SCIENCE CHINA
Information Sciences

June 2022, Vol. 65 160106:1–160106:2
https://doi.org/10.1007/s11432-021-3447-9

• NEWS & VIEWS •
Special Focus on Deep Learning for Computer Vision

Comprehensive benchmark datasets for Amharic scene text detection and recognition

Wondimu DIKUBAB[1], Dingkang LIANG[2], Minghui LIAO[1] & Xiang BAI[2*]

[1]*School of Electronic Information and Communication, Huazhong University of Science and Technology, Wuhan* 430074*, China;*
[2]*School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan* 430074*, China*

Received 18 August 2021/Accepted 15 January 2022/Published online 26 April 2022

Citation Dikubab W, Liang D K, Liao M H, et al. Comprehensive benchmark datasets for Amharic scene text detection and recognition. Sci China Inf Sci, 2022, 65(6): 160106, https://doi.org/10.1007/s11432-021-3447-9

Ethiopic/Amharic script is one of the oldest African writing systems, which serves at least 23 languages (e.g., Amharic, Tigrinya) in East Africa for more than 120 million people.

The Amharic writing system, Abugida, has 282 syllables, 15 punctuation marks, and 20 numerals. The Amharic syllabic matrix is derived from 34 base graphemes/consonants by adding up to 12 appropriate diacritics or vocalic markers to the characters. Unlike Latin alphabets, each Amharic character constitutes conjugation of consonants and vowels as a single syllable. The syllables with a common consonant or vocalic markers are likely to be visually similar and challenge text recognition tasks. Moreover, visual complexity, poor image quality, and intermittent text appearance cause failures of Amharic scene text detection and recognition.

Recently, detecting and recognizing Latin and Chinese characters in natural scenes have progressed tremendously. However, the discussion on Amharic script detection and recognition is insufficient mainly due to the lack of public datasets. Recently, Addis et al. [1] presented the first private dataset for Ethiopic/Amharic scene text recognition, which contains 2500 text images and lacks robustness.

In this study, we presented the first comprehensive public datasets for Amharic script detection and recognition in the natural scene to address the abovementioned problem.

Datasets for text detection. We construct Amharic scene text detection datasets: the Amharic real-world scene text (HUST-ART) and the Amharic SynthText (HUST-AST).

HUST-ART contains 2200 natural scene images: 1500 for the training and 700 for the testing. Specifically, it includes 11254 cropped text instances. The HUST-ART pictures are collected across Ethiopia by mobile phones, professional cameras, and a few from the Internet. This dataset comprises diversified scenes, including signboards, posters, indoors, and streets. We use quadrilateral coordinates to represent the ground truth of the text instance: $G = [x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4]$, and word regions are categorized as easy or difficult. The easy regions will be used for the recognition task. HUST-ART is robust and challenging because it contains multi-orientation text, small and large scale text, various illumination, and complex backgrounds, as shown in Figure 1(a). Moreover, HUST-ART has more text instances than the popular text detection dataset [2].

HUST-AST contains 75904 images with 829394 cropped synthetic text instances, and it is generated by SynthText [3] tool. The text sample is rendered upon natural images with random transformations and effects according to the local surface adaptation, as shown in Figure 1(b).

We implemented SOTA methods DCLNet [4], DB [5] to evaluate their performance on the proposed datasets. Firstly, we use HUST-AST to pretrain the models, and then, we finetune the models on HUST-ART. Eventually, we select their final epoch for evaluation. As illustrated in Figure 1(e), we measure text detection performance by precision (P), recall (R), and F1-measure (F). DCLNet [4] achieves the best F1-measure of 84.86%. Yet, we can see room for further improvement in the future.

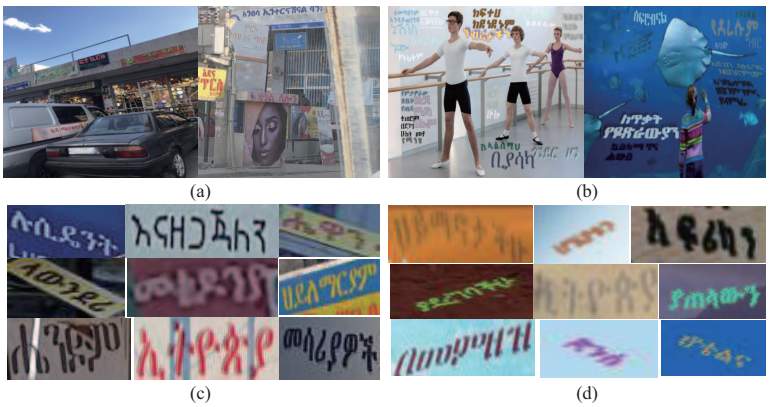
Datasets for text recognition. Besides cropped word images from HUST-ART and HUST-AST datasets, we constructed two text recognition datasets of real-world and synthetic text, ABE and Tana, respectively.

ABE contains 12839 real-word text images: 7621 for training and 5218 for testing. It is obtained by phone camera from Ethiopia and some from the Internet. The samples are shown in Figure 1(c). Compared with some previous datasets [1, 2], the proposed ABE contains more text images.

Tana consists of 2851778 synthetic word images, including the 829394 HUST-AST cropped text images. Besides HUST-AST, the text images are generated: applying random color, font rendering, blurring randomly, skewing the text arbitrarily, and blending with real-world images, as shown in Figure 1(d).

We adopt SOTA methods MASTER [6] and SATRN [7] to evaluate their Amharic scene text recognition performance on the proposed datasets ABE and HUST-ART. We use the Tana dataset as the training data, the union of ABE and HUST-ART training sets as validation data, and the ABE and HUST-ART testing sets as evaluation data. We

* Corresponding author (email: xbai@hust.edu.cn)



(a) (b) (c) (d)

Network	E2E (%)	P (%)	R (%)	F (%)
DB [5]	-	95.31	74.62	83.71
DCLNet [4]	-	93.82	77.47	84.86
PAN++ [8]	30.31	93.38	30.06	45.48
MTSV3 [9]	71.23	88.31	80.82	84.40

(e)

Model	Average accuracy (%)	
	ABE	HUST-ART
SATRN [7]	85.66	87.54
MASTER [6]	86.50	87.70

(f)

Figure 1 (Color online) Images from (a) HUST-ART, (b) HUST-AST, (c) ABE, (d) Tana; (e) text detection and spotting results; (f) text evaluation recognition results. E2E, P, R, and F refer to the end-to-end recognition rate, precision, recall, and F1-measure, respectively.

measure the average accuracy rate by the success rate of word predictions per image. We only evaluate 302 character classes of syllables and Amharic numerals.

As the evaluation results in Figure 1(f) show, MASTER [6] outperforms both on ABE and HUST-ART datasets archiving 86.50% and 87.70%, respectively. The common causes of scene text recognition failure can be long text, blurred and distorted images, and uncommon fonts. Additionally, the Amharic scene text recognition failure can be caused by visual similarity among the characters that share a common consonant, the same kind of vocalic markers, or similar graphical structure. Therefore, the recognition of Amharic scripts requires more robust methods that can handle the visual similarity among the syllables.

End-to-end text spotting. We train PAN++ [8] and Mask TextSpotter v3 (MTSV3) [9] on joint HUST-AST and HUST-ART to evaluate their end-to-end text detection and recognition performance. We evaluate text spotting performance by P, R, F, and end-to-end recognition accuracy (E2E). The end-to-end text spotting performance evaluation results are presented in Figure 1(e). MTSV3 [9] outperforms PAN++ [8] achieving 71.23% end-to-end recognition accuracy and 84.4% F1-measure.

Generally, the end-to-end text detection and recognition failure can be caused by inaccurate detection results, complex background with text-like patterns, the presence of irregular fancy text, low-resolution or blurred text, and false recognition results. Moreover, the evaluation results suggest that end-to-end Amharic text spotting demands more robust models.

Conclusion. In this study, we presented the first comprehensive public datasets named HUST-ART, HUST-AST, ABE, and Tana for Amharic script detection and recognition in the natural scene. We have also conducted extensive experiments to evaluate the performance of the state-of-the-art methods in detecting and recognizing Amharic scene text on our datasets. The evaluation results demonstrate the robustness of our datasets for benchmarking and their potential of promoting the development of robust Amharic script detection and recognition algorithms. Consequently, the outcome will benefit people in East Africa, including diplomats from several countries and international communities.

According to the quantitative results, we observed that the text detection and recognition performance demand a new attempt to design robust models that can address a unique feature of the Amharic script. We will dedicate ourselves to investigating the challenges and improving the detection and recognition performance in the future.

Access methods. The datasets and more detailed information can be obtained from the website[1)].

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2018YFB1004600) and National Natural Science Foundation of China (Grant No. 61733007).

References

1 Addis D, Liu C-M, Ta V-D. Ethiopic natural scene text recognition using deep learning approaches. In: Proceedings of International Conference on Advances of Science and Technology. Berlin: Springer, 2019. 502–511
2 Karatzas D, Gomez-Bigorda L, Nicolaou A, et al. ICDAR 2015 competition on robust reading. In: Proceedings of International Conference on Document Analysis and Recognition, 2015. 1156–1160
3 Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2315–2324
4 Bi Y, Hu Z. Disentangled contour learning for quadrilateral text detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021. 909–918
5 Liao M, Wan Z, Yao C, et al. Real-time scene text detection with differentiable binarization. In: Proceedings of American Association for Artificial Intelligence, 2020. 34: 11474–11481
6 Lu N, Yu W, Qi X, et al. MASTER: multi-aspect non-local network for scene text recognition. Pattern Recognition, 2021, 117: 107980
7 Lee J, Park S, Baek J, et al. On recognizing texts of arbitrary shapes with 2D self-attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020. 546–547
8 Wang W, Xie E, Li X, et al. PAN++: towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. IEEE Trans Pattern Anal Mach Intell, 2021. doi: 10.1109/TPAMI.2021.3077555
9 Liao M, Pang G, Huang J, et al. Mask TextSpotter v3: segmentation proposal network for robust scene text spotting. In: Proceedings of the European Conference on Computer Vision (ECCV), 2020

1) https://dk-liang.github.io/HUST-ASTD/.