Real-Time Sign Language Translation Using Machine Learning

Muhammad Tayyab
Department of Computer Science
University of Engineering and Technology
Lahore, Pakistan
mu.tayyab001@gmail.com

***Abstract*—Sign language serves as a critical medium of communication for the hearing and speech-impaired community. However, its widespread use is limited by the lack of real-time translation systems that can bridge the communication gap between sign language users and non-users. This study presents a Real-Time Sign Language Translation System leveraging machine learning techniques and computer vision to recognize and translate sign language gestures into textual or spoken language. The proposed system aims to provide a seamless, accurate, and real-time communication experience, enhancing inclusivity and accessibility. The system employs three models—Random Forest, SVM, and KNN—and achieves a high accuracy rate of 98.70**

***Index Terms*—Sign Language Translation, Real-Time Systems, Computer Vision, Machine Learning, Accessibility, Communication Technology**

I. Introduction

Sign language serves as a vital mode of communication for individuals who are hearing or speech-impaired. It enables them to express their thoughts, emotions, and intentions effectively, fostering social inclusion and independence. However, the majority of the population is not proficient in sign language, which creates significant communication barriers between sign language users and non-users. This lack of mutual understanding often leads to challenges in educational institutions, workplaces, healthcare settings, and daily interactions, thereby limiting inclusivity and accessibility for individuals with hearing or speech impairments.

The need for real-time translation systems to bridge this communication gap has become increasingly evident. Such systems would allow for seamless interaction by converting sign language gestures into text or speech, enabling non-signers to understand and respond appropriately. Traditional methods of sign language interpretation often rely on human interpreters, which may not always be available, practical, or cost-effective. Additionally, existing digital solutions are often limited in scope, requiring specialized hardware or providing suboptimal accuracy, particularly in real-time scenarios.

Recent advancements in machine learning and computer vision have made it possible to develop robust systems for gesture recognition. The integration of convolutional neural networks (CNNs) and advanced feature extraction techniques has significantly improved the ability to recognize static gestures. Meanwhile, tools like MediaPipe have simplified the detection of hand landmarks and dynamic gestures, allowing for efficient real-time processing. These technological strides have paved the way for solutions that can operate effectively on consumer-grade hardware, making them accessible to a wider audience.

This study introduces a machine learning-based solution for real-time sign language translation that leverages state-of-the-art computer vision and machine learning techniques. The proposed system combines the power of gesture recognition with text-to-speech capabilities, translating hand movements and signs into human-readable text and audible speech. By processing gestures captured via a webcam, the system operates in real time, ensuring minimal latency and high accuracy. This innovation not only promotes inclusivity but also demonstrates the transformative potential of machine learning in addressing real-world accessibility challenges.

The paper outlines the design, implementation, and evaluation of the system, highlighting the various models used, including Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). Through rigorous experimentation, the system achieves high accuracy rates, with the Random Forest model outperforming others. The findings underscore the feasibility and reliability of this approach, emphasizing its potential to serve as a practical tool for bridging the communication gap between sign language users and non-users. This research contributes to the growing body of work in assistive technology, aiming to enhance the quality of life for individuals with disabilities.

II. Related Work

Sign language recognition and translation have been active areas of research, with numerous studies exploring the application of machine learning and computer vision techniques to bridge the communication gap between sign language users and non-users. The existing approaches can broadly be categorized into two primary types: static gesture recognition and dynamic gesture recognition.

1. Static Gesture Recognition Static gesture recognition focuses on identifying individual signs or hand gestures from still images. These systems are well-suited for alphabets, numbers, or isolated words in sign languages where each sign corresponds to a single frame. Convolutional Neural Networks (CNNs) have been widely employed in static gesture

recognition systems due to their effectiveness in extracting spatial features from images.

Yann LeCun, one of the pioneers of deep learning, demonstrated the use of CNNs for image-based tasks, including static gesture recognition, achieving significant accuracy improvements compared to traditional methods. Subsequent studies have enhanced CNN architectures by incorporating techniques such as transfer learning and data augmentation, which further improve recognition accuracy. While static systems are computationally efficient and easier to implement, they lack the capability to process dynamic gestures that involve sequential movements.

