A Historical Handwritten Dataset for Ethiopic OCR with Baseline Models and Human-level Performance

Birhanu Hailu Belay*, Isabelle Guyon+*†, Tadele Mengiste†, Bezawork Tilahun†, Macus Liwicki‖, Tesfa Tegegne†, and Romain Egele*

*LISN, Université Paris-Saclay, France + Google Brain, USA ‡ChaLearn,USA
†Bahir Dar University, Ethiopia ‖Luleå University of Technology, Sweden
birhanu-hailu.belay@upsaclay.fr, guyon@chalearn.org
marcus.liwicki@ltu.se, romain.egele@inria.fr
{tadele.mengiste,bezawork.tilahun,tesfa.tegegne}@bdu.edu.et

Abstract. This paper introduces a new OCR dataset for historical handwritten Ethiopic script, characterized by a unique syllabic writing system, low-resource availability, and complex orthographic diacritics. The dataset consists of roughly 80,000 annotated text-line images from 1700 pages of 18^{th} to 20^{th} century documents, including a training set with text-line images from the 19^{th} to 20^{th} century and two test sets. One is distributed similarly to the training set with nearly 6,000 text-line images, and the other contains only images from the 18^{th} century manuscripts, with around 16,000 images. The former test set allows us to check baseline performance in the classical IID setting (Independently and Identically Distributed), while the latter addresses a more realistic setting in which the test set is drawn from a different distribution than the training set (Out-Of-Distribution or OOD). Multiple annotators labeled all text-line images for the HHD-Ethiopic dataset, and an expert supervisor double-checked them. We assessed human-level recognition performance and compared it with state-of-the-art (SOTA) OCR models using the Character Error Rate (CER) and Normalized Edit Distance (NED) metrics. Our results show that the model performed comparably to human-level recognition on the 18^{th} century test set and outperformed humans on the IID test set. However, the unique challenges posed by the Ethiopic script, such as detecting complex diacritics, still present difficulties for the models. Our baseline evaluation and dataset will encourage further research on Ethiopic script recognition. The dataset and source code can be accessed at `https://github.com/ethopic/hhd-ethiopic-I`.

Keywords: Historical Ethiopic script · Human-level recognition performance · HHD-Ethiopic · Normalized edit distance · Text recognition

1 Introduction

The gathering of historical knowledge heavily relies on analyzing digitized historical documents [30]. In order to process a large number of these document

images, automated tools that can convert images of the original handwritten documents into its digital format (e.g., with Unicode or ASCII texts) are necessary [12,55]. One such tool is Optical Character Recognition (OCR), which enables computers to extract textual information contained in images to then provide editing, translation, or search capabilities [16,53]. OCR systems often face difficulty in accurately recognizing historical documents, particularly those written in Ethiopic scripts, due to a shortage of suitable datasets for training machine learning models and the unique complexities of orthography [10,40].

In literature, the Ethiopic writing system is also known by various names, including Abugida, Amharic, Ge'ez, and Fidel. It is one of the oldest writing systems in the world, with a history dating back to the 4^{th} century AD [27]. It is used to write several languages in Ethiopia and Eritrea, including Amharic, Tigrinya, and Ge'ez. As depicted in Figure 1 and demonstrated by character samples in Figure 2, the script has a distinct visual appearance, characterized by its curved and geometric shapes, making it visually distinctive. It is written and read, as English, from left to right and top to down [12,4,38,36]. Ethiopic script contains about 317 graphemes, including 231 basic characters arranged in a 33 consonants by 7 vowels matrix, one special (1×7) character, 50 labialized characters, 9 punctuation marks, and 20 numerals (refer Appendix A for an extended discussion).

Ethiopic script has been a significant cultural and linguistic heritage of the region, playing a vital role in preserving the rich history and traditions of Ethiopia. Given this significance, the Ethiopian National Archive and Library Agency (ENALA) has collected numerous non-transcribed historical Ethiopic manuscripts from various sources, covering different periods and geographical regions [56]. These documents are manually cataloged with some being digitized and stored as scanned copies. In addition, some manuscripts have been officially registered in UNESCO's Memory of the World program [11,41].

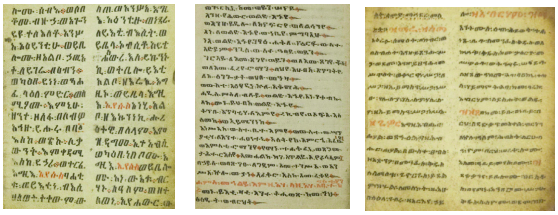

Fig. 1. Sample historical handwritten document image from HHD-Ethiopic dataset: two-column 19^{th}-century manuscript (left), one-column 20^{th}-century manuscript (middle), two-column 18^{th}-century manuscript (right).

Despite the long history of the Ethiopic script, it has encountered numerous challenges in the digital world due to its low-resource nature [18,46]. Issues such as limited digitized fonts, linguistic tools, and datasets have posed obstacles in

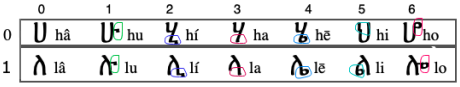

Fig. 2. The first two rows of Fidel-Gebeta (a matrix structure of Ethiopic characters): The first column shows the consonants, while the following columns (1-6) illustrate syllabic variations (obtained by adding diacritics or modifying parts of the consonant, circled in color) (see Appendix A)

the fields of natural language processing and document image analysis technologies. To address the scarcity of suitable datasets for machine learning tasks in historical handwritten Ethiopic text-image recognition, we aim to prepare a new dataset that can advance research on the Ethiopic script and facilitate access to knowledge from these historical documents by various communities, including paleographers, historians, librarians, and researchers.

The main contributions of this paper are stated as follows.

- We introduce the first sizable dataset for historical handwritten Ethiopic text-image recognition, named Historical Handwritten Dataset for Ethiopic (HHD-Ethiopic).
- We assess the human-level performance of multiple participants in historical handwritten recognition to establish a baseline for comparison with machine learning models.
- We evaluate several state-of-the-art Transformer, attention, and Connectionist Temporal Classification (CTC)-based methods.
- We compare the prediction results of machine learning model with human-level performance in predicting the sequence of Ethiopic characters in text-line images.

The rest of the paper is organized as follows: Section 2 reviews the relevant methods and related works. Settings of human-level recognition performance and OCR models are described in section 3. Section 4 presents results obtained from the experiment and comparative analysis between the model and human-level recognition performance. Finally, in Section 5, we conclude and suggest directions for future works.

2 Related work

In this section, we briefly review related work in optical character recognition and highlight challenges we are facing in OCR of historical Ethiopic manuscripts.

2.1 Optical character recognition

Machine Learning techniques have been extensively applied to the problem of optical character recognition, see [14,55,58,15,57] for a review. These achievements

are possible due to the availability of numerous datasets designed for various document image analysis tasks across a variety of scripts:

Among these, we can mention IAM-HistDB [24], DIDA [29], IMPACT [43], GRPOLY-DB [25], DIVA-HisDB [51], ICDAR-2017 Dataset [45], SCUT-CAB [16] and HJDataset [47] as examples of historical and handwritten datasets. There are other datasets that can be used for printed and scene text-image recognition, including OmniPrint datasets [52], UHTelPCC [28], COCO dataset [54], and TextCaps [50], in addition to the historical and handwritten datasets mentioned previously.

Similarly, there are datasets created for Ethiopic script recognition such as the ADOCR for printed text by [10], TANA for scene text by [19], HARD-I, and DEHR for modern handwritten text by [2] and [5] respectively(refer appendix B.1 for detail statistics). Recently, a dataset containing 50k words of historical handwritten Ethiopic was utilized to train a model in Transkribus [3]. However, the data is not currently accessible. Relying on subscription-based libraries like Transkribus raises ethical concerns due to data uploading, and these models still require training data and a stable internet connection.

Nowadays, segmentation-free OCR approaches [6,42,58] based on Connectionist Temporal Classification (CTC) [12,20,14,37,55,48] attention mechanisms [32,44,49,57], and transformer-based models [8,23,39] have become a popular choice among researchers and are widely used for text-image recognition (in both well-known and low-resourced scripts), as opposed to the traditional segmentation-based OCR approaches.

With these approaches remarkable recognition performances have been reported for many well-known scripts, such as Latin-based and Chinese, ranging from historical to modern [8], and from handwritten to machine-printed [34]. However, Ethiopic scripts have remained among the underresearched writing system, failing on taking on these advantages and consequently lacking functional OCR systems. In the following section, we briefly discuss the features and the different collecetions of historical Ethiopic manuscripts.

