Manu Narula et al. International Journal of Recent Research Aspects ISSN: 2349-7688,
Vol. 5, Issue 1, March 2018, pp. 178-180

Text Extraction From Digital Images

Manu Narula[1] , Neha Chawla[2] ,Anurag[3] , Sagar[4]
[1,4] Department of Information Technology [1][4]
[2] [3] Department of Electronics and Communication Engineering
Northern India Engineering College, New Delhi, India
[1]Manunarula1996oct@gmail.com[1]
[2]Nehachawla599@gmail.com[2]
[3]Anuragkumar29997@gmail.com[3]
[4]Sagarparjapati97@gmail.com[4]

***Abstract* - This research aims at providing the best possible way to extract information associated with images with the help of character description, web mining and artificial intelligence. Various State-Of-The-Art Techniques have been discussed in this paper which include Optical Character Recognition which is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, source may originate from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example Screenshot of a video). Also, Omnifont Recognition has been stated which is commonly used term in conjunction with OCR software. Omnifont recognition is the ability of computer software to recognize nearly every font. We have introduced some efficient algorithms for extracting text from images which are then communicated only to the authorized person using IOT technology and android platform.**

***Keywords* - web mining, IOT, AI, avr microcontroller, algorithm, entropy.**

I. INTRODUCTION

Data is the basic means of representation of facts in an understandable manner which can be globally accepted. Data is required for the flow of information such that, it provides a platform for expressing the views and specific perspectives of an individual. It is the essential unit of communication over the generations which has facilitated in the need for preserving and utilizing the data. Over the years, there has been a rapid escalation in the information gradient supported by convenient methods of data storage and representation based on the underlying application. From this abundant data the main challenge encountered by a user is to capture data of his interest. Retrieving a part of the data or extracting a relevant section from a given document is an active area of research and lead to the introduction of search techniques based on Key terms and pattern matching. The image is captured from any source and fed as input to the Text Extractor. This image is corrupted by many types of noises like Gaussian noise, salt & pepper noise, Poisson noise and speckle noise. Using MATLAB software this distorted image is preprocessed to remove all these noise using various filters. The preprocessed image is then character segmented with the help of efficient algorithms. The effect of shadows in the text can be removed by using contrast variations, tilted images or tilted characters can be detected using rotation technique. Each character is properly extracted using stroke width variation method. All the extracted characters are matched with a predefined set of library. Hand written texts and incomplete words are identified using AI. This way text is extracted from images more efficiently. This text can be read in language like English, Hindi, and Punjabi by using set of predefined libraries. This text can only be accessed by the authorized person[10].

II. PREVIOUS ATTEMPTS

OCR: - Optical character recognition (OCR) is process of classification of optical patterns which are in a digital image. The character recognition is achieved through feature, optical scanning, pre-processing and location segmentation, representation, training, post processing and classification. It is the electronic conversion of images (handwritten or printed) into machine encoded text. OCR technology is used to convert different types of documents such as scanned paper documents, pdf files or images captured by a digital camera into editable and searchable data. There are special circuit boards and computer chips designed especially for OCR to speed up the recognition process. Though many commercial systems for performing OCR exist for a wide variety of applications, the available machines are still not able to compete with human reading capabilities with desired accuracy levels [5,10].

Omnifont Recognition:-Omnifont Recognition term is used in conjunction with OCR software. It can contribute tremendously to the advancement of automation process, improve man-machine interfaces, and have many applications in office automation, data entry, and especially as a major component in information systems (e.g. Optical archiving system) . Many approaches do exist for OCR. However, they usually lack generality andare limited to some fonts, styles or even sizes [5]. OTR approach is to avoid particularities of fonts and styles,

thusaiming ata multi-font, multi-style, multi-size algorithm, which can evolve towards a more sophisticated system for handwritten script. OTR consist of any optical recognition system which is composed of the pretreatment or character filtering, feature extraction and decision procedure [1]. Character recognition methods can be classified into two distinct types: statistical and syntactical (The latest ones are usually presented as a process of description and interpretation with several levels).

III. SYSTEM ARCHITECTURE

Objective of this system is the extraction of text from any image and then displaying its related information on the mobile screen or any Device through IOT. Main goal of this system is that if a person doesn't have or know any specific thing then he/she could get its information with the help of this software (Android application).
Different modules used in this system are as follows:

A. Text Extraction

In text extraction feature text is being extracted from the natural scene or an image. Here text extraction is done with the help of character description and stroke configuration. The image from any source (like space agency) is fed as input to this text extractor. This image is preprocessed to remove any type of noise associated with it. After preprocessing, efficient techniques like segmentation, stroke width variations and color gradient are used to extract text from images.
This research paper aims even at extracting the incomplete words or information present on images to correctly recognize with the help of Artificial Intelligence.

B. Searching

The text which is extracted from images are searched over the internet or database. This searching is entirely rank based means search appear according to area of interest. It basically derives meta data information about the item of interest by extending the user's given interest. This technique helps in the arrangement of extracted text under different categories (like name of place, landmark) to be further used by the government, space and regional agencies.

C. Web Mining

In this mining process required information is retrieved from the web or from database in an efficient manner. This is done at low entropy using Semantic and Synaptic web mining. This process can make the text extractor more efficient for extracting information from images in a more elaborated manner. After retrieving the information successfully it is displayed on the mobile screen.

