Natural Scene Text Localization by Convolution Neural Network with SVM

[1]Harmanpreet Kaur, [2]Rakesh Singh
[1]Student, [2]Asstistant Professor
[1,2]Department. of Computer Science and Engineering,
[1,2]Punjabi University, Patiala, India

Abstract: In the past few years, text in natural scene images has a effective feature for content based retrieval of images. It can be extracted and used in search engines, providing effective information about the images. In this study two main challenges, one of the text position and second is prediction of box size instead of growing edges approach.In Prediction approach use convolution neural network architecture(CNN) of Deep learning with classification by support vector machine (SVM). In Experimental results show that the proposed method performs as well existing method on ICDAR 2013 Dataset.In experiment result analysis on two aspects detection of text localization and second is classification both are show significance improve Performance.

Index Terms: Edge detection, Convolution Neural Network(CNN), Support Vector Machine(SVM).

I. INTRODUCTION

The use of digital technologies and the modern gadgets such as mega-pixel cameras, latest smart technologies equipped with digital devices like android phones, PDA etc. captures huge amount of data which further increases the demand of data extraction or information retrieval. Such an extensive form of data requires the process of analyzation. The texts that is embedded behind a natural scene object or image represents one of the most challenging method for extraction of the hidden information. The use of natural-scene text extraction strategy aims to obtain the text embedded behind the natural scene image. These methods are helpful in many practical applications such as helping the visually impaired to perceive the environment, improving the text-based retrieval techniques, designing an alarm for the car driver to get safe from road accidents. The process of text extraction plays a significant role in searching an important and valuable source of information. Text extraction basically involves localization, recognition, detection, tracking, extraction, enhancement of the text from an image i.e. known. These texts in embedded form are difficult to be recognized and very hard to extract due to deviation in font, style, size, alignment, orientation, highly complexed texture and colouring. Due to rapid developments in the multimedia-based documents and a large requirement for indexing, identification, and the process of data retrieval, the researchers have worked upon large extraction methodologies in images [1, 2]. Various techniques have been generated to extract the text hidden behind an image. Text extraction denotes a big challenging problem, as it may cause lot of variation such as the distinct conditions of shadows and lighting that affects the illumination properties of an image, Outdoor-images with complex backgrounds etc. poses a great difficulty in extraction the useful information i.e. in embedded form. The text extraction methodology is divided into three of its significant sub-parts namely, detection and localization, extraction and enhancement and text-based recognition subsystem [7]. Firstly, the process of detection involves the text presence in a known input form of image whereas, the localization is basically used to find the text location in the image and to generate boxes (bounded) around the text. Secondly, the extraction and enhancement processes includes the segmented text components from its background, and involves small text magnification at a high level resolution respectively. Thirdly, the text-based recognition subsystem is recognized as a string based text and this method plays an important step among the other two subsystems as if the system gets failed, the other two subsystems automatically gets rejected as it forms a prerequisite condition of the two subsystems. In other words if this text-based subsystem fails, the other two fails automatically [8, 9]. A major challenge in scene-text detection is the ability to model a system that is of flexible nature, handling all variations and inconsistent models in the conditions of imaging. Various forms of natural scene-text localization strategies based on English language have been developed the field of literature. But the Farsi language scene-text localization lags behind the English scene text localization. The cursive and elongated format of Farsi language faces lots of challenges, therefore it offers large similarities and these similarities, and the detection method for the Farsi texts is not as successful as the English text detection. In most of the cases, the English-based methods are not at all suitable for Farsi. But it needs various modifications in the Farsi script. The figure below shows the text which works very well for the English-based text, but it won't work well for the Farsi script. The fact represents that the English and Farsi scripts have distinct attributes [10].

Fig.1 (a) (b) (c) (d)

In fig. 1 (a-b) English text sample image from the dataset of ICDAR 2013 (c-d) Farsi sample images. The English text regions are detected as red-rectangles whereas the Farsi undetected regions are distinctly represented by green rectangles (dashed). A new method has been developed for the process of text-localization in case of natural-scene images working both for English and Farsi scripts employing a new operator known as Edge colour transform operator which, assigns a colour for edge-pixels. These edge-pixels are significantly used for finding candidate-text regions that may be classified as text or non-text processes.

