Volume XXVI 2023
ISSUE no.2
MBNA Publishing House Constanta 2023

Scientific Bulletin of Naval Academy

SBNA PAPER • OPEN ACCESS

The use of Scene Text Detection methods for news crawl extraction

To cite this article: Milutinovici Sorin and Răcuciu Ciprian, Scientific Bulletin of Naval Academy, Vol. XXVI 2023, pg. 165-170.

Submitted: 20.01.2024
Revised: 20.02.2024
Accepted: 18.03.2024

Available online at www.anmb.ro

ISSN: 2392-8956; ISSN-L: 1454-864X

doi: 10.21279/1454-864X-23-I2-020

The use of Scene Text Detection methods for news crawl extraction

Milutinovici Sorin, Răcuciu Ciprian

Titu Maiorescu University, Faculty of Informatics,
Military Technical Academy, Doctoral School,

Milutinovici Sorin, sorin.milutinovici@prof.utm.ro

Abstract. News crawl extraction is a subset of media monitoring. It is not directly related to the broadcast itself and not spoken by any anchor or guest - therefore its content cannot be inspected via speech-to-text tools. News crawl content is only available as text in video frames so its extraction can be done using Scene Text Detection techniques. This work tests the most used deep learning Scene Text Detection methods (both regression and semantic segmentation approaches) to see if they can be used as a base in news crawl extraction. Our investigation shows that regression based model CTPN performs better for our particular purpose that the semantic segmentation based EAST and CRAFT models.…

1. Introduction

Video monitoring refers either to techniques of video surveillance or to information extraction from video broadcasts. There are many types of information that can be extracted – object detection (e.g. [1,2,3]), motion prediction (ex: [4,5]), video instance segmentation [27], video summarization [6,7,8], activity recognition [9, 10], speech-to-text [11,12], and others. But many video materials have text superimposed on the images (such as news broadcasts crawl systems, tables with information, titles, section names, etc). To extract this information from video we need to correctly isolate the text regions in the video frames with the subsequent possibility to use an OCR system to convert it to text. There are many models that can correctly isolate the text regions in images, known usually as "scene text detection". Scene text detection is a hot research area lately, but its focus has shifted from simple horizontal texts [13,14] to deformed texts with multiple orientations e.g. [25], even curved texts e.g. [26]. The most successful methods are those based on various deep learning models.

Detecting the text in video broadcasts (that is, the text that has been placed in the frames by the video authors and not the one filmed accidentally) should be a simpler problem. This text is almost all the time horizontal, and it is placed so that it can be easily read. To test if this is true, we selected three scene text detection methods and tried to see how well they perform on frames extracted from Romanian tv broadcast. The first methos is a regression-based object detection network (CTPN) [15], the second is a word level detector based on semantic segmentation (EAST) [16] and the third is a character level semantic segmentation method (CRAFT) [17].

2. The models used in the tests.

All three methods that we tested are based on neural networks supervised training. Each method has a different neural network architecture and a different training/detecting process. Regardless of the inner workings of each model, a final phase of postprocessing, usually based on geometric calculations, will generate a series of text lines that will be sent to OCR. We aim to obtain a complete description of the text lines with good coverage.

2.1. Connectionist Text Proposal Network (CTPN)

Proposed by Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao in 2016 CTPN is inspired by object detection frameworks, such as Region Proposal Network technique proposed by the authors of R-CNN object detector [18].

The Faster R-CNN paper introduced a unified framework for object detection that integrated the region proposal step into the detection pipeline. Prior to Faster R-CNN, region proposal methods such as Selective Search [28] or EdgeBoxes [29] were commonly used to generate potential object regions in an image. However, these methods were separate from the object detection network, resulting in slower and less efficient detection systems. Region Proposal Network covers the image with a network of densely packed boxes having different dimensions called anchor boxes. The network will incorporate the location and size of the anchor boxes both the training process and the result.

CTPN extends RPN to text detection by adding a vertical anchor mechanism that predicts location and text/non text score of each anchor. Anchor boxes have the same size (unlike RPN) and there are connected by a recurrent neural network, thereby increasing the probability of correct detection of ambiguous regions such as spaces between words, punctuation marks, etc.

