Network: A Review “, International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064.

[2] R. Seiler M. Schenkel E Eggimann, ”Off-Line Cursive Handwriting Recognition Compared with On-Line Recognition”, 1015-4651/96 $5.00 0 1996 IEEE Proceedings of ICPR '96.

[3] Nafiz Arica and Fatos T. Yarman-Vural, ” An Overview of Character Recognition Focused on Off-Line Handwriting”, IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 31, no. 2, May 2001.

[4] B. Verma, M. Blumenstein & S. Kulkarni, ”Recent Achievements in Offline Handwriting Recognition Systems”.

[5] Rejean Plamondon, and Sargur N. Srihari, “On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey”, IEEE transactions on pattern analysis and machine intelligence. Vol. 22, no. 1. January 2000

[6] J. Pradeep, E. Srinivasan, S. Himavathi, ” Diagonal based feature extraction for handwritten alphabets recognition system using neural network”, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[7] N. Arica and F. Yarman-Vural, “An Overview of Character Recognition Focused on Off-line Handwriting”, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 31, no. 2, (2001), pp. 216 - 233.

[8] Kauleshwar Prasad, Devvrat C. Nigam, Ashmika Lakhotiya and Dheeren Umre, ” Character Recognition Using Matlab’s Neural Network Toolbox” International Journal of u- and e- Service, Science and Technology Vol. 6, No. 1, February, 2013.

[9] J. pradeep, E. Srinivasan, S. Himavathi, ” Neural Network Based Recognition System Integrating Feature Extraction and Classification for English Handwritten “, International journal of Engineering. Vol. 25, No. 2, (May 2012) 99-106.

[10] Rokus Arnold, Poth Miklos, ” Character Recognition Using Neural Networks”, 11th IEEE International Symposium on Computational Intelligence and Informatics • 18–20 November, 2010.

[11] Anshul Gupta, Manisha Srivastava, Chitralekha Mahanta, ” Offline Handwritten Character Recognition Using Neural Network, ” International Conference on Computer Applications and Industrial Electronics(ICCAIE-2011).

[12] Ankit Sharma, Dipti R Chaudhary, ” Character Recognition Using Neural Network”, International Journal of Engineering Trends and Technology (IJETT) - Volume4Issue4- April 2013.

[13] Hong Lee and BrijeshVerma, ”A Novel Multiple Experts and Fusion Based Segmentation Algorithm for Cursive Handwriting Recognition”, 978-1-4244-1821-3/08/$25.00 c 2008 IEEE.

[14] Tasweer Ahmad, Ahlam Jameel, Dr. Balal Ahmad, ”Pattern Recognition using Statistical and Neural Techniques”, 978-1-61284-941-6/11/$26.00 <S> 20 11 IEEE.