TRANSLATING SIGN LANGUAGE INTO TEXT AND SPEECH FOR THE HEARING IMPAIRED

Enes Karaağaç & Murathan Akgenç & Selin İrem Özdoğan
Artificial Intelligence Engineering Program
Department of Computer Engineering
Hacettepe University
{eneskaraagac,murathanakgenc,selinozdogan}@cs.hacettepe.edu

ABSTRACT

Sign language serves as a vital mode of communication for the deaf and hard-of-hearing communities, yet real-time translation into spoken or written language remains a technological challenge. In this study, we propose an end-to-end sign language translation system that converts video inputs into both text and speech, using a skeleton-based approach centered on spatio-temporal dynamics. Hand and body landmarks are extracted from videos via the Mediapipe framework, resulting in structured keypoint sequences used to train and evaluate four deep learning models: LSTM, LSTM with Attention, 1D CNN, and 1D CNN with Attention.

All models were trained to classify 45 isolated Turkish Sign Language (TİD) gestures. Our experiments reveal that convolutional models outperform recurrent ones in this context, with the 1D CNN + Attention model achieving the highest accuracy of 88.10%. While the attention mechanism improved performance slightly in CNNs, its impact on LSTM-based models was negligible. These findings suggest that attention mechanisms are more effective when combined with convolutional architectures, which better capture local spatial-temporal features from skeletal data.

Overall, the system demonstrates the viability of skeleton-based representations in sign language recognition and offers a comparative analysis of temporal and spatial modeling approaches. This work contributes toward the development of lightweight, real-time translation systems aimed at enhancing communication accessibility.

1 INTRODUCTION

Sign language is a rich, structured, and fully expressive visual language used predominantly by deaf and hard-of-hearing individuals. Despite its depth and widespread use, signers often face significant barriers in communicating with the broader hearing population due to the absence of real-time, accessible translation technologies. As a result, there is a growing need for automated systems that can bridge this communication gap and promote inclusivity.

Among the emerging techniques in automatic sign language recognition (ASLR), skeleton-based representation using body and hand keypoints has proven to be a promising direction. These representations are lightweight, robust to visual noise, and capable of capturing essential motion dynamics needed for gesture understanding.

In this study, we introduce a system that utilizes Mediapipe to extract skeletal landmarks from sign language videos, converting them into structured numerical arrays. We then conduct two separate experiments using this data as input to distinct neural network architectures: a Long Short-Term Memory (LSTM) model that captures temporal dependencies, and a Convolutional Neural Network (CNN) that focuses on spatial relationships. Unlike comparative analysis, our goal is not to benchmark these models against each other, but rather to evaluate which approach is more suitable for the task of gesture-to-language translation in our specific context.

The outputs of each model are mapped to corresponding textual labels, which are then vocalized using a text-to-speech engine. Through this work, we aim to identify an effective backbone architecture for real-time sign language translation and contribute to the development of accessible communication technologies.

2 RELATED WORK

The task of translating Turkish Sign Language (TİD) into spoken or written language has gained increasing attention with the advent of deep learning and pose estimation technologies. Our review of the literature draws from a wide range of studies indexed on platforms such as Google Scholar, IEEE Xplore, ResearchGate, and arXiv, using focused search queries including "Turkish Sign Language deep learning," "AUTSL dataset sign recognition," and "sign language transformer Turkish." This section summarizes recent and relevant contributions, especially those that address the use of machine learning for gesture recognition and sign-to-text translation in the context of the Turkish language.

One of the foundational resources in this domain is the Mercanoglu & Keles (2020). This paper introduces the AUTSL dataset, a significant benchmark in TİD research comprising over 38,000 isolated sign video samples from 43 signers. The dataset includes RGB, depth, and skeletal data, enabling multi-modal analysis. The authors evaluate baseline models using CNNs for spatial feature extraction and LSTMs for temporal modeling. Notably, their results underscore the difficulty of signer-independent recognition, with a drop from 95.95% (random splits) to 62.02% (user-independent setting).

Building on this foundation, the Öztürk & Keles (2024) extends research into continuous sign language translation. The dataset comprises nearly 24 hours of annotated videos and targets Turkish's morphologically rich structure, which includes high rates of rare and singleton words. Baseline models include a Pose-to-Text Transformer (P2T-T) and a GNN-based Transformer (GNN-T), with modest BLEU scores (19.% BLEU-1 and 3.28% BLEU-4), emphasizing the complexity of continuous, real-world translation tasks in Turkish.

