International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.1 (Jan-2021)
http://dx.doi.org/10.12785/ijcds/100120

Hybrid Technique: Text Detection Using a Neural Network and Boxes

Hamsa D. Majeed[1] and Reem M. Ibrahim[2]

[1] *College of Science and Technology of the University of Human Development, Sulaimania, Iraq*
[2] *Department of Control & System Engineering of the University of Technology, Baghdad, Iraq*

Received 29 May 2020, Revised 16 Jul. 2020, Accepted 29 Jul. 2020, Published 1 Jan. 2021

Abstract: Text Recognition is one of the challenging tasks of computer vision with considerable practical interest. This paper presents a text detector system, named NNBoxes, which detects text with both high accuracy and efficiency and move with the topic forward, NNBoxes is a hybrid technique between the Neural network (Noise reduction and letters prediction) and (Lockbox and hitbox) method created by the Author, both boxes draw grids and lines to recreate the shape of the letters and draw a path between the boxes to support the decision algorithm, in some cases the boxes technique detect text pattern in the image that the language text is not supported in the algorithm Dataset and train data.

Keywords: Text detection, Text Recognition, Neural Network (NN), AI.

1. INTRODUCTION

With emerging of useful technologies that facilitate people's lives, Text detection and recognition in many different languages, fonts and sizes were one of these technologies, this emerging technique came after many types of research and studies in this field all led to an improvement on algorithms, applications and approaches to detect and recognize text in the images.

This study is a Hybrid Technique between Neural Network technology and the proposed technique called "Boxe's", the name came from the term “search and find”, which means the search for the letters, find it and recognize it. Also, the name came to show that the method will not leave an area without searching and findings. The Hybrid Technique came to add improvement on the previous studies in text detection techniques fields, with no depending on one technique or algorithm, the NN will reduce the noise in the image with an early prediction (there is no prediction or text recognition in this part, this phase came to save time and resources before taking the selected area for further investigation) if the specific area contains a possible text the Boxes technique will start dividing the area in phase one into grids we will call it the LockBox will be divided to the grid of lines called the hitbox, in this phase the hitboxes will start drawing lines to connect the pixels to make an estimated shape or find a possible shape in the picture (text or letters), the final drawing shape will be sent again to a pre-train NN algorithm for the final call the final phase will be drawing the result box on the text or letters in the investigated area.

2. RELATED WORK

Tian et al. [1] Propose a segmenting frame that extracts each text instance as an individual linked portion and handles text-detection as instance segmentation. System [1] maps pixels in the embedding region, enabling pixels belonging to the same text to look closer together and to look back. A Shape-Aware loss to provide training adapts to different aspect ratios of the test instances and the small gaps between them and a different post-processing pipeline to produce accurate predictions for bounding boxes. The findings of our work on our three difficult datasets (ICDAR15, MSRA-TD500, and CTW1500).

Liao et al. [2] Text recognition approach scene from a two-dimensional point of view. A basic but powerful model is devised for recognizing the texts in arbitrary types called Character Attention Fully Convolutional Network (CA-FCN). Scene text recognition is accomplished through a semantic segmentation network that adopts a character focus mechanism. CA-FCN can

E-mail: hamsa.al-rubaie@uhd.edu.iq , 11362@uotechnology.edu.iq

simultaneously detect the script and predict each character's position combined with a word-formation module. Experiments reveal that both normal and unusual text datasets use the new algorithm rather than ever.

Lyu et al. [3] Investigate the issue of text spot in a scene that aims to detect and recognize simultaneous text in natural images. An end-to-end neural network model is introduced for scenario text spotting. The proposed model, known as Mask TextSpotter, is inspired by Mask R-CNN's recently published article. Mask TextSpotter uses a simple, smooth end to end learning algorithm, which is distinct from previous approaches that often conduct text spotting using deep neural end-to-end trainable networks. In comparison, the treatment of irregular messages, for example, bent text, is superior to previous approaches. ICDAR2013, ICDAR2015, and Total-Text tests demonstrate that the suggested approach helps in both text scene identification and end-to-end text recognition function.

Lu et al. [4] Propose a new approach that combines a corner response map and transfers deep coevolutionary neural networks for video text detection and recognition. To detect strong recall and division of candidate text regions into candidate text lines by screening analysis using the two alternative approaches, using a corner response map. A prediction analysis. To eliminate false-positive effects, the authors build classification networks transferred to VGG16, ResNet50, and InceptionV3. By creating a novel c-fuzzy, you can obtain a clean layer of text from complicated backgrounds such that the text is correctly identified by a program for industrial, optical character recognition.