2.2 Features of historical Ethiopic manuscripts

There are various collections of ancient Ethiopic manuscripts in museums and libraries in Ethiopia and other countries. For example, the ENALA collection contains 859 manuscripts, the Institutes of Ethiopian Studies has 1500 manuscripts [41,1], and the collections in Rome (Biblioteca Apostolica Vaticana), Paris (Bibliothèque nationale de France), and London (British Library) contain a total of 2700 manuscripts [41]. These manuscripts were typically written on a material called Brana, which could vary in quality depending on the intended purpose or function of the book [35,41].

Black and red were the most commonly used inks, with black reserved for the main text and red reserved for religious headings and names of significance. A shown in Figure 3 historical Ethiopic manuscripts' layout can also vary and include formats such as three columns in the Synaxarion, one column for Psalms and prayer books, and two columns in liturgical books [9,41]. The materials

Fig. 3. Examples of historical Ethiopic Manuscripts: (a) Two-column writing in liturgical books with decorated heading[1], (b) Two-column writing in liturgical books without decoration[2], (c) Three-column writing in the Synaxarion[3], (d) One column for Psalms and prayer books[4].

used for writing, including the pen and ink and the writing style can also vary depending on the time period and region in which the manuscripts were produced (see Appendix A, Figure 11 for an extended discussion).

Historical documents, such as Ethiopic manuscripts, often have artifacts like color bleed-through, paper degradation, and stains, making them more challenging to work with than contemporary, well-printed documents [22]. Lack of labeled historical handwritten datasets makes it even harder to develop functional OCR system for Ethiopic text image recognition.

Therefore, in this paper, we aim to tackle the challenges in recognizing the Ethiopic script by creating a new dataset called HHD-Ethiopic which is composed of manuscripts dating from the 18^{th} to 20^{th} centuries. We also evaluate various state-of-the-art OCR models and compare their performance against human-level benchmarks.

3 Dataset and baseline methods

In this section, we provide an overview of our work, focusing on two key aspects: the detailed characteristics of our new dataset (subsection 3.1) and the settings for benchmark methods including human-level recognition performance and baseline OCR models (subsection 3.2). Our dataset serves as a valuable resource for evaluating historical handwritten Ethiopic OCR.

3.1 HHD-Ethiopic dataset

The HHD-Ethiopic dataset consists of 79,684 text-line images with their corresponding ground-truth texts that are extracted from 1,746 pages of Ethiopic manuscripts dating from 18^{th} to 20^{th} centuries. The dataset includes 306 unique characters (including one blank token), with the shortest text comprising two characters and the longest containing 46 characters. These 306 characters are

[1] https://expositions.nlr.ru/eng/ex_manus/efiopiia/efiopiia_letter.php
[2] https://upload.wikimedia.org/wikipedia/commons/2/2f/Sample_of_Ge%27ez_writing.jpg
[3] https://elalliance.files.wordpress.com/2013/11/world-history2.jpg
[4] https://www.w3.org/TR/elreq/images/kwk-mashafa-sawasew-page-268-typeface-change-for-emphasis.jpg

Table 1. Summary of the training and test text-line images

Type-of-data	Pub-date-of-manuscript	#text-line images
Training-set	90% of (A+B+C)	57,374
Test-set-I (IID)	10% of (A+B+C)	6,375
Test-set-II (OOD)	100% of (D)	15,935

A= Unknown pub. date, B= 20^{th} century, C= 19^{th} century, D= 18^{th} century manuscript

Fig. 4. Sample historical handwritten Ethiopic text-line images from HHD-Ethiopic

not distributed equally; some occur more frequently due to the nature of the script, being widely used in the writing system. For example characters such as ወ፣ ነ፣ ስ፣ በ፣ ት፣ ይ፣ ም፣ ለ፣ ር፣ ተ፣ ም፣ ብ፣ ከ፣ ል , etc are among the most frequent characters, whereas characters like ፐ፣ ፓ፣ ዣ፣ ኝ፣ ኺ፣ ጿ፣ ቨ፣ ቿ፣ ሯ፣ ቮ , etc are notably infrequent, occurring almost below a count of 10. In response to this issue of underrepresentation, we generated a separate synthetic text-line images from these characters (see the Appendix section B.1 for an extended discussion)

The training set includes text-line images from recent manuscripts, primarily from the 19^{th} and 20^{th} centuries. We created two test set: the first one consists of 6375 images that are randomly selected using a sklearn train/test split protocols[5], from a distribution similar to the training set, specifically from 19^{th} and 20^{th} century books. The second one, with 15,935 images, is drawn from a different distribution and made of 18^{th} century manuscripts (see Table 1 for the splitting processes and size of the each set). The goal of the first test set is to evaluate the baseline performance in the IID (Independently and Identically Distributed) setting, while the second test aims to assess the model's performance in a more realistic scenario, where the test set is OOD (Out-Of-Distribution) and different from the training set.

To perform preprocessing and layout analysis tasks, such as text-line segmentation, we utilized the OCRopus[6] framework. For text-line image annotation, we developed a simple tool with a graphical user interface, which displays an image of a text-line and provides a text box for typing and editing the corresponding ground-truth text. In adddition, we employed this tool to collect predicted text during the evaluation of human-level performance.

A team of 14 people participated in creating the HHD-Ethiopic datasets, with 12 individuals tasked with labeling and the remaining two individuals responsible for reviewing and ensuring the accuracy of the alignment between the ground-truth text and text-line images, making any necessary corrections as needed. To ensure the accuracy of the annotations, participants were provided with access

to reference materials for the text-lines, and all of them were familiar with the characters in the Ethiopic script. Table 1 and Figure 4 provide a summary of the dataset and show sample text-line images of the HHD-Ethiopic dataset, respectively (see Appendix B for an extended discussion).

3.2 Settings for human-level performance and baseline models

To establish a baseline for evaluating the performance of models on the HHD-Ethiopic OCR dataset, we propose two approaches: (i) **Human-level performance** and (ii) **Sequence-to-sequence models**.

The human-level performance serves as a benchmark for evaluating and comparing the recognition performance of machine learning models on historical handwritten Ethiopic scripts and provides insights for error analysis. To calculate the human-level recognition performance, 13 independent annotators were hired and divided into two groups. It is important to note that these individuals are different from those mentioned in section 3.1. The first group transcribed text-line images from the first test set, which consisted of 6375 randomly selected images from the training set. The second group transcribed the second test set of 15935 images from the 18^{th} century. Each text-line image was predicted by multiple people (i.e nine for Test-set-I and four for Test-set-II). The annotators were already familiar with the Ethiopic script, and they were explicitly instructed to carry out the task without using any references. The predicted texts by each annotator, along with comprehensive details of the data collection and annotation process, is documented as metadata for future reference.

The second reference point involves various state-of-the-art OCR models, which includes methods based on CTC-loss, attention mechanism and transformer model. The CTC-based models employ a combination of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) as an encoder and CTC as a decoder, and is trained end-to-end with and without an Attention mechanism (see Appendix C for an extended discussion). In addition, for the attention-based baseline, we utilize ASTER [49], and for the transformer-based baselines, we adopt ABINet as proposed by [23].

Moreover, we use Bayesian optimization (see e.g., [7,21] for a review) to optimize the hyperparameters of the CTC-based models. Optimizing hyperparameters involves finding an optimal setting for the model hyperparameters that could result in the best generalization performance, without using test data. Considering the trade-offs between model performance and computational cost, we use a small subset of the training set to optimize the hyperparameters of models (see, e.g., [13] for a review), and then train the model on the full training set using the optimal hyperparameter settings.

We used the Character Error Rate (CER) [12,26] and Normalized Edit Distance (NED) [17] as our evaluation metric for both the OCR models and human-level recognition performances (see appendix C, equation 3& 4 for an extended discussion).

[5] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
[6] https://github.com/ocropus/ocropy

4 Experimental results

Our objective is to perform a fair comparison between human and machine performance on historical handwritten Ethiopic scripts recognition task. This comparison is intended to showcase the utility and value of our new HHD-Ethiopic dataset, evaluate human recognition capabilities, and highlight any advancements made by baseline OCR methods.