A practical and performance-oriented contribution is offered by the paper **?**, which introduces a YOLOv4-CSP-based region-focused CNN model. This system emphasizes real-time performance by isolating hand regions to improve classification accuracy, achieving high accuracy and low latency suitable for real-world deployment scenarios such as mobile apps.

Another significant direction is static alphabet recognition, as shown in Aksoy et al. (2021). This study combines classical image processing techniques with CNNs to classify the 29 letters of the Turkish sign language alphabet. The hybrid approach, using features extracted from over 10,000 images, demonstrates the effectiveness of combining handcrafted features with deep learning for specific static gestures.

In terms of broader system perspectives, the paper Najib (2024) provides a comprehensive review of AI-based sign language interpretation. It covers multi-modal integration, including gesture, facial expression, and lip reading, and underscores remaining challenges such as context understanding and signer generalization. It also discusses potential future directions such as mobile deployment and hybrid AI-rule-based architectures.

An important community benchmark is described in Mercanoglu et al. (2021). This paper details the 2021 challenge focused on signer-independent recognition using the AUTSL dataset. Top-performing models achieved over 96% accuracy, yet the challenge exposed gaps such as the need for better generalization and multi-modal fusion techniques. This work highlights the importance of robustness in real-world deployment.

From a language-structural perspective, Celik & Reza (2024) discusses the difficulties in mapping spoken Turkish to TİD due to differences in word order and grammar. The authors propose a hybrid system combining deep learning and rule-based transformations to account for syntactic and semantic mismatches, concluding that current systems still fall short in handling linguistic complexity.

Efficiency-focused systems are also represented in Karacı et al. (2021). The authors bypass computationally intensive deep learning methods by leveraging handcrafted features such as Histogram

of Oriented Gradients (HOG) and Local Binary Patterns (LBP), combined with a cascade voting classifier, demonstrating competitive real-time performance on resource-constrained devices.

Addressing dynamic two-handed gestures, Katılmış & Karakuzu (2023) utilizes the YOLOv5x model with integrated attention mechanisms. The model excels in recognizing complex two-handed motions in video streams, achieving 98.9% accuracy — a promising result for interactive, high-fidelity translation systems.

Another real-time focused system is introduced in Güney & Erkuş (2022). This study simplifies CNN architecture for rapid classification of static signs and achieves strong performance with minimal computational overhead, suggesting practicality for embedded or mobile devices.

A more novel skeleton-based method is presented in Laines et al. (2023). Here, skeleton joint data is encoded as RGB images through a Tree Structure Skeleton Image (TSSI) transformation, which is then processed by a DenseNet-121 CNN. Results on the AUTSL dataset show that this representation is both compact and effective, further validating skeleton-based approaches in reducing computational load while maintaining accuracy.

Finally, Kayahan & Gungor (2019) proposes a hybrid rule-based/statistical method to translate spoken Turkish into avatar-performed sign language. The system leverages grammatical parsing to preserve TİD's non-linear structure and suggests that hybrid approaches can better manage real-time translation quality, especially in educational or assistive contexts.

In summary, the literature reveals a broad spectrum of approaches, ranging from deep learning to hybrid models, static to continuous translation, and lightweight handcrafted methods to transformer-based architectures. Several themes emerge:

- (i) the importance of signer-independence and dataset diversity,
- (ii) the growing effectiveness of skeleton-based input representations, and
- (iii) the continued exploration of attention mechanisms;

— though, as our own findings suggest, the benefits of attention may vary by architecture. Specifically, in our experiments, the addition of attention layers resulted in a marginal gain for CNN-based models and even a slight drop in performance for LSTM-based networks. This indicates that attention is not universally beneficial and must be applied contextually based on data type and model structure.

Our system contributes to this ongoing body of work by leveraging the simplicity and structure of skeletal keypoints extracted with Mediapipe and testing their efficacy through both LSTM and CNN architectures, ultimately aiming for an efficient and scalable framework for gesture-to-language translation in Turkish Sign Language.

