Supplementary Document for "*Comprehensive Benchmark Datasets for Amharic Scene Text Detection and Recognition* "

1. Introduction

Text is among the most significant invention of humankind that plays an essential role in our daily communications. Texts in natural scene images contain sophisticated semantic information, which is useful to analyze the related environment. Text identification can be considered the primary building block to extract valuable information from the natural scene. The principal aim of text detection and text localization is to generate a bounding box around the text that emerged in an image or video. At the same time, the purpose of the scene text recognition task is to transform detected text regions into characters or words. Thus, scene text detection and recognition methods detect, locate and transform text in complex scene images into words.

At present, text detection and recognition in scene images have become a popular research area due to hands-on applications such as robotics, analysis of image or video contents, visual search, intelligent transportation systems, and industrial automation. Recently, detecting and recognizing Latin and Chinese characters in natural scenes have achieved remarkable progress. However, the research on Amharic scripts detection and recognition is insufficient mainly due to the lack of public datasets. Thus, the efforts to develop Amharic scene text detection and recognition methods are not enough.

Amharic serves as an official working language of the Federal Democratic Republic of Ethiopia. It is the second-largest spoken Semitic language family next to Arabic globally. It is also used in Eritrea, Djibouti, Sudan, Somali Land, USA, Israel, Sweden as a business and second language. Amharic became one of the six non-English languages in Washington DC in the Language Access Act of 2004, allowing government services and education in Amharic[1].

The Amharic/Ethiopic script is adapted from the Ethiopic syllabary, used for Ge'ez, and developed in Ethiopia sometime during the 4th-century[2]. The Ethiopic script has been adapted to write at least 20 different languages in Ethiopia, such as Tigrinya, Argobba, Awngi, Chaha, Harari, Sebat Bet, etc. It has conventionally been used for Tigrinya, Tigre, and Bilen in Eritrea.

The Amharic writing system is called Fidäl, Ethiopic or Abugida interchangeably. It has 282 syllables, 15 punctuation marks, and 20 numerals. The syllables of Abugida are derived from 34 base graphemes/consonants, which transformed into 248 syllabic symbols by adding appropriate diacritics or vocalic markers to the characters, as illustrated in Figure 1.

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th
Pron.	e/ä	u	i	a	ē	ə	o	oue	ui	wu	wua	uē
1	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ					
2	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ				ሏ	
3	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ				ሗ	
4	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ				ሟ	
5	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ				ሧ	
6	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ				ሯ	
7	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ				ሷ	
8	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ				ሿ	
9	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቈ	ቊ	ቋ	ቌ	ቍ
10	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ				ቧ	
11	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ				ቯ	
12	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ				ቷ	
13	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ				ቿ	
14	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	ኈ	ኊ	ኋ	ኋ	ኌ
15	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ				ኗ	
16	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ				ኟ	
17	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ					
18	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኰ	ኲ	ኳ	ኳ	ኴ
19	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ					
20	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ					
21	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ					
22	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ				ዟ	
23	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ				ዧ	
24	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ					
25	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ				ዷ	
26	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ				ጇ	
27	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ	ጐ	ጒ	ጓ	ጓ	ጔ
28	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ				ጧ	
29	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ				ጯ	
30	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ				ጿ	
31	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ				ፏ	
32	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ					
33	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ				ፏ	
34	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ				ፗ	

Figure 1: Amharic Syllabary (ፊደል ገበታ) Matrix.

The Ethiopic writing system is a featural syllabary, i.e., each Amharic character constitutes conjugation of consonants and vowels as a single syllable. The first 34 by seven Amharic Syllabary matrix is the core syllables. The others are known as Labiovelars and Labialized syllables. Labiovelar syllables (columns 8,9,10,12) are pronounced with the rounding of the lips, which are special Amharic characters. Labialized syllables (column 11) involve the lips while the remainder of the oral cavity produces consonant sound plus “wa” vocal. The Amharic character “ኧ” is not presented in the matrix but is used as the “አ” family. As illustrated in Figure 2, every Amharic character pronunciation represents a union of consonant and vowel sounds as an individual syllable. The pronunciation of each row is almost uniform, with few exceptions. The exceptions are the syllables

"H" families ሀ, ሐ and ኀ and "A" families አ ኣand ዐ. The exceptional first column is pronounced with a vowel "Ä," while the rest first column is pronounced with a vowel " E."

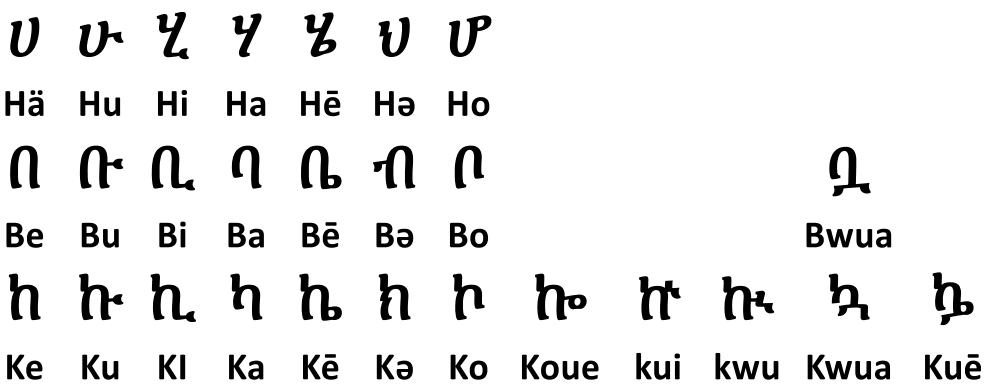


Figure 2: Example of Amharic syllabic pronunciation.

The Ethiopic writing system is univocal, and combining characters is not common. Unlike Latin, there is no upper and lower case distinction for Amharic characters. The Amharic script is written from left to right in horizontal lines. As shown in Fig 3, the Amharic writing system has 20 numeric digits. Unlike the Western text number system, it is additive; for instance, 267 is expressed as ፪፻፷፯.

፩	፪	፫	፬	፭	፮	፯	፰	፱	፲
1	2	3	4	5	6	7	8	9	10
፳	፴	፵	፶	፷	፸	፹	፺	፻	፼
20	30	40	50	60	70	80	90	100	10000

Figure 3: Ethiopic numeric digits

However, the Abugida does not have a symbol for zero, negative, basic mathematical operators' signs and decimals. The current Amharic writing system generally uses spaces to separate words; however, sometimes, the Ethiopic word-space character (፡) is still used in handwriting.

The unique nature of Abugida demands the datasets that fit its exhaustive featural syllabary. As we can see in figure 1, there is graphic similarity and structural correlation within characters that share a common consonant or vowel sound. The visual similarity challenges the task of Amharic script recognition. This work introduces the first comprehensive datasets for Amharic scene detection and recognition, both the real-world and SynthText datasets.

2. The Proposed Datasets

We construct the first publicly available comprehensive datasets to address the absence of extensive datasets and promote the development of robust algorithms for Ethiopic scene text detection and recognition. We develop the real-world and synthetic text datasets for text detection and recognition tasks.

2.1. Text Detection

We construct Amharic detection datasets: the Amharic real-world (HUST-ART) and the Amharic SynthText Dataset (HUST-AST).

A. HUST-ART

HUST-ART contains 2,200 natural scene images: 1,500 for the training and 700 for the testing. Specifically, it includes 11,254 cropped text instances. The HUST-ART pictures are collected across Ethiopia by mobile phone, professional cameras, and a few from the Internet. This dataset comprises diversified scenes, including signboards, posters, indoors, streets, etc.

Figure4: Sample Amharic Images from the HUST-ART Dataset

Besides Amharic, our dataset also consists of Tigrinya, a widely spoken language in Eritrea and the northern part of Ethiopia, mainly in the Tigray Region. Figure 4 shows samples of Tigrinya scene text images captured from Mekele and neighboring towns.

Figure 5: Sample Tigrinya Scene Text Images from the HUST-ART Dataset

We use quadrilateral coordinates to represent the ground truth of the text instance:

G = [x1, y1, x2, y2, x3, y3, x4, y4, text-transcription], where x1, y1, x2, y2, x3, y3, x4, y4 represent the x, y coordinates of the top left, top right, bottom right and bottom left corners of each word.

As illustrated in figure 5, word regions were categorized as either difficult or easy. The difficult words transcribed as" ###" include text in non-Amharic scripts and text that the ground-truther regarded as non-readable. The easy word regions will be cropped for the recognition task.

```
876,713,1046,717,1046,774,865,770,ሙዘሚል
1058,713,1137,714,1136,773,1055,772,ሱቅ
1655,789,1953,770,1958,830,1659,846,ሙከረም
1972,772,2115,765,2122,823,1981,829,ጁስ
1798,1026,1881,1027,1877,1063,1793,1055,በት
269,1145,306,1147,300,1182,259,1176,ናሲ
381,1153,494,1160,485,1195,371,1186,ሬስቶራንት
86,1156,157,1162,165,1183,91,1176,አልፈራም
159,1165,197,1170,201,1187,162,1186,ድንግልን
196,1171,217,1175,226,1192,203,1192,ይዤ
137,1207,209,1212,225,1258,163,1258,ማርያምን
1635,1368,1648,1354,1660,1363,1648,1379,###
1653,1353,1687,1320,1698,1330,1666,1364,###
1367,1388,1394,1388,1393,1398,1366,1396,###
1394,1389,1425,1390,1425,1400,1393,1400,###
1392,1380,1417,1381,1417,1390,1391,1388,###
```

Figure 5: Sample Scene Text Images and its annotations from the HUST-ART dataset

B. HUST-AST

HUST-AST contains 75,904 images with 829394 cropped synthetic text instances, and it is generated by SynthText[3]. The text sample is rendered upon natural images with random transformations and effects according to the local surface adaptation, as shown in Figure 6. The vocabularies are collected from various

sources (e.g., media, regulations, newspapers, etc.), and the ground truth type is similar to the HUST-ART dataset.

Figure 6: Sample SynthText Images of HUST-AST datasets

C. Evaluation of Text Detection

We implemented recent text detection methods such as DB [4], PSENET[5], PAN[6]and East [7] to evaluate the proposed datasets. All models are trained on a local machine with 4 NVIDIA TITAN V GPUs. We first train them on the HUST-AST dataset for 100k iteration and then finetune them on the HUST-ART dataset up to 1,200 epochs.

Table 3: The detection performance on the HUST-ART dataset. P, R, and F refer to the Precision, Recall, and F1-measure, respectively.

	Network	**Backbone**	**Precision**	**Recall**	**Hmean/F-Score**
1	EAST	Resnet50	79.67	79.1	79.38
2	PSENET	Resnet50	94.79	72.21	81.97
3	PAN	Resnet18	95.21	73.52	82.97
4	DB [4]	Resnet18	96.61	73.67	83.6
5	DB[4]	Resnet50	95.31	74.62	83.71

As illustrated in Table 1, DB[4] achieves the best F1-measure of 83.71\%. As we can observe in Figure 7, there is still a lot of room for improvement in the future.

Figure 7: Examples of text detection results on the benchmark dataset.

2.2 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.

A. ABE

ABE contains 12,872 real-word images: 7,621 for training and 5,251 for testing. It is obtained by phone camera from Ethiopia and some from the Internet. As illustrated in Table 2, ABE consists of more word images than the other. The samples are shown in Figure 8. a.

Table 2: The comparisons of ABE and other datasets

	Cropped Word Images	**Test set**	**Training set**
D. Addis et al.	2,500	2,500	0
ABE	12,839	5,218	7,621
HUST-ART	11,254	4039	7215

a. ABE

b. Tana

Figure 8: (a) Images from ABE. (b) Images from Tana.

B. Tana

Tana consists of 2851778 synthetic word images, including the 829394 HUST-AST cropped text images. Besides HUST-AST images, the text images are generated: random color, font rendering, projection distortions, blurring randomly, skewing the text arbitrarily, and blending with real-world images, as shown in Figure 8. b.

C. Amharic Text Recognition

We evaluate the proposed ABE and HUST-ART adopting SOTA methods such as MASTER [8] and SATRN [9] and recent methods ASTER [10], RARE [11] and CRNN [12]. 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.

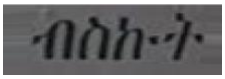
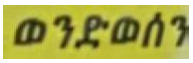
Table 3: The recognition performance on the ABE and HUST-ART datasets

Model	**Average Accuracy**	
	ABE	HUST-ART
CRNN [12]	75.91	80.26
RARE [11]	78.13	82.05
ASTER [10]	81.4	85.3
SATRN [9]	85.66	87.54
MASTER [8]	86.5 &	87.7

We measure the accuracy by the success rate of word predictions per text image. The decoder recognizes 302 character classes, including syllables and digits; however, punctuation marks are excluded. The results are presented in Table 3; MASTER outperforms both on ABE and HUST-ART datasets. However, the results unveil the demand for more robust recognition methods for the Ethiopic/Amharic scripts.

We have presented some failure cases in visuals due to visual similarities among the characters or poor image qualities such as blurred and distorted images. Prediction errors are marked by blue and red for missing characters.

Table 4: Failure cases of recognition results. Prediction errors are marked by blue and red for missing characters.

Word Images	Ground Truth	Prediction
	ብስኩት	በስኩት
	ሣሮን	ሣርን
	ሸቀጣሸቀጥ	ሽቀጣሸቀጥ
	ሀይለማሪያም	ሀይለማርያም
	የሸገር	የሸገር
	ቢራቢሮ	ቤራቢሮ
	ወንድወሰን	ወንድወሰን
	ኤንድ	አንድ
	መቄዶንያ	መቱዶንያ
	ቪዳ	ኪዳ

2.3. End-To-End Text Spotting

We train Mask TextSpotter v3 (MTSV3) [13] and PAN++[14]on joint HUST-AST and HUST-ART datasets. We implemented according to their original paper, except MTSV3 character masking. Since we use 302-character classes, which is greater than 256; thus, character masking cannot be performed. Their end-to-end text detection and recognition performance evaluation results are presented in Table 5; MTSV3 [13] outperforms PAN++ [14] in all evaluation parameters with large margins.

Table 5: End to end text spotting quantitative results of the models on the HUST-ART dataset. E2E, P, R, F and FPS refer to the End-to-End recognition rate, Precision, Recall, and F1-measure, respectively.

Method	E2E	P	R	F
PAN++[14]	30.31	93.38	30.06	45.48
MTSV3 [13]	81.71	88.31	80.82	84.40

The qualitative end-to-end detection and recognition results in Figure 8 show that MTSV3 [13] performance seems satisfactory while PAN++[14] performance is insufficient.

a. MTSV3 Results

b. PAN++ Results

Figure 8: End-to-end text spotting qualitative results examples.

Reference:

[1] "What is the D.C. Language Access Act?," no. 202, p. 20001, 2007.

[2] D. Appleyard, "Amharic," in *Encyclopedia of Language & Linguistics (Second Edition)*, Second Edi., K. Brown, Ed. Oxford: Elsevier, 2006, pp. 193–197.

[3] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic Data for Text Localisation in Natural Images," 2016.

[4] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time Scene Text Detection with Differentiable Binarization," *AAAI 2020 - 34th AAAI Conf. Artif. Intell.*, pp. 11474–11481, 2020, doi: 10.1609/aaai.v34i07.6812.

[5] W. Wang, E. Xie, and X. Li, "Shape Robust Text Detection with Progressive Scale Expansion Network †."

[6] W. Wang *et al.*, "Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 8440--8449, 2019.

[7] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 2642–2651, 2017, doi: 10.1109/CVPR.2017.283.

[8] N. Lu *et al.*, "Master: Multi-aspect non-local network for scene text recognition," *Pattern Recognit.*, vol. 117, p. 107980, 2021.

[9] J. Lee, S. Park, J. Baek, S. J. Oh, S. Kim, and H. Lee, "On Recognizing Texts of Arbitrary Shapes with 2D Self-Attention," *CoRR*, vol. abs/1910.0, 2019, [Online]. Available: http://arxiv.org/abs/1910.04396.

[10] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: An Attentional Scene Text Recognizer with Flexible Rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2035–2048, 2019, doi: 10.1109/TPAMI.2018.2848939.

[11] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4168–4176.

[12] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern*

Anal. Mach. Intell., vol. 39, no. 11, pp. 2298–2304, 2016.

[13] M. Liao, G. Pang, J. Huang, T. Hassner, and X. Bai, “Mask textspotter v3: Segmentation proposal network for robust scene text spotting,” in *European Conference on Computer Vision*, 2020, pp. 706–722.

[14] W. Wang *et al.*, “{PAN++:} Towards Efficient and Accurate End-to-End Spotting of Arbitrarily-Shaped Text,” *CoRR*, vol. abs/2105.0, 2021, [Online]. Available: https://arxiv.org/abs/2105.00405.

[15] C. K. Ch’Ng and C. S. Chan, “Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 1, pp. 935–942, 2017, doi: 10.1109/ICDAR.2017.157.

[16] L. Yuliang, J. Lianwen, Z. Shuaitao, and Z. Sheng, “Detecting Curve Text in the Wild: New Dataset and New Solution,” 2017, [Online]. Available: http://arxiv.org/abs/1712.02170.