Text Extraction from Image and Displaying its Related Information

N.K. Gundu[1], S.M. Jadhav[2], T.S. Kulkarni[3], A.S. Kumbhar[4]

[*] Computer Engineering, MMIT, Pune.

Abstract- In this project we combine the best ideas from the text extraction withthe help of character description and stroke configuration, web context searchand web mining with the help of semantic web and synaptic web at low entropy. First, we design a discriminative character descriptor. Second, we model characterstructure at each character class by designing stroke configuration maps.With the help of web context search, the extracted text is searched over thenet. In the proposed approach we have extended the context of user's interestand developed an unsupervised algorithm to find the items of interest for theuser. Web mining is the application of data mining technique to automaticallydiscover and gathered information from web documents. It is used to find outthe relevant and efficient results from the web. Semantic-Synaptic web mining interlinks the web of data to different data sources at low entropy.

Index Terms- Character descriptors, stroke configuration, SWT, interest-item matching, Semantic Web, Synaptic Web, meta-data, entropy.

I. INTRODUCTION

This paperaims at making a Mobile Application Which recognize the text from an imagecapture from a mobile camera and gives information related to it. This is done usingCharacter Descriptor and Structure Configuration.Extracting text directly from natural scene image or videos is a challenging task becauseof diverse text patterns and variant background interferences. Previously proposed algorithmsare applied to obtain text regions from scene image only. The best word recognitionrate for scene images was only about 41.2%.To extract text information by mobile devices from natural scene, automatic and efficientscene text detection and recognition algorithms are essential. However, extractingscene text is a challenging task due to two main factors:

1) Cluttered backgrounds with noise and non-text outliers, and
2) Diverse text patterns such as character type, fonts, and size.

To solve these problems, scene text extraction is divided into two processes: text detectionand text extraction. Extracting text from natural scene and displaying information related to it willhelp to know the surrounding environment. This application will help tourist who did notknow local languages of particular country/area he will get information with the help ofthis application. This application will also help to know the information related to anyhand-held objects. As we will embed this application in a smart phone, there will be moremobility.

II. RELATED WORK

Extracting text from image is a difficult task. To perform this task various techniques have been implemented before. Cluster classification [1] is one of the techniques which have high accuracy in detecting text area and non-text area. There is new trend towards content based document image retrieval technique without going through OCR process [2]. There is another technique named as sliding window detection which has high accuracy of detecting text in natural scene.

This paper uses different techniques for text extraction, web context search and web mining. Different techniques used are character descriptor and stroke configuration [4] for text detection and extraction, item ranking based on users interest [6] for web context search and semantic and synaptic web mining at low entropy for retrieving most relevant data from the web[7].
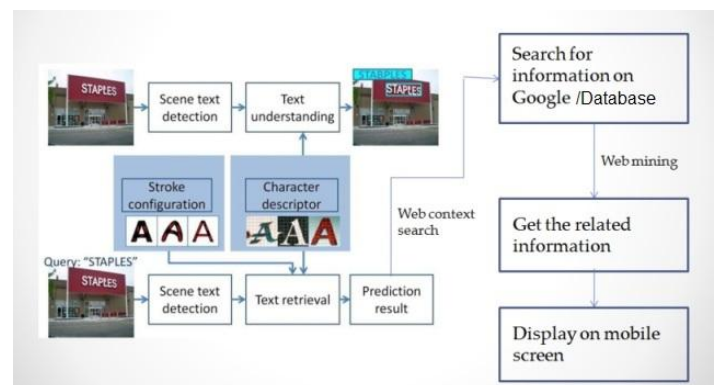
III. SYSTEM ARCHITECTURE

Figure 1: System Architecture

Objective of this system is the extraction of text from any image and then displaying itsrelated information on the mobile screen. Main goal of this system is that if a persondoesn't have or know any specific thing then he/she could get its information with the helpof this 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. Heretext extraction is done with the help of character description and stroke configuration [1].Firstly the text will be detected, understood and then recognized.

B. Searching:

In searching process extracted text is being searched over net or in database. Here searchingis done with the help of item ranking according to the item of interest.It basically derives meta data information about the item of interest by extending the user'sgiven interest.

