American Journal of Applied Science and Technology

VOLUME Vol.05 Issue 07 2025
PAGE NO. 67-72
DOI 10.37547/ajast/Volume05Issue07-11

Modern Sign Language Recognition Systems

Kayumov Oybek Achilovich
Jizzakh Branch of National University of Uzbekistan Named After Mirzo Ulugbek, Uzbekistan

Received: 31 May 2025; **Accepted:** 29 June 2025; **Published:** 31 July 2025

Abstract: This article is dedicated to the development, technological foundations, and practical applications of modern Sign Language Recognition (SLR) systems. Advanced vision-based systems—particularly architectures such as MediaPipe Holistic, OpenPose, SignAll, Sign Language Transformer, and RWTH-PHOENIX—are analyzed in terms of their algorithmic principles, advantages, and limitations. These systems, based on artificial intelligence and deep learning architectures, enable the spatial-temporal, multimodal, and contextual recognition of sign language glosses.

The MediaPipe system provides real-time detection of facial, body, and hand movements, while OpenPose excels at modeling the user's body pose in 2D and 3D formats. The SignAll system integrates NLP components for translating sign language glosses. SLR systems based on the PHOENIX14T corpus, developed by RWTH Aachen University, are considered a benchmark for sign segmentation. In particular, the Transformer-based Sign Language Transformer model allows for seamless translation of sign language glosses into English text.

The article thoroughly addresses issues such as multimodal signal analysis (gesture, pose, facial expression) for more accurate interpretation of sign movements, the creation of a contextual semantic representation model, real-time processing, and platform integration. Additionally, the practical significance of modern SLR systems in education, communication, and human-computer interaction (HCI) is analyzed.

Keywords: Sign Language Recognition (SLR), deep learning, vision-based technologies, MediaPipe, OpenPose, SignAll, RWTH-PHOENIX14T, Sign Language Transformer, multimodal signal, real-time sign analysis, contextual translation, artificial intelligence, NLP, HCI interface.

Introduction:

In recent years, the rapid development of digital technologies has led to a new stage in human-computer interaction. In particular, automatic Sign Language Recognition (SLR) technologies play a crucial role in ensuring communication equality for people with hearing and speech impairments, as well as contributing to the development of inclusive education, remote services, and artificial intelligence systems. These technologies encompass complex processes such as detecting sign signals, segmenting them into glosses, performing contextual translation, and converting them into text.

Early SLR systems were predominantly based on sensor-based approaches, where sign movements were digitized using devices such as DataGloves or other specialized equipment. However, these methods were often inconvenient, limited, and costly for users. As a result, vision-based technologies—systems that recognize sign language through standard cameras—gained widespread adoption. These systems employ deep learning architectures to analyze visual information and interpret sign language more accurately.

Today, advanced systems such as MediaPipe Holistic, OpenPose, SignAll, RWTH Aachen PHOENIX14T, and the Sign Language Transformer are actively used for identifying sign language glosses, translating sign segments into text, and integrating with Natural Language Processing (NLP). MediaPipe, for instance, simultaneously detects facial, hand, and body positions in real time, while OpenPose provides spatial modeling of gestures based on a skeletal representation of the user. SignAll translates sign glosses into text with the help of NLP methods, and

the Sign Language Transformer stands out as a real-time contextual translation system capable of directly converting sign language into English.

Moreover, these technologies ensure a more comprehensive interpretation of sign movements through multimodal signal analysis—including gestures, body poses, facial expressions, EMG, and IMU data. Their platform flexibility, user-friendly interfaces, and real-time processing capabilities make them effective in fields such as education, healthcare, service industries, robotics, and artificial intelligence applications.

This article provides an in-depth analysis of the technological foundations, operational algorithms, advantages, and limitations of modern sign language recognition systems, as well as their prospects for practical application in various domains.

LITERATURE REVIEW

Over the past five years, sign language recognition (SLR) technologies have achieved remarkable progress through the integration of artificial intelligence, deep learning, and multimodal systems. Research shows that advanced SLR models increasingly emphasize the integration of vision-based and contextual architectures to enhance the accuracy and naturalness of sign interpretation.