A. EDGE DETECTION

Edge detection is technique which consists of different variety of mathematical tools aiming to identify the digital image points where the brightness of the image changes formally, or very sharply and has large discontinuities. The identified points where the brightness of the image gets sharply changed into a set of line-segments (curved) represents the 'edges'. So, it designates an elementary tool for the processing of an image and mostly in the applications of features based extraction and detection [6]. The process of text extraction is the most impulsive and natural way to understand naturally-based scene images. The first basic step is the step of localization. A specific form of operator named ECT i.e. Edge Colour Transform is introduced to solve the problems related to natural scene images. After the extraction of input image edge-map, this ECT operator usually follows the gradient or the opposite direction to allot nearby colour to each of its edge-pixels. In its next phase, each of its channel based on Red, Green, and Blue pixel-colour gets normalized to improve the robustness which intensifies the distinct changes occurring in the system [4]. Finally, the process of grouping the coloured pixels takes place with the help of extended version of region-based growing algorithm. With the use of such customized algorithm, the region pixels do not get spatially connected and the connectivity of these is observed or decided based on a pre-defined neighbourhood structure (oval-shaped). After the regions get forged, each of region pixel forms a text-based candidate region. Eventually, all the candidates formed are applicable to the classifier in order to obtain or extract the texted regions.

Fig.2 (a) ECT: Sample Image (b) Output of ECT

B. SCENE TEXT

There has been a constant research and development of scene text detection methods. The main problem of finding the text in a random image describes complexity in its working operation. Firstly, the input image contents provides variation from an urban-structure to natural-subjects for example, trees. In scene images, the text regions are not properly bounded as they are in documents. It generally represents text in scenes with a few words at one place [3]. There are no such long lines or paras for analysing the process, though the text is modelled or designed to be readable in its basic form algorithm, etc. These techniques comprises of their own restrictions and benefits. Various schemes have been used for text extraction procedure and this study provides the comparison based on other techniques that are already developed.

II. LITERATURE SURVEY

Yipeng Wang, et.al [3] conducted a study on natural scene images. The use of stroke width histogram was done for the purpose of improving the edge detection methods in critical conditions. In order to generate super-pixel series. Secondly, a novel method of using and distance transform was used for character-based extraction process. Then there was a skeleton used to improve the accuracy of stroke-width. This method was evaluated on two standardized datasets named as ICDAR 2011 and ICDAR 2005 showing better analysis as compared to existing methods.

Soman, et.al [4] conducted a study on an approach using an algorithm on connected components CC-based text detection that involved a MSER (Edge-Enhanced) i.e. Maximally Stable Extremal Regions along with a transform using a stroke Width parameter. The researchers usually worked on natural images where the component connection was provided by using edge-enhanced MSER algorithm. In the next step these candidates were filtered using stroke width filtering and CC-based analysis. For the process of non-horizontal text it performed a rotation procedure in order to change it in a horizontal form of text and the further localization (boundary box) was performed. The text that was detected could be easily recognized with the help of recognizer known as Optical-character. The experimental analysis was done over ICDAR 2005 robust-read proposed dataset method.

Ye, Qixiang, et.al [5] analyzed, compared, and provided contrasting challenges, strategies, and the performance analysis based on text-detection and its recognition in colored images. The techniques that already exist are classified as stepwise or integrated techniques and their inbuilt problems gets highlighted that includes the process of verification, text localization, recognition, and segmentation. In this research specific problem have been discussed that are associated with video text processing, degraded text enhancement and the processing of video text, multi-oriented text, and multi-lingual form of text are The survey involves the comparative study of the basic representative strategies or methods.

Lu, Shijian, et.al [6] conducted a survey that presented a technique of scene (natural) text extraction which detects and segments the texts from the natural-scene images automatically. Over the edges of the image three of the text-specific features were modelled or designed and with the help of this model the candidate text boundaries is first detected. For each detected boundary of candidate text was firstly detected. Further many more candidate attributed were extracted using a local-threshold value that was approximated or estimated using the surrounded image-pixels. Finally, original words and the characters were identified using a model of regression based on support vector that got trained with a representation basically known as 'bags-of-words'. The set that was used in this technique was the modern developed ICDAR-2013 Robust-Reading-Competition data-set. The analysis resulted in calculation of F-measure comprising of 78.19% and 75.24 represents an atom level for detection and the task of segmentation.