The key aspect of the typical material spotting pipeline is the detection of contents techniques [5]: the distribution of separating or skipping words on the signature scene images. A fundamentally meaningless activity is the identifying phenomenon of terms in uproarious and jumbled images, and the strategies used for this reason depend on each characteristic region ([6]; [7]; [8]; [9, 10]; [11, 12, 13, 14]; [15]; [16]).

The district character strategies require pixels to be separated into characters and characters to be translated into words. [7] can consider locals with a clear stroke width – the difference between two parallel edges – by adjusting stroke width (SWT). Characters are regions of the comparable width of the stroke and thus structure characters are grouped into pixels and characters obtained by geometry. [14], revert to the notion of characters as strokes and use biases to consider the strokes rather than the SWT. In comparison to places where the stroke width is constant, [11, 12, 13] use the local 123 Int J Comput Vis as a character [17]. The Maximally Stable Extreme Regions are being established by [10] by incorporating a strong CNN definition, to effectively prune trees in extreme regions, encouraging fewer deceptive identifiers.

3. METHODOLOGY

In this part, a Hybrid technique with NN and the lockbox and hitbox technique (Boxes) will be used, this technique created by the author (the technique will be explained later in this section). The aim is to build an accurate and fast model to detect the text from images using NN, the Author will tackle the problem with a Hybrid technique between NN and Box's method.

We implement our method in Python and conduct all experiments on a regular workstation with Nvidia Titan Xp GPUs. The model is trained in parallel and evaluated on a single GPU.

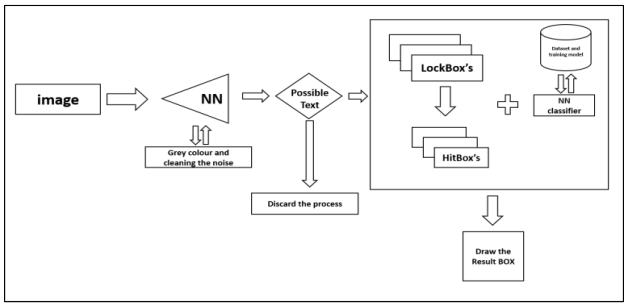

Figure1. NNBoxes workflow

The boxes will upload the picture and start the following process:

1. The image will be converted to grey style and noise reduction, NN Algorithm will be used in this process for early call wither this image contains possible text or not. After the NN process, the NNBoxes will determine where the possible area that contains a text or not blank areas.
2. The NNBoxes will search for differences in image color locating a pattern that can be a text, the NNBoxes will mark it in the box, the box will be sent for deep investigation in square 3.
3. The marked box will be named lockbox, if lockbox > 0, the box will be divided into the grid (100*100), (Fig. 2), the inner squares will be also divided to another grid (100*100), the NNBoxes will start drawing lines inside each square and connecting lines in the neighbor squares for a possible hit or what the process here called connecting the "hitbox" (Fig. 3), the NNBoxes will arrange the Hitboxes together from the inner grid to the upper grid.

The new network consists mainly of two sections: extraction of features and multistage fusion. The VGG16 network is a backbone network that is pre-trained on ImageNet. The last pooling layer and the following layers are discarded fully connected. As sizes of text can vary considerably, small text instances with just coarse

features are difficult to detect. Therefore, features from various stages are combined to capture text instances of multiple scales. In particular, we use the VGG16 backbone network role maps from stage3, stage4, and stage5. Such multi-level characteristics are sampled to the same size as the stage3 characteristic and are then combined with concatenation. The result is a two-channel map that predicts the direction field given by Eq, followed by three convolution layers. Finally, the forecast path field is sampled in the original dimension. For all upstream sampling operations, we follow bilinear interpolation.

To reverse the proposed direction region, and an instance-balanced Euclidean loss, network parameters are optimized. In particular, the loss function is a weighted sum of each pixel of the Ω domain, the average squared error. The following are given by:

$$L = \sum_{p \in \Omega} w(p) * \|V_{gt}(p) - V_{pred}(p)\|_2, \qquad (1)$$

where the predicted direction field is Vpred and w(p) indicates the pixel p weight coefficient. As text sizes that vary significantly in the images on a scene, large text instances in loss computation will be dominant while smaller text instances are ignored if all text pixels contribute equally to the loss function. We adopt a balanced strategy to address this problem. More specifically, weight w for a given pixel p is defined as follows for an image containing N text instances:

$$w(p) = \begin{cases} \frac{\sum_{T \in \mathbb{T}} |T|}{N * |T_p|}, & p \in \mathbb{T} \\ 1, & p \notin \mathbb{T} \end{cases} \qquad (2)$$

where indicates the total pixel number for a text T instance and Tp refers to the text pixel p instance. Every text instance of any size is equally weighted and contributes to Eq's loss function. (2). This is in line with the current program for text detection, which is equally critical for each test case.