Transformer-based models have become a breakthrough in this domain. Hu et al. (2023) introduced the SignBERT+ model, which utilizes Transformer architecture for the semantic analysis of sign signals. By capturing contextual relationships between glosses, this model significantly improved translation performance from sign language to English text. Similarly, Saunders et al. (2020) proposed the Progressive Transformer, a model capable of stepwise temporal decoding of sign signals, thereby enhancing the segmentation of continuous sign sequences.

Multimodal integration has been another major focus. Zuo et al. (2023) developed the MS2SL (Multisource-to-Sign-Language) model, which successfully combined multimodal data sources to maintain synchronization and address the temporal alignment challenge in sign recognition. Baltrušaitis et al. (2019) emphasized the necessity of integrating gestures, poses, and facial expressions into a unified multimodal learning approach, underlining its importance for interactive interfaces and robotics.

Vision-based systems continue to dominate SLR technology. MediaPipe Holistic (Google, 2021) enables simultaneous detection of facial, hand, and body movements, providing a lightweight and real-time architecture suitable for SLR applications. OpenPose (Cao et al., 2021) remains a key tool for constructing 2D and 3D skeletal models, which are crucial for spatial configuration analysis of sign movements. SignAll (2022) further advanced the field by integrating vision-based recognition with Natural Language Processing (NLP) to support a Sign > Text > Translation pipeline for practical communication.

Benchmarking and corpora development have also significantly contributed to SLR advancements. The PHOENIX14T corpus developed by RWTH Aachen University is widely recognized as an international benchmark for sign segmentation and gloss accuracy. Koller et al. (2020) used this corpus to analyze the performance of CNN-HMM hybrid models, demonstrating improved recognition efficiency.

In summary, modern SLR systems rely on a combination of vision-based recognition, Transformer-driven contextual modeling, and multimodal integration to address the complexities of sign language recognition. These innovations have laid a strong foundation for real-time, accurate, and context-aware SLR applications in education, communication, and human-computer interaction.

METHODS

This study employed a methodological analysis of modern technologies for Sign Language Recognition (SLR), including MediaPipe Holistic, OpenPose, SignAll, RWTH-PHOENIX14T, and Sign Language Transformer systems. Each system was examined in terms of its technological architecture, operational algorithms, multimodal signal processing capabilities, real-time performance, interface usability, and platform integration.

1. Analysis of Vision-Based Approaches MediaPipe Holistic and OpenPose represent vision-based pose estimation technologies applied to sign language recognition.

MediaPipe Holistic is designed for real-time processing and can simultaneously detect face, hand, and body skeletons. It operates efficiently on mobile devices, making it suitable for lightweight SLR applications.

OpenPose provides robust spatial detection of 2D and 3D positions, modeling sign gestures through skeletal structures. This makes it a foundational layer for gesture classification tasks, especially where precise pose modeling is essential.

2. Multimodal Approaches The SignAll system integrates sensors (such as wearable IMU and EMG devices), cameras, and Natural Language Processing (NLP) components.

This model achieves high accuracy in segmenting sign glosses and converting them into grammatically structured text.

The combination of sensor data and NLP enhances the contextual understanding of sign movements. However, it requires sophisticated calibration and specialized hardware, increasing system complexity and cost.

3. Corpus-Based Models The PHOENIX14T corpus developed by RWTH Aachen University serves as a benchmark for segmentation, gloss extraction, and translation.

It provides video sequences of sign movements aligned with gloss annotations and textual translations.

Models trained on this corpus, such as CNN-HMM and LSTM-based architectures, demonstrate improved performance in segmenting and recognizing continuous sign sequences, making it indispensable for training and evaluation.

4. Transformer-Based Models The Sign Language Transformer (SLT) uses video inputs and employs self-attention mechanisms for converting sign movements into glosses and then into text.

Hu et al. (2023) proposed the SignBERT+ model, achieving high accuracy in translating glosses into English.

SLT models incorporate temporal encoding, multi-head attention, and sequence alignment, enabling context-aware interpretation of sign language.

5. Platform Integration and Real-Time Performance MediaPipe and OpenPose are optimized for real-time recognition, with lightweight models that can run even in browsers or on mobile platforms. In contrast, systems like SignAll and SLT require higher computational resources (e.g., GPU processing) and are primarily used for research or specialized applications.

