CESLR: A Multi-Signer Benchmark and Spatio-Temporal End-to-End Framework for Continuous Ethiopian Sign Language Recognition

Anteneh Yehalem Tegegne

Department of Data Science

Bahir DarUniversity

Bahir Dar, Ethiopia

anteneh.yehalem@bdu.edu.et

Yohannes AyanaEjigu*

Development Team, AI research

Amplitude Ventures AS

Stavanger, Norway

yohannesayana10@gmail.com

Surafel Amsalu Tadesse*

Development Team, AI research

Amplitude Ventures AS

Stavanger, Norway

surafelamsalu2013@gmail.com

***Abstract*—Continuous Ethiopian Sign Language Recognition (CESLR) remains an under explored challenge due to the language's rich visual dynamics and variability among signers. This work introduces CESLR, the first large-scale, multi-signer video corpus for continuous EthSL recognition, consisting of 1,320 sentence recordings performed twice by 22 participants. To process these visual sequences, we design an end-to-end deep learning architecture that jointly models spatial and temporal cues through 2D convolutional feature extraction, 1D temporal convolution, and bidirectional recurrent encoding. A Connectionist Temporal Classification (CTC) layer enables sequence alignment without frame-level annotation, allowing the network to learn directly from sentence-level supervision. Experimental evaluation shows that the system attains a Word Error Rate (WER) of 8.82% in signer-independent testing and 47.02% on unseen sentence evaluation, revealing strong cross-signer generalization while highlighting the difficulty of novel sentence recognition. The proposed dataset and model establish a foundational benchmark for advancing automatic EthSL translation and assistive communication technologies in multilingual settings. Our dataset and source codes are available at here.**

Keywords—Ethiopian Sign Language; Sign Language Recognition; Deep Learning; Continuous SLR, EthSL

I. INTRODUCTION

Sign languages are rich visual–gestural languages that convey meaning through coordinated hand movements, facial expressions, and body postures, each governed by its own linguistic structure and grammar [1][2][3]. For the Deaf community, sign language serves as the primary medium of everyday interaction. However, communication barriers persist between signers and non-signers due to the time and effort required to learn sign languages. To mitigate this challenge, Sign Language Recognition (SLR) technologies aim to automatically interpret visual gestures captured in video streams and translate them into corresponding textual or spoken forms, thereby enhancing accessibility and inclusion.

SLR research is typically categorized into **isolated** and **continuous** recognition tasks [4]. The isolated form deals with identifying individual signs or words, where each video clip corresponds to a single gesture. In contrast, continuous SLR involves recognizing entire sequences of signs within unsegmented videos, reflecting how natural communication occurs in real life. This task is inherently more difficult because there are no explicit temporal boundaries between consecutive signs, and sentence-level annotation is often limited or unavailable. Consequently, continuous SLR has evolved as a **weakly supervised sequence learning** problem, requiring models capable of capturing both visual and temporal dynamics.

Despite global progress in continuous SLR, **Ethiopian Sign Language (EthSL)** research has largely remained confined to finger-spelling and isolated word recognition [5]–[12]. These approaches overlook the linguistic complexity and temporal fluidity of continuous signing. To address this gap, the present work focuses on developing an **end-to-end deep learning approach** for continuous EthSL recognition. The proposed system simultaneously learns spatial representations from video frames and temporal dependencies across sequences, without requiring frame-level segmentation. Furthermore, this study introduces a **multi-signer video dataset** designed to support future research on sentence-level EthSL recognition.

The remainder of this paper is organized as follows: Section II reviews related literature, Section III describes dataset preparation and collection protocols, Section IV details the proposed methodology, Section V presents experimental settings and results, and Section VI concludes the paper with key insights and directions for future work.

II. RELATED WORKS

Research on Ethiopian Sign Language (EthSL) recognition has evolved through several methodological phases. Early approaches primarily relied on probabilistic models and handcrafted features. For instance, Hidden Markov Models (HMMs) were widely applied to capture temporal dependencies in sequential gestures [8][13]. Similarly, Kinect-based systems were employed to extract 3D motion cues and depth information, enhancing the modeling of manual and non-manual components of signing [13]. While effective for small vocabularies, these traditional methods struggled to generalize across signers and lighting variations due to their dependence on manually engineered features.

With the emergence of deep learning, researchers began to adopt data-driven architectures capable of learning richer visual representations directly from raw video input. In the context of EthSL, convolutional neural networks (CNNs) and hybrid CNN–LSTM models have been introduced to

recognize isolated words and short phrases [9][14]. These models substantially improved classification accuracy by jointly learning spatial and temporal features, yet their applicability to continuous sign recognition remained limited.