4.1 Human-level performance

As previously discussed in Section 3.1, the ground-truth text was annotated by multiple people and double-checked by supervisors who were familiar with Ethiopic scripts. For this phase, new annotators who were also familiar with Ethiopic characters were selected and instructed not to use any reference materials. Unlike the training set, the test sets were reviewed by an expert in historical Ethiopic manuscripts. We assessed the agreement between the supervisors and the expert for the test sets by comparing the annotation texts before and after the expert's review. This resulted in CER of 7.05% for Test-set-I and 6.72% for Test-set-II.

Table 2. The human-level recognition performance in CER and NED

Type-of-test data	Year-of-Pub	Annotator-ID	CER	NED
IID	19^{th} & 20^{th}	Annot-I	29.02	27.67
		Annot-II	27.87	25.89
		Annot-III	29.93	28.16
		Annot-IV	29.16	27.80
		Annot-V	26.56	24.56
		Annot-VI	**25.39**	**23.78**
		Annot-VII	29.26	28.08
		Annot-VIII	25.95	24.78
		Annot-IX	51.03	46.25
		Average	30.46	28.59
OOD	18^{th}	Annot-X	**33.20**	**30.77**
		Annot-XI	54.33	52.20
		Annot-XIII	39.96	35.90
		Annot-XIV	45.06	39.89
		Average	43.13	39.69

To evaluate the human-level recognition performance, multiple annotators were asked to predict the text in the images and then their character recognition rates were recorded. The best annotator on Test-set-I scored a CER of 25.39% and an NED of 23.78% on Test-set-I, and a CER of 33.20% and an NED of 30.77% on Test-set-II. The average human-level recognition performance was a CER of 30.46% and an NED of 28.59% on Test-set-I, and a CER of 43.13% and

an NED of 39.63% on Test-set-II. We used the best human-level performance as a baseline for comparison with SOTA OCR models' performance throughout this paper. Table 2, shows the human-level performance on both test sets, based on assessments from nine annotators on Test-set-I and four on Test-set-II.

4.2 Baseline OCR models

This section presents the results obtained from the experimental setups detailed in Section 3. Firstly, we present the results of the CTC-based OCR models previously proposed for Amharic script recognition [12,11], followed by the results of other state-of-the-art models [20,23,48,49] validated in Latin, Amharic and/or Chinese scripts.

The experiments conducted using the CTC-based models previously proposed for Amharic script were categorized into four groups:

- **HPopt-Plain-CTC**: plain-CTC (optimized hyper-parameters)
- **Plain-CTC**: Plain-CTC
- **HPopt-Attn-CTC**: Attention-CTC (optimized hyper-parameters)
- **Attn-CTC**: Attention-CTC

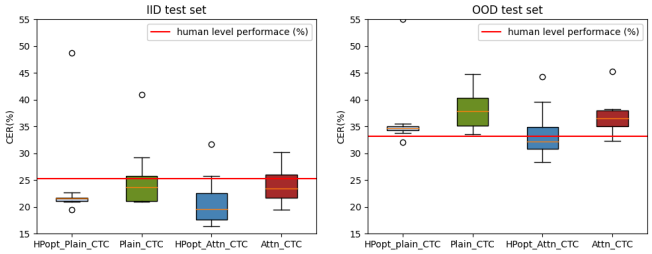

Fig. 5. Box Plot comparison of variance in the recognition performance of CTC-based models and human level performance from ten experiments with varying random weight initialization and training sample orders on Test-set-I (IID) (left) and Test-set-II(OOD) (right). The results demonstrate that HPopt-attn-CTC outperforms all other CTC-based methods and surpassing human-level recognition on both test sets. Since the second group of models is too complex, we conducted individual experiments. Therefore, instead of a Box Plot, a learning curve is presented in Figure 6.

In all the CTC-based setups, to minimize computational costs during training, we resized all the text-line images to 48 by 368 pixels. We used 10% of the text-line images randomly drawn from the training set for validation. As previously discussed, in Section 3, we have two test sets: (i) Test-set-I, which includes 6375 text-line images randomly selected from 19^{th}, 20^{th} century manuscripts and other manuscripts with unknown publication dates, and (ii) Test-set-II, a text-line images that are drawn from a different distribution other than the

training set, which includes 15935 text-line images from 18^{th} century Ethiopic manuscripts only. The HPopt-Attn-CTC baseline model achieved the best CER of 16.41% and 28.65% on Test-set-I and Test-set-II, respectively (see Table 3 for details).

The results depicted in Figure 5 demonstrate that the CTC-based OCR models outperform human-level performance on Test-set-I in all configurations. However, only the HPopt-Attn-CTC model can surpass human-level performance, while the other configurations achieve comparable or worse results compared to human recognition on Test-set-II. Test-set-I was randomly selected from the training set, while Test-set-II consisted of 18^{th} century manuscripts and represented out-of-distribution data. This disparity in performance is to be expected, as machine learning models typically perform better on samples that are independently and identically distributed rather than those in an out-of-distribution setting. The repeat experiments aimed to capture the variability in the performance of the models due to random weight initialization and sample order.

HPopt-plain-CTC exhibits consistent variability across the 10 experiments due to the benefits of hyper-parameter optimization and a simplified architecture without attention mechanisms. The systematic fine-tuning of hyper-parameters, coupled with a simpler model architecture, resulted in stable and predictable performance throughout the experiments. In contrast, HPopt-attn-CTC achieved the lowest error despite some variability in certain trials, demonstrating its robustness across ten trials (see Table 3). The optimized hyperparameter configuration significantly improved recognition accuracy compared to non-optimized settings on both test sets, highlighting the importance of hyperparameter tuning for superior performance beyond relying solely on prior knowledge or trial-and-error approaches.

The second category of baseline OCR models assessed using our HHD-Ethiopic dataset comprises state-of-the-art models, including CRNN [48], ASTER [49], ABINet [23], and SVTR [20]. Considering our available computing resources, all these models were trained for 25 epochs. The learning curve illustrating the recognition performance of these models on the IID and OOD test sets is depicted in Figure 6. In this group, the SVTR and ABINet models achieved the

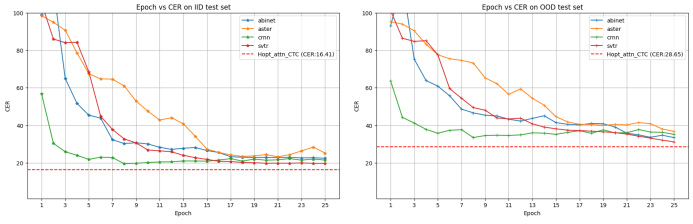

Fig. 6. Learning curve on IID and OOD test data. CER on IID test set (left), CER on OOD test set (right) across 25 epochs for ASTER, ABInet, SVTR, and CRNN models. In all plots, the red horizontal line represents the CER value of the Hopt-attn-CTC on IID and OOD data respectively.

highest performance, with both models showing nearly equivalent results within a 1% difference during evaluation. As shown in Table 3, compared to the CTC-based models, the attention and transformer-based models exhibit larger number of parameter (see Appendix C for an extended discussion).

Table 3. A summary of baseline models and recognition performance on Test-set-I (IID, 6k) and Test-set-II (OOD, 16k) using CER and NED. The table lists model parameters in millions (M).

Methods	**#Model-Parms**	**Type-of-test data**	**CER**	**NED**
Plain-CTC [12]	2.5M	IID	20.88	19.09
		OOD	33.56	31.9
Attn-CTC [11]	1.9M	IID	19.42	21.01
		OOD	33.07	32.92
HPopt-Plain-CTC	4.5M	IID	19.42	17.77
		OOD	32.01	29.02
HPopt-Attn-CTC	2.2M	IID	**16.41**	**16.06**
		OOD	**28.65**	**27.37**
TrOCR [31]	333.9M	IID	35.0	33.0
		OOD	45.0	43.87
CRNN [48]	8.3M	IID	21.04	21.01
		OOD	29.86	29.29
ASTER [49]	27M	IID	24.43	20.88
		OOD	35.13	30.75
SVTR [20]	6M	IID	19.78	17.98
		OOD	30.82	28.00
ABINet [23]	23M	IID	21.49	18.11
		OOD	32.76	28.84
Human-performance		IID	25.39	23.78
		OOD	33.20	30.77

Based on Figure 7 and our experimental observations, we observed distinct error patterns between humans and models: both exhibit substitution errors, but the model tends to make a higher number of insertions and deletions. This highlights the imperfection of the baseline OCR models in terms of sequence alignments. Furthermore, our study found that the evaluated baseline OCR models were highly effective, surpassing human-level recognition performance on Test-set-I. However, only a few models achieved better recognition performance on Test-set-II. Compared to other methods, the HPopt-Attn-CTC model has achieved the best recognition accuracy on both datasets.