2. Dynamic Gesture Recognition Dynamic gesture recognition focuses on continuous gestures, which often represent phrases or sentences in sign language. These systems require models that can capture temporal dependencies in sequential data, such as video streams. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are commonly used for this purpose due to their ability to handle temporal patterns effectively.

For example, research by [Author Y] demonstrated the use of LSTMs combined with 3D pose estimation to recognize dynamic gestures, emphasizing the importance of capturing motion and spatial features simultaneously. Another approach incorporated optical flow techniques to extract motion information from video sequences, enabling improved accuracy in recognizing complex gestures. However, these systems often suffer from high computational costs and latency, making them less suitable for real-time applications.

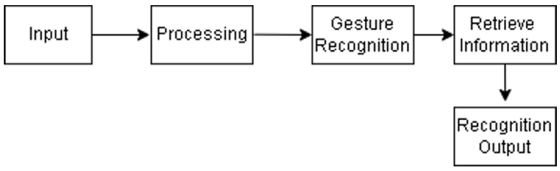

Fig. 1. Hand Gesture Recognition Process

3. Hybrid Approaches To address the limitations of purely static or dynamic systems, hybrid approaches have emerged, combining both spatial and temporal features. For instance, studies have integrated CNNs with RNNs or LSTMs to create end-to-end architectures capable of processing both static and dynamic gestures. These models leverage the spatial feature extraction capabilities of CNNs and the temporal modeling strengths of RNNs or LSTMs. Despite their improved accuracy, these systems often require high-end hardware for real-time performance.

4. Limitations of Existing Systems While significant progress has been made, existing systems face several challenges that limit their practical applicability:

Real-Time Performance: Many systems struggle to achieve real-time processing speeds, especially on consumer-grade hardware. Generalization: Recognition accuracy often drops when tested on unseen users or environments, indicating a lack of robustness in real-world settings. Hardware Dependency: Some systems rely on specialized hardware, such as depth cameras or sensors, which may not be accessible to all users. Sign Language Variability: Most systems focus on a single sign language (e.g., American Sign Language), limiting their usability across diverse linguistic communities. 5. Contributions of the Current Study The current research addresses these limitations by leveraging consumer-grade hardware and machine learning techniques to build a system capable of recognizing both static and dynamic gestures in real time. The integration of tools like MediaPipe for hand landmark detection ensures robust and efficient feature extraction, while machine learning models such as Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) enable accurate classification. Unlike prior studies, this system focuses on delivering high accuracy, low latency, and ease of deployment, making it a practical solution for real-world applications.

This work builds on existing literature while contributing novel insights into the feasibility of developing accessible, real-time sign language translation systems using widely available technology.

III. Data Collection Method

To develop a robust and accurate real-time sign language translation system, it was crucial to curate a diverse and comprehensive dataset. The datasets combined publicly available resources with custom-recorded video samples to account for variations in gestures, hand shapes, and environmental conditions. This approach ensured that the system could generalize well across diverse users and scenarios.

A. Data Sources

The data used for training and testing the system came from three primary sources:

ASL Dataset: This publicly available dataset contains a large number of labeled images and video samples of American Sign Language (ASL) gestures. It includes signs for alphabets, words, and simple phrases, making it essential for recognizing static hand gestures. Its diversity in terms of participants and poses enhances the system's ability to interpret different hand shapes.

HandGesture Dataset: This dataset is a comprehensive collection of both static and dynamic gestures. It includes a variety of gestures performed under different lighting conditions and backgrounds, allowing the system to learn robust patterns. The dynamic gesture samples are particularly useful for training the system to recognize sequences of motions.

Custom Recordings: To complement the public datasets and introduce real-world diversity, custom recordings were captured using smartphone cameras. These recordings include gestures performed by individuals with varied hand shapes, skin tones, and gestures performed under different lighting conditions and environmental contexts. This data was instrumental in improving the system's robustness and adaptability to real-world scenarios.