In the broader field of continuous sign language recognition (CSLR), substantial progress has been achieved through the development of large-scale datasets such as RWTH-PHOENIX-Weather 2014 for German Sign Language [15] and the Chinese Sign Language corpus [16]. Early CSLR techniques relied on handcrafted features in conjunction with HMM-based temporal modeling [18][19]. To enhance robustness, several studies later integrated CNNs and LSTMs with HMMs [21][22], allowing frame-wise alignment between video sequences and gloss labels. Although these hybrid systems demonstrated improvements, their reliance on explicit segmentation limited scalability to longer, sentence-level data.

A major advancement came with the introduction of Connectionist Temporal Classification (CTC) by Graves et al. [23], which enabled alignment-free training of neural sequence models. This innovation allowed modern CSLR frameworks to learn direct mappings from video sequences to gloss strings without predefined boundaries, paving the way for fully end-to-end architectures that unify spatial, temporal, and sequential modeling.

Despite global success in CSLR, video-based continuous recognition for EthSL has remained largely unexplored. Existing studies focus predominantly on isolated or finger-spelling tasks, lacking resources for sentence-level analysis. To bridge this gap, the current work presents the first multi-signer EthSL video dataset, covering 30 sentences recorded from 22 signers of varying age and gender. This dataset forms the foundation for developing robust deep learning models tailored to continuous EthSL recognition and future benchmarking efforts.

III. DATA PREPARATION

Developing a robust continuous Ethiopian Sign Language (EthSL) dataset requires careful planning, participant diversity, and standardized recording conditions. Because no public dataset previously existed for EthSL sentence-level recognition, a new multi-signer video corpus was created to establish a reproducible research benchmark for the community.

Figure 1: Sample Images Representing Sentences from the CESL Dataset.

A. Participants and Data Overview

The dataset contains 1,320 annotated video samples recorded from 22 participants (14 male and 8 female), whose ages range between 10 and 60 years. Each signer performed 30 unique sentences twice, resulting in two repetitions per sentence. The participants were recruited from Fasilo Secondary School and Yekatit 23 Primary School in Bahir Dar, Ethiopia, representing both native (deaf since birth) and non-native signers. This diversity ensures the dataset captures a variety of signing speeds, styles, and motion dynamics, which is crucial for training generalizable recognition models.

B. Sentence Selection

The sentence set was curated from introductory EthSL learning materials designed for beginners. The selected phrases cover essential topics such as family relations, occupations, manners, emotions, and daily necessities. These sentences were chosen to reflect common real-world communication contexts while maintaining manageable linguistic complexity for annotation and modeling. A few examples of the selected sentences are summarized in Table 1.

C. Recording Setup and Environment

Data recording sessions were conducted in controlled outdoor environments at the participating schools. To maintain consistent visual conditions, all signers wore black clothing against a solid green backdrop to facilitate background subtraction and feature extraction (as shown in Figure 1). A Canon Mark IV DSLR camera was positioned approximately two meters from the participant, aligned to capture upper-body movements and facial expressions clearly. The camera recorded videos in Full HD resolution (1920×1080 pixels) at 25 frames per second (FPS) to ensure smooth motion representation.

Prior to data collection, ethical approval was obtained from Bahir Dar University's Research Ethics Committee. Informed consent forms translated into Amharic were collected from all participants and legal guardians in the case of minors. The final dataset, named CESLR (Continuous Ethiopian Sign Language Recognition), serves as the first publicly available EthSL corpus supporting sentence-level recognition research. The dataset composition and statistical summary are presented in Table 2, while sample video frames illustrating different sentence recordings are shown in Figures 1 and 2.

Table 1 :Sample Selected Sentences from CESL dataset

No	Sentence
1	አባቴ ዘመድ ይወዳል
2	ወንድሜ ዳቦ በማር በላ
3	እህቴን ጸሎት ማድረግ ትወዳለች
4	ልጆች ፍቅር ይፈልጋሉ
5	የእናቴን ሻሽ እወድዋለሁ

D. Suggested splits

Following Guo et al. [24], two data splits were designed:

- Split I – Signer-independent**:** 16 signers for training, 3 for validation, and 3 for testing. This split evaluates model generalization to unseen signers, important due to variability in age, appearance, and signing style.
- Split II – Unseen sentence split: 25 sentences used for training and 5 for testing, with overlapping signers. This

tests the model's ability to generalize to novel sentence structures.

IV. METHOD

Our proposed end-to-end framework for Continuous Ethiopian Sign Language Recognition (CESLR) is illustrated in Figure 2. It integrates spatial, temporal, and sequence alignment modules into a unified pipeline.

We use 2D-CNNs (ResNet-18) to extract spatial features from each video frame, foll owed by 1D-CNNs to capture short-term temporal patterns. These spatio-temporal features are then processed by a Bidirectional LSTM (Bi-LSTM), which models long-range temporal dependencies across the video sequence.