Simranjit Singh, et.al [7] conducted a study on the techniques of edge detection based on distinct coloring. Edge detection outperforms in many research applications resulting in black or white image where the object is differentiated using the color based lines either black or white. In most of the cases it was found that the color based edges was neglected. So, the study worked over the facts that most of the techniques i.e. existing once failed in most of the operation in context of text detection methods.

Cho, Hojin, et.al [8] proposed a hardback study on scene-text detection algorithm named Canny Text Detector, which is benefited with a similarity between text localization and the image-edge processes with Enhanced recall rate. As image-edge pixels that are closely related to each other helps in constructing an object information using structural methods. It was observed that the characters (cohesive) compose a sentence or word with its proper meaning which shared its homogenous properties like size, colour, spatial location, and its stroke width.

Most of the approached based on scene text detection did not worked well over such kind of similarity, but most of these relied the classified characters resulting in low-recalling rate. The performed analysis on such datasets demonstrates that the used algorithm performs best in detection terminology.

Siddiqui, et.al [9] proposed a newly developed technique that helps in detecting the texts-based on random directions in natural-scene images or objects. The proposed method involves a set of two specially designed characteristics for the process of capturing both the texts using the MSER-regions based on OSTU method. The data-sets used are ICDAR, MSRA, and Proposed Dataset. The experimental study was done based on the standardized and proposed data sets that showed that the algorithm used provides a positive connection with the latest modern algorithms accomplishing an improved performance on texts with random orientations in context of natural scenes images of composite nature.

Kumuda, T., et.al [10] proposed a work using an algorithm to efficiently analyse and extract the text inside complexed form of scene images. Initially, the edges are searched and detected using a DWT. Then connected-component (CC) based clustering and a classifier named Ada-Boost used for the process of localizing texted regions. Further, in the next step, the heuristic rules and morphological operations were used for extracting the characters. The final step involves the analysis of extracted texts using OCR. The algorithm (proposed) was evaluated on distinct forms of data-base images that achieved fine upgraded results.

III. THE PROPOSED METHOD

In this section, the proposed flowchart for the process of edge detection has been explained which extends it capability to handle the operation or the tasks related to text scene localization approach. In general, the process of localizing the texted regions has been illustrated in the given flow chart in fig.4.

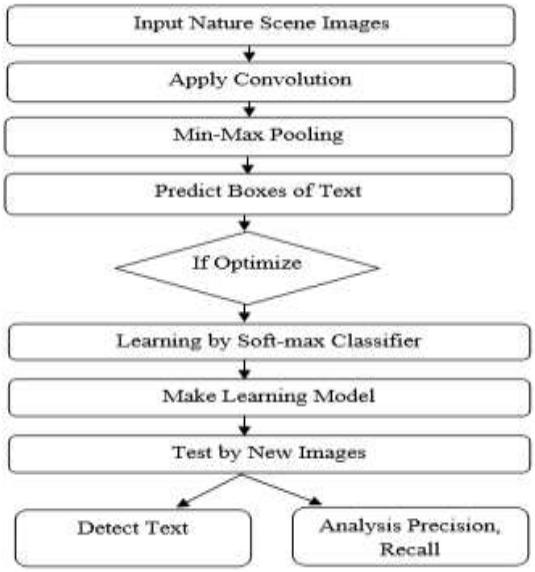

Fig.4 Proposed Flowchart

The flowchart consists of the mentioned steps that perform a specific function as per the requirement or the operation conditions of the process. The operation is presented in its general way. The first step involves the method of input in the form of nature scene text image. The second and the third steps are interlinked to each other where the convolutional property is applied and the step of maximum pooling is done. Here, the main function of this property is to compress the natural scene images such that the parameters and computational work of these images reduces for simple operation. After, the max pooling, next step is to predict the text boxes and perform the function of optimization over that. The process of optimization is followed by the learning the methodology with the help of a classifier named Soft- max classifier. Further, a model is prepared and the process of testing is done which is further followed by the final step of detecting the text and the recalling method based on precision analysis.

IV. RESULTS

In order to get the formalized desired form of results, the experimental analysis has been done based on certain parameters as presented below.

(a) Proposed Model: Methodology

1. Convolution Network: The process of convolution network uses 18 convolutions and 5 max-pool layers, which uses 3*3 convolution filter, pooling step, and adds 1*1 filter to compress 3*3 filters. So, the calculation is based on the following model analysis $\left(\frac{W}{32} \times \frac{H}{32} \times 512\right)$ where W and H denote source image width and height.