The baseline models evaluated in this study comprise CTC-based models previously proposed for the Amharic script, alongside five state-of-the-art attention and transformer-based models validated using English and Chinese scripts. These models could serve as references for evaluating the effectiveness of ad-

[7] https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.boxplot.html

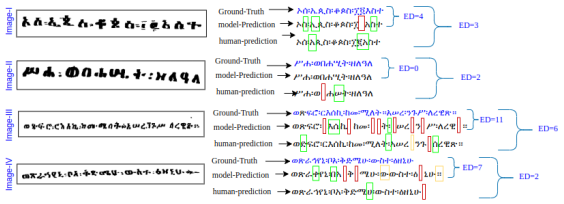

Fig. 7. Sample human-machine recognition errors per text-line image from the Test-set-I. Deleted characters are marked in red, while substituted and inserted characters are marked by green and yellow boxes, respectively. The inner ED denotes the Edit distance between the ground-truth and model prediction, while the outer ED denotes ground-truth to human prediction Edit distance.

vanced models in recognizing historical handwritten Ethiopic scripts. Each of the CTC-based models previously proposed for Amharic script underwent ten experiments. In contrast, the other models, although trained for only single experiments and fewer epochs, achieved comparable results. In addition, among the CTC-based models, the optimized hyperparameters model demonstrates superior performance, benefiting from fine-tuning and reduced overfitting. The reported results and dataset serve as a benchmark for future research in machine learning, historical document image recognition, while the analysis of human-level recognition performance enhances our understanding of the dataset.

5 Conclusion

In this paper, we presented a novel dataset for text-image recognition research in the field machine learning and historical handwritten Ethiopic scripts. The dataset comprises 79,684 text-line images obtained from manuscripts ranging from the 18^{th} to 20^{th} centuries and includes two test sets for evaluating OCR systems in both the IID (Independent and Identically Distributed) and OOD (Out-of-Distribution) settings. We provided human-level performance and baseline results using CTC, attention and transformer based models to aid in the evaluation of OCR systems. To the best of our knowledge, this is the first study to offer a sizable historical dataset with human-level performance in this domain.

In addition to the human-level performance, we also demonstrated the utility of our dataset in tackling text-image recognition challenges. Our evaluation involved the previously proposed smaller-size models for the Amharic script and SOTA models that has been validated with Latin-based and Chinese scripts. Our experiments demonstrate that both the trained SOTA methods and the smaller networks yield comparable results. Notably, the SOTA models produce equivalent outcomes even with smaller iterations but larger parameter size, which shows the potential for smaller networks to be suitable for low-resource computing infrastructure while still achieving comparable results. The dataset and source code can be accessed at `https://github.com/ethopic/hhd-ethiopic-I`. As part of our future work, we plan to expand the dataset and incorporate language models

and contextual information to enhance recognition performance. In addition, we aim to refine the baseline models and conduct further experiments to enable a more systematic and conclusive evaluation of the different methods.

Acknowledgments

This work was partially supported by ANR Chair of Artificial Intelligence HUMANIA ANR-19-CHIA-0022, TAILOR EU Horizon 2020 grant 952215, ChaLearn, and ICT4D research center of Bahir Dar Institute of Technology. We are also grateful to the the Ethiopian National Archive and Library Agency (ENALA) staffs who provided valuable assistance with data collection, and allowing us access to necessary documents, as well as to Tsiyon Worku, Tariku Adane, Gizaw Wakjira, and Lemma Kassaye for their help with data collection approval and validation.

References

1. A. Wion, C.B.T., Derat, M.L.: Inventory of libraries and catalogues of ethiopian manuscripts (2008), http://www.menestrel.fr/
2. Abdurahman, F., Sisay, E., Fante, K.A.: Ahwr-net: offline handwritten amharic word recognition using convolutional recurrent neural network. SN Applied Sciences **3**, 1–11 (2021)
3. Alemayehu, H.M.: Handwritten text recognition best practice in the betamasaheft workflow. Journal of the Text Encoding Initiative (2022)
4. Amha, A.: On loans and additions to the fidäl (ethiopic) writing system. In: The Idea of Writing, pp. 179–196. Brill (2009)
5. Assabie, Y., Bigun, J.: Offline handwritten amharic word recognition. Pattern Recognition Letters **32**(8), 1089–1099 (2011)
6. Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H.: What is wrong with scene text recognition model comparisons? dataset and model analysis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4715–4723 (2019)
7. Balaprakash, P., Salim, M., Uram, T.D., Vishwanath, V., Wild, S.M.: Deephyper: Asynchronous hyperparameter search for deep neural networks. In: 2018 IEEE 25th international conference on high performance computing (HiPC). pp. 42–51. IEEE (2018)
8. Barrere, K., Soullard, Y., Lemaitre, A., Coüasnon, B.: Transformers for historical handwritten text recognition. In: 16th International Conference on Document Analysis and Recognition (ICDAR) (2021)
9. Bausi, A.: La tradizione scrittoria etiopica. Segno e testo **6**, 507–557 (2008)
10. Belay, B., Habtegebirial, T., Liwicki, M., Belay, G., Stricker, D.: Amharic text image recognition: database, algorithm, and analysis. In: 2019 International conference on document analysis and recognition (ICDAR). pp. 1268–1273. IEEE (2019)
11. Belay, B., Habtegebrial, T., Liwicki, M., Belay, G., Stricker, D.: A blended attention-ctc network architecture for amharic text-image recognition. In: ICPRAM. pp. 435–441 (2021)
12. Belay, B., Habtegebrial, T., Meshesha, M., Liwicki, M., Belay, G., Stricker, D.: Amharic ocr: An end-to-end learning. Applied Sciences **10**(3), 1117 (2020)
13. Bottou, L.: Stochastic gradient descent tricks. Neural Networks: Tricks of the Trade: Second Edition pp. 421–436 (2012)
14. Breuel, T.M.: High performance text recognition using a hybrid convolutional-lstm implementation. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1, pp. 11–16. IEEE (2017)
15. Chen, X., Jin, L., Zhu, Y., Luo, C., Wang, T.: Text recognition in the wild: A survey. ACM Computing Surveys (CSUR) **54**(2), 1–35 (2021)
16. Cheng, H., Jian, C., Wu, S., Jin, L.: Scut-cab: A new benchmark dataset of ancient chinese books with complex layouts for document layout analysis. In: International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 436–451. Springer (2022)
17. Chng, C.K., Liu, Y., Sun, Y., Ng, C.C., Luo, C., Ni, Z., Fang, C., Zhang, S., Han, J., Ding, E., et al.: Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1571–1576. IEEE (2019)
18. Destaw, T., Yimam, S.M., Ayele, A., Biemann, C.: Question answering classification for amharic social media community based questions. In: Proceedings of the 1st

Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages. pp. 137–145 (2022)
19. Dikubab, W., Liang, D., Liao, M., Bai, X.: Comprehensive benchmark datasets for amharic scene text detection and recognition. arXiv preprint arXiv:2203.12165 (2022)
20. Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., Du, Y., Jiang, Y.G.: Svtr: Scene text recognition with a single visual model. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. pp. 884–890. International Joint Conferences on Artificial Intelligence Organization (7 2022). https://doi.org/10.24963/ijcai.2022/124, https://doi.org/10.24963/ijcai.2022/124, main Track
21. Egele, R., Gouneau, J., Vishwanath, V., Guyon, I., Balaprakash, P.: Asynchronous distributed bayesian optimization at hpc scale. arXiv preprint arXiv:2207.00479 (2022)
22. Esposito, F.: Symbolic machine learning methods for historical document processing. In: Proceedings of the 2013 ACM symposium on Document engineering. pp. 1–2 (2013)
23. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
24. Fischer, A.: Iam-histdb a dataset of handwritten historical documents. Handwritten historical document analysis, recognition, and retrieval-state of the art and future trends (2020)
25. Gatos, B., Stamatopoulos, N., Louloudis, G., Sfikas, G., Retsinas, G., Papavassiliou, V., Sunistira, F., Katsouros, V.: Grpoly-db: An old greek polytonic document image database. In: 2015 13th international conference on document analysis and recognition (ICDAR). pp. 646–650. IEEE (2015)
26. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning(ICML. pp. 369–376 (2006)
27. Jeffrey, J.: The 4th century art that died out across the world and the ethiopian scribes trying to preserve it. https://www.globalissues.org/news/2014/05/08/18652 (2004)
28. Kummari, R., Bhagvati, C.: Uhtelpcc: a dataset for telugu printed character recognition. In: International Conference on Recent Trends in Image Processing and Pattern Recognition. pp. 24–36. Springer (2018)
29. Kusetogullari, H., Yavariabdi, A., Hall, J., Lavesson, N.: Digitnet: a deep handwritten digit detection and recognition method using a new historical handwritten digit dataset. Big Data Research **23**, 100182 (2021)
30. Lenc, L., Martínek, J., Král, P., Nicolao, A., Christlein, V.: Hdpa: historical document processing and analysis framework. Evolving Systems **12**, 177–190 (2021)
31. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models. In: AAAI 2023 (February 2023)
32. Lin, Q., Luo, C., Jin, L., Lai, S.: Stan: A sequential transformation attention-based network for scene text recognition. Pattern Recognition **111**, 107692 (2021)
33. Liwicki, M., Graves, A., Fernàndez, S., Bunke, H., Schmidhuber, J.: A novel approach to on-line handwriting recognition based on bidirectional long short-term