At the end of the detection phase, we get a list of anchor boxes that have different probabilities of having a text inside and a height and a vertical position. Post-processing these boxes will allow us to obtain text lines.

2.2. EAST: An Efficient and Accurate Scene Text Detector

EAST was proposed by by Xinyu Zhang, Zhanzhan Cheng, Qiuyu Zhu, and Fan Bai in 2017 [16] and it is a derivation of the semantic segmentation techniques. Unlike CTPN, that has a series of bounding boxes as results, EAST will output a series of maps that mark every pixel in the image with a value. EAST uses an FCN (Fully Convolutional Network), inspired by the U-net [19] that incorporates a multi-scale Feature Fusion mechanism to support text of various sizes.

The geometry output of EAST can be either rectangular boxes or quads (general quadrilateral polygons that have arbitrary angles). We only implemented the rectangular version and this one will generate five maps for any image: four of them with distances from the bounding box boundaries and one with the text rotation. EAST can generate word-level and text line level predictions. We only implemented the word level predictions.

At the end of the detection phase, we obtain a set of rectangles that, in the best-case scenario, will be correct bounding boxes for every word in the frame. Post-processing is necessary to get full text lines that can then be sent to OCR software.

2.3. Character Region Awareness for Text Detection (CRAFT)

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee proposed CRAFT in 2019, as a method that should work best with arbitrary-oriented text that all previous ones. Still based on semantic-segmentation technique [17], CRAFT is a character-level detector that generates two maps for each image, a character map and an affinity map. The character map will show the probability of each pixel to be part of a character in image whereas the affinity map will show for each pixel the probability to be placed between two existing characters, that is, to be part of a region that join the median vertical lines of the two adjoining characters.

Training is a problem for character level segmentation techniques since most available databases are annotated at word level. CRAFT uses a weakly supervised technique to predict the character bounding boxes during training based on the number of characters in each word (annotations contains the actual text).

At the end of CRAFT based on the probability maps, post-processing cand get word level bounding boxes that are joined to form the text lines.

3. Testing Procedure

The testing procedure aim is to determine which of the three approaches works best within the constraints of video broadcast text. Video broadcast text has good contrast, and is presented as horizontal lines, characters are in most instances evenly spaced.

Each of the three models was implemented using PyTorch library [22] and subsequently trained using ICDAR 2015 [23] and ICDAR 2017 [24] datasets.

All three models have been used to detect text boxes in the prepared dataset and the results are obtained by automatically comparing the result of the detection process with the annotations on the pictures. The detection was run at three different resolutions: SM – 320x200 pixels, MD 640x480 pixels and LG = 1280x720 pixels.

The evaluation was done by comparing the marked boxes to the detected boxes. The following parameters are used to quantify the detection:

Horizontal coverage. Horizontal coverage starts from the observation that a marker box can be partially "covered" by one or more detection boxes. Thus, for each marker box, a horizontal coverage factor is calculated, which can vary between 0 s, and 1. Based on this factor, the marker box is classified as:

Real positive – marked boxes that are covered more than 80% (0.8) by detection boxes. These are the boxes that have the highest probability of intelligible text extraction.

Partial positive – annotated boxes that are covered between 50% (0.5) to 80% (0.8). The probability that some text can be extracted from these boxes is small, but partial text fragments can still be obtained.

Useless positives – annotated boxers covered less than 30% by detection boxes. Extraction of text from these boxes is very unlikely.

Images in our dataset have more than one text line and therefore, each image will have an average coverage value.

Split – a measurement of how fragmented the detection of each marker box is. Each image will have an average split – the larger the split value, the more “fragmented” the detection process. As a rule, CTPN has a lower split than EAST and CRAFT since the post-processing of the former is geared towards full lines of text.