2. Region Prediction: The region is predicted by the convolution proposed above. The bounding box position and the dimension is calculated with the help of starting box using the SVM (Support Vector Machine). Further, the process of bounding box (x, y) and dimension (w, h) in the source image is done.

$x = \sigma(\Omega_x) + C_x$ ..... (1)

$y = \sigma(\Omega_y) + C_y$ ..... (2)

$w = \alpha_w \, exp\,(\Omega_w)$ ..... (3)

$h = \alpha_h \, exp\,(\Omega_h)$ ..... (4)

$\theta = \Omega_\theta$ ..... (5)

Here, C_x and C_y denotes the offset of all in the last convolutional layer and α_h tells the predefined height and the width of starting box.

3. Prediction Region by SVM: Based on the given region (above), the step of predicting the region by CNN feature is done

$U = R^{w \times h \times c}$

$U = R^{\frac{wh\prime}{h} \times h\prime \times C}$

$P^{C}_{x',y'} = \sum_{x=1}^{w} \sum_{y=1}^{h} U^{C}_{x',y'} \; K\left(x - T_x\,(x')\right) K\left(y - T_y\,(y')\right)$ ..... (6)

Where, T← coordinate point of y

$P_{x',y'}$← Predicted Region

U← CNN feature.

Based on the equations analysis, the results are analysed as follows using table.1.

Table.1 Proposed and Existing Parameters

Parameters	Proposed	Existing
Accuracy	96.26	62.23
Precision	92.45	34.23
Recall	98.45	37.23

The table.1 represents the proposed and the existing parameters based on the accuracy, precision, recall methods and the graphical analysed have been presented in fig.5, 6, 7 and fig.8 represents the combined analysis of all.

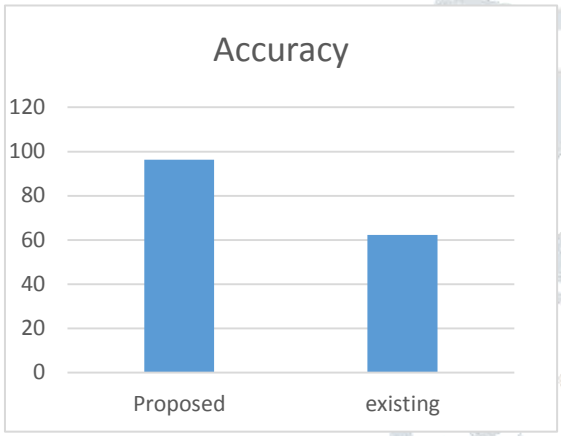

Fig.5 Accuracy

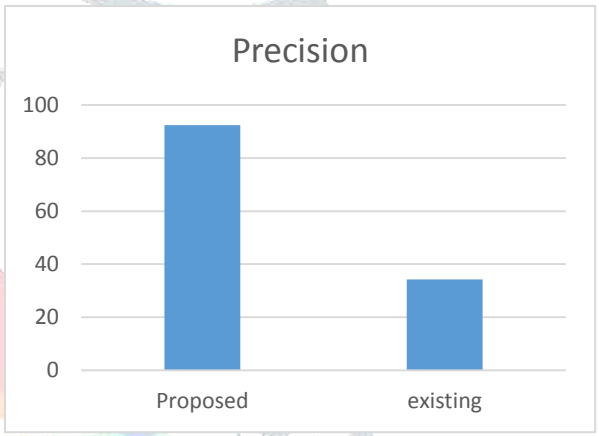

Fig.6 Precision

Fig.7 Recall

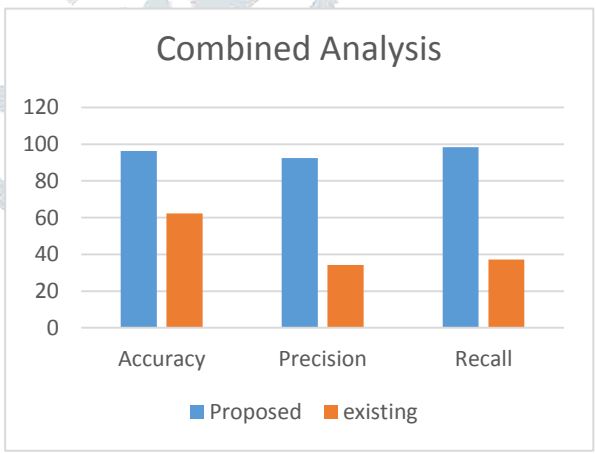

Fig.8 Combined Analysis

Analysis Of Results

In Above analysis under Results Heading Done in two Parts

(A) Classification

(B) Detection

(A)Classification:

In Fig 6 comparison of precision. Precision depict correctness of Text region predict in proposed and grow in existing method.In fig 6 comparitive analysis , which indicate proposed approach significant improve because of prediction based approach and abstract features generate by CNN and then learn by SVM. In Fig 5 comparison of Accuracy. accuracy depict correctness of Text region and not text region

predict in proposed and grow in existing method.In fig 5 comparitive analysis , which indicate proposed approach significant improve because of prediction based approach and abstract features generate by CNN and then learn by SVM.

In Fig 7 comparison of recall. recall depict correctness of not Text region predict in proposed and grow in existing method.In fig 7 comparitive analysis , which indicate proposed approach significant improve because of prediction based approach and abstract features generate by CNN and then learn by SVM.

Based on the experimental analysis done, here are some of the natural scene imaged with their text detected images as shown in fig, 9,10,11,12

Fig.9, 10, 11, 12 (Images and Text Detected)

In fig 9 ,10,11 and 12 compartive analysis of detection by proposed and existing approach.In fig 9 Detected text subscription depict the existing approach and other are proposed approach.

V. CONCLUSION

In this paper,analysis challenges in scene text analysis.In textlocalization find the effective box on the text.Proposed approach do this process by CNN with SVM and existing approach use region finding in text. text regions that their background is not mostly darker or lighter than the foreground the proposed method cannot detect the region.In experiment work on ICDAR 2013 with 500 images and classified by neural network and CNN with SVM .CNN with SVM show significance improved performance in precision,recall and accuracy which show in figure 5,6,7 and 8

REFERENCES

[1] Liu, Xiaoqian, Ke Lu, and Weiqiang Wang. "Effectively localize text in natural scene images." In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 1197-1200. IEEE, 2012.

[2] Sumathi, C. P., T. Santhanam, and G. Gayathri Devi. "A survey on various approaches of text extraction in images." *International Journal of Computer Science and Engineering Survey* 3, no. 4 (2012): 27.

[3] Zhou, Yu, Shuang Liu, Yongzheng Zhang, Yipeng Wang, and Weiyao Lin. "Text localization in natural scene images with stroke width histogram and superpixel." In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pp. 1-4. IEEE, 2014.

[4] Soman, Sreeja K., and M. P. Sindhu. "Scenetext Detection and Recognition in Natural Images with Maximally Stable Extremal Regions and Stroke Width Transform." *International Journal of Computer Science and Mobile Computing IJCSMC, Vol. 3, Issue. 9, September 2014, pg.663 – 669* (2014).

[5] Ye, Qixiang, and David Doermann. "Text detection and recognition in imagery: A survey." *IEEE transactions on pattern analysis and machine intelligence* 37, no. 7 (2015): 1480-1500.

[6] Lu, Shijian, Tao Chen, Shangxuan Tian, Joo-Hwee Lim, and Chew-Lim Tan. "Scene text extraction based on edges and support vector regression." *International Journal on Document Analysis and Recognition (IJDAR)* 18, no. 2 (2015): 125-135.

[7] Walia, Simranjit Singh, and Gagandeep Singh. "Color based Edge detection techniques–A review." *International Journal of Engineering and Innovative Technology (IJEIT)* 3, no. 9 (2014): 297-301.

[8] Cho, Hojin, Myungchul Sung, and Bongjin Jun. "Canny text detector: Fast and robust scene text localization algorithm." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3566-3573. 2016.

[9] Siddiqui, Imran, and Varsha Namdeo. "Cohesive multi-oriented text detection and recognition structure in natural scene images regions has exposed." *Int. J. Distrib. Parallel Syst.* 7, no. 6 (2016): 01-15.

[10] Kumuda, T., and L. Basavaraj. "Edge Based Segmentation Approach to Extract Text from Scene Images." In *Advance Computing Conference (IACC), 2017 IEEE 7th International*, pp. 706-710. IEEE, 2017.

[11] Stag log, "Comparison of Text Extraction Techniques-", Available at: http://www.rroij.com/openaccess/pdfdownload.php?download=open-access/comparison-of-text-extraction-techniques-areview.pdf&aid=44548