memory networks. In: Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR 2007 (2007)
34. Ly, N.T., Nguyen, C.T., Nakagawa, M.: An attention-based end-to-end model for multiple text lines recognition in japanese historical documents. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 629–634. IEEE (2019)
35. Mellors, J., Parsons, A.: Ethiopian Bookmaking: bookmaking in rural Ethiopia in the twenty-first century. New Cross Books (2002)
36. Meshesha, M., Jawahar, C.: Indigenous scripts of african languages. Indilinga African Journal of Indigenous Knowledge Systems **6**(2), 132–142 (2007)
37. Messina, R., Louradour, J.: Segmentation-free handwritten chinese text recognition with lstm-rnn. In: 2015 13th International conference on document analysis and recognition (ICDAR). pp. 171–175. IEEE (2015)
38. Meyer, R.: the ethiopic script: linguistic features and socio-cultural connotations. Oslo Studies in Language **8**(1) (2016)
39. Mostafa, A., Mohamed, O., Ashraf, A., Elbehery, A., Jamal, S., Khoriba, G., Ghoneim, A.S.: Ocformer: A transformer-based model for arabic handwritten text recognition. In: 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC). pp. 182–186. IEEE (2021)
40. Nikolaidou, K., Seuret, M., Mokayed, H., Liwicki, M.: A survey of historical document image datasets. IJDAR 25 p. 305–338 (2022)
41. Nosnitsin, D.: Ethiopian manuscripts and ethiopian manuscript studies. a brief overview and evaluation. Gazette du livre médiéval **58**(1), 1–16 (2012)
42. Pal, U., Sharma, N., Wakabayashi, T., Kimura, F.: Off-line handwritten character recognition of devnagari script. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). vol. 1, pp. 496–500. IEEE (2007)
43. Papadopoulos, C., Pletschacher, S., Clausner, C., Antonacopoulos, A.: The impact dataset of historical document images. In: Proceedings of the 2Nd international workshop on historical document imaging and processing. pp. 123–130 (2013)
44. Qiao, Z., Zhou, Y., Yang, D., Zhou, Y., Wang, W.: Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13528–13537 (2020)
45. Sanchez, J.A., Romero, V., Toselli, A.H., Villegas, M., Vidal, E.: Icdar2017 competition on handwritten text recognition on the read dataset. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1, pp. 1383–1388. IEEE (2017)
46. Seyoum, B.E., Miyao, Y., Mekonnen, B.Y.: Morpho-syntactically annotated amharic treebank. In: CLiF. pp. 48–57 (2016)
47. Shen, Z., Zhang, K., Dell, M.: A large dataset of historical japanese documents with complex layouts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 548–549 (2020)
48. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2298–2304 (2016)
49. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: An attentional scene text recognizer with flexible rectification. IEEE transactions on pattern analysis and machine intelligence **41**(9), 2035–2048 (2018)
50. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: European conference on computer vision. pp. 742–758. Springer (2020)

51. Simistira, F., Seuret, M., Eichenberger, N., Garz, A., Liwicki, M., Ingold, R.: Divahisdb: A precisely annotated large dataset of challenging medieval manuscripts. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 471–476. IEEE (2016)
52. Sun, H., Tu, W.W., Guyon, I.M.: Omniprint: A configurable printed character synthesizer. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021), https://openreview.net/forum?id=R07XwJPmgpl
53. Thuon, N., Du, J., Zhang, J.: Improving isolated glyph classification task for palm leaf manuscripts. In: International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 65–79. Springer (2022)
54. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140 (2016)
55. Wick, C., Zöllner, J., Grüning, T.: Rescoring sequence-to-sequence models for text line recognition with ctc-prefixes. In: International Workshop on Document Analysis Systems (DAS). pp. 260–274. Springer (2022)
56. Wion, A.: The national archives and library of ethiopia: six years of ethio-french cooperation (2001-2006). In: The National Archives and Library of Ethiopia: six years of Ethio-French cooperation (2001-2006). pp. 20–p (2006)
57. Zhang, J., Du, J., Dai, L.: A gru-based encoder-decoder approach with attention for online handwritten mathematical expression recognition. In: 2017 14th IAPR International conference on document analysis and recognition (ICDAR). vol. 1, pp. 902–907. IEEE (2017)
58. Zhong, Z., Jin, L., Xie, Z.: High performance offline handwritten chinese character recognition using googlenet and directional feature maps. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 846–850. IEEE (2015)

Appendix

This appendix comprises three sections: Ethiopic writing system, dataset collection and baseline model training details. The Ethiopic writing system section explores background of Ethiopic script, scripting structure and its historical significance. The dataset collection section outlines the data collection, preprocessing and annotation steps, and statistically information about samples in HHD-Ethiopic dataset. Lastly, the model baseline training process section presents insights into baseline model training strategies and sample results.

A Ethiopic writing systems

Ethiopic script is an ancient writing system used primarily in Ethiopia and Eritrea. With its origins dating back to the 4^{th} century AD [27]. The script is characterised by its unique syllabic structure, which combines consonants and vowels to form complex characters. The Ethiopic writing system can be also named as "Abugida", "Amharic", "Ge'ez", and "Fidel" and it is primarily used for writing over 27 languages including the Amharic and Tigrinya languages, among others. Sample document of the Ethiopic script are depicted in Figure 8.

Fig. 8. Sample historical handwritten Ethiopic manuscripts

The Ethiopic script poses unique challenges for machine learning due to the scarcity of available resources [18,46]. This script is characterized by its complex orthographic identities and visually similar characters. Comprising over 317 distinct characters, including approximately 280 characters organized in a 2D matrix format known as Fidel-Gebeta (Figure 9), along with 20 digits and 8 punctuation marks (Figure 10).

As depicted in Figure 9, the Ethiopic script consists of 34 consonant characters, which serve as the base for deriving additional characters using diacritics. These diacritics can be found as small marks placed on the top, bottom, left,

or right sides of the base character. Furthermore, specific vowel characters are formed by shortening either the left or right leg of consonant characters, as demonstrated in columns 4 (shortening left leg) and 7 (shortening right leg) of the fidel-Gebeta. The vowels, derived from these consonants, span from 1 to 12 and correspond to the respective columns.

For example, in the second row of the fidel-Gebeta, the consonant character ለ represents the sound "le" in Ethiopic. From this base character, various vowel characters emerge, such as:

- ሉ is formed by adding a horizontal diacritic at the middle left side of the base character and represents the sound "lu".
- ሊ is formed by adding a horizontal diacritic at the bottom left leg of the base character and represents the sound "li".
- ላ is formed by shortening the left leg of the base character and represents the sound "la".

These examples showcase the versatility of the Ethiopic script, where modifying the diacritics or leg lengths of consonant characters allows for the representation of different vowel sounds.

Ethiopic numerals also called Ge'ez numerals, are a numeric system traditionally used in Ethiopic writing. These numeral system has its own distinct symbols for representing numbers, which are different from the Arabic or Roman numerals commonly used in many other parts of the world. The system has a base of 10, with unique characters for each digit from 1 to 9, as well as special symbols for tens, hundreds, and thousands (Figure 10). For example:

- Ethiopic symbol ፩ is similar to the Arabic numeral 1.
- symbol ፵፬ is similar to the Arabic numeral 44.
- symbol ፱፼፺፱፼፺፱ similar to the Arabic numeral 99999.
- symbol ፼፪ is similar to the Arabic numeral 10002.
- symbol ፲፪፻፴፫ similar to the Arabic numeral 1233.