C. Web Mining:

In this mining process required information is retrieved from the web or from database inan efficient manner. This is done with the help of Semantic and Synaptic web mining at low entropy [7]. After retrieving the information successfully it is displayed on the mobile screen.
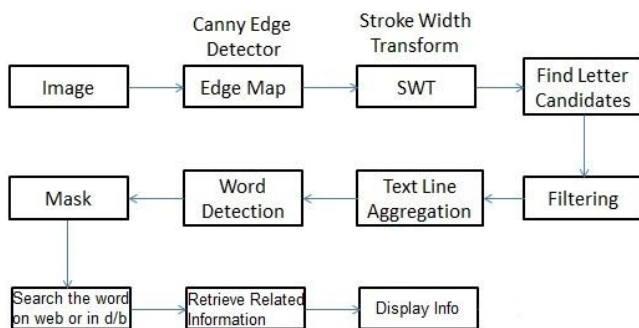
IV. ALGORITHM AND FLOWCHART

Figure 2: System Flowchart

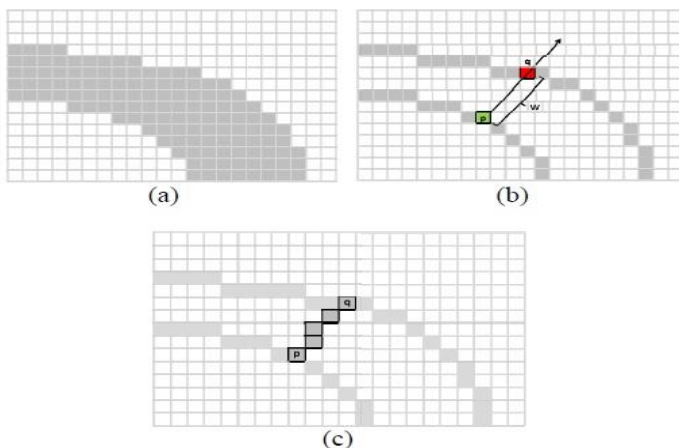**SWT:** Computes per pixel, width of the most likely stroke containing pixel [8].

Figure 3: Stroke Width Transform

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 ± Π/6** then **SWT= |p - q|** and dp++ else discard the ray.
5. If **SWT ratio <= 3** then group neighbouring 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. TECHNICAL SPECIFICATION

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.

Disadvantages:
1)Handwritten text cannot be accurately recognized.

Application:
1)Analysis of documents can be easily done.
2) Industrial automation.

VI. CONCLUSION

Thus this paper achieves the objective of text extraction from image and displaying its information on android platform, with the help of text extraction algorithm, web context search algorithm and web mining algorithm.

APPENDIX

GLOSSARY:

1) HOG - It models character structure feature and cross correlation analysis of character similar from text recognition and detection.

2) Meta-data - A set of data that describes and gives information about other data.

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

4) 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 objects (Content/ Information) are more important than the object themselves makes the smarter web.

5) 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.

always having the time to see us.

We have furthermore to thank Prof. P. M. Daflapurkar, Head of the Department of Computer Engineering, to encourage us to go ahead and for continuous guidance. We also want to thank Prof. S. A. Agrawal for all his assistance on Text Extraction and guidance for preparing report.

We would like to thank Prof. D. B. Satre and our colleagues who helped us time to time from preparing report and giving good suggestions. We also extend sincere thanks to all the staff members of Department of Computer Engineering and Information Technology for helping us in various aspects.

Last but not least we are grateful to our parents for all their support and encouragement.

REFERENCES

[1] 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.

[2] IEEE Li Linlin Chew Lim Tan Senior Member IEEE Shijian Lu, Member. Documentimage retrieval through word shape coding. pages 11–15, 2008.

[3] Anand Mishra Karteek Alahari C. V. Jawahar. Top-down and bottom-up cues for scene text recognition. pages 1–3.

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

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

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

[7] 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.

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

AUTHORS

First Author – Nilesh K. Gundu, BE Computer (Appeared), MMIT and Email ID: gundu.nilesh@gmail.com.
Second Author – Shwetarani M. Jadhav, BE Computer (Appeared), MMIT and Email ID: shwetaranijadhav@gmail.com.
Third Author – Tushar S. Kulkarni, BE Computer (Appeared), MMIT and Email ID: tskulkarni25@gmail.com.
Fourth Author – Avinash S. Kumbhar, BE Computer (Appeared), MMIT and Email ID: avinashkumbhar95@gmail.com.

Correspondence Author – Tushar S. Kulkarni,Email ID: tskulkarni25@gmail.com,nilesh.gundu@gmail.com.
Contact number: 8983064518.