Text Detection and Recognition from Scene Images using MSER and CNN

Savita Choudhary, Nikhil Kumar Singh, Sanjay Chichadwani
Department of Computer Science and Engineering
Sir MVIT
Bangalore, India
choudhary7.mvit@gmail.com

***Abstract*—Detection and recognition of text from natural images is very important for extracting information from images but is an extensively challenging task. This paper proposes an approach for detection of text area from natural scene images using Maximally Stable Extremal Regions (MSER) and recognizing the text using a self-trained Neural Network. Some preprocessing is applied to the image then MSER and canny edge is used to locate the smaller areas that may more likely contain text. The text is individually isolated as single characters by simple algorithms on the binary image and then passed through the recognition model specially designed for hazy and unaligned characters.**

Keywords—Detection; Recognition; Classification; Neural Network; Optical Character Recognition;

I. Introduction

There has been a rapid surge in generation of images in the last decade due to ease in availability of gadgets. Billions out of these are stored or shared on the web. We have Exabyte of visual data available, many out of which contain textual information. Problem emerges in attempting to understand and interpret these images [1]. The most trivial way is to use the metadata available with the image. Image search engines have been using this approach to classify and tag images for their image search feature. Several kind of information can be extracted from images with varying levels of computation, metadata being the most basic. We can mine the text in the image, classify objects and even understand facial expressions. Human driven approach in this regard is very useful but requires a lot of man hours [2] and also this cannot be done manually for images. We require an automated system to mine this information without human supervision. A real-time system that extracts text from scene images can prove to be very useful [1] [3]. These systems can be used to automatically read and understand hoarding and posters, finding shops and addresses from images, solving captchas etc.

This paper proposes an efficient and improvised approach to detect text region from scene images using Maximally Stable Extremal Regions and recognize it using a self-trained model.

Section II presents the related work. The proposed system with design methodology is shown in Section III. Section IV presents the observations and results. Conclusion and future scope is given in Section V.

II. Previous Work

There have been many successful attempts in this direction using various different approaches and algorithms. The text area first needs to be isolated from the whole scene image. Researchers use various methods to segregate text area from the whole image involving contrast manipulation, converting to binary image [1] and thresholding. Feature detection [4] is then performed on the image which results in probable text regions. On later stage edge and texture detection and canny edges [1] are used. Stroke Width Transform [5] [1] is also used to find text based on the fact that text has constant width strokes on pixel level. Once the regions are found, it is fed to Optical character Recognition system to recognize the text present [4]. Often off-the-shelf solutions used for recognition as training a model is complex and already very efficient character classifiers are available.

An approach involving the combination of multiple algorithms gives better results in finding the text regions, as the contrast and text font vary largely from image to image. Unlike scanned documents, the obtained regions often do not have clear edges and curves.

An Off-the-shelf OCR [6] [7] often fails to read characters from such images. Self-trained classifier is required to be trained on hazy and abruptly aligned characters to read such text.

III. Proposed approach

As the characters found in scene images are not very clear and well defined, we have taken a different approach. This approach broadly involves two tasks:

- Text region extraction
- Character recognition using CNN

A. Text region Extraction

The text area extraction is necessary to find the segment of the image that contains the text so that feeding the image to an OCR becomes easier. As shown in fig. 1, the image is first resized and converted into gray scale. Unwanted noise is removed and the image is processed to set a threshold for the binary conversion.

978-1-5386-3785-2/18/$31.00 ©2018 IEEE

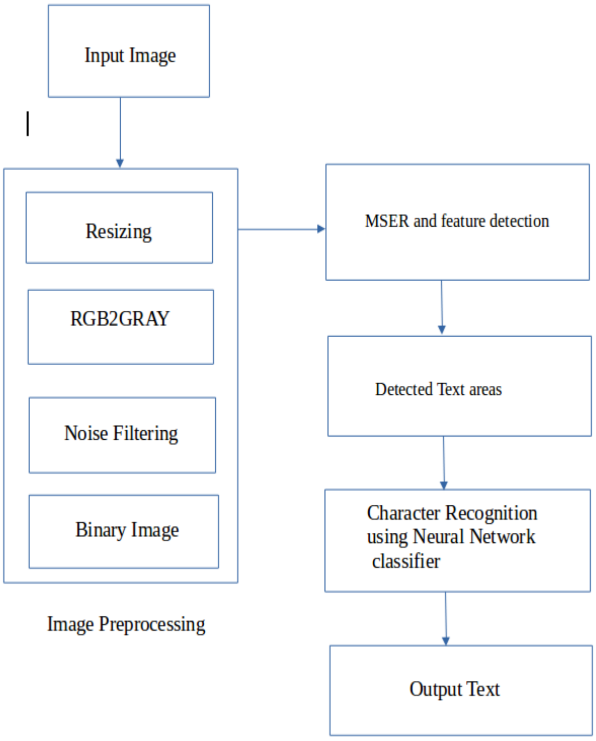

Fig. 1: Proposed Model

The threshold value varies for different images. On the obtained binary images, MSER is applied followed some basic feature extraction. Maximally Stable Extremal Regions (MSER) [4] [3] is a feature detector used as a method of blob detection.