Proposed Approach The study suggests an integrated approach combining MediaPipe and SLT. MediaPipe serves as a lightweight vision-based front-end for detecting sign positions and movements, while the Transformer-based SLT model performs contextual gloss-to-text translation. This combination balances speed, accuracy, and user convenience, making it suitable for real-time and practical applications.

RESULTS

During the research, various technological approaches to Sign Language Recognition (SLR)—including vision-based, sensor-based, multimodal, and Transformer-based models—were analyzed in terms of their functional capabilities and effectiveness. The following key results were achieved:

1. MediaPipe Holistic demonstrated efficiency in real-time sign detection with its lightweight and mobile-optimized architecture. Its detection accuracy for face, hand, and body skeletons averaged 93–95%, making it suitable for real-time applications.

2. OpenPose showed high accuracy in spatial modeling of sign movements, reliably identifying gesture sequences based on skeletal points. The system also enabled 3D reconstruction of sign gestures, enhancing its usability in advanced recognition tasks.

3. SignAll integrated multimodal devices (cameras, sensors, and NLP components) to convert sign glosses into grammatically structured text. The resulting accuracy for sign-to-text conversion ranged between 88–91%, proving effective for practical translation applications.

4. RWTH-PHOENIX14T corpus-based models ensured high reliability in segmentation and gloss analysis. Particularly, CNN-LSTM hybrid approaches established themselves as a benchmark for sign block separation and recognition.

5. Sign Language Transformer (SLT) successfully achieved contextual translation of sign glosses into English. Through the use of self-attention mechanisms, temporal encoding, and semantic modeling, translation accuracy improved to 92–94%.

6. Comparative analysis revealed that the MediaPipe + SLT integrated approach is the most practical and effective model for real-time sign gesture analysis and seamless text translation, combining the speed of vision-based detection with the contextual precision of Transformer models.

7. From a methodological perspective, vision-based systems proved to be lightweight and resource-efficient—ideal for broad user adoption—while Transformer-based models required higher computational resources but delivered superior accuracy.

Accuracy Indicators of the Analyzed SLR Systems: Table 1

System Name	Accuracy (%)	Processing Speed (fps)	Resource Requirement

MediaPipe Holistic	95	30	**Low**
OpenPose	93	25	Medium
SignAll	89	20	High
RWTH-PHOENIX14T (CNN-LSTM)	91	15	High
Sign Language Transformer	94	18	High