All subnetworks of our model can be trained both synchronously and end-to-end, unlike previous methods of text spotting using two independent models (the detector and the recognizer) or alternate training strategy. The entire training process includes real-world data pre-trained.

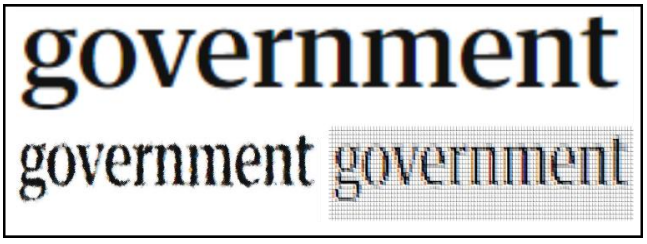

Figure2. lockbox for possible text supported by NN

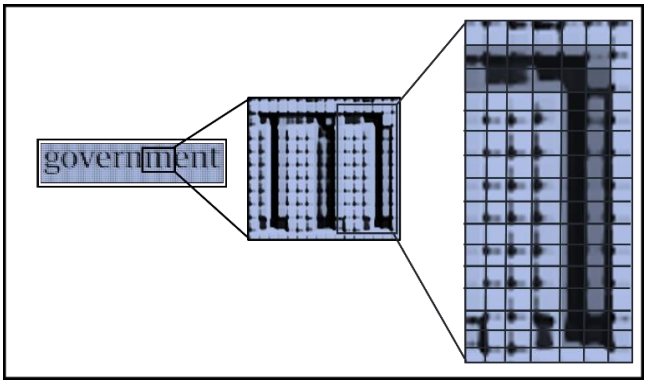

Figure3. NNBoxes draws the line in the hitbox

4. For more certainty, the neural network technique will be used again to check the results from the Hitboxs, in this phase, there is no recognition. The goal of this phase is to increase confidence that the results contain text, not an icon or image that resembles a letter or text.

5. After scanning the image supported by NN and (Lockbox and Hitbox), the NNBoxes will draw the result box on the detected text on the resulted image without any change in the real image. In the next section, the results and discussion several images will be tested using the methodology, and the resulted image and the table of the summary will be presented and discussed.

4. RESULTS AND DISCUSSION

In this section, four images have been chosen to be tested on the model that been proposed, the images are different from shape, size, and text content. To test the model with maximum letters from languages we choose the images in Fig. 4 and Fig. 6, this approach came to test the limit of the proposed model and masseurs the weakness. Fig. 8 and Fig. 10 an article recent articles from the Washington Post and Aljazeera testing the model with the matter dealing with a lot of letters in the single images and what is the error percentages that can be appeared. In the end, a summary table will be presented with the results and accuracy rates from the results and comparison in table I.

In Fig. 4, a Multilingual 'Hello' Poster has been used for the first test; in this test, the aim is to monitor the results from the model and the type of the languages that the system is capable to deal with. As the results after the run of the model in Fig. 5. The results show that the system is capable to detect with the most letters with a different type of font and colors.

Figure4. Multilingual 'Hello' Poster

Figure5. Multilingual 'Hello' Poster Results

In Fig.6, Google Languages box has been used to be tested in the model, the aim of this test to show that the model is capable to detect the most languages letters without the extra noise in the images and the font type, this challenge has been showing in Fig.5 and the model shows more than 90% accuracy of detecting the words and letters, Fig.7 shows the results after running the model with a detection rate 95%, the words in (Urdu and Korean) has not been selected by the model.