3 METHODOLOGY

3.1 DATA

In this study, we utilize a subset of the AUTSL (Turkish Sign Language) public dataset, which contains videos of 266 distinct sign gestures and over 30,000 samples for training and more than 7,000 samples for testing. Due to computational and resource constraints, we opted to use a reduced version of the dataset. Specifically, we selected 45 of the most frequently occurring gestures, focusing on those that are commonly used in daily communication.

This filtered subset comprises approximately 2,700 training samples and 750 testing samples. To ensure a manageable data volume and to maintain balance across gesture classes, we further reduced the dataset by randomly discarding approximately half of the samples for each selected gesture. This step was taken to streamline the model training process while retaining sufficient data for effective learning.

3.2 PREPROCESSING

Following the selection of gestures and samples, we employed Mediapipe's holistic model to extract skeletal keypoints representing hand and body movements from the videos. This model provides spatial coordinates of landmarks for both hands, the upper body, and facial features. For our purposes, only hand and body keypoints were utilized, as they contain the most critical motion information for sign gesture recognition.

To reduce computational overhead and remove redundant information, we did not process every frame of the video. Instead, we extracted frames at regular intervals—specifically, we selected every 6th frame from each video. This downsampling approach preserves the overall temporal dynamics of the gesture while significantly decreasing the number of frames per sample, thus optimizing the data size and training time. Each selected frame was processed to extract hand and body landmarks, which were then converted into structured arrays encoding the spatio-temporal configuration of the signer's movements. These arrays were saved and used as input for training and evaluating the LSTM and CNN models.

Figure 1: Mediapipe Skeletal Keypoints on Different Signers and Signs

3.3 MODELS

3.3.1 LSTM (WITH AND WITHOUT ATTENTION MECHANISM)

The first model implemented is based on a multi-layer Long Short-Term Memory (LSTM) network designed to capture the temporal structure of gesture sequences. The architecture consists of three stacked LSTM layers, with the final layer being bidirectional to capture both past and future context. An attention mechanism is applied after the LSTM layers to allow the model to focus on the most informative time steps.

Interestingly, while the attention mechanism is designed to enhance performance by emphasizing important parts of the sequence, its inclusion slightly reduced the model's accuracy in this case—from 81.75% without attention to 81.08% with attention. This suggests that either the LSTM alone was sufficient for modeling temporal dependencies, or the attention mechanism may have introduced complexity that didn't yield gains on this dataset.

Model structure includes: • Three LSTM layers (the last one bidirectional) • Attention mechanism after LSTM (optional) • Two fully connected layers with dropout • Final softmax output for classification

3.3.2 1D Convolutional Neural Network (with and without Attention Mechanism)

The second model uses a one-dimensional Convolutional Neural Network (1D CNN) to process gesture sequences by capturing local temporal patterns through convolutional filters. The architecture includes three convolutional layers with batch normalization and max pooling to progressively reduce temporal resolution and enhance feature extraction.

Incorporating an attention mechanism after the final convolution layer led to a slight improvement in accuracy—from 87.83% to 88.10%. While this indicates that attention helped the model focus on slightly more informative regions of the sequence, the performance gain was minimal. This suggests that the convolutional layers were already effective at extracting discriminative features, and attention only offered marginal refinement. Model structure includes:

- Three 1D convolutional layers (kernel sizes 5 and 3)
- Batch normalization and max pooling after each convolution
- Optional attention mechanism over the temporal dimension
- Two fully connected layers with dropout
- Final softmax output for classification

4 RESULTS

To evaluate the performance of the proposed models, we conducted experiments using the preprocessed subset of the AUTSL dataset, consisting of 45 selected sign gestures. The evaluation metric used was classification accuracy on the test set.

4.1 LSTM with Attention

The LSTM-based model was first trained without any attention mechanism, achieving a test accuracy of 61%. Upon integrating an attention layer before the fully connected layers, the model's performance improved significantly. The attention mechanism allowed the model to focus on the most relevant frames in each sequence, resulting in an increased accuracy of 68%. This demonstrates the effectiveness of temporal modeling and the added benefit of attention in highlighting key information within gesture sequences.

4.2 1D Convolutional Neural Network

The 1D CNN model, which uses temporal convolutional filters to extract local features from the sequence data, outperformed the LSTM-based approach. This architecture achieved a test accuracy of 81%, indicating that convolutional layers are well-suited for capturing short-term dependencies in structured skeletal data.