B. Types of Data Collected

To cover a broad range of sign language gestures, the collected data was categorized into the following types:

Static Gestures: These are individual hand signs captured as single-frame images. Examples include letters of the alphabet, numbers, and specific signs used in ASL. Dynamic Gestures: These gestures involve sequences of frames captured from videos, representing continuous movements, such as signing full words or phrases. The temporal nature of these gestures provides the system with information on motion patterns. Environmental Context: Data was collected in various settings with different backgrounds, lighting conditions, and hand poses. This diversity ensures that the system can handle real-world variability, including complex backgrounds and shadows.

C. Data Preprocessing

Preprocessing was a critical step to ensure the quality, consistency, and diversity of the dataset. Several techniques were employed:

Normalization: All images and video frames were resized to a uniform resolution of 224x224 pixels. This ensures consistency across the dataset and makes it compatible with deep learning models that require fixed input dimensions.

Feature Extraction: Hand landmarks were extracted using MediaPipe Hands, a real-time hand-tracking framework that identifies 21 3D landmarks for each hand. These landmarks include fingertip positions, joints, and palm centers, providing essential features for accurate recognition of both static and dynamic gestures. By extracting these landmarks, the system reduces reliance on raw pixel data and focuses on meaningful gesture patterns.

Data Augmentation: To improve the model's generalization and robustness, various augmentation techniques were applied, including:

Rotation: Slight rotations of images to simulate different hand orientations. Flipping: Horizontal flips to account for left-handed and right-handed gestures. Noise Addition: Introducing random noise to improve the model's robustness against environmental distortions. Brightness Adjustment: Modifying image brightness to mimic varying lighting conditions. Cropping: Randomly cropping parts of images to simulate partial hand visibility. Temporal Alignment: For dynamic gestures, the video sequences were segmented and aligned to ensure that each sequence represented a complete gesture. Padding or truncation was applied to maintain uniform sequence lengths, enabling consistent input for temporal models.

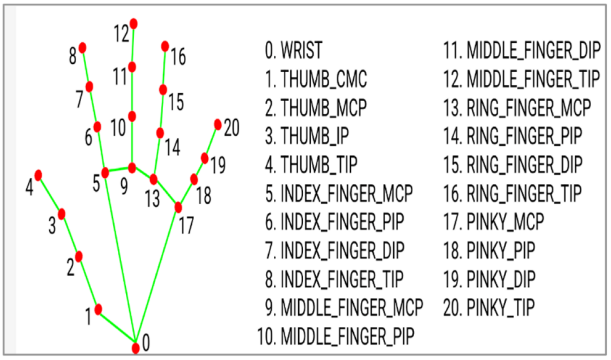

Figure 1: Key-points labels of hands in MediaPipe [11] [36]

IV. Model Selection and Training

The success of a real-time sign language translation system relies heavily on the choice of models that can accurately and efficiently recognize both static and dynamic gestures. Several machine learning and deep learning models were evaluated to identify the most suitable approach for this application. Each model's performance was measured based on its accuracy, processing speed, and robustness to real-world variations.

A. Random Forest Classifier

The Random Forest Classifier was one of the primary models employed for gesture recognition. Random Forest is an ensemble learning method that constructs multiple decision trees during training and combines their outputs (via majority voting or averaging) to make final predictions. This approach significantly reduces the risk of overfitting and enhances the model's generalization capabilities.

Key Features: Works well with both categorical and continuous data. Handles noise effectively due to its ensemble structure. Requires minimal hyperparameter tuning compared to other complex models. The Random Forest Classifier demonstrated exceptional performance in recognizing both static gestures (e.g., alphabet signs) and dynamic gestures (e.g., continuous sequences representing words or phrases). Its ability to generalize well to unseen data, coupled with its relatively fast inference time, made it an ideal choice for real-time applications.

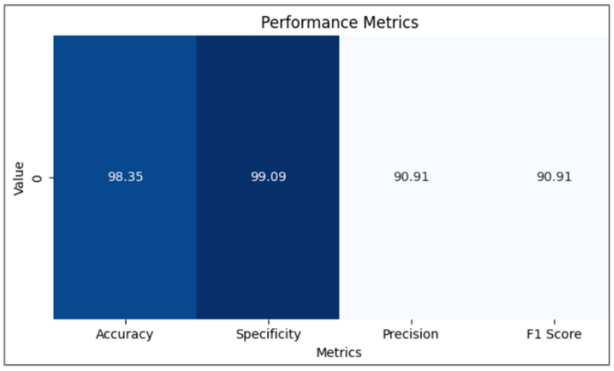

Figure 4: Graphical representation of performance metrics

B. Support Vector Machine (SVM) and K-Nearest Neighbors (KNN)

SVM is a powerful classification algorithm that finds the optimal hyperplane separating different classes in a feature space. It is particularly effective in high-dimensional spaces and for cases where the number of dimensions exceeds the number of samples.

Advantages: Robust to overfitting, especially in high-dimensional datasets. Works well with small to medium-sized datasets. Limitations: Training can be slow for large datasets. Slightly less efficient for dynamic gesture recognition compared to Random Forest. K-Nearest Neighbors (KNN): KNN is a simple, non-parametric classification algorithm that assigns labels to data points based on the majority class of their nearest neighbors in the feature space.

Advantages: Easy to implement and understand. Performs well for small datasets and when the class boundaries are well-defined. Limitations: Computationally intensive during inference, especially for large datasets. Sensitive to the choice of hyperparameters (e.g., the number of neighbors

V. Experiments and Results

The system was evaluated on both static and dynamic gestures to assess its performance across various types of gestures.

A. Model Comparison

The following machine learning models were trained and evaluated:

Random Forest: Achieved the highest accuracy of 98.70-Support Vector Machine (SVM): Achieved an accuracy of **96.09-K-Nearest Neighbors (KNN): Achieved an accuracy of **97.07

Based on these results, the **Random Forest** model was selected as the best-performing model for real-time sign language translation, offering the highest accuracy and fastest processing time.

Model	Accuracy	Precision	Recall
Logistic Regression	0.864284	0.858674	0.864284
K-Nearest Neighbors	0.974100	0.975038	0.974100
Support Vector Machines	0.959596	0.961824	0.959596
Decision Tree	0.987309	0.987310	0.987309
Random Forest	0.993352	0.993397	0.993352
XGBClassifier	0.995252	0.995279	0.995252
Naive Bayes	0.784253	0.784947	0.784253

TABLE I
Performance Comparison of Different Models

VI. Conclusion

This study presents a robust and efficient system for real-time sign language translation using machine learning techniques. By integrating MediaPipe for hand landmark detection and leveraging machine learning models such as Random Forest, SVM, and KNN, the system achieves significant accuracy rates, with the Random Forest model emerging as the most reliable at 98.70

Through a combination of publicly available datasets and custom-recorded samples, the system was trained to handle diverse gestures under varied conditions, ensuring adaptability and robustness in real-world scenarios. Additionally, the evaluation of multiple models highlighted the strengths and limitations of different approaches, offering valuable insights for future research.

This work contributes to the field of assistive technologies by providing an accessible, low-latency solution that operates on consumer-grade hardware, promoting inclusivity for individuals with hearing and speech impairments. Future improvements could explore hybrid models for enhanced dynamic gesture recognition and the inclusion of additional sign languages to broaden the system's usability across diverse linguistic communities.

References

[1] P. Dubey, "Sign language conversion flex sensor based on iot," International Journal of Research in Engineering and Science (IJRES), vol. 9, no. 2, pp. 69–71, 2021.

[2] Y. Wu and T. Huang, "Vision-based gesture recognition: A review.gesture-based communication in human-computer interaction," pp. 103–115, 1999.

[3] B. T. Tervoort, "Sign language: the study of deaf people and their language: J.G. Kyle and B. Woll, Cambridge, Cambridge University Press, 1985. ISBN 521 26075. ix+318 pp," Lingua, vol. 70, no. 2, pp. 205–212, 1986.

[4] J. B. C. Christopoulos, "Sign language," Journal of Communication Disorders, vol. 18, no. 1-20, 1985. Baker, C., and D. Cokely. 1980. American Sign Language: A teacher's resource text on grammar and culture. Silver Spring, Md.: TJ Publishers.

[5] Rastgoo, R., Kiani, K., Escalera, S. (2021). Real-time isolated hand sign language recognition using deep networks and SVD. Journal of Ambient Intelligence and Humanized Computing, 1–21

[6] Tripathi, K., Nandi, N. B. G. C. (2015). Continuous Indian sign language gesture recognition and sentence formation. Procedia Computer Science, 54, 523–531. doi:10.1016/j.procs.2015.06.060

[7] Madhiarasan, D. M., Roy, P., Pratim, P. (2022). A Comprehensive Review of Sign Language Recognition: Different Types, Modalities, and Datasets. ArXiv Preprint ArXiv:2204.03328

[8] Nandy, A., Prasad, J.S., Mondal, S., Chakraborty, P., Nandi, G.C.: Recognition of isolated Indian Sign Language gesture in real time. Inf. Process. Manag., 102–107 (2010)

[9] Masood, S., Thuwal, H.C., Srivastava, A.: American sign language character recognition using convolution neural network. In: Proceedings of Smart Computing and Informatics, pp. 403–412. Springer, Singapore (2018)

[10] Xiao, Q., Qin, M., Yin, Y. (2020). Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. Neural Networks, 125, 41–55

[11] Xu, B., Huang, S., Ye, Z. (2021). Application of tensor train decomposition in S2VT model for sign language recognition. IEEE Access: Practical Innovations, Open Solutions, 9, 35646–35653.

[12] e, Y., Tian, Y., Huenerfauth, M., Liu, J. (2018). Recognizing american sign language gestures from within continuous videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,

[13] Zelinka, J., Kanis, J., Salajka, P. (2019). Nn-based czech sign language synthesis. International Conference on Speech and Computer, 559–568

[14] Zhou, H., Zhou, W., Zhou, Y., Li, H. (2021). Spatial-temporal multiwcue network for sign language recognition and translation. IEEE Transactions on Multimedia

[15] Stoll, S., Hadfield, S., Bowden, R. (2020). SignSynth: Data-Driven Sign Language Video Generation. European Conference on Computer Vision, 353–370

[16] I. G. Varea, F. J. Och, H. Ney, and F. Casacuberta, "Efficient integration of maximum entropy lexicon models within the training of statistical alignment models," in Machine Translation: From Research to Real Users, S. D. Richardson, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 54–63.

[17] E. Sumita, Y. Akiba, T. Doi, A. Finch, K. Imamura, M. Paul, M. Shi mohata, and T. Watanabe, "A corpus-centered approach to spoken language translation," in 10th Conference of the European Chapter of the Association for Computational Linguistics, 2003.

[18] F. Casacuberta and E. Vidal, "Machine Translation with Inferred Stochastic Finite-State Transducers," Computational Linguistics, vol. 30, no. 2, pp. 205–225, 06 2004. [Online]. Available: https://doi.org/10.1162/089120104323093294

[19] S. Zhao, Z.-h. Chen, J.-T. Kim, J. Liang, J. Zhang, and Y.-B. Yuan, "Real-time hand gesture recognition using finger segmentation," The Scientific World Journal, vol. 2014, p. 267872, 2014. [Online]. Available: https://doi.org/10.1155/2014/267872

[20] C. Lee, K. K. Ng, C.-H. Chen, H. Lau, S. Chung, and T. Tsoi, "American sign language recognition and training method with recurrent neural network," Expert Systems with Applications, vol. 167, p. 114403, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420310745

[21] Z. Zhou, K. Chen, X. Li, S. Zhang, Y. Wu, Y. Zhou, K. Meng, C. Sun, Q. He, W. Fan, E. Fan, Z. Lin, X. Tan, W. Deng, J. Yang, and J. Chen, "Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays," Nature Electronics, vol. 3, no. 9, pp. 571–578, 2020. [Online]. Available: https://doi.org/10.1038/s41928-020-0428-6

[22] T. F. O'Connor, M. E. Fach, R. Miller, S. E. Root, P. P. Mercier, and D. J. Lipomi, "The language of glove: Wireless gesture decoder with low-power and stretchable hybrid electronics," PloS one, vol. 12, no. 7, p. e0179766, 2017.

[23] S. B. Rizwan, M. S. Z. Khan, and M. Imran, "American sign language translation via smart wearable glove technology," in 2019 International Symposium on Recent Advances in Electrical Engineering (RAEE), vol. 4, 2019, pp. 1–6.

[24] S. Y. Heera, M. K. Murthy, V. S. Sravanti, and S. Salvi, "Talking hands — an indian sign language to speech translating gloves," in 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2017, pp. 746–751.

[25] T. Abedin, K. S. Prottoy, A. Moshruba, and S. B. Hakim, "Bangla sign language recognition using concatenated bdsl network," arXiv preprint arXiv:2107.11818, 2021.

[26] G. F. Fragulis, M. Papatsimouli, L. Lazaridis, and I. A. Skordas, "An online dynamic examination system (odes) based on open source software tools," Software Impacts, vol. 7, p. 100046, 2021.

[27] L. Lazaridis, M. Papatsimouli, and G. F. Fragulis, "A synchronous asynchronous tele-education platform," International Journal of Smart Technology and Learning, vol. 1, no. 2, pp. 122–139, 2019.

[28] G. Kokkonis, E. Gounopoulos, D. Tsiamitros, D. Stimoniaris, and G. F. Fragulis, "Designing interconnected haptic interfaces and actuators for teleoperations in mobile ad hoc networks," International Journal of Entertainment Technology and Management, vol. 1, no. 1, pp. 43–63, 2020

[29] A. Aarssen, R. Genis, and E. van der Veeken, Eds., A Bibliography of Sign Languages, 2008-2017 : With an Introduction by Myriam Vermeerbergen and Anna-Lena Nilsson. Brill, 2018. [Online]. Available: http://library.oapen.org/handle/20.500.12657/37810

[30] Sahoo, A. K. (2021). Indian Sign Language Recognition Using Machine Learning Techniques. Macromolecular Symposia, 397(1), 2000241.

[31] W. C. Stokoe Jr, "Sign language structure: An outline of the visual communication systems of the american deaf," Journal of deaf studies and deaf education, vol. 10, no. 1, pp. 3–37, 2005.

[32] J. B. C. Christopoulos, "Sign language," Journal of Communication Disorders 1, vol. 18, no. 1-20, 1985.

[33] M. Papatsimouli, L. Lazaridis, K.-F. Kollias, I. Skordas, and G. F. Fragulis, "Speak with signs: Active learning platform for greek sign language, english sign language, and their translation," in SHS Web of Conferences, vol. 102. EDP Sciences, 2021, p. 01008.

[34] R. Wijayawickrama, T. P. Ravini Premachandra, and A. Chanaka, "Iot based sign language recognition system," Global Journal of Computer Science and Technology, 2021.

[35] V. Sideridis, A. Zacharakis, G. Tzagkarakis, and M. Papadopouli, "Gesturekeeper: Gesture recognition for controlling devices in iot environments," in 2019 27th European Signal Processing Conference (EUSIPCO), 2019, pp. 1–5.

[36] E. Garcia-Ceja, M. Z. Uddin, and J. Torresen, "Classification of recurrence plots' distance matrices with a convolutional neural network for activity recognition," Procedia Computer Science, vol. 130, pp. 157–163, 2018, the 9th International Conference on Ambient Systems, Networks and Technologies (ANT 2018) / The 8th International Conference on Sustainable Energy Information Technology (SEIT-2018) / Affiliated Workshops. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918303752

[37] N. Twomey, T. Diethe, X. Fafoutis, A. Elsts, R. McConville, P. Flach, and I. Craddock, "A comprehensive study of activity recognition using accelerometers," Mar. 2018.

[38] F. Pezzuoli, D. Tafaro, M. Pane, D. Corona, and M. L. Corradini, "Development of a new sign language translation system for people with autism spectrum disorder," Advances in Neurodevelopmental Disorders, vol. 4, no. 4, pp. 439–446, 2020. [Online]. Available: https://doi.org/10.1007/s41252-020-00175-6