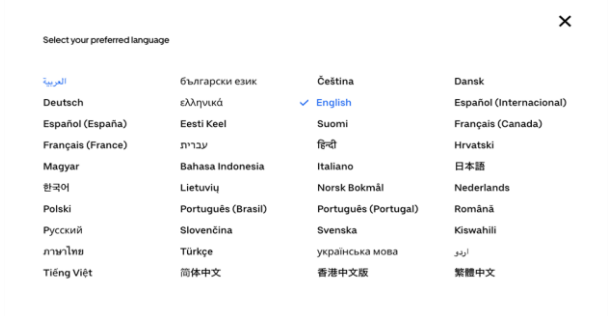

Figure6. Google Choose languages Box

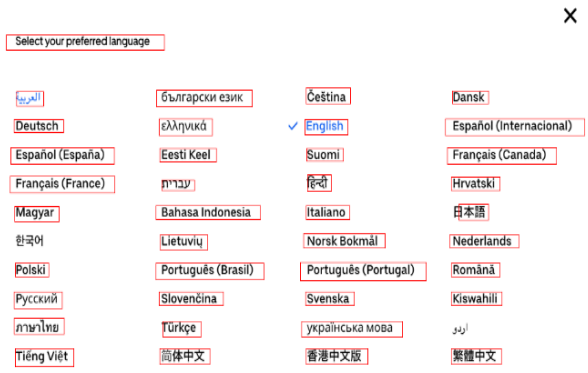

Figure7. Google Choose languages Box results

In Fig.8, the model has been tested on the Aljazeera English article, the model detected the words in English as shown in Fig.9.

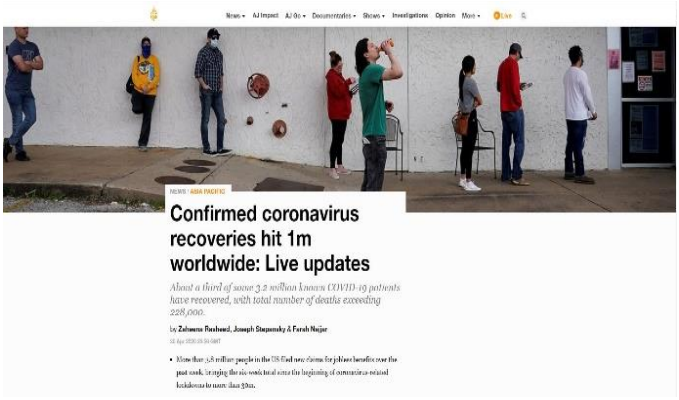

Figure 8. Aljazeera English article

Figure 9. Aljazeera English article results

In Fig.10, the model has been tested on the Aljazeera Arabic article; the model shows accurate rate detection

words in Arabic as shown in Fig.11 as the model can detect Arabic sentences in the article.

Figure10. Aljazeera Arabic article

Figure11. Aljazeera Arabic article Results

Table I summarizes the results from running the images and the detection rate and the total for the accuracy of the model.

TABLE I. RESULTS OF THE IMAGES

Image	Accuracy
Multilingual 'Hello' Poster	97%
Google Choose languages Box	95%
Aljazeera English article	98%
Aljazeera Arabic article	94%
Total:	**96%**

Fig .12 shows a chart presenting a comparison in the time difference in detecting text (with and without using the Box's technique). X here presenting Text blocks (words) and the time was measured in seconds. The results show that using the Box's technique made the process faster due to the support from the Box's to find the text and pattern of the text.

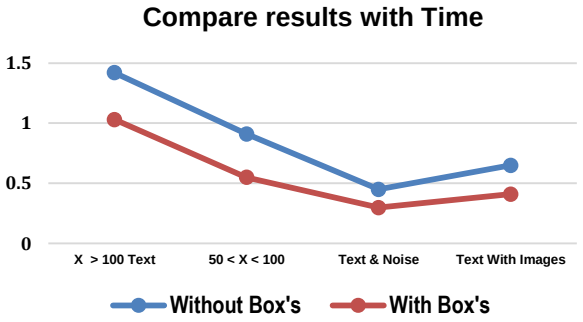

Figure12. Comparison charts in time between (with Box's or without Box's) technique

Table II offers speed comparisons. While our model is not the fastest, it is possible to equate the speed of our system with previous models. Specifically, the input scale can be 3.8 FPS, 3.1 FPS and 2.0 FPS, 720 AD1280, 1000 AD1778, and 1600 AD2844. It needs approximately 0,20 seconds for the input scale of 720 x 1280 and average 0,06 seconds for the detection.

TABLE II. RESULTS WITH OTHER METHODS

Image	Accuracy
Baseline OpenCV3.0 + Tesseract [18]	**14.7**
TextSpotter [19]	**37.0**
TextBoxes++ [20] MS	**76.5**
He et al. [21]	**85.0**
Deep text spotter [22]	**58.0**
The Proposed Method	**96%**

Comparing the results with the previous methods, the results show that the proposed method had a better resulting in the detection and spotting the letters in the Images and had a better and strong True labeling and detection of the letters.

