Autonomous Character Region Score Fusion for Word Detection in Camera-captured Handwriting Documents

Sidra Hanif
sidra.hanif@temple.edu
Temple University
Philadelphia, PA, USA

Longin Jan Latecki
latecki@temple.edu
Temple University
Philadelphia, PA, USA

ABSTRACT

Word detection is considered an object detection problem. However, characters are the basic building block in words, and the presence of characters makes word detection different from general object detection problems. Character region scores identification performs consistently for handwritten text in low-contrast camera-captured images, But detecting words from characters poses a challenge because of variable character spacing in words. Nevertheless, considering the only character and ignoring a word's entirety does not cope with overlapping words in handwriting text. In our work, we propose the fusion of character region scores with word detection. Since the character level annotations are not available for handwritten text, we estimate the character region scores in a weakly supervised manner. Character region scores are estimated autonomously from the word's bounding box estimation to learn the character level information in handwriting. We propose to fuse the character region scores and images to detect words in camera-captured handwriting images. Fusion of character region score with image has a higher recall of 88.4(+1.2) and outperforms the state of the state-of-the-art object detector with 92.2(+0.4) mAP@0.5 and 64.0(+0.4) mAP@0.5:0.95. The code and trained models are shared at the link: http://github.com/sidrahhanif/KDD-DI-Word_detection-2022.

CCS CONCEPTS

• **Applied computing** → **Document metadata**; • **General and reference** → *General conference proceedings*; • **Computing methodologies** → *Matching*; *Visual inspection*; **Image representations**; *Neural networks*; *Neural networks*; *Object recognition.*

KEYWORDS

Handwritten text, Camera-captured images, Character scores, Object detection, Multi-channel input

1 INTRODUCTION

Recently, detecting and recognising a handwritten text has gained much attention from the research community. However, word detection from unconstrained low-contrast camera-captured images is still an open problem in document analysis. Word detection from handwritten text plays a crucial role in the success of subsequent applications such as word recognition or reconstruction.

In the previous work, the words are segmented in the top-down approach where lines are segmented, then in the lines, the words and characters are segmented [10]. This approach uses a variable-sized window which is not robust to variation of word size in camera-captured images. [9] used confidence scores of word hypotheses from word recognition and lexical modeling to improve word detection. But if the lexicon method perform poorly then the word detection is also effected.

Character detection shows promising results for natural image scene text detection. Nevertheless, the scene text is very different from the handwritten text. Words in natural image scene text are separated from one another in an arbitrary shape and mainly in a typed format. However, words in handwritten text often overlap with one another in adjacent lines. Furthermore, the spacing between characters in words may vary depending on the individual handwriting style. [16, 20] parameterized the word detection with the character gap, but these approaches are not effective against the overlapping words in handwritten text.

Moreover, scene text detection aims to localize the words in natural scenes from varying perspectives. However, in handwriting, the objective of word detection is to cope with different handwriting styles. Specifically, varying character/word spacing in handwritten text with uncontrolled camera conditions makes word detection challenging. In the scene text detection domain, character detection is used to localize word instances. [8] construct word detector based on characters. The character region score detector is trained on word bounding boxes. Similarly, [2] localized the individual character and linked them to a text instance. The character region scores are trained in a weakly supervised manner with synthetic [5] and real dataset. Despite the success of character region scores in scene text detection, it cannot be directly applied to handwritten text since bounding box construction from character region scores cannot handle overlapping text frequently seen in handwriting.

Previous research explore various object detection frameworks for word detection in handwriting images.

Another work [24] proposed to use a Cascade R-CNN [4] for handwriting detect. The invoice datsests is used in [24] with printed and handwritten parts. It detects the words from both handwritten and printed text. A two-stage framework is build in [22] where the first stage generates a region proposal for words and the second stage classifies the bounding box centered on a word. Also, [23] searched the word in historical handwritten documents by initializing the search using region proposals and embedding the proposals into word embedding space. These methods rely on a two-stage framework and proposal generation network. However, the presence of region pooling for region proposals in a two-stage network gives unsatisfactory results with handwriting images.

