

SP25: DATA-255 Sec 11 - Deep Learning Technologies
"Deep Learning for Real-Time Sign Language Detection and Caption Translation"

Instructor:

Dr. Mohammad Masum

Team Members:

Hamsalakshmi Ramachandran 017423666

Himani Shah 017411407

Kush Bindal 017441359

Manjot Singh 017557462

Saqib Chowdhury 017514978

Sugandha Chauhan 017506190

Introduction

Sign language serves as an essential means of communication for people with hearing disabilities, allowing them to convey their thoughts and feelings efficiently. Nonetheless, the limited awareness of sign language creates major obstacles in everyday life, hindering access to education, job opportunities, and vital services. Thanks to progress in deep learning and computer vision, there's a chance to create automated solutions that can close this communication gap.

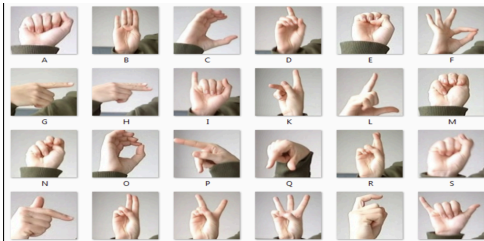
This project aims to utilize action recognition methods to effectively detect sign language gestures. In contrast to conventional gesture recognition systems that mainly depend on static hand shape identification, this project will examine motion patterns to better interpret sign language. Through the incorporation of advanced deep learning models, the system seeks to improve communication accessibility and inclusivity worldwide. Creating such a system not only aids the hearing-impaired community but also adds to the wider realm of AI-powered accessibility innovations.

Dataset:

After researching a number of datasets on multiple platforms, we filtered the following dataset to use for this project.

Sign Language MNIST (Kaggle - [Link](#))

# label	# pixel1	# pixel2	# pixel3	# pixel4	# pixel5
3	107	118	127	134	139
6	155	157	156	156	156
2	187	188	188	187	187



The Sign Language MNIST dataset contains images of hand signs representing the 26 letters of the American Sign Language (ASL) alphabet, formatted similarly to the classic MNIST dataset. It consists of 27,455 training images and 7,172 test images, each in grayscale (28x28 pixels) with labels from 0 to 25 (excluding 'J' and 'Z' due to their dynamic motion).

Project Summary

- **Objective & Significance:**

This project seeks to create a deep learning-driven system for recognizing actions in sign language detection, allowing for real-time recognition of sign language gestures in videos. Once signs are identified, the system will produce text captions and improve accessibility via multilingual translation. We aim to reduce the communication gap for those with hearing impairments by utilizing computer vision models to enable smooth sign-to-text and sign-to-voice translation. The importance of this project is found in its capacity to enhance accessibility and inclusivity for individuals with hearing impairments by offering an automated, scalable, and efficient solution that removes the dependence on human interpreters, who might not always be accessible. Its influence goes beyond personal communication to sectors like education, healthcare, and professional engagements, where accessibility is still a vital issue. By enabling real-time processing and offering multilingual support, the system can serve a worldwide audience, promoting a more connected and inclusive community. Moreover, developments in computer vision applied in this project enhance the wider domain of AI, creating opportunities for additional research and progress in sign language recognition and assistive technologies.

- **Problem Statement:**

Millions of individuals around the globe depend on sign language as their main form of communication, but a limited general awareness results in a considerable communication gap between those with hearing disabilities and the broader community. The lack of easily accessible interpreters restricts access in critical fields like education, employment, and healthcare, frequently resulting in social isolation. Although current technologies emphasize gesture recognition, they have difficulty accurately understanding the intricate and variable aspects of sign language, which includes hand motions, facial cues, and spatial orientation. This initiative aims to tackle this issue by creating a real-time action recognition system capable of precisely identifying and converting sign language from video feeds into written captions. Moreover, we aim to incorporate caption translation. Utilizing deep learning technologies like computer vision, this study seeks to offer a scalable and automated approach that improves accessibility, encourages inclusivity, and supports independent communication for those with hearing disabilities.

- **Proposed Approach/Framework**

Our approach builds a real-time ASL recognition system that extracts 3D hand and facial keypoints from 15-frame video clips using MediaPipe, processes them with a MobileNetV3 CNN and a Transformer pretrained on unlabeled data, translates the output into multiple languages via a fine-tuned mT5 model.

- **Novelty: -**

The novelty lies in combining facial and hand gesture recognition for better accuracy, using a Transformer pretrained on unlabeled videos to leverage more data, fine-tuning mT5 for in-house multilingual translation, and adding a live user feedback loop for model adaptation—features that go beyond the typical hand-only, single-language, static systems.

- **Expected Impact & Contributions:**

This project aims to improve accessibility for the deaf and hard-of-hearing community by enabling real-time sign language recognition, captioning, and translation. Advanced deep learning models like transformers and hybrid CNN-LSTM architectures enhance accuracy,

support multiple sign languages, and integrate with assistive technologies such as video conferencing. Additionally, it has the potential to significantly contribute to artificial intelligence research by enabling advanced, scalable models for sign language processing.

- **Relevance To The Field:**

The project aligns with advancements in artificial intelligence for accessibility, helping bridge communication gaps between signers and non-signers. It addresses challenges in dynamic sign language recognition and multilingual support, expanding beyond traditional static recognition approaches. With applications in digital communication, healthcare, and social inclusion, it holds strong industry relevance and potential for policy impact, promoting inclusivity and compliance with global accessibility standards.

Project Background:

A real-time sign language detection and translation system that leverages computer vision and natural language processing to convert American Sign Language (ASL) gestures into text and translate them into user-selected languages (e.g., Spanish, French, Hindi). The approach begins with collecting a hybrid dataset from existing sources like the ASL Alphabet and WLASL, supplemented by 500-1000 custom video clips of 20 signs, followed by preprocessing with MediaPipe to extract 3D hand and facial keypoints from 15-frame sequences. A lightweight MobileNetV3 CNN will encode spatial features, feeding into a 2-layer Transformer to classify dynamic and static signs, pretrained self-supervised on unlabeled sign videos using contrastive loss, then fine-tuned on labeled data. The recognized text will be translated using a fine-tuned mT5-small model. This pipeline ensures efficiency, accuracy, and adaptability, targeting 95% accuracy on static signs and 85% on dynamic ones.

Up to now, we have discovered six segmentation models that we intend to explore for our project: U-Net, DeepLabV3, Mask R-CNN, YOLOv8-Seg, HRNet, and the Segment Anything Model (SAM). These models provide a combination of traditional and contemporary methods, harmonizing precision, velocity, and flexibility for sign language recognition. Although this list serves as a solid foundation, it is not definitive—we might enhance our choices, modify model settings, or add further specifics as we advance. Our objective is to assess their efficiency in action recognition and enhance performance according to our dataset and application needs.

Existing research and improvement:

Current advancements in sign language recognition and action recognition have been fueled by deep learning and computer vision innovations. Contemporary models like 3D Convolutional Neural Networks (3D-CNNs), Recurrent Neural Networks (RNNs), and Transformers have greatly enhanced the precision of gesture and movement recognition. Spatio-temporal models such as CNN-LSTM and Vision Transformers (ViTs) enhance comprehension of sequential movements of hands and bodies in videos. Furthermore, platforms such as MediaPipe and OpenPose enable real-time pose estimation, enhancing the accessibility of sign language detection.

Unlike existing systems that primarily focus on hand gestures and output text in a single language (typically English), our approach incorporates facial landmarks alongside hand key points to enhance recognition accuracy for signs reliant on expressions, a feature often overlooked. The use of a Transformer with self-supervised pre-training on unlabeled data sets it apart from traditional CNN-LSTM models, leveraging abundant video resources to improve generalization with limited labeled data. Additionally, integrating a fine-tuned mT5 for multilingual translation directly within the system, rather than relying on external APIs, offers a standalone, customizable solution. The interactive learning mode, allowing real-time user corrections to adapt the model, introduces a personalization aspect absent in static systems. Collectively, these innovations position my project as

a technically advanced and user-centric advancement over current sign language recognition frameworks.

Literature Review:

Research on sign language recognition and translation has made significant strides over the past two decades, especially since the rapid development and resurgence of "deep learning" in 2012. The initial steps toward automated sign language recognition were multidisciplinary. Vogler and Metaxas explored the fusion of artificial intelligence with linguistic and computer vision approaches to process sign language efficiently (Vogler & Metaxas, 2005). This early research provided the conceptual framework for subsequent advancements in computational models and neural networks.

The next major leap came with deep learning models designed explicitly for sign language processing. Koller et al. introduced one of the first deep learning-based systems for static sign recognition, which leveraged convolutional neural networks (CNNs) to improve classification accuracy (Koller, Zargaran, Ney, & Bowden, 2020). Around the same time, Pu et al. extended these efforts by developing scalable recognition systems that could process multiple sign languages, significantly enhancing the applicability of deep learning techniques in real-world scenarios (Pu, Zhou, & Li, 2020). By 2020, transformer models began reshaping natural language processing and were soon applied to sign language translation. Camgoz et al. introduced Sign Language Transformers (SLT), an end-to-end framework that utilized self-attention mechanisms to improve recognition and translation efficiency (Camgoz, Koller, Hadfield, & Bowden, 2020). Hanke et al. further refined these approaches by integrating design science research methodologies to enhance the usability and implementation of artificial intelligence in sign language translation systems (Hanke, Storz, & Wagner, 2020).

Kumar and Sharma explored AI techniques for sign language processing. They portrayed the strength of their proposed framework, especially in feature extraction and model generalization (Kumar & Sharma, 2021). In parallel, the need for real-time sign language recognition prompted the development of novel techniques. Jana et al. suggested an advanced deep-learning framework that could perform recognition tasks with less latency, enabling more seamless communication (Jana, Paul, & Bhandari, 2022). Mishra et al. also focused on bridging audio and sign language translation, creating a multimodal framework for converting spoken words into Indian Sign Language (Mishra, Sharma, Qureshi, & Chaudhary, 2022). Samonte et al. presented an innovative deep learning-based approach for sign language translation into text. Their work focused on an end-to-end model that integrated deep neural networks (DNNs) and RNNs to capture the sequential nature of sign language gestures. The study demonstrated that combining temporal modeling and deep learning techniques significantly enhances translation accuracy, making real-time text generation from sign language more feasible. However, the authors noted the need for a larger, more diverse dataset to improve model generalization across different sign languages (Samonte, Guingab, Relayo, Sheng, & Tamayo, 2022). The Design framework developed by Kothadiya et al. introduced a novel pipeline for sign language detection and recognition using deep learning. The study combined CNNs with bidirectional long short-term memory (BiLSTM) networks to enhance spatial and temporal recognition. Integrating BiLSTM allowed the system to analyze motion patterns and context in signing, improving the system's ability to differentiate between similar gestures. The authors emphasized that integrating real-time video processing and optimizing computational efficiency remain essential for deploying such models in practical applications (Kothadiya, Bhatt, Sapariya, & Corchado Rodríguez, 2022).

As deep learning continued to evolve, researchers began focusing on hybrid architectures that combined CNNs with sequential learning models like long short-term memory (LSTM) networks. Pathan et al. demonstrated the efficacy of such hybrid models in improving the accuracy of sign language recognition by integrating both spatial and temporal information (Pathan et al., 2023). Pathan et al. explored this sensor data with machine learning, achieving enhanced recognition accuracy through sensor-based input processing (Pathan et al., 2023). Similarly, as Jana et al. demonstrated, deep learning models incorporating YOLO object detection algorithms were employed

to generate real-time sign language captions (Jana, Paul, & Bhandari, 2023). Triwijoyo et al. explored deep learning approaches for sign language recognition, emphasizing the application of CNN for feature extraction and classification. Their study demonstrated how CNNs can effectively identify hand gestures and movements, leading to improved accuracy in sign language recognition systems. The research further highlighted the challenges associated with variations in signing speed, hand orientations, and background noise, suggesting that future work should incorporate more robust preprocessing techniques and real-time processing capabilities (Triwijoyo, Karnaen, & Adil, 2023).

A comprehensive review by Li et al. analyzed various machine learning and AI techniques for the same. The study categorized existing approaches into vision-based, sensor-based, and hybrid methods, evaluating their strengths and drawbacks. Their review identified the growing role of deep learning, especially in sign language processing, and they also highlighted the advancements in transfer learning, multimodal integration, and transformer-based architectures. The authors concluded that while significant progress has been made, challenges such as model adaptability, computational cost, and dataset standardization need further exploration to ensure more robust and scalable sign language interpretation systems (Li, Wang, & Zhang, 2024). Recent advancements have continued to push the limits of real-time sign language processing. Zhang and Jiang extensively reviewed recent deep-learning applications, identifying key trends and future research directions (Zhang & Jiang, 2024). Zhang et al. introduced EvSign, a framework incorporating streaming event-based data processing for more scalable sign language recognition and translation (Zhang et al., 2024). Sharma and Balpande developed region-specific solutions for Indian sign language, ensuring that translation systems cater to diverse linguistic needs (Sharma & Balpande, 2024). endeavors like SignBridge leverage the integration of sign language recognition for people who are deaf or hard of hearing and mute into a video calling application that demonstrates and highlights the potential of deep learning technologies in fostering communication, emphasizing inclusivity and accessibility (Kumar & Singh, 2024). Similarly, Patel and Desai introduced a deep-learning web application designed to raise awareness and improve accessibility for sign language users (Patel & Desai, 2024).

As deep learning for sign language processing continues to evolve, the fusion of multimodal technologies, real-time captioning, and scalable translation frameworks promises to revolutionize communication for the deaf and hard-of-hearing communities. The progress in this field attests to the power of deep learning and the advancements in artificial intelligence technologies in breaking communication barriers and enhancing inclusivity.

Performance Evaluation: The model will be evaluated on the Sign Language MNIST test set using accuracy, F1-score, and AUC-ROC as primary metrics. Based on existing research, which reports 85-95% accuracy in general and 85-90% accuracy for Sign Language MNIST dataset specifically, and ~0.90 AUC-ROC scores. We aim to achieve ~95% accuracy for static signs, ~85% on dynamic signs and >0.90 AUC-ROC. Additionally, we target an F1-score exceeding 0.88 to ensure a balance between precision and recall. For caption translation, we will use the BLEU score, aiming for >0.75 to indicate high-quality translations.

To assess the model's effectiveness, we will compare its performance against baseline CNN and Vision Transformer models. We will perform comparative analysis on the effect of different techniques used, such as Hyperparameter tuning, fine-tuning and prompt engineering by removing them and evaluating the resulting performance metric differences. This will help identify the most effective strategies for improving the model.

Efficiency metrics, including memory usage and inference time, will also be measured to ensure the model is practical for deployment on resource-constrained devices. By rigorously evaluating the model using this framework, we aim to achieve state-of-the-art performance while maintaining efficiency for real-time applications.

Work Division & Timeline:

Milestones	Task Description	Responsible Member(s)	Deadline
Milestone 1 (Weeks 1-2)	Data Collection & Preprocessing Project proposal submission Identify and collect sign language datasets. Perform preprocessing (frame extraction, noise reduction, augmentation). Ensure the dataset is structured for model training.	Sugandha Chauhan, Saqib Chowdhury (Data Collection) Hamsalakshmi Ramachandran, Himani Shah, Kush Bindal, Manjot Singh (Preprocessing & Augmentation)	Feb 22 - Mar 7
Milestone 2 (Weeks 3-5)	Model Development - Action Recognition Implement action recognition model for sign language detection. Train deep learning and segmentation models (CNN-LSTM, 3D CNN, or Transformer-based models). Perform initial model evaluation.	Hamsalakshmi Ramachandran, Himani Shah (Model Architecture & Implementation) Kush Bindal, Manjot Singh (Training & Optimization) Sugandha Chauhan, Saqib Chowdhury (Assist in Training & Initial Testing)	Mar 8 - Mar 28
Milestone 3 (Weeks 6-7)	Model Testing & Evaluation Test trained models and compare performance metrics (accuracy, precision, recall, F1-score). Identify improvements and fine-tune models. Prepare Project Update Presentation for April 8.	Hamsalakshmi Ramachandran, Himani Shah (Model Testing) Kush Bindal, Manjot Singh (Hyperparameter Tuning & Optimization) Sugandha Chauhan, Saqib Chowdhury (Evaluation & Documentation)	Mar 29 - Apr 7
Milestone 4 (Week 8)	Project Update Presentation Present progress on data preprocessing, model training, and initial results. Demonstrate evaluation metrics and discuss next steps.	All Members (Presentation & Discussion)	April 8

Milestone 5 (Weeks 9-10)	Integration of Language Translation & Real-Time Deployment Implement translation for generated captions. Develop a real-time processing system for user interaction. Optimize performance for deployment.	Sugandha Chauhan, Saqib Chowdhury (Translation Model Integration) Hamsalakshmi Ramachandran, Himani Shah (Text Processing & Formatting) Kush Bindal, Manjot Singh (Real-Time Processing & Deployment)	Apr 9 - Apr 28
Milestone 6 (Week 11)	Final Testing & Project Submission Preparation Conduct final testing on all features. Prepare the final project report and presentation slides. Ensure all deliverables are ready.	All Members (Final Review & Testing)	Apr 29 - May 5
Milestone 7 (Week 12)	Final Project Presentation Present the final project, demonstrating sign language detection, translation, and real-time application. Discuss key results, impact, and future scope.	All Members (Final Presentation)	May 6

References:

1. **Vogler, C., & Metaxas, D. (2005).** Sign language recognition and translation: A multi disciplined approach from the field of artificial intelligence. *Journal of Deaf Studies and Deaf Education*, 11(1), 94-101. Retrieved from <https://academic.oup.com/jdsde/article/11/1/94/410770>
2. **Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020).** Sign language transformers: Joint end-to-end sign language recognition and translation. *arXiv preprint arXiv:2003.13830*. Retrieved from <https://arxiv.org/abs/2003.13830>
3. **Hanke, T., Storz, J., & Wagner, S. (2020).** Artificial intelligence for sign language translation – A design science research approach. *Communications of the Association for Information Systems*, 53(1), Article 22. Retrieved from <https://aisel.aisnet.org/cais/vol53/iss1/22/>
4. **Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2020).** Deep learning-based sign language recognition system for static signs. *Neural Computing and Applications*, 32, 15167–15178. Retrieved from <https://link.springer.com/article/10.1007/s00521-019-04691-y>
5. **Pu, J., Zhou, W., & Li, H. (2020).** Sign language recognition: High performance deep learning approach applied to multiple sign languages. *Proceedings of the International Conference on Computational Linguistics*. Retrieved from <https://www.proquest.com/docview/2671968253>
6. **Kumar, P., & Sharma, S. (2021).** Artificial intelligence technologies for sign language. *Journal of Artificial Intelligence Research*, 70, 765-792. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8434597/>
7. **Triwijoyo, B. K., Karnaen, L. Y. R., & Adil, A. (2023).** Deep learning approach for sign language recognition. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, 9(1), 12-21. Retrieved from https://www.researchgate.net/publication/367461727_Deep_Learning_Approach_For_Sign_Language_Recognition
8. **Pathan, R. K., Biswas, M., Yasmin, S., Khandaker, M. U., Salman, M., & Youssef, A. A. F. (2023).** Deep learning in sign language recognition: A hybrid approach using CNN and LSTM. *Mathematics*, 11(17), 3729. Retrieved from <https://www.mdpi.com/2227-7390/11/17/3729>
9. **Samonte, M. J. C., Guingab, C. J. M., Relayo, R. A., Sheng, M. J. C., & Tamayo, J. R. D. (2022).** Using deep learning in sign language translation to text. *Proceedings of the International Conference on Industrial Engineering and Operations Management*. Retrieved from <https://ieomsociety.org/proceedings/2022istanbul/758.pdf>
10. **Mishra, H., Sharma, M., Qureshi, M. A., & Chaudhary, S. (2022).** Audio to Indian sign language translation. *International Journal for Research in Applied Science & Engineering Technology*, 10(V), 3904-3908. Retrieved from <https://www.ijraset.com/best-journal/audio-to-indian-sign-language-translator>
11. **Jana, U., Paul, S., & Bhandari, D. (2022).** Real-time sign language recognition using deep learning techniques. *Proceedings of the IEEE International Conference on Information Technology, Electronics and Communications (ICITEC)*. Retrieved from <https://ieeexplore.ieee.org/document/9825192>
12. **Kothadiya, D., Bhatt, C., Sapariya, K., & Corchado Rodríguez, J. M. (2022).** Deepsign: Sign language detection and recognition using deep learning. *ProQuest Dissertations &*

- Theses. Retrieved from <https://www.proquest.com/docview/2674332265/fulltextPDF?pq-origsite=primo>
13. **Pathan, R. K., Biswas, M., Yasmin, S., Khandaker, M. U., Salman, M., & Youssef, A. A. F. (2023).** Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network. *Scientific Reports*, 13, Article 16975. Retrieved from <https://www.nature.com/articles/s41598-023-43852-x>
 14. **Jana, U., Paul, S., & Bhandari, D. (2023).** Real-time caption generation for the American sign language using YOLO and LSTM. *Proceedings of the IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*. Retrieved from https://www.researchgate.net/publication/383326962_Real-Time_Caption_Generation_for_the_American_Sign_Language_Using_YOLO_and_LSTM
 15. **Zhang, Y., & Jiang, X. (2024).** Recent advances on deep learning for sign language recognition. *Computer Modeling in Engineering & Sciences*, 139(3), 2399–2450. Retrieved from <https://www.techscience.com/CMES/v139n3/55626>
 16. **Zhang, P., Yin, H., Wang, Z., Chen, W., Li, S., Wang, D., Lu, H., & Jia, X. (2024).** EvSign: Sign language recognition and translation with streaming events. *Lecture Notes in Computer Science*. Retrieved from <https://arxiv.org/abs/2407.12593>
 17. **Li, H., Wang, J., & Zhang, Y. (2024).** Sign language interpretation using machine learning and artificial intelligence: A review. *Neural Computing and Applications*. Retrieved from <https://link.springer.com/article/10.1007/s00521-024-10395-9>
 18. **Sharma, M., & Balpande, S. (2024).** Indian sign language recognition and translation: Text to sign language using deep learning technique. *AIP Conference Proceedings*, 3112(1), 020027. Retrieved from <https://pubs.aip.org/aip/acp/article-abstract/3112/1/020027/3295724>
 19. **Kumar, S., & Singh, R. (2024).** SignBridge: Bridging communication gaps with a video calling app for inclusive conversations – Integrating sign-language recognition for the deaf and mute. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*. Retrieved from <https://ijarsct.co.in/Paper19045.pdf>
 20. **Patel, A., & Desai, M. (2024).** Promoting sign language awareness: A deep learning web application for sign language recognition. *Proceedings of the ACM Conference on Human Factors in Computing Systems*. Retrieved from <https://dl.acm.org/doi/pdf/10.1145/3695719.3695723>