The final output is decoded using Connectionist Temporal Classification (CTC), which enables alignment between input frames and output gloss labels without needing frame-level annotation. CTC introduces a special blank token to separate repeated or adjacent signs and computes all valid alignment paths to optimize sequence prediction.

V. EXPERIMENT

A comprehensive set of experiments was carried out to evaluate the proposed end-to-end framework for continuous Ethiopian Sign Language (EthSL) recognition. All tests were conducted using the CESLR dataset, which contains 1,320 annotated video clips featuring 22 signers and 30 unique sentences. Two evaluation protocols were designed: a signer-independent split, where the model was trained and tested on disjoint signer groups performing the same sentences, and an unseen sentence split, where identical signers performed different sentence sets to assess the model's ability to generalize to novel linguistic compositions.

Performance was measured using the Word Error Rate (WER) metric, widely adopted in Continuous Sign Language Recognition (CSLR) tasks [15]. WER quantifies the minimum number of substitutions, deletions, and insertions required to convert the predicted sequence into the reference transcript. Formally, it is defined as:

WER = (#substitutions + #deletions + #insertions) / #words in reference

The experimental design followed the configurations of state-of-the-art CSLR systems [26][27], employing ResNet18 [19] pretrained on ImageNet [28] as the 2D CNN feature extractor. Temporal representations were learned using a 1D convolutional network structured as {K5, P2, K5, P2}, where Kσ and Pσ denote convolution and pooling operations with kernel size σ, respectively. For sequence modeling, both Bidirectional Long Short-Term Memory (Bi-LSTM) and Bidirectional Gated Recurrent Unit (Bi-GRU) networks were tested. Each recurrent module consisted of two layers with a hidden dimension of 1024, followed by a fully connected output layer for gloss sequence prediction.

During training, frames were resized to 256×256 pixels, randomly cropped to 224×224, and augmented through horizontal flipping (50% probability) and 20% temporal rescaling. At inference time, a centered 224×224 crop was applied. The model was trained for 25 epochs using the Adam optimizer [28] with a learning rate of 0.001, reduced by a factor of five after the 20th epoch, a batch size of 2, and a weight decay of 0.001. All experiments were performed on NVIDIA Quadro RTX 6000/8000 GPUs.

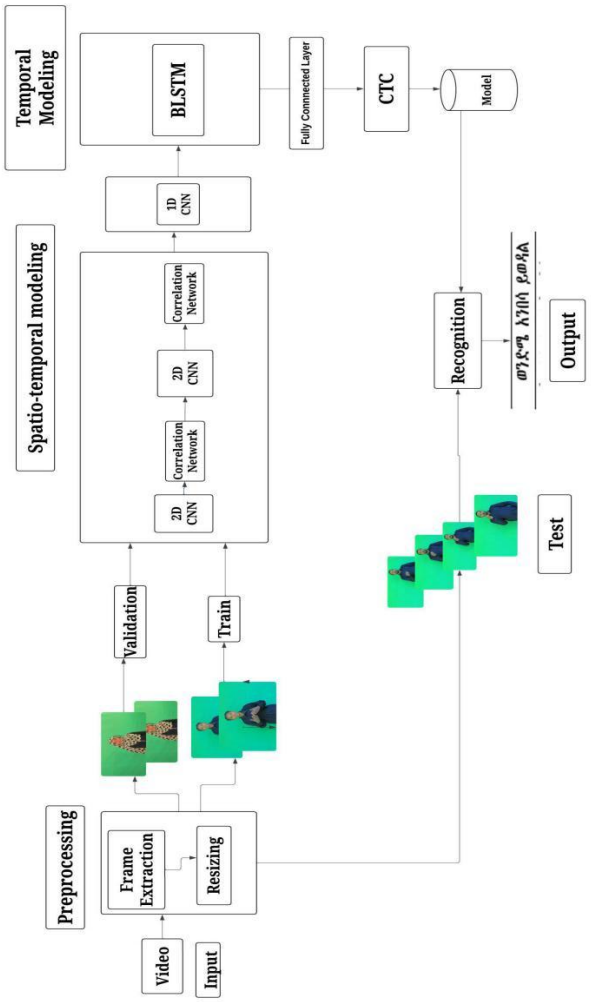

Figure 2 Proposed Architecture

The comparative results of both recurrent models are summarized in Tables 3 and 4. The Bi-LSTM network achieved a WER of 8.82% in the signer-independent test and 58% on unseen sentences, while validation scores were 4.09% and 1.54%, respectively. The Bi-GRU model, on the other hand, recorded 16% WER for signer-independent testing and 47% for unseen sentences, with corresponding validation results of 5.81% and 1.79%. These outcomes indicate that Bi-LSTM provides superior performance when generalizing across different signers, whereas Bi-GRU demonstrates better adaptability when encountering unseen sentence structures. The overall findings confirm that unseen sentence recognition remains a more challenging setting, emphasizing the need for models with stronger linguistic generalization in continuous EthSL recognition.