In recent years, the single-stage object detection algorithm has improved object detection accuracy and speed. In [12] a single-stage word detector [13] detect words and grade the examination automatically. In [17], [13] detect and recognize Kawi characters on copper inscriptions. These works are evaluated either on scanned images or printed text. However, the proposed framework for word

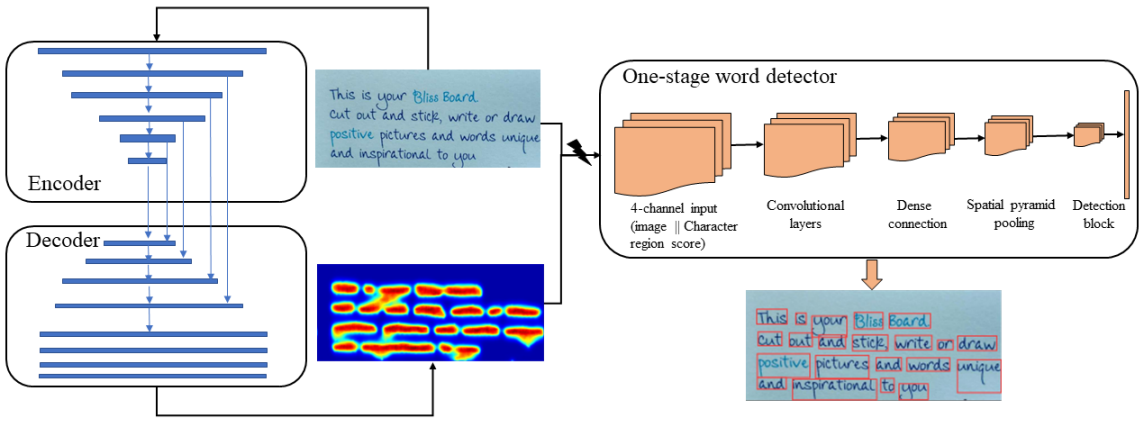


Figure 1: The overall block diagram of the convolution network comprising encoder, decoder and detection branches. Encoder-decoder pair is used to learn character region scores and detection branch learn the multi-channel word detection for fused character region scores and handwriting image. The break symbol between encoder-decoder branch and detection branch shows the autonomy of both branches.

detection is evaluated on more realistic low-contrast camera captures images.

The proposed word detector has the following contributions: 1) We explore the character region score for word detection in handwriting images. 2) Fuse the character region, affinity score, and input image for multi-channel word detection. 3) Our work is designed for low-contrast camera-captured handwriting images. 4) Our proposed character region score and input fusion outperform the state-of-the-art object detector for word detection in handwritten text.

2 METHODOLOGY

In our work, we propose to fuse character region and affinity score with the input image to determine the word localization for camera-captured handwritten text. The character region and affinity scores give information about the character's existence and probability of character belonging to the same word, respectively. Characters are fundamental building blocks for camera-capturing handwritten images. However, the handwritten text lacks character annotation for words. So, we adapt the character region score from [2]. In [2], an encoder-decoder pair is trained for character region score with 80k synthetic images having character bounding box annotations. Encoder in character region score network consist of VGG-16 [18] backbone and decoder with skip connections similar to U-Net [15]. The encoder-decoder pair learns character region and affinity scores for handwritten text. The labels for characters in the handwritten text are generated in a weakly supervised manner. The predicted character bounding boxes from handwritten text and ground-truth character bounding boxes from synthetic datasets [5] are used to train the encoder-decoder pair, which predicts character region scores for handwriting images. The loss function for character region score and affinity map is given in eq. 1.

$$L = \sum_{p} S_c(p) \cdot \left(\left\| S_r(p) - S_r^*(p) \right\|_2^2 + \left\| S_a(p) - S_a^*(p) \right\|_2^2 \right) \quad (1)$$

where $S_r^*(p)$ and $S_a^*(p)$ denote the pseudo-ground truth region score and affinity map, respectively, and $S_r(p)$ and $S_a(p)$ denote the predicted region score and affinity score, respectively.

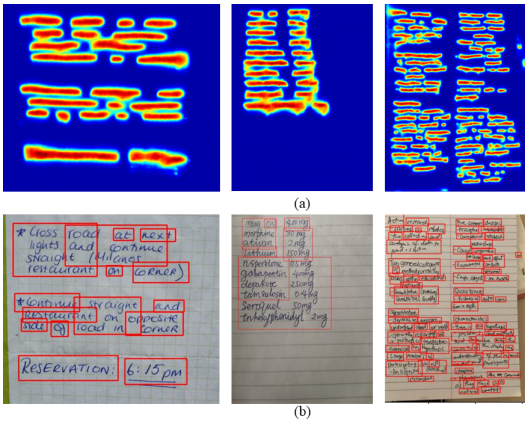


Figure 2: (a) character region + affinity score (b)word detection based on (a)

The two fundamental challenges in designing a word detection algorithm for practical application are the diversity in handwriting styles and the low-contrast of camera-captured handwriting images. Fig. 2(a) shows the normalized sum of character region scores and affinity scores for GNHK handwriting datasets [11]. The character map predicted by weakly supervised learning is a good indicator of the character's presence. However, Fig. 2(b) shows that the deterministic method to construct bounding boxes on these character region scores [2] is not able to detect the word for handwritten text. Therefore, in our work, we propose to train a multi-channel object detector with these character region scores described in the Sec 2.1.