IV. ALGORITHM

SWT: Computes per pixel, width of the most likely stroke containing pixel.
1. Initially set SWT = ∞
2. Find edge by canny edge detector.
3. Follow the ray $r=p+n \cdot dp$, $n>0$ until another edge is found.
4. If $dq = -dp \pm \Pi/6$ then $SWT = |p - q|$ and dp++ else discard the ray.
5. If SWT ratio <= 3 then group neighboring pixels.
6. If two letters are having similar stroke width, they can be grouped.
7. The output is a set of rectangles designating bounding boxes for detected words.
8. Search the text on web or in database.
9. Match the word, and retrieve the related information.
10. Display retrieved information on mobile screen.

V. ADVANTAGES

1) Tilt text is detected.
2) High accuracy in natural scene.
3) Requires less text extraction database.
4) Most relevant and accurate data is retrieved from the web.
5) Reduces human effort.
6) self-assessing and develop own its own.

VI. DISADVANTAGE

1) Handwritten text cannot be accurately recognized.

VII. APPLICATIONS

1) Analysis of documents can be easily done.
2) Industrial automation.
3) Satellite analysis (ISRO)

VIII. SECURITY CONCERN IN TEXT EXTRACTOR

Security is a big issue in any field. This project allows only the authorized user to access the images and information associated with it. This is achieved using a sophisticated program that asks for password each time you want to use the system.

At mega 16 microcontroller is used along with a 4X4 keyboard which acts an input device. If the entered password is correct then one can use the system otherwise an alarm will be generated if you enter wrong password more than thrice.

IX. CONCLUSION

This paper achieves the objective of text extraction from images. The extracted information is sent to the authorized user with the help of IOT and android platform.

Integration of MATLAB features proposed in this paper helps in achieving better efficiency and overall performance of the system. Effect of shadows from the images / texts, tilted images / texts can be sort out by using contrast variations and rotation techniques. Handwritten texts and incomplete words can be extracted using Artificial Intelligence. The information so extracted can be converted to regional languages thereby making this technique applicable for different parts of the world. Security for accessing images and its associated information is provided using highly secure password. This text extractor uses efficient techniques of extraction algorithm, cloud and web mining algorithm to properly implement this technology.

X. GLOSSARY

1) Semantic Web - It is a technique to manage content and process with creation and use of semantic metadata.

2) Synaptic Web - Synapse is a biological term, it is the connection between different neurons in the brain, same as in the synaptic web like the human brain the synaptic connections between the content or information are more important than the content or information themselves makes the smarter web.

3) Entropy - In information theory the term generally refers to the Shannon entropy, is a measurement of uncertainty and inconsistency in random variable, which evaluate the information content in a message.

XI. ACKNOWLEDMENT

We would like to express our gratitude to those who helped us to complete this work. We want to thank our teacher, Ms. Tanvi Dhingra and seniors Ms. Arushee Mehan and Ms. Ishika Mridul Moitra for their continuous effort and guidance. She helped in a broad range of issues from giving us direction, helping to find the solutions, outlining the requirements and always having the time to see us.

We have furthermore to thank Prof. Rajiv Sharma, Head of the Department of Electronics and communication Engineering, to encourage us to go ahead and for continuous guidance.

REFERENCES

[1] Gaurav Kumar, Pradeep Kumar Bhatia, "A Detailed Review of Feature Extraction in Image Processing Systems", IEEE 4th International Conference on Advanced Computing & Communication Technologies, pp. 5-12, Feb. 2014.

[2] Sunil Kumar Rajat Gupta Nitin Khanna, Santanu Chaudhury and Shiv Dutt Joshi. Text extraction and document image segmentation using matched wavelets and mrf model. IEEE TRANSACTIONS ON IMAGE PROCESSING, 16:1–8, 2007.

[3] Anand Mishra, Karteek Alahari, C. V. Jawahar, "Top-down and bottom-up cues for scene text recognition", IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–3, 2012.

[4] Yonatan Wexler Boris Epshtein, Eyal Ofek. Detecting text in natural scenes with stroke width transform. pages 2–5.

[5] Gaurav Kumar, Pradeep Kumar Bhatia, Indu, "Analytical Review of Preprocessing Techniques for Offline Handwritten Character Recognition", International Journal of Advances in Engineering Sciences, Vol. 3, No. 3, pp. 14-22, July 2013.

[6] Qiuhau Lin Tong Zhang, Derek Ma. Mobile camera based text detection and translation. IEEE, 7(2):87–105, 2010.

[7] J. Chand S. Mukherjee, J. V. Bhayani and R. N. Raj. Keyword recommendation forinternet search engines. 2004.

[8] A. Hotho B. Berendt and G. Stumme. Towards semantic web mining. Proceeding of the First International Semantic Web Conference: The Semantic Web (ISWC 2002), Sardinia, Italy, vol. 2342, pages 264–278, 2002.

[9] E. Ofek B. Epshtein and Y. Wexler. Detecting text in natural scenes with stroke width transform. in Proc. CVPR, page 29632970, 2010.

[10] Gaurav Kumar, Pradeep Kumar Bhatia, "Neural Network based Approach for Recognition of Text Images", International Journal of Computer Applications, Vol. 62, No. 14, pp. 8-13, Jan. 2013.