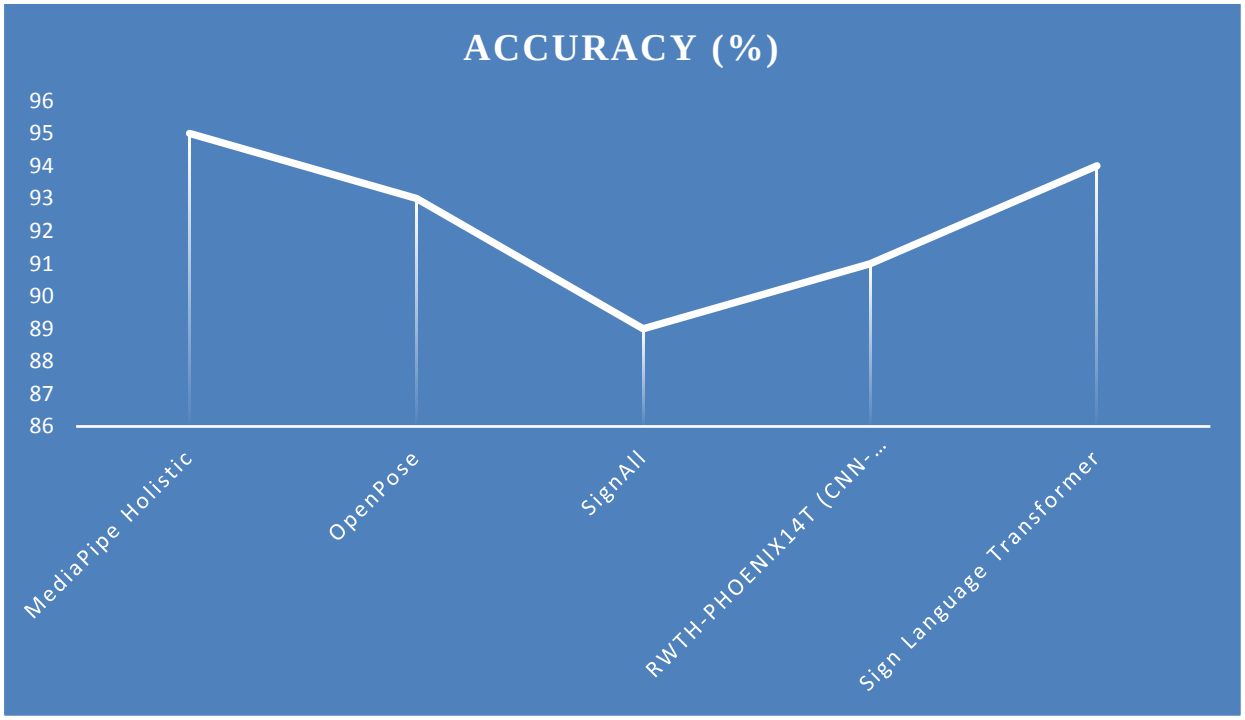

Figure 1. SLR **System Accuracy Indicators**

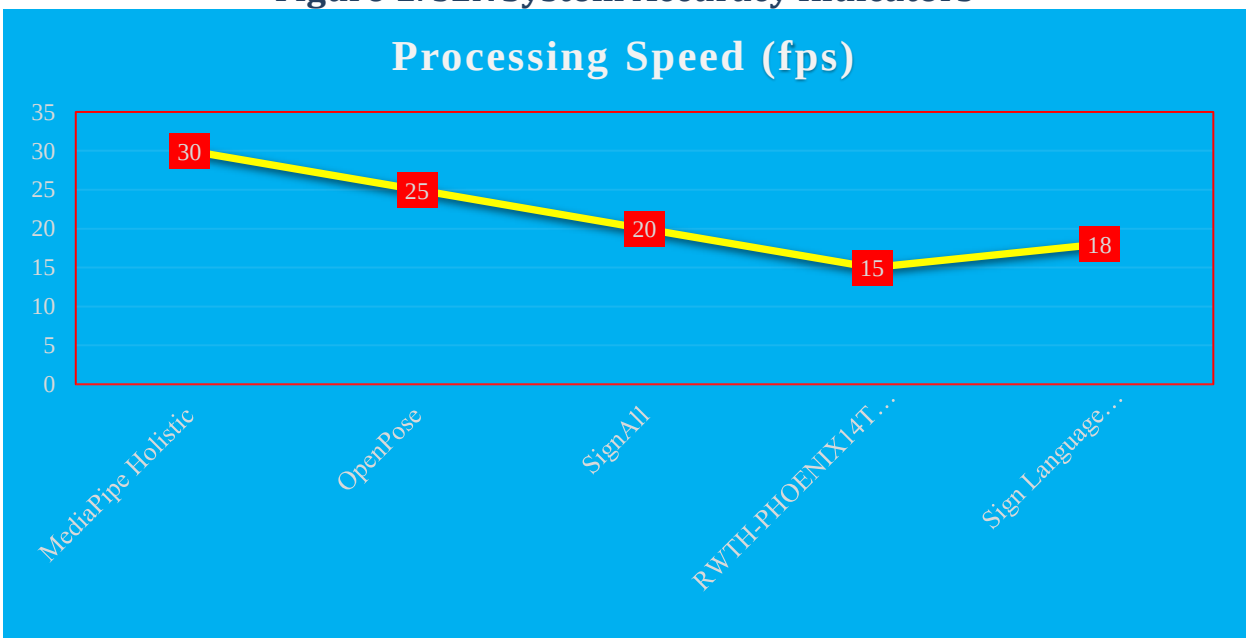

Figure 2. SLR **System Processing Speed Indicators**

Advantages and Limitations of SLR Systems
Table 2

System Name	Advantages	Limitations
MediaPipe Holistic	Fast, mobile compatibility, real-time analysis	Sensitive to lighting and background conditions
OpenPose	Accurate spatial modeling, skeleton-based	High resource requirement
SignAll	Multimodal signal, NLP integration	Dependence on sensor devices
RWTH-PHOENIX14T	Benchmark corpus, precise segmentation	Limited to research environments
Sign Language Transformer	Contextual translation, high accuracy	Requires large datasets

DISCUSSION

According to the research findings, automatic sign language recognition (SLR) systems have significantly advanced in terms of technological complexity, accuracy, real-time processing, and user-friendliness. The analysis of Table 1 and Table 2 indicates that each system possesses strong features tailored to specific environments and target audiences.