Though modern Arabic numerals dominate daily life and official documents, understanding Ethiopic numerals is vital for deciphering historical texts and preserving cultural heritage.

In the Ethiopic writing system, punctuation marks convey meaning and guide text interpretation (see Figure 10). Understanding their usage is vital for clear and effective written communication in Ethiopic script.

The complexities of symbols within the Ethiopic script present significant challenges for machine learning tasks, requiring attentive approaches to achieve accurate recognition and analysis. An example of these challenges is the non-standardized usage of punctuation marks 11 and variations in writing styles, as depicted in Figure 8. These factors contribute to the difficulties encountered in the development of Ethiopic OCR systems.

B Data collection and annotation process

The Ethiopic script, one of the oldest in the world, is underrepresented in the fields of document image analysis (DIA) and natural language processing (NLP).

		1	2	3	4	5	6	7	8	9	10	11	12
		ä/e	***u***	***i***	***a***	***ē***	***ə***	***o***	***ʷä/ue***	***ʷi/u***	***ʷa/ua***	***ʷē/uē***	***ʷə***
1	h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ					
2	l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ			ሏ		
3	ḥ	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ			ሗ		
4	m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ			ሟ		
5	ś	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ			ሧ		
6	r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ			ሯ		
7	s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ			ሷ		
8	š	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ			ሿ		
9	q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቈ	ቍ	ቋ	ቌ	ቍ
10	b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ			ቧ		
11	v	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ			ቯ		
12	t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ			ቷ		
13	č	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ			ቿ		
14	ẖ	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	ኈ	ኍ	ኋ	ኌ	ኍ
15	n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ			ኗ		
16	ñ	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ			ኟ		
17	'	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ			ኧ		
18	k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኰ	ኵ	ኳ	ኴ	ኵ
19	x	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ	ዀ	ዅ	ዃ	ዄ	ዅ
20	w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ					
21	'	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ					
22	z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ			ዟ		
23	ž	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ			ዧ		
24	y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ					
25	d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ			ዷ		
26	ǧ	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ			ጇ		
27	g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ	ጐ	ጕ	ጓ	ጔ	ጕ
28	ṭ	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ			ጧ		
29	č	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ			ጯ		
30	p	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ			ጷ		
31	ṣ	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ			ጿ		
32	ś	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ					
33	f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ			ፏ		
34	***p***	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ			ፗ		

Fig. 9. Fidel-Gebeta: the row-column matrix structure of Ethiopic characters. The first column shows the consonants, while the following columns (1-12) illustrate syllabic variations (obtained by adding diacritics or modifying parts of the consonant).

This is due to the lack of attention from researchers in these fields and the absence of annotated datasets suitable for machine learning. However, in recent times, there has been a significant increase in interest from individuals involved in computing and digital humanities. As part of this growing attention, we have contributed by preparing this first sizable historical handwritten dataset for Ethiopic text-image recognition. The primary source of these documents is the Ethiopian National Archive and Library Agency (ENALA), spanning from the 18^{th} to the 20^{th} century. To ensure privacy, each page is randomly sampled from about seven different books covering cultural and religious related contents. After obtaining scanned copies of the documents from ENALA, we utilize the

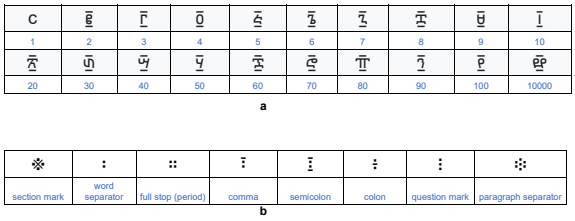

Fig. 10. Numbering system (a) and punctuation marks (b) in Ethiopic script

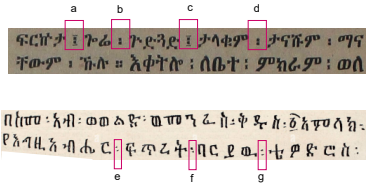

Fig. 11. Examples of punctuation usage and writing Styles: As shown by the red rectangle and labeled by [a, b, c, d], there is typically a space before and after the punctuation mark. In contrast, the punctuation marks labeled by [e, f, g] do not have any space before or after them. The punctuation marks labeled by a and c serve as list separators and are distinct from the other punctuation marks, which are used as word separators.

OCRopus[2] OCR framework and the ground-truth text annotation process is described as follows:

The annotation process can be grouped in three phase:

- **Phase-I**: In this phase, we hired 14 individuals who are familiar with the Ethiopic script. Out of the 14, 12 were assigned the task of annotation, while the remaining two served as supervisors responsible for follow-up the annotation process and ensuring the completeness of each annotation submission. In ddition, the supervisors were responsible for multiple tasks, including monitoring the progress of each annotator, providing assistance when issues arose, making decisions to address any problems encountered during the annotation process, checking alignment consistency between images and ground-truth at each phase of the annotator's submission, and making necessary corrections in case of errors. Throughout the annotation process, all annotators and supervisors had the freedom to refer to any necessary references.
- **Phase-II**: Once we have all the annotated text-line images from phase-I, we divide the text-image into training and test sets. For the training set, we reserve all text line images from the 19^{th} and 20^{th} centuries, as well as a few documents with unknown publication dates. The test set is exclusively composed of text line images from the 18^{th} century. In addition,

[2] https://github.com/ocropus/ocropy

we randomly sample another test set, which constitutes 10% of the training set. We call this randomly selected set as **Test-set-I**, which allows us to evaluate the baseline performance in the classical IID (Independently and Identically Distributed) setting.

On the other hand, the test set that is drawn from a different distribution than the training set, known as Out-Of-Distribution (OOD), is called **Test-set-II**. This setup enables us to assess the performance in real scenarios where the test set differs from the training distribution.

– **Phase-III**: In this phase, we hired approximately 20 individuals who are familiar with the Ethiopic script, along with one historical expert for the second round of annotation and request them to submit within 5 weeks. This annotation phase has the following two objectives:

  - to ensure the quality of the test set.
  - to evaluate the human-level performance in historical Ethiopic script recognition, which serves as a baseline for comparison with machine learning models.

  Out of the 20 individuals hired, only 13 annotators successfully completed the annotation task within the specified submission deadline, while the remaining individuals failed and resigned from the task. Among the 13 successful annotators, the first group comprised 9 people who transcribed text-line images from the first test set, which consisted of 6,375 randomly selected images from the training set. The second group consisted of 4 people who transcribed the second test set, consisting of 15,935 images from the 18^{th} century.

  With the exception of the expert reviewer, who was allowed to use external references, all annotators in this phase were instructed to perform the task without the use of references. Detailed data from each annotator was documented as metadata for future reference and can be accessed from our GitHub repository. One observation we made during this annotation process was that some annotators anonymously shared information, despite our efforts to ensure data confidentiality. However, despite this limitation, we have successfully compute the human-level performance for each annotator and have reported the results accordingly.

Considering the resources available to the annotators, including computing infrastructure and internet access, we developed a simple user-friendly tool with a easy to use Graphical User Interface (GUI) for the annotation process. The tool is depicted in Figure 12 and sample text-line images with the corresponding ground truth are shown in figure 13.

Each annotator's machine was equipped with this tool, enabling them to work offline when internet access was unavailable. In addition, we provided them with a comprehensive *README* file and instructed them on how to utilize the annotation tool.

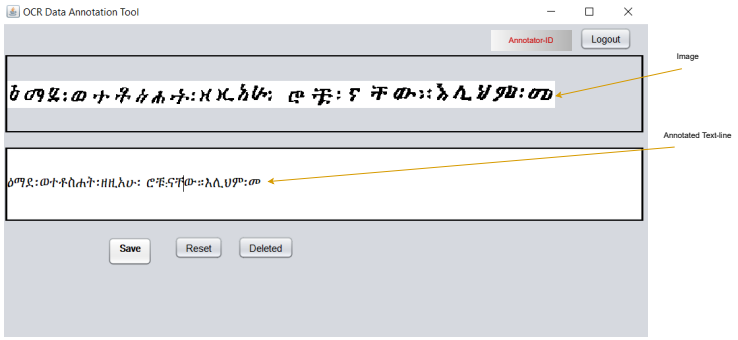

Fig. 12. Text-line image annotation tool

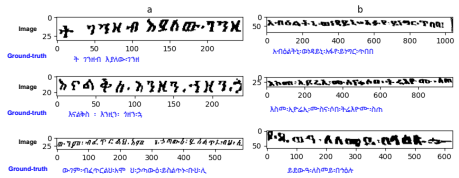

Fig. 13. Sample text-line images and ground-truth for HHD-Ethiopic: a) Training-set. b) Test-set.

B.1 Dataset statistical overview and comparisons

This section provides a detailed description of the characteristics of the HHD-Ethiopic dataset. These characteristics include the diversity of content, variations in image quality, distribution of image sizes in the trainin and test sets, the number of samples per class, and a comparison with related datasets. Examples of sample page images are illustrated in Figure 14, showcasing pages from various publication years (categorized as $18^{th}, 19^{th}$, 20^{th}, and unknown date of publication). In addition, Figure 15 displays page images categorized by image quality, which ranges from bad to medium and good. It's important to note that documents of insufficient quality, falling below the "bad" threshold, are excluded during the process of text line extraction.

The histogram in Figure 17 illustrates the distribution of text-line image sizes (width and height) across the training set and two test sets. In addition, access to the distribution of characters for each class (i.e., the frequency of characters within the 306 unique characters) in both the training and test sets.

To better represent characters that are infrequent or absent in the training set, we have employed a solution involving the generation of synthetic images. Each character is incorporated into synthetic images approximately 200 times on average.

Though it may not be fair to directly compare datasets from distinct settings, we provide a comparisons between our historical handwritten (HHD-Ethiopic)

Fig. 14. Sample page images ranging from 18^{th}, 19^{th}, 20^{th} centuries, as well as images of unknown publication dates, arranged from top left, top right, bottom left and bottom right respectively.

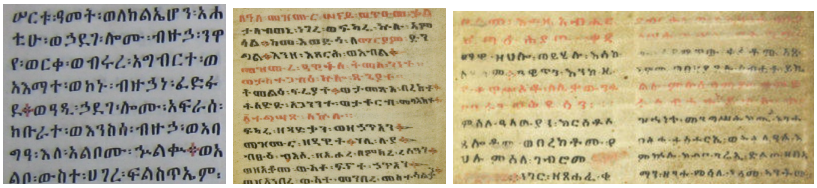

Fig. 15. Sample page image images with good(left) , medium(middle) and bad (right) quality.

dataset and the existing collections of modern printed, modern handwritten, and scene text datasets for the task of Ethiopic script recognition. The summary of comparisons is is given in Table B.1.

C Baseline models and implementation details

In this section, we provide additional details of models implemented and evaluated on our HHD-Ethiopic OCR dataset. We evaluate several state-of-the-art methods, which can be broadly grouped as CTC-based, Attention, and Transformer-based. However, our primary focus in this section is on the CTC-based model, which is designed to work effectively in lower resource settings. This is because the other CTC, attention and Transformer-based model (evaluated on this new datasets) are validated from previous works [20,23,48,49] and involves larger parameter size, making it more suited for higher-resource environments. These SOTA methods are implemented based on the open-source toolbox, mmocr: `https://github.com/open-mmlab/mmocr`.

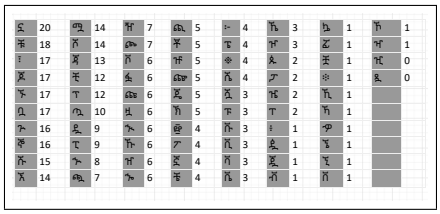

Fig. 16. Frequency distribution of underrepresented characters occurring 20 times or less in the training set. zero in the frequency column refers to the characters that exit in the test set but not in the training set.

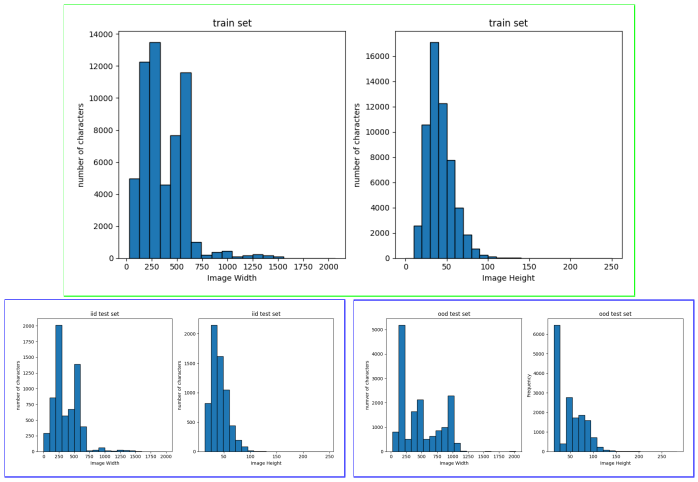

Fig. 17. A histogram for distribution of image sizes in the HHD-Ethiopic dataset: a) Training-set (top). b) IID test-set (bottom left), c) OOD test set(bottom right).

C.1 Baseline models

The implementation of the CTC-based model follows a typical pipeline depicted in Figure 18. In case of Plain-CTC, initially, the preprocessed images are passed through a convolutional neural network (CNN) backbone, which extracts relevant image features using a series of convolutional and pooling layers.

The output features from the CNN backbone are reshaped and subsequently fed into a a Long Short-term Memory (LSTM) network with connectionist temporal classification (CTC) network. This combination enables the model to effectively capture the temporal dependencies between the image features and the corresponding text labels. The RNN layer incorporates two Bi-directional LSTM units to learn sequential patterns and generate a $[(c+1) \times T]$ matrix of Softmax probabilities for each character at each time-step, where c and T denote the number of characters and the length of maximum time-step. Finally, a the CTC converts the intermediate representations into the final output text predictions.

Table 4. Summary of publicly available datasets for Ethiopic script

Dataset-type	image-type	# images	# uniq-chars	# test-sample	annotations
Printed [10]	real	40,929	280	2,907	line-level
	synthetic	296,408	280	15724	line-level
Scene [19]	real	15,39	302	9,257	word-level
	synthetic	2.8M	302	-	word-level
Handwritten [2]	real/modern	12,064	300	1,264	word-level
	Augmented	33,672	-	*	word-level
Handwritten [5]	real/modern	10,932	265	-	word-level
Our (HHD-Ethiopic)	real/historical	79,684	306	22,310	line-level
	synthetic	1200	64	*	line-level

the synthetic data generated for our new dataset denotes the number of underrepresented characters

\- denotes information that is unavailable/ not given
* denotes data that has not been utilized for testing

The alternative CTC-based approach, referred to as Attn-CTC within this paper and previously introduced for Amharic text recognition [11], extends the Plain-CTC methodology by incorporating an attention mechanism into the CTC layers. The rationale behind incorporating the attention layer lies in leveraging its capacity to derive a more potent hidden representation through a weighted contextual vector. This model comprises a combination of CNN and LSTM as the encoding module. The output of this module feeds into the attention module, and subsequently, the decoded output string is obtained through the CTC layer.

During training, the CTC algorithm calculates the likelihood of the output sequence given the input sequence and uses it as the objective function [26,33]. The training process maximizes this likelihood, which, in turn, maximizes the probability of the correct output sequence. The loss that is minimized during training is the negative of this likelihood, which can be defined as:

$$CTC_{loss} = -log \sum_{(y,x)\in S} p(y/x) \tag{1}$$

where x and y denote pair of input and output sequences in sample dataset S respectively and the probability of label sequence for a single pair p(y/x) is computed by multiplying the probability of labels along a specific path π for the overall time steps T and it can be defined as:

$$P(y/x) = \prod_{t=1}^{T} p(a_t, \pi) \tag{2}$$

where a is a character in the specified path and p(a) is its probability on each time-step on that path.

Once training and evaluating the OCR model with network settings proposed in [12,11], we employed Bayesian optimization for the selection of hyperparameters, with the CTC validation loss serving as the criteria for optimization.

Bayesian optimization captures the relationship between the hyperparameters and the CTC validation loss, iteratively updating and refining the model as it explores different hyperparameter configurations (see ref [7]for details) that yields lower CTC validation loss values. This approach allowed us to effectively tune our model and enhance its performance, contributing to the overall success of our text-image recognition model.