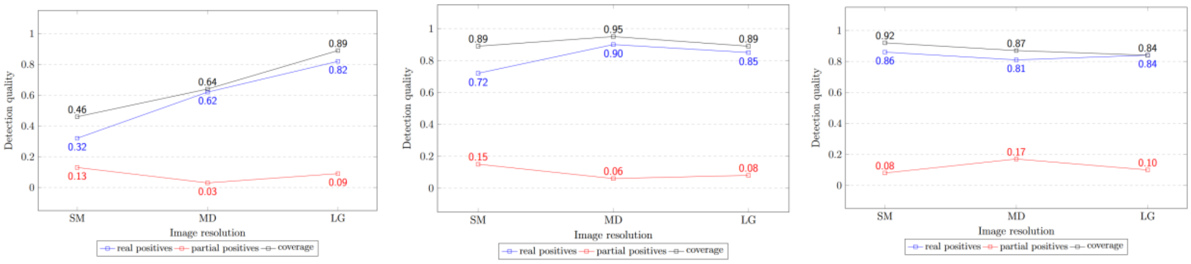
Total OCR represents the number of characters detected by extracting the text boxes and applying Optical Character Recognition using the tesseract package [21].

4. Results

4.1. The importance of resolution of the images/videoframes

Variations of the main testing parameters with the image resolution can be seen in Fig. 1 a, b and c. The most sensitive to the resolution decrease was CTPN that was almost unusable for small resolution images (320x200). Medium resolution (640x480) will lead to a significant increase in the coverage and real positives number, but CTPN cannot really work until we reach the 1280x720 resolution.

EAST and CRAFT are less sensitive to image dimensions, CRAFT seemingly being even better at lower resolutions. We have to mention that the subsequent OCR process is also extremely dependent of the image resolution and the lower resolution images (SM and MD) will not be usable for OCR anyways.



a) CTPN b) EAST c) CRAFT

Fig. 1 Influence of image resolution on the detection process

4.2. Detection results

Detection results can be seen in Table 1. The coverage does not differ much (CRAFT being slightly lower) and the real positives and partial positives are also similar. That shows that the detection process for high resolution images is almost identically successful. However, the OCR extraction process is markedly better with CTPN. The post-processing of CTPN will yield much larger text boxes, almost complete lines, while EAST and CRAFT remain at or near word level. This will influence the OCR quality since single word OCR will not benefit from Tesseract language models [20].

Parameter Model	Real positives	Partial positives	Split	Coverage	OCR Characters
EAST	0.85	0.08	5.64	0.89	15.60
CRAFT	0.84	0.1	5.03	0.84	18.10
CTPN	0.82	0.09	0.77	0.89	67.61

Table 1. Detection results for high resolution images

5. Conclusions and future developments.

Scene text detection techniques can be used for text extraction from broadcast video materials. Image resolution is important, both for CTPN and for the subsequent OCR process. Of all the three methods tested, CTPN performed best. CTPN is specially designed to extract horizontal text lines and the recurrent neural network helps in extending detection boxes over multiple words. CTPN has a real problem with low resolution images, but low-resolution images are a real problem for OCR also.

EAST and CRAFT, although more performant in actual detection of text areas, being significantly better at lower resolutions, will output mostly word-based detection boxes, which makes the OCD process harder and therefore the results are much less accurate.

CRAFT and EAST are more advanced models but the performance for small resolution and for angled / curved text are not important for our purpose.

The extracted text may differ slightly from one video frame to another. Future research is needed to use the multiple appearance of the same text, on multiple frames, for correction purposes.

References

[1] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10781–10790, 2020.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[3] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430, 2021.

[4] S. Lefèvre, D. Vasquez, and C. Laugier. A survey on motion prediction and risk assessment for intelligent vehicles. ROBOMECH journal, 1(1):1–14, 2014.

[5] G. Welch, G. Bishop, et al. An introduction to the kalman filter. 1995.

[6] Zhou, K., Qiao, Y., & Xiang, T. (2018). Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward. *ArXiv [arXiv:1801.00054v3]*.