Method	mAP@0.5	mAP@0.5:0.95
Two-stage object detector [14]	78.0	56.5
Character region score [2]	60.3	56.5

Table 1: The detection accuracy for two-stage object detector [14] and Character region score [2]

2.1 Multi-channel word detector

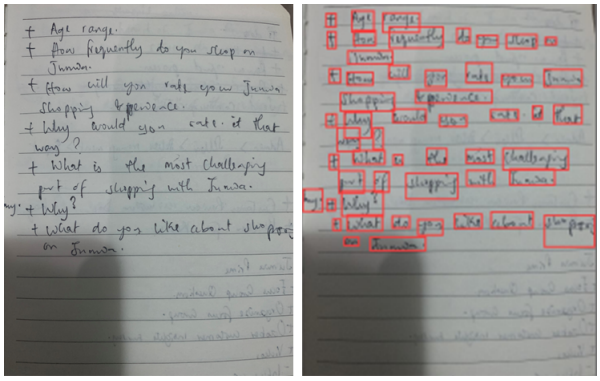
Though character region scores are a good indicator for words, as shown in 1(a), the bounding box estimation on character region scores cannot handle a handwriting style. Fig. 1(b) highlighted that character region scores alone are insufficient to perform well for overlapping words in adjacent lines.

To keep the advantages offered by the character region score and to prevent the problem shown in 1(b), we propose to fuse the character region and affinity scores with the input image to learn word detection from handwritten text. Word detection from handwritten text is a single-class detection problem with multiple steams of input information. The detection network consists of convolutional layers with dense connections and pyramid pooling. It is a multi-channel word detection framework for handwritten text.

In previous research, multi-channel input is utilized for object detection in satellite images [19] and outdoor scenes [21]. However, these researches used additional information bands such as infrared or depth maps, which are readily available with datasets. However, we propose learning the character region score in weakly supervised manner without any character level annotations. We fuse the character region score and affinity scores for word detection into an object detection framework [3, 6, 14]. Single-stage objects detector [14] performs better on word detection than a two-stage object detectors [3]. The work in [1] used region score to detect word. It estimates the heat map of words and generates the region proposal on the estimated heat map. The heat map and regional proposals are fed into the filter network to learn if the region proposal envelops a word. This work is very different from our proposed approach. First, it does not learn any information about the character regions and is limited to estimating word region scores. Secondly, their region proposal generation is also limited to a heat map of words. However, in our work, we propose combining the cons of both handwriting image and its character region scores. The character region scores break the word entity into basic building blocks (character), and the affinity map provides the probability of them belonging to the same word. Our proposed method is independent of vocabulary and considers words' character region and affinity scores in handwritten text. Therefore, it can be easily scalable to any document type and vocabulary.

3 EXPERIMENTS

In the next section, we briefly describe the low-contrast camera-captured GNHK dataset [11] and evaluate the performance of word detection on it.



(a) (b)

Figure 3: Sample images from GNHK datasets [11] with bounding boxes

Input	Prec	Recall	mAP@0.5	mAP@0.5:0.95
Word detector [3]				
image	90.1	87.2	91.8	63.6
Multi-channel word detector with character region scores [2, 3]				
image || RS	**90.3 (+0.2)**	87.7	91.8	64.0
image || (RS + AS)	89.8	**88.4 (+1.2)**	**92.2 (+0.4)**	**64.0 (+0.4)**

Table 2: The detection statistics for image, character region and affinity score on multi-channel object detector network [3]. In the table, image stands for 3-channel handwriting image, *RS* stands for character region score and *AS* stands for character affinity scores.