The recognition performance of all human-level and baseline models evaluated in this work is reported using the character error rate (CER) and Normalized Edit Distance (NED) metrics. All results reported with these two metrics are converted to 100%. The CER metric can be computed as follows,

$$CER(T,P) = \left(\frac{1}{c}\sum_{m\in T, n\in P} ED(m,n)\right) \times 100, \tag{3}$$

where c denotes the total number of characters in the ground-truth, t and p denote the ground-truth labels and predicted respectively, and $ED(m,n)$ is the Levenshtein edit-distance between sequences m and n.

while the NED metric is computed as:

$$NED = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{ED(m_i,n_i)}{\max(l_i,\hat{l}_i)}\right) \times 100 \tag{4}$$

where N is the maximum number of paired ground truth and prediction strings, ED is the Levenshtein edit distance, m_i and n_i denote the predicted text and the corresponding ground truth (GT) string, respectively, and l_i and $\hat{l}_i$ are their respective text lengths.

C.2 Training details and configurations

During our experiments, we employed various hyperparameter settings, including those selected by Bayesian Optimization [7] specifically for the CTC-based models. Training and evaluation were performed on a single NVIDIA RTX A6000 GPU for all the baseline models.

For the CTC-based baseline models, we trained them multiple times with different hyperparameter values, including epochs ranging from 10 to 100, employing a trial-and-error approach and utilizing the hyperparameters suggested by Bayesian Optimization. In this paper, we report the results obtained from the two CTC-based models (without attention) achieving better CER in just 15 epochs. In addition, the attention-CTC models showed improved performance as we trained them for more epochs. The reported results, for attention-CTC models, in the main paper were trained for 100 epochs. The dataset and the code can be accessed at `https://github.com/ethopic/hhd-ethiopic-I`

Considering our focus on low-resource settings, we prioritize optimizing our time and resources effectively. Hence, as it is not suitable for training in resource-constrained environments, we do not recommend utilizing complex models for

[6] https://deephyper.readthedocs.io/en/latest/index.html

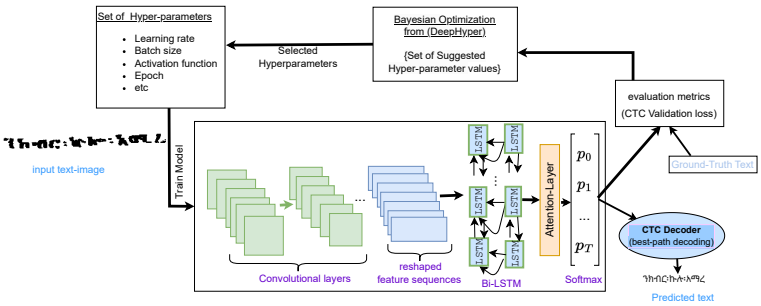

Fig. 18. A typical view of the proposed model and set of best hyper-parameters value selection using Bayesian optimization from DeepHyper[6]. The output denoted by $p_0, p_1 ... p_T$, is a matrix of Softmax probabilities with dimensions $[(c+1) \times T]$, where c is the number of unique characters in the ground-truth text and T is the length of the input time-step to the LSTM layers. The validation loss was utilized as the metric for tuning the hyperparameters. To obtain the final output sequence from the predicted probabilities produced by the model, we use the best-path decoding strategy.

Ethiopic text recognition. Instead, we prioritize exploring alternative models (such as the smaller CTC-based methods discussed in the main paper) which balance between computational efficiency and performance to ensure the feasibility of the OCR system in limited resources. However, if you possess significant computing resources, using synthetic data and conducting more extensive training iterations on those models could lead to an improvement in recognition performance for historical handwritten Ethiopic manuscripts.

We also evaluated various SOTA models [20,23,48,49] using our HHD-Ethiopic dataset. Although these models still have a relatively high number of parameters in comparison to the CTC-based models (the plain and Attn-CTC), they remain more manageable in low-resource settings. Despite the increased parameter count, we run these models for 25 epochs using limited computational resources. We achieved an improved recognition performance compared to the results presented in the TrOCR paper. By balancing performance and resource demands, the models [20,23,48,49] present a viable option for practical deployment and utilization, especially in situations where computational resources are constrained.

Due to the limited number of experimental runs conducted for [20,23,31,48,49] baseline models, we decided not to include box plots for all baseline models in the main paper. Box plots are commonly used to visualize results distribution

[1] Please note that the CER can exceed 100% when the predicted text is much longer than the ground truth. Excessive length leads to an edit distance surpassing the ground truth's character count. For instance, if the ground truth is 'ab' and the prediction is 'abced' the edit distance is 3 compared to the ground truth's 2 characters. This results in a ratio of 1.5*100=150 (see equations 3). In contrast, NED ranges from 0 to 100%, where values close to 0 are better, while values closer to 100% are indicative of poorer performances in both metrics.

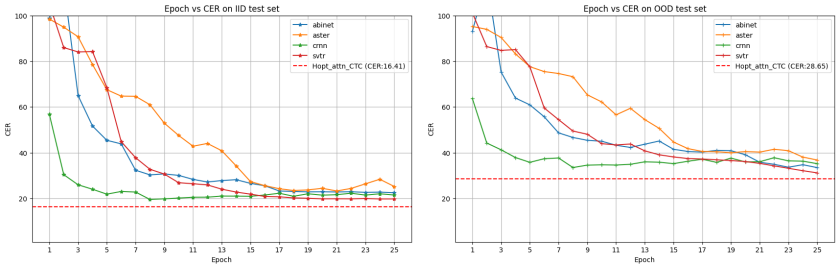

Fig. 19. Learning curve on IID and OOD test data. CER[1] on IID test set (left), CER on OOD test set (right) across 25 epochs for ASTER, ABInet, SVTR, and CRNN models. In all plots, the red horizontal line represents the CER value of the Hopt-attn-CTC network on IID and OOD data respectively.

across multiple runs, allowing for the assessment of variations and identification of outliers. Since a box plot is not suitable for representing a single experiment, we have illustrated the learning curve of the four models (ABINet, ASTER, SVTR and CRNN) in Figure 19. This learning curve illustrates the recognition performance on both IID and OOD test sets using the CER and metric across 25 epochs.

Based on learning curve depicted in Figure 19, we can conclude that all models would perform better as we train for longer epochs. Within the first 25 epochs, SVTR outperforms the others, while ASTER is the least performer. We are limited to running for 25 epochs due to time and computational resources. The red horizontal line in both the right and left plots represents the CER for Hopt-attn-CTC model. This line serves as our benchmark, as it represents the best-performing model.

C.3 Sample predicted texts

Sample images with the corresponding ground truth, model prediction and the edit distance between the ground truth and the prediction at line level is shown in Figure20

In text lines where characters with low occurrence rates appear in the ground truth of the training set often leads to an increased edit distance between the ground truth and the predicted texts during test time. This pattern is demonstrated by sample examples depicted in Figure.21

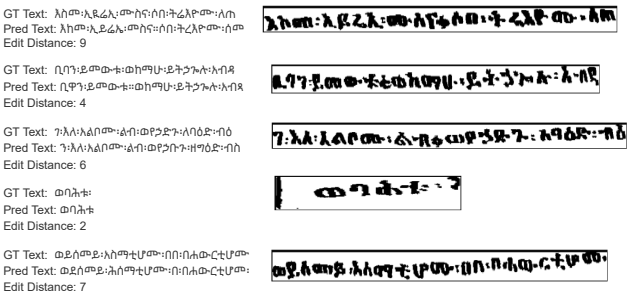

Fig. 20. Sample text-line images with their corresponding ground-truth and prediction texts

GT Text: እስከ፡ማዕዜኑ፡ትኴንኑ፡ዓመፃ፡
Pred Text: እስከ፡ማዕዚኑ፡ትፍገኑ፡ዓመዊ
Edit Distance: 5

GT Text: ዣሞ፡መዓጅሞ፡መዓጅ
Pred Text: ገርም፡መዓድ፡ም፡መዓድ
Edit Distance: 6

GT Text: መጿ፡እበክ፡ሃይማኖት
Pred Text: መጺኡ፡እንከ፡ሃይመኖት
Edit Distance: 6

GT Text: ጰ፡ፖፒሮስ፡ኢየአክርዕ፡
Pred Text: ክ፡ምጥሮስ፡ኢየአክሮስ
Edit Distance: 7

Fig. 21. Examples of prediction errors for underrepresented characters. The characters marked in red within the ground-truth text are less frequent characters and are wrongly predicted.