These blobs are regions having properties like brightness and color, different from other regions. A set of different thresholds are applied to the image, the regions that maintain consistent intensity across these thresholds are the regions of interest. Generally text maintains a consistent contrast with the background, thus MSER is an ideal approach for text region detection.

It extracts from the image a number of such covariant regions, called MSER. MSER is quite sensitive to image blur. Canny edge detection is used to overcome this problem. The original image is processed by both the algorithms MSER and canny edge detection; the common regions from both are marked. This gives many non-text regions detected alongside the text. As shown in fig. 2, multiple overlapping regions of detected text and features are obtained. These regions are converted into rectangles and the overlapping ones are removed. The small regions that cannot contain text are excluded by putting simple size constraints; it is shown in fig. 3.

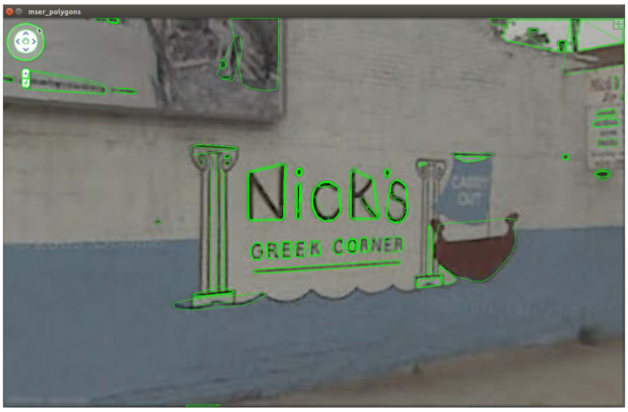

Fig. 2: Probable Text regions detected after applying MSER

The found regions are isolated as single images in binary and saved as bmp files. The files are saved with special sequential naming convention that proves to be useful during word compilation. These files then fed directly to the Neural Network character classifier.

Textual characters in natural scene images do not follow any font, they can be hand drawn in freehand and may contain unconventional styles. The noise and contrast also varies from image to image.

Thus, the detected binary text regions are not in perfect shape and font as in the results obtained from documents and scanned images (Fig. 4). This raises a problem for simple OCRs, as they are designed to read printed text from scanned images and clear texts that follow conventional fonts.

Hence we used our own character classifier using Neural Networks.

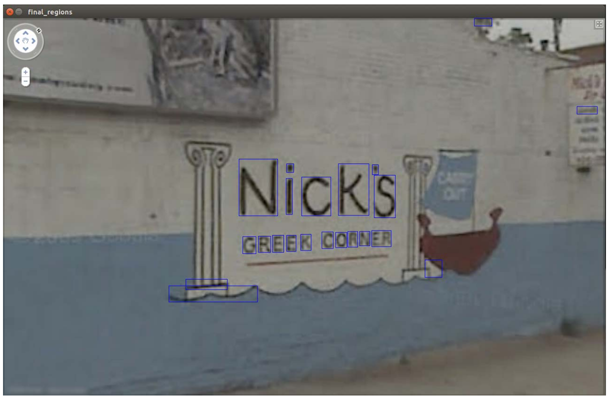

Fig. 3: Final Text regions after eliminating overlapping.

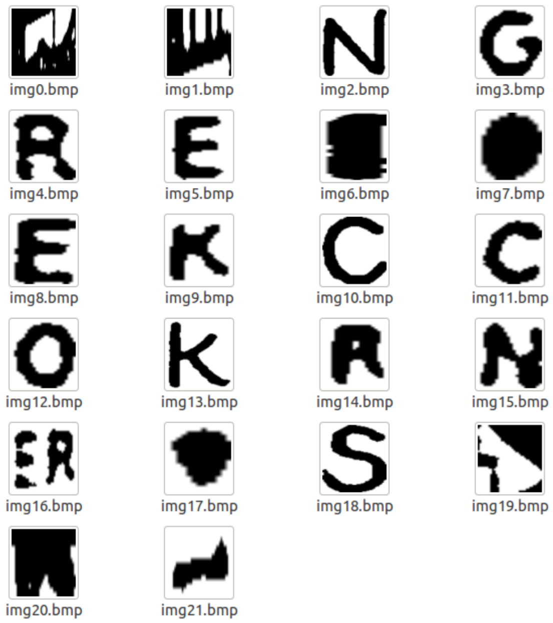

Fig. 4: Resulted individual images after character extraction

B. Character recognition with CNN

The paper proposes an innovative character classifier model based on Neural Networks to read the text regions obtained from the extraction step. The classifier is trained on 28 x 28 pixel size images of individual English alphabets. The dataset contains images of characters taken from signboards, banners, posters, graffiti and other natural scenes. Thus it contains very vivid and freestyle drawn fonts of each character which is perfect for our requirement. The datasets IIIT 5K-word [2] was obtained from CVIT, IIIT Hyderabad. There were not enough images for each character, hence we manually cropped images. Apart from that Chars74K [8] dataset was also used.