5. CONCLUSION

The presented NNBoxes is a Hybrid Text Detection software and algorithm, the NNBoxes showed reliable and accuracy in detecting text in English (Latin) letters and less accuracy with Arabic and Asian letters, also Box's technique supported NN to detect text in the images with reduce in process time and increase accuracy. The Accuracy rate that we obtained is better than the compared studies that been used in this paper,

With the previous results, the research will continue to enhance the capabilities for the software and the hybrid technique.

6. LIMITATION

Even with Box's support for detecting the text and the pattern, the NNBoxes still not that accurate to detect other types of letters especially the Asian also the big letters and huge spaces between the letters made it hard to detect that these letters belong to one word or sentence.

REFERENCES

[1] Z. Tian, M. Shu, P. Lyu, R. Li, Ch. Zhou, X. Shen, and J. Jia; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4234-4243,(2019).

[2] M. Liao, J. Zhang, Z. Wan, F. Xie, J. Liang, P. Lyu, C. Yao and X. Bai; IEEE Scene Text Recognition from Two-Dimensional Perspective, (2018)

[3] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai; The European Conference on Computer Vision (ECCV), pp. 67-83,(2018)

[4] L. Lu, Y. Yi, F. Huang, K. Wang, and Q. Wang, "Integrating Local CNN and Global CNN for Script Identification in Natural Scene Images", Access IEEE, vol. 7, pp. 52669-52679, (2019).

[5] X. Chen and A. L. Yuille, "Detecting and reading the text in natural scenes". Computer Vision and Pattern Recognition, (CVPR) (Vol. 2, pp. II-366). Piscataway, NJ: IEEE,(2004).

[6] H. Chen, S. Tsai, G. Schoch, D. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge enhanced maximally stable extremal regions ", In Proceedings of the international conference on image processing (ICIP) (pp. 2609– 2612), (2011).

[7] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform", In Proceedings of the IEEE (pp.2963– 2970, IEEE), (2010).

[8] L. Gomez, and D. Karatzas, "Multi-script text extraction from natural scenes", In 12th International conference on document analysis and recognition (ICDAR) (pp. 467–471). (2013) IEEE.

[9] L. Gomez, and D. Karatzas, " A fast hierarchical method for multi-script and arbitrary oriented scene text extraction", arXiv:1407.7504, (2014)

[10] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees", In D. J. Fleet, T. Pajdla, B.Schiele, and T.Tuytelaars (Eds.), (2014).

[11] L. Neumann, and J. Matas, "A method for text localization and recognition in real-world images", In Proceeding of the Asian conference on computer vision (pp. 770–783) Springer,(2010).

[12] L. Neumann, and J. Matas, "Text localization in real-world images using efficiently prune exhaustive search", In Proceedings of ICDAR (pp. 687–691) IEEE, (2011).

[13] L. Neumann, and J. Matas, "Real-time scene text localization and recognition", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (CVPR) (2012)

[14] L. Neumann, and J. Matas, "Scene text localization and recognition with oriented stroke detection", In Proceedings of the international conference on computer vision (pp. 97–104) (2013).

[15] C. Yi, and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping", IEEE Transactions on Image Processing, 20(9), 2594–2605 (2011).

[16] X. C. Yin, X. Yin, and K. Huang, "Robusttextdetectioninnatural scene images", CoRRarXiv:1301.2628, (2013)

[17] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions", In Proceedings of the British Machine Vision Conference, Cardiff, UK, pp. 384–393 (2002).

[18] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. K. Ghosh, A. D. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valley. ICDAR 2015 competition on robust reading. In Proc. ICDAR pages 1156– 1160, 2015.

[19] L. Neumann and J. Matas. Real-time lexicon-free scene text localization and recognition. IEEE Trans. Pattern Anal. Mach. Intell., 38(9):1872–1885, 2016.

[20] M. Liao, B. Shi, and X. Bai. Textboxes++: A single-shot oriented scene text detector. IEEE Trans. Image Processing, 27(8):3676–3690, 2018.

[21] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, and C. Sun. An end-to-end textspotter with explicit alignment and attention. In Proc. CVPR, pages 5020–5029, 2018.

[22] M. Busta, L. Neumann, and J. Matas. Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In Proc. ICCV pages 2223–2231, 2017.

Hamsa D. Majeed, Master in Electronic Engineering form University of Technology. Research area is image processing and Artificial Intelligent.

Reem M. Ibrahim, Master in Computer Science form University of Technology. Research area is image processing and Artificial Intelligent