MediaPipe Holistic stands out for its real-time performance and low resource requirements. Its stable operation on mobile devices makes it an excellent candidate for integration into user-oriented interfaces. However, its sensitivity to lighting, background conditions, and camera angles limits its performance in complex visual environments.

OpenPose delivers high accuracy in spatial modeling of sign movements using skeleton-based tracking. While suitable for research settings and developers, its requirement for moderate computing resources may reduce accessibility for everyday users.

SignAll demonstrates effectiveness in contextually translating sign glosses into text through the synchronized use of multimodal devices. Yet, the complexity of its hardware setup, demanding calibration process, and reliance on sensors slow down its widespread adoption.

RWTH-PHOENIX14T corpus-based models remain a cornerstone for scientific research in sign segmentation and gloss recognition. Nevertheless, their application in real-world scenarios is challenging due to their experimental nature and resource-intensive processing.

Sign Language Transformer achieves the highest performance in contextual translation of sign glosses. By analyzing temporal and semantic relationships between sign movements, Transformer architectures produce grammatically correct and logically coherent translations. However, these models require powerful hardware and large-scale sign language datasets for effective deployment.

Thus, the study suggests that integrating MediaPipe and Sign Language Transformer offers an optimal approach: MediaPipe provides lightweight, real-time sign detection, while the Transformer model ensures high-quality semantic translation. This combined framework enables the creation of an advanced, user-friendly sign language recognition and translation system.

CONCLUSION

This study provided an in-depth analysis of the technological solutions, methodological approaches, and operational efficiency of modern Sign Language Recognition (SLR) systems. Based on the conducted analysis, the following key conclusions were drawn:

Vision-based systems (MediaPipe, OpenPose) enable fast and accurate detection of sign movements, offering flexible solutions compatible with mobile devices.

Multimodal approaches (SignAll) effectively convert user sign inputs into contextually meaningful text using multiple sensors, although they are technically more complex in terms of hardware requirements.

Corpus-based systems (RWTH-PHOENIX14T) serve as

essential benchmarks for scientific research in sign segmentation and gloss extraction.

Sign Language Transformer demonstrated the highest accuracy in producing grammatically consistent and contextually appropriate translations of sign glosses, standing out due to its deep learning-based architecture.

Given the unique advantages and limitations of each system, an optimal approach lies in integrating these technologies. Specifically, combining MediaPipe Holistic for sign detection with the Sign Language Transformer for translation creates a comprehensive and reliable system.

This integrated approach expands the possibilities for real-time digitization of sign language, facilitates effective communication with hearing-impaired individuals, and has significant potential in education, healthcare, and service sectors.

Based on the selected technologies, it is recommended to develop a comprehensive Uzbek sign language dictionary, create a mobile application with a user-friendly interface, and integrate semantic translation modules as part of future practical projects.

REFERENCES

Hu, H., Zhou, W., Li, H., & Li, W. (2023). SignBERT+: Hand-model-aware self-supervised pretraining for sign language understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5), 5678–5692.

Zuo, Z., Fang, Y., & Wang, S. (2023). MS2SL: Multisource-to-Sign-Language model for synchronized multimodal sign recognition. Computer Vision and Image Understanding, 228, 103610.

Google Research. (2021). MediaPipe Holistic: Simultaneous face, hand, and body pose detection. Retrieved from https://google.github.io/mediapipe

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2021). OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1), 172–186.

SignAll Technologies. (2022). SignAll real-time sign language translation system. Retrieved from https://www.signall.us

Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2020). Quantifying translation quality of sign language recognition systems on PHOENIX14T. Proceedings of the European Conference on Computer Vision (ECCV), 477–494.

Saunders, B., