A Survey on Text Recognition from Natural Scene Images

Revathy A S
M-Tech Scholar
Dept. of Computer Science and Engg
College of Engineering Kidangoor
Kerala, India

Anitha Abraham
Assistant Professor
Dept. of Computer Science and Engg
College of Engineering Kidangoor
Kerala, India

Jyothis Joseph
Assistant Professor
Dept. of Computer Science and Engg
College of Engineering Kidangoor
Kerala, India

***Abstract* - Natural image has many features, and the text in natural scene image has different meanings. In complex environments such as light, low visual acuity, dim fonts, font distortion, and multiple colours, it is very difficult to distinguish text from a natural scene image. This paper gives a survey on different text recognition methods in complex backgrounds. Different types of methods are used to extract text from complex natural scenes. This survey also includes a comparative study of different text recognition methods based on the accuracy and data sets used. There are mainly two types of classification used in this paper: image processing-based methods and deep learning-based methods. This comparative study is helpful for beginners in research fields.**

Keywords - Image processing; Deep learning; CNN; RNN

I. INTRODUCTION

Text that appears in images contains important, useful and rich semantic information. This information is of great value for image interpretation. Text localization and recognition of natural scene images are based on the analysis and identification of scanned documents and images. Reading text from natural images is a challenging problem mainly in complex backgrounds. Deep learning is a type of machine learning and it is a main part of data science including statistics and predictive modeling. A neural network is a series of algorithms that are designed to recognize different patterns. It seeks to identify the underlying relationships in a set of data through a process that mimics the way the human brain works. It currently provides excellent solutions to many problems such as speech recognition, image recognition and natural language processing.

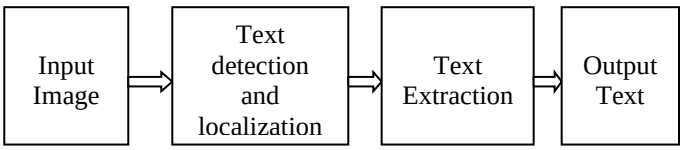

Fig. 1. Text Recognition Steps

Image processing is a method of performing certain actions on an image to obtain a modified image or to extract some meaningful information from images. Different image processing methods like morphological operations and MSER (Maximally Stable Extremal Regions) detectors are used for text extraction from scene images. There are many applications for locating and recognizing text from images, such as vehicle number plate recognition, keyword based image exploration, objects recognition, and visually impaired assistance. The text recognition process is mainly divided into the following steps (fig. 1)

- *Input Image:* It is the natural scene image contains text with complex backgrounds.
- *Text Detection and Localization:* Text detection means detecting the text in input image and localization means locating the text areas by eliminating the background regions.
- *Text Extraction:* The detected text in natural scene images is extracted to words or strings.
- *Output Text:* Output is in the form of words, strings or characters

For better comparative study, the papers considered in this survey are mainly divided into two categories that is the text detection methods are categorized into Neural Network in deep learning based methods and Image processing based methods (fig. 2).

- *Neural Networks in Deep Learning:* Different deep learning based methods are used for text detection from natural scene images. Convolutional Neural Networks (CNN), Feature Pyramid Networks (FPN) and Recurrent Neural Networks (RNN) are some examples.
- *Image Processing based methods:* Image binarization, morphological operations like erosion and dilation, MSER detector are some image processing methods used for detection and extraction of text from natural scene images.

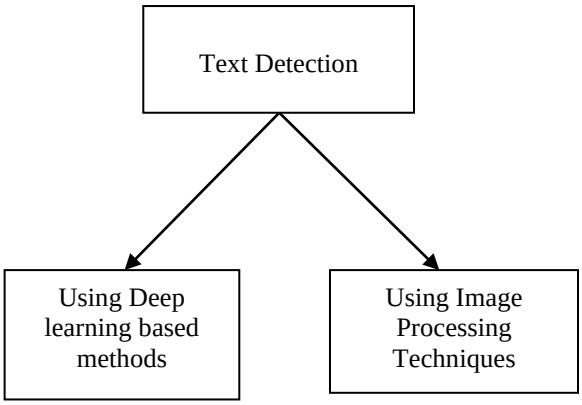

Fig. 2. Classification of literature

II. LITERATURE SURVEY

A. Text Extraction using Neural Network in Deep Learning

Many researchers are using deep learning based methods to extract text features from natural scene images. Liu et al. [5] proposed a Feature Pyramid Networks (FPN) that combines CNN and RNN to perform multi-scale and multi-oriented scene text detection and it applied in the CNN part.

Zuo et al. [1] proposed an Encoder-Decoder framework for scene text recognition which combines connection time classification (CTC) and attention mechanism that converts the natural text into a sequence mark. The feature extraction of input image is carried out by CNN and extracted features are encoded using Bidirectional Long Short term Memory (BiLSTM) and generate the feature maps. CTC and attention mechanism is used in the decoder layer to decode into the output. The text recognition model includes three parts: feature extraction, sequence analysis, and CTC-Attention joint mechanism decoding. The performance of the proposed system is evaluated using four datasets: SVT (Street View Text), IIIT5K Words, ICDAR2003 and ICDAR2013. The system gets more accuracy by using the dataset ICDAR2013.

Chernyshova et al. [2] proposed Two-step CNN framework for text line recognition. The method is for the ID recognition of camera captured images. Text line recognition is the part of the field recognition step in the ID recognition. The two ANNs are used here known as segmenter and classifier. Representation of a text line includes an Ascender line, Ascent line, Cap line, Mean line, Baseline, Descender line. The performance of the proposed system is evaluated using the datasets: MIDV-500, MNIST. The evaluation metrics used here is OCR accuracy and the proposed method have the 96.69% OCR accuracy.

Gao et al. [3] proposed a detection and verification model based on Single Shot Multibox Detector (SSD) and Encode-Decoder network which consist of initial text detection with text localization neural network and eliminating background regions with text verification model. The text localization neural network is based on SSD which detect only horizontal text. The text verification model is mainly based on Encoder-Decoder framework which deletes the non-text areas detects from initial detection. The BiGRU (Bidirectional Gated Recurrent Unit) network in the encoder layer encodes context information with text image features. The GRU (Gated Recurrent Unit) network with attention mechanism in decoder layer that decodes feature sequences into words. The performance of the proposed system is evaluated using the dataset: ICDAR 2017. The evaluation metrics used are Precision, Recall and F-Measures.

Liang et al. [4] proposed a RNTR-Net: A Robust Natural Text Recognition Network. RNTR is the combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). CNN is used for feature extraction and RNN is used for sequence recognition. The method of text recognition includes mainly three steps. Feature extraction, Sequence Recognition and Transcription. The input field image first passed through convolutional layer with residual block and generates the feature maps. The feature sequences from feature extraction are passed through two BLSTM's (Bidirectional Long Short term Memory) and these are very effective for modeling and predicting sequential data. The transcription layer which translates the frame sequences to final output. The performance of the proposed system evaluated using the datasets: ICDAR 2003, ICDAR 2013, SVT and IIT5K. The method has 96.7% accuracy by using the dataset IIIT5K.

Liu et al. [5] proposes scene text detection with Feature Pyramid Based Text Proposal Network (FTPN). The CNN section uses a feature pyramid network to extract the multi-scale feature of an image. Bidirectional Long Short Term Memory (Bi-LSTM) is used for encoding. Region of Interest (RO) pooling is used to speed up the training and testing process. Text connector is used in the last stage to construct the final output. The Performance of the proposed system evaluated by using the datasets: ICDAR 2013, ICDAR2015. The evaluation Metrics used are Precision, Recall and F-measure.

B. Image processing based methods for Text Extraction

Su et al. [6] proposed image processing technologies for text recognition with the combination of Optical Character Recognition (OCR) technologies. The proposed method includes a character recognition system for cosmetic related advertisement images. Also a text detection and recognition system is used for natural scenes. The first system converts the input image into a gray scale image and simplifies the image, and then searches for counters using the sobel filter. The next step is the binarization process. This process detects the limit value of the gray scale image and checks whether a pixel has a specific value or not. Dilation adds pixels to the object boundaries and erosion removes pixels on object boundaries. In next step redilation is performed and seeks the character contours and then Region of Interest (ROI) is lassoed. The last operation is character recognition which

identifies the individual characters in an image. The performance of the system is evaluated by using the dataset ICDAR Robust Reading Competition used as test images. The evaluation metrics used are Accuracy, Recall and F1-measures. The system achieved 93% accuracy.