Table 3: Performance Results of the CESL Recognition System Using the Bi-LSTM Model.

Experiment	WER	
	Validation	Test
Split I	4.09%	8.82%

Split II	1.54%	58%

Table 4: Performance Results of the CESL Recognition System Using the Bi-GRU Model.

Experiment	WER	
	Validation	Test
Split I	5.81%	16.%
Split II	1.79%	47%

VI. DISCUSSION

Our evaluation revealed key differences between the signer-independent and unseen sentence splits. The model performed well in the signer-independent setting, with Bi-LSTM achieving 8.82% WER (Table 3), demonstrating strong generalization across different signers. In contrast, the unseen sentence split posed greater challenges; Bi-GRU outperformed Bi-LSTM with a WER of 47% vs. 58% (Table 4), highlighting difficulty in recognizing new sentence structures These results emphasize the need for more diverse and balanced training data to improve generalization to rare signs and varied sentence structures. Expanding the dataset with broader signer styles and vocabulary can further enhance model robustness for real-world applications.

VII. CONCLUSION

This study introduces the first continuous Ethiopian Sign Language (EthSL) dataset and an end-to-end recognition model combining CNNs, Bi-LSTM, and CTC loss. Results show strong performance in signer-independent settings (WER 8.82%, Table 3), highlighting the model's robustness across signers. However, unseen sentence tests reveal limitations in generalizing to new sign combinations, emphasizing the need for more diverse data and exploration of advanced temporal models to improve recognition accuracy.

REFERENCES

[1] *Ong, S.C. & Ranganath, S., Automatic sign language analysis, TPAMI, 2005.*

[2] *Dreuw, P. et al., Speech recognition techniques for sign language, 2007.*

[3] *Sandler, W. & Lillo-Martin, D.C., Sign Language and Linguistic Universals, Cambridge Univ. Press, 2006.*

[4] *Cui, R. et al., RCNNs for continuous sign recognition, CVPR, 2017.*

[5] *Tamiru, N.K. et al., Amharic sign recognition using ANN and SVM, Vis. Comput., 2021.*

[6] *Abeje, B.T. et al., EthSL recognition with CNN, Multimed. Tools Appl., 2022.*

[7] *Salau, A.O. et al., Number sign recognition using SVM, Springer, 2022.*

[8] *Gimbi, T., Isolated EthSL recognition, MSc Thesis, AAU, 2014.*

[9] *Teferi, G.M., Phrase-level sign recognition using DL, PhD Thesis, AAU, 2021.*

[10] *Yigzaw, N. et al., Generic Amharic sign recognition, AHCI, 2022.*

[11] *Abebe, S.T., Isolated word-level EthSL recognition, PhD Thesis, AAU, 2013.*

[12] *Kuma, D. et al., Skeleton-based EthSL using DL, PhD Thesis, HU, 2023.*

[13] *Haji, A.J., Translation of continuous EthSL to Amharic, MSc Thesis, AAU, 2017.*

[14] *Kassa, B.M. & Alemu, G., Amharic sign recognition with DL, MSc Thesis, AAU, 2013.*

[15] *Koller, O. et al., Continuous SLR for large vocabularies, CVIU, 2015.*

[16] *Huang, J. et al., Sign language recognition without segmentation, AAAI, 2018.*

[17] *Gao, W. et al., Chinese SLR using SOFM/SRN/HMM, PR, 2004.*

[18] *Han, J. et al., Subunit modeling in SLR, PRL, 2009.*

[19] *He, K. et al., Deep residual learning, CVPR, 2016.*

[20] *Simonyan, K. & Zisserman, A., Very deep CNNs, arXiv:1409.1556, 2014*

[21] *Koller, O. et al., Deep Sign: Hybrid CNN-HMM, BMVC, 2016.*

[22] *Koller, O. et al., Re-sign: End-to-end CNN-HMM, CVPR, 2017.*

[23] *Graves, A. et al., Connectionist Temporal Classification, ICML, 2006.*

[24] *Guo, D. et al., Dense temporal conv. for SLT, IJCAI, 2019. 2022.*

[25] *Deng, J. et al., ImageNet: A large-scale database, CVPR, 2009.*

[26] *Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11 542–11 551. 5*

[27] *R. Zuo and B. Mak, "C2slr: Consistency-enhanced continuous sign language recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022,*

[28] *J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009*