[7] Fajtl, J., Sokeh, H. S., Argyriou, V., Monekosso, D., & Remagnino, P. (2018). Summarizing Videos with Attention. *CoRR*, *abs/1812.01969*. Retrieved from http://arxiv.org/abs/1812.01969

[8] Jadon, S., & Jasim, M. (2019). Video Summarization using Keyframe Extraction and Video Skimming. *CoRR*, *abs/1910.04792*. Retrieved from http://arxiv.org/abs/1910.04792

[9] Piergiovanni, A. J., & Ryoo, M. S. (2018). Representation Flow for Action Recognition. *CoRR*, *abs/1810.01455*. Retrieved from http://arxiv.org/abs/1810.0145

[10] Ma, C.-Y., Chen, M.-H., Kira, Z., & AlRegib, G. (2017). TS-LSTM and Temporal-Inception: Exploiting Spatiotemporal Dynamics for Activity Recognition. *CoRR*, *abs/1703.10667*. Retrieved from http://arxiv.org/abs/1703.10667

[11] Wang, C., Tang, Y., Ma, X., Wu, A., Popuri, S., Okhonko, D., & Pino, J. (2022). fairseq S2T: Fast Speech-to-Text Modeling with fairseq. *ArXiv [2010.05171]*. Retrieved from http://arxiv.org/abs/2010.05171

[12] Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., … Wu, Y. (2019). Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. *ArXiv [arXiv:1806.04558]*. Retrieved from http://arxiv.org/abs/1806.04558

[13] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In Thirty-First AAAI Conference on Artificial Intelligence, 2017

[14] Zheng Zhang, Wei Shen, Cong Yao, and Xiang Bai. Symmetry-based text line detection in natural scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2558–2567, 2015.

[15] Tian, Zhi & Huang, Weilin & Tong, He & He, Pan & Qiao, Yu. (2016). Detecting Text in Natural Image with Connectionist Text Proposal Network. 9912. 56-72. 10.1007/978-3-319-46484-8_4.

[16] Zhou, Xinyu & Yao, Cong & Wen, He & Wang, Yuzhi & Zhou, Shuchang & He, Weiran & Liang, Jiajun. (2017). EAST: An Efficient and Accurate Scene Text Detector.

[17] [Baek, Youngmin & Lee, Bado & Han, Dongyoon & Yun, Sangdoo & Lee, Hwalsuk. (2019). Character Region Awareness for Text Detection. 9357-9366. 10.1109/CVPR.2019.00959.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15). MIT Press, Cambridge, MA, USA, 91–99.

[19] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 234–241. Springer, 2015

[20] Mubeen, Suraya & Brahmani, Jally & Kalyan, Datha & Jagirdar, Ayesha & Kumar, A. (2022). Optical Character Recognition Using Tesseract. International Journal for Research in Applied Science and Engineering Technology. 10. 672-675. 10.22214/ijraset.2022.47414.

[21] Ooms J (2023). *tesseract: Open Source OCR Engine*. https://docs.ropensci.org/tesseract/ (website) https://github.com/ropensci/tesseract (devel).

[22] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[23] D. Karatzas *et al.*, "ICDAR 2015 competition on Robust Reading," *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, 2015, pp. 1156-1160, doi: 10.1109/ICDAR.2015.7333942.

[24] N. Nayef *et al.*, "ICDAR2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification - RRC-MLT," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, Japan, 2017, pp. 1454-1459, doi: 10.1109/ICDAR.2017.237.

[25] Ma, Jianqi & Liang, Zhetong & Zhang, Lei. (2022). A Text Attention Network for Spatial Deformation Robust Scene Text Image Super-resolution. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[26] Zhang, S., Zhu, X., Yang, C., Wang, H., & Yin, X. (2021). Adaptive Boundary Proposal Network for Arbitrary Shape Text Detection. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 1285-1294.

[27] Wu, J., Liu, Q., Jiang, Y., Bai, S., Yuille, A.L., & Bai, X. (2022). In Defense of Online Models for Video Instance Segmentation. European Conference on Computer Vision.

[28] Uijlings, Jasper & Sande, K. & Gevers, T. & Smeulders, A.W.M.. (2013). Selective Search for Object Recognition. International Journal of Computer Vision. 104. 154-171. 10.1007/s11263-013-0620-5.

[29] Zitnick CL, Dollár P. Edge Boxes: Locating Object Proposals from Edges. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, eds. Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I. Springer; 2014:391-405.