Handwriting size	Prec	Recall	mAP@0.5	mAP@0.5:0.95
image				
Large	82.0	89.6	89.5	64.8
Med	91.5	**92.2**	94.1	**68.0**
Small	**89.5**	86.1	91.1	**62.2**
image || (RS + AS)				
Large	**84.9**	**90.7**	**91.1**	**65.7**
Med	**92.0**	91.7	**94.4**	67.1
Small	89.2	**86.9**	**91.3**	62.1

Table 3: The detection statistics for image and character region scores as input for the object detector network [3] for large, medium, and small handwriting sizes

3.1 Datasets

The images in GNHK datasets are sourced from Europe, North America, Africa, and Asia. It is a diverse dataset as penmanship varies in different parts of the world. The dataset consists of 687 images containing different types of handwritten text, such as shopping lists, sticky, and diaries notes. Mobile phone cameras captured images under unconstrained settings. Therefore, it may contain

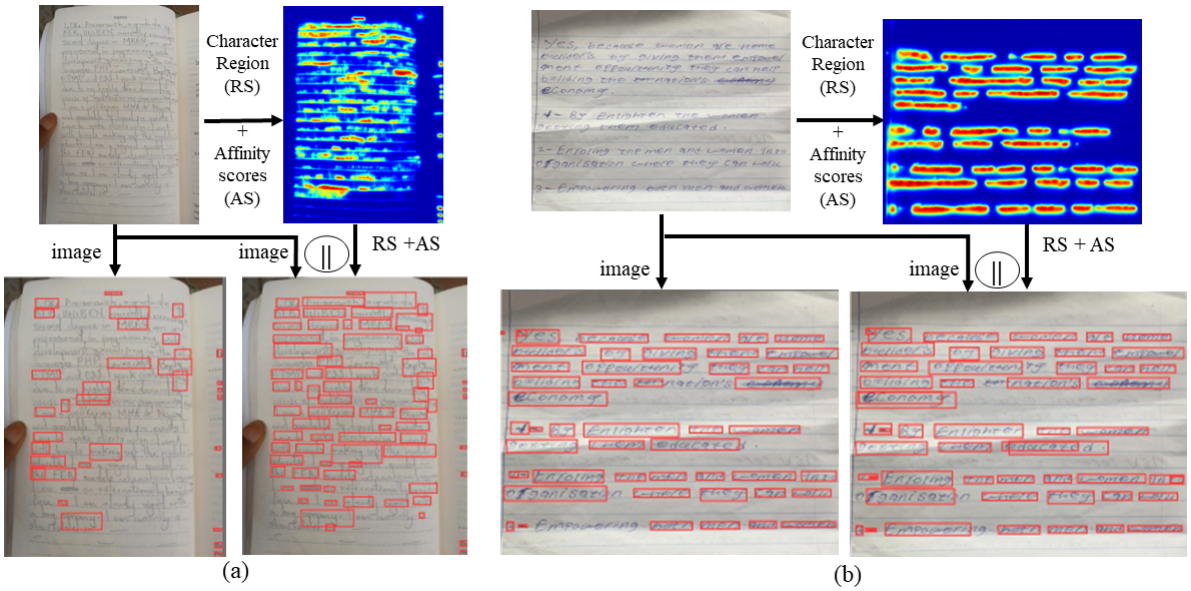


Figure 4: The qualitative results for word detection for low-contrast camera -captured handwritten text

shadows from mobile devices, and handwriting has very low contrast with the background, as visible in Fig. 3. There is a corresponding JSON file for each handwritten-text image containing the annotations of words in the images. Fig. 3(a) shows the image of handwritten text, and Fig. 3(b) shows the ground truth bounding boxes of words for each word in handwritten text.

3.2 Results and discussion

In our work, a multi-channel object detector is proposed for word detection for GNHK datasets [11]. In [11] the baseline is established with two-stage word detector [7, 14]. Table. 1 shows the performance of the two-stage object detector and character region score. It can be seen in Table. 1 that the object detector and character region score have the same mAP@0.5:0.95 accuracy (0.565), however for mAP@0.5 two-stage object detector (0.780) have higher accuracy than character region score (0.603). The low accuracy of the character region score is because of a deterministic bounding box estimation for words [2]. For overlapping text in handwriting, the deterministic estimation does not work for handwriting text, as shown in Fig. 1.

In our work, we propose to perform multi-channel word detection leveraging character region and affinity score along with handwriting text image. The single-stage object detector outperforms the two-stage detector by a large margin. Table. 2 shows the quantitative results for image, character region (*RS*), and affinity scores (*AS*) for the word detector network [3]. In Table. 2, we can see that the multi-channel information consisting of handwriting image, character region scores, and affinity map outperforms the word detector without character region and affinity scores [3]. The multi-channel word detector increases the recall by 1.2%, mAP@0.5, and mAP@0.5:0.95 by 0.4%. Therefore, additional information from weakly supervised character region scores beats the state-of-the-art word detector.

Fig. 4 show the qualitative results for challenging examples with low-contrast and overlapping words. Nevertheless, our method still produces a reasonable detection compared to state-of-the-art word detector [3]. In that case, the character region and affinity score provide the word clues shown in character scores map in Fig. 4.

We also validated in Table. 3 that character region and affinity scores gives better performance for large and medium-size words. It gives approximately 1% improvement in mAP@0.5 with the single-stage word detector. On the other hand mAP@0.5:0.95 declines for small handwriting text as the quality of character region scores declines for very small word sizes. In Table 3, we illustrate the results for large, medium, and small handwritten text. Character region and affinity scores for large and medium handwriting outperform input images for word detection. Character region and affinity scores as word clues with RGB images of handwritten text improves the word detection accuracy on camera-captured handwriting images.

4 CONCLUSION

In our work, we propose the multi-channel word detection that leverage the character region scores trained in a weakly supervised manner for handwritten text. The character region and affinity scores improve the qualitative and quantitative results. The state-of-the-art word detector struggles to detect words in low-contrast camera-captured handwriting images. However, the proposed multi-channel word detector performs well also on challenging examples.

REFERENCES

[1] Gregory Axler and Lior Wolf. 2018. Toward a dataset-agnostic word segmentation method. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2635–2639.
[2] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9365–9374.
[3] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*

(2020).

[4] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6154–6162.

[5] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic Data for Text Localisation in Natural Images. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[6] Sidra Hanif, Chao Li, Anis Alazzawe, and Longin Jan Latecki. 2019. Image Retrieval with Similar Object Detection and Local Similarity to Detected Objects. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 42–55.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[8] Han Hu, Chengquan Zhang, Yuxuan Luo, Yuzhuo Wang, Junyu Han, and Errui Ding. 2017. Wordsup: Exploiting word annotations for character based text detection. In *Proceedings of the IEEE international conference on computer vision*. 4940–4949.

[9] Sana Khamekhem Jemni, Yousri Kessentini, and Slim Kanoun. 2019. Out of vocabulary word detection and recovery in Arabic handwritten text recognition. *Pattern Recognition* 93 (2019), 507–520.

[10] Rajiv Kumar and Amardeep Singh. 2010. Detection and segmentation of lines and words in Gurmukhi handwritten text. In *2010 IEEE 2nd International Advance Computing Conference (IACC)*. IEEE, 353–356.

[11] Alex WC Lee, Jonathan Chung, and Marco Lee. 2021. GNHK: A Dataset for English Handwriting in the Wild. In *International Conference on Document Analysis and Recognition*. Springer, 399–412.

[12] Mingliang Lu, Weili Zhou, and Ruijie Ji. 2021. Automatic Scoring System for Handwritten Examination Papers Based on YOLO Algorithm. In *Journal of Physics: Conference Series*, Vol. 2026. IOP Publishing, 012030.

[13] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[16] Jewoong Ryu, Hyung Il Koo, and Nam Ik Cho. 2015. Word segmentation method for handwritten documents based on structured learning. *IEEE Signal Processing Letters* 22, 8 (2015), 1161–1165.

[17] Rachmat Santoso, Yoyon Kusnendar Suprapto, and Eko Mulyanto Yuniarno. 2020. Kawi Character Recognition on Copper Inscription Using YOLO Object Detection. In *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*. IEEE, 343–348.

[18] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition.

[19] Kazuki Uehara, Hidenori Sakanashi, Hirokazu Nosato, Masahiro Murakawa, Hiroki Miyamoto, and Ryosuke Nakamura. 2017. Object detection of satellite images using multi-channel higher-order local autocorrelation. In *2017 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE, 1339–1344.

[20] Tamas Varga and Horst Bunke. 2005. Tree structure for word extraction from handwritten text lines. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*. IEEE, 352–356.

[21] Li Wang, Ruifeng Li, Hezi Shi, Jingwen Sun, Lijun Zhao, Hock Soon Seah, Chee Kwang Quah, and Budianto Tandianus. 2019. Multi-channel convolutional neural network based 3D object detection for indoor robot environmental perception. *Sensors* 19, 4 (2019), 893.

[22] Tomas Wilkinson and Anders Brun. 2015. A novel word segmentation method based on object detection and deep learning. In *International Symposium on Visual Computing*. Springer, 231–240.

[23] Tomas Wilkinson, Jonas Lindstrom, and Anders Brun. 2017. Neural Ctrl-F: segmentation-free query-by-string word spotting in handwritten manuscript collections. In *Proceedings of the IEEE International Conference on Computer Vision*. 4433–4442.

[24] Yuli Wu, Yucheng Hu, and Suting Miao. 2021. Object Detection Based Handwriting Localization. In *International Conference on Document Analysis and Recognition*. Springer, 225–239.