Ghoshal and Banerjee [7] proposed an improved scene text and document image binarization scheme. This method is more effective than other binarization methods for the natural image, which contains lower resolution, noise, lower visual acuity, and unbalanced light. This is a new approach to natural scene text image binarization by tracking the text border based on edge and gray level variant information. In addition, the broken boundaries are connected to form a complete boundary map. An adaptive threshold is set here based on the boundary edge information for effective binarization of the image. The proposed method first read the image and computes the variance matrix. To obtain binarized image different adaptive thresholding techniques are applied. The performance of the system is evaluated by using Precision, Recall and F-measures.

Rampurkar et al. [8] proposed text detection from complex images using morphological techniques like erosion and dilation. There is a color based partition is performed on the input image having text. Next step is the connected component labeling which is used to detect the connected regions in color and digital images. In the adjacent character grouping method, the sibling groups of each character candidate are treated as string segments and then the fragmented sibling groups are merged into the text. Text line grouping methods is used to locate text strings with arbitrary orientations. The evaluation metrics used are Precision and Recall.

Chidiac et al. [9] proposed a robust algorithm for text extraction from images. Here Maximally Stable Extremal Regions (MSER) method is used instead of performing simple thresholding method. It is a type of blob detection in images and extracting a comprehensive number of corresponding image elements contributes to wide baseline matching. The MSER enhancer is enhances the detected region. Stroke Width Transform (SWT) is a technique used to extract text from noisy natural scene image, by isolating shapes it share a constant stroke width and produced more reliable results. The resulting system is to detect text regardless of its scale, font and direction. Then filtering is applied on it which is technique that modify or enhancing an image. After filtering the text line is formed with similar height, constant spacing and similar stroke width by grouping connected component. The performance of the proposed system evaluated by using the datasets: ICDAR and KAIST Scene Text database. The evaluation metrics used are Precision Rate, Recall Rate and F-measure.

Ling et al. [10] proposed a model for automatic recognition of vertical texts in natural scene images. The method includes mainly two processes Text localization and segmentation then the second process is Text Recognition. Gray scaling operation is performed on input scene image containing vertical text then a gray scaled scene image containing vertical text is formed. By using MSER Detector gray scaled image with possible candidate characters are formed then the binarization process applied and the output is binary image with possible candidate characters. After dilation binary image with possible candidate characters regions are formed. The last step is connected component segmentation which eliminates false positives. Segmented text regions obtained from text localization and segmentation in binary image undergoes correction and orientation determination and Optical Character Recognition (OCR). In orientation determination there are three types of vertical texts, Horizontal-stacked, Top-to-Bottom and Bottom-to-Top. Before OCR recognition Top-to-Bottom and Bottom-to-Top vertical texts are rotate 90 degree. The output of OCR process is a vertical text, then string formation operation is performed on it and the final output is a character string. The performance of the proposed system is evaluated by using the datasets: SVT, MSRA-TD 500. The evaluation metrics used are Precision, Recall and F-measures.

III. COMPARATIVE STUDY

TABLE 1. COMPARISON OF LITERATURE (DEEP LEARNING METHODS)

SL NO	AUTHOR AND YEAR	DATASET USED	ACCURACY
1	Zuo et al (2019)	ICDAR 2003	98.2%
2	Chernyshova et al (2020)	1961 census sample	96.69%
3	Gao et al (2019)	IIIT-5K	98.7%
4	Liang et al (2019)	ICDAR 2017	Precision=83% Recall=78.4% F-measure-=80.6%
5	Liu et al (2019)	ICDAR 2013 ICDAR 2015 USTB-1K	Precision=93.2% Recall=91.9% F-measure=92.5%

TABLE 2. COMPARISON OF LITERATURE (IMAGE PROCESSING METHODS)

SL NO	AUTHOR AND YEAR	DATASET USED	ACCURACY
1	Su et al (2019)	ICDAR	93%
2	Ghoshal and Banerjee (2018)	ICDAR 2011 ICDAR 2003	Precision=92% Recall=88% F-measure=89.50%
3	Rampurkar et al (2018)	Its own dataset	Recall=77% Precision=70%
4	Chidiac et al (2016)	ICDAR KAIST	Precision=90% Recall=88% F-measure-=89%
5	Ling et al (2018)	MSRA-TD500 SVT	Precision=89% Recall=87% F-measure=88%

There are three main types of evaluation measures used: Precision, Recall and F-measures. Accuracy is also used in some papers as an evaluation measure. Precision and recall are the two main model evaluation measures. Precision measures the number of positive class predictions that include the positive class. The recall counts the number of positive class predictions made from all the positive examples in the dataset. F-measurements give a single score, which compares the accuracy and recall of a number. Accuracy is the measure of closeness to an actual value.