It is a sequential model that takes input of 28x28 pixels in binary. First layer is 2D convolution with 5x5 size and activation as 'RELU'. It is followed by a Max Pooling layer of size 2x2. Third layer is dropout of 0.2 and followed by flattening. The model is then compiled after two dense layers with activations 'RELU' and 'SOFTMAX'. The output is a number that refers to each character in the English alphabet. This model can be used for any image classification task by making small changes in the input and output layers. Other well-designed CNN models like ALEXNET and VGG can also be used to train the character classifier. After training on the character dataset, the trained model can be used in any application involving recognition of non-trivial English alphabets fonts.

The result for each text region is finally compiled together using the sequential naming convention used to save the text region images. This helps maintain the integrity of the words.

IV. Observation and Results

This approach successfully detects the text regions and there has been an appreciable increase in accuracy while recognizing the characters. In the traditional approach wherein they are directly feed (whole text region) to the OCR, some of the characters are unclear. Thus, it results in dropping of characters and misrecognition. This approach eliminates the chance of dropping the characters as it individually classifies each of them. Conventional approaches sometimes missed few characters in images with varying contrast and high blur. Since this approach uses MSER [9], the problem with inconsistent contrast and color variations do not create trouble in detecting text regions. Images of any kind can be fed to this system regardless of resolution, contrast, blur etc. and it gives comparatively better results.

The CNN character classifier is highly efficient with an accuracy of around 85-90% on individual characters. The overall accuracy of text recognition output from images is around 70-75%. Due to this approach no character is left out, although there are some wrong character predictions which can be eliminated using a word predictor.

V. Conclusion and Future Work

In this paper, we proposed a new methodology to detect and recognize text from scene images. The system works very well and produces good results over a range of scene images with appreciable accuracy. In the first step, the system preprocesses the image and then finds out MSERs. Based on that text regions are found and saved individually. The character classifier specially trained for this purpose successfully predicts the characters with very less error rate. Due to this new approach no character is left out or dropped in the prediction process.

In future, there can be some changes that will help in the prediction process. The classifier can be trained on more wide design of characters that will help read complex graffiti and calligraphic posters. The characters are predicted in the sequence followed by the text regions, later it is compiled together to form words. In multiple words, it becomes difficult to obtain the correct word. We can integrate a language based word predictor that can efficiently predict the words from the recognized characters. This will solve the problem of character dropout with ease.

Practical use of this system can be in mapping work where addresses and landmarks need to be read from images. The banners and posters on roads can be perfectly read by this system. It can also read graffiti and sign boards without much difficulty.

Blurred images still create trouble for text recognition systems with no solid solution. Some more feature detection methods can be added to include extremely blurred text present in the images.

References

[1] Anand Mishra, Karteek Alahari, C.V. Jawahar, "An MRF model for Binarization of Natural Scene Texts", International Conference on Document Analysis and Recognition (ICDAR), 2011.

[2] Anand Mishra, Karteek Alahari, C. V. Jawahar, “Scene Text Recognition using Higher Order Language Priors”, In Proceedings British Machine Vision Conference 2012. Pages 127.1--127.11.

[3] Udit Roy, “Text Recognition and Retrieval in Natural Scene Images”, CVIT, International Institute of Information Technology, Hyderabad, 2015.

[4] Kethineni Venkateswarlu, Sreerama Murthy Velaga, “Text Detection On Scene Images Using MSER”, International Journal of Research in Computer and Communication Technology, Vol 4, Issue 7 , July-2015.

[5] G. Werner, "Text Detection in Natural Scenes with stroke Width Transform" in , Israel:Ben Gurion University, 2013.

[6] Hiral Modi, M.C. Parikh, “A Review on Optical Character Recognition Techniques”, International Journal of Computer Application, Volume 160, No 6, February 2017.

[7] “A Complete Optical Character Recognition Methodology for Historical Documents”, The Eighth IAPR International Workshop on Document Analysis Systems, 2008 (DAS '08), Japan, 2008.

[8] T.E. de Campos. The Chars74K dataset. http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/

[9] Md. Rabiul Islam, Chayan Mondal, Md. Kawsar Azam, Abu Syed Md. Jannatul Islam, “Text detection and recognition using enhanced MSER detection and a novel OCR technique”, 5th International Conference on Informatics, Electronics and Vision (ICIEV), Bangladesh, 2016.