4.3 Summary

The results suggest that, while both models can learn from skeletal representations of sign gestures, convolutional architectures offer stronger performance in this context. Their ability to capture localized patterns efficiently makes them a promising direction for real-time sign language translation systems.

Model	Attention	Accuracy
LSTM	No	81.75%
LSTM + Attention	Yes	81.08%
1D CNN	No	87.83%
1D CNN + Attention	Yes	88.10%

Table 1: Model Result Summary

5 CONCLUSION

In this study, we proposed a system for translating sign language videos into text and speech by using skeletal keypoints extracted with Mediapipe. We evaluated two different deep learning architectures—LSTM with attention and 1D CNN—to determine their effectiveness in modeling the temporal and spatial aspects of sign gestures.

Our experiments demonstrated that while the LSTM-based approach benefits from attention mechanisms to better capture sequential dependencies, the 1D CNN model achieved higher overall accuracy. This suggests that convolutional methods, which are capable of efficiently extracting local features from structured time-series data, are particularly well-suited for gesture recognition tasks.

The results also highlight the potential of lightweight skeletal data as an input modality for sign language translation, offering both robustness and computational efficiency compared to raw video or image-based approaches.

6 FUTURE WORK

In future work, we aim to:

- Expand the dataset to include more gesture classes and real-world variability.
- Explore hybrid architectures that combine CNN and LSTM layers.
- Integrate facial expressions and hand shape details for more nuanced recognition.
- Develop a real-time interface that provides immediate feedback through speech synthesis.

This research represents a step toward more accessible and intelligent sign language translation systems that can help bridge the communication gap between the hearing and deaf communities.

REFERENCES

Bekir Aksoy, Osamah Salman, and Özge Ekrem. Detection of turkish sign language using deep learning and image processing methods. *Applied Artificial Intelligence*, 35, 09 2021. doi: 10.1080/08839514.2021.1982184.

Ozer Celik and Pinar Reza. Application of ai in turkish sign language translation: A case study of its use and purpose. *Nafath*, 9, 10 2024. doi: 10.54455/MCN2701.

Selda Güney and Mehmet Erkuş. A real-time approach to recognition of turkish sign language by using convolutional neural networks. *Neural Computing and Applications*, 34, 03 2022. doi: 10.1007/s00521-021-06664-6.

Abdulkadir Karacı, Kemal Akyol, and Mehmet Turut. Real-time turkish sign language recognition using cascade voting approach with handcrafted features. *Applied Computer Systems*, 26:12–21, 05 2021. doi: 10.2478/acss-2021-0002.

Zekeriya Katılmış and Cihan Karakuzu. Double handed dynamic turkish sign language recognition using leap motion with meta learning approach. *Expert Systems with Applications*, 228:120453, 2023. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.120453. URL https://www.sciencedirect.com/science/article/pii/S0957417423009557.

Dilek Kayahan and Tunga Gungor. A hybrid translation system from turkish spoken language to turkish sign language. pp. 1–6, 07 2019. doi: 10.1109/INISTA.2019.8778347.

David Laines, Miguel Gonzalez-Mendoza, Gilberto Ochoa-Ruiz, and Gissella Bejarano. Isolated sign language recognition based on tree structure skeleton images. pp. 276–284, 06 2023. doi: 10.1109/CVPRW59228.2023.00033.

Ozge Mercanoglu and Hacer Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 01 2020. doi: 10.1109/ACCESS.2020.3028072.

Ozge Mercanoglu, Julio Junior, Sergio Escalera, and Hacer Keles. Chalearn lap large scale signer independent isolated sign language recognition challenge: Design, results and future research. pp. 3467–3476, 06 2021. doi: 10.1109/CVPRW53098.2021.00386.

Fatma Najib. Sign language interpretation using machine learning and artificial intelligence. *Neural Computing and Applications*, 37:841–857, 11 2024. doi: 10.1007/s00521-024-10395-9.

Şükrü Öztürk and Hacer Yalim Keles. E-tsl: A continuous educational turkish sign language dataset with baseline methods. In *2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1–7, 2024. doi: 10.1109/HORA61326.2024.10550648.