Different types of datasets have been used in these survey papers. They are IIIT-5K: It is a word dataset that contains query words like billboards, movie posters, signboards and also it has 5000 cropped word images from scene texts and born digital images. SVT (Street View Text) contains 647 text image and are corrupted by low resolution, blur and noise. ICDAR (International Conference on Document Analysis and Recognition) that consist of ICDAR 2003 (IC03), ICDAR 2013 (IC13), ICDAR 2017 (IC17), ICDAR 2015(IC15) datasets. ICDAR 2003 contains 860 cropped images and 251 text images. The test dataset of IC13 contains 1015 labeled text images and this dataset harvested from IC03 dataset. The background of IC15 dataset is hard and confused. MIDV 500 (Mobile Identity Document Video) dataset contains ID document images. KAIST scene text dataset consist of 3000 images captured in complex backgrounds like different lightning conditions. MSRA-TD500 is a text detection dataset consist of 200 test images and 300 training images. Each dataset gives different types of results. Some datasets give more accuracy and some gives less accuracy.

There are several applications for the text detection and recognition such as object or product identification, Automated reading, Web indexing, Retrieval and indexing of video images, Document processing and Screenshot OCR. The survey had to deal with a variety of issues, such as blurred background images, multilingual environments, uneven lightning and noisy images. In this types of situations the text detection and recognition is a very challenging problem. Therefore, this survey finds which gives the most accuracy when comparing different types of methods.

IV. CONCLUSION

Text recognition from natural scene image is a difficult task due to complex backgrounds. The papers in this survey have been evaluated according to two types of classification. Therefore, comparative study has been facilitated. Deep learning methods are more accurate and give better results than image processing methods. From the comparative study of all the methods, it was found that the deep learning method RNTR-Net (A Robust Natural Text Recognition Network) gives more accuracy. Accuracy, precision, recall, and F-measure are the evaluation measurements adopted to determine the performance of methods. The performance of each system varies according to the datasets used.

ACKNOWLEDGMENT

The author would like to thank Dr. Ojus Thomas Lee (HOD & Associate Prof. Dept. of CSE, CE Kidangoor) , Mrs. Rekha K.S (Assistant Prof. Dept. of CSE, CE Kidangoor) for their valuable suggestions.

REFERENCES

[1] Zuo, L. Q., Sun, H. M., Mao, Q. C., Qi, R., & Jia, R. S. (2019). Natural scene text recognition based on encoder-decoder framework. *IEEE Access*, *7*, 62616-62623

[2] Chernyshova, Y. S., Sheshkus, A. V., & Arlazarov, V. V. (2020). Two-step cnn framework for text line recognition in camera-captured images. *IEEE Access*, *8*, 32587-32600.

[3] Gao, X., Han, S., & Luo, C. (2019). A Detection and Verification Model Based on SSD and Encoder-Decoder Network for Scene Text Detection. *IEEE Access*, *7*, 71299-71310.

[4] Liang, Q., Xiang, S., Wang, Y., Sun, W., & Zhang, D. (2020). RNTR-Net: A robust natural text recognition network. *IEEE Access*, *8*, 7719-7730.

[5] Liu, F., Chen, C., Gu, D., & Zheng, J. (2019). FTPN: scene text detection with feature pyramid based text proposal network. *IEEE Access*, *7*, 44219-44228.

[6] Su, Y. M., Peng, H. W., Huang, K. W., & Yang, C. S. (2019, November). Image processing technology for text recognition. In *2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)* (pp. 1-5). IEEE.

[7] Ghoshal, R., & Banerjee, A. (2018, March). An improved scene text and document image binarization scheme. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)* (pp. 1-6). IEEE.

[8] Rampurkar, V. V., Shah, S. K., Chhajed, G. J., & Biswash, S. K. (2018, January). An approach towards text detection from complex images using morphological techniques. In 2018 2nd International Conference *on Inventive Systems and Control (ICISC)* (pp. 969-973). IEEE.

[9] Chidiac, N. M., Damien, P., & Yaacoub, C. (2016, June). A robust algorithm for text extraction from images. In *2016 39th International Conference on Telecommunications and Signal Processing (TSP)* (pp. 493-497). IEEE..

[10] Ling, O. Y., Theng, L. B., Chai, A., & McCarthy, C. (2018, November). A Model for Automatic Recognition of Vertical Texts in Natural Scene Images. In *2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)* (pp. 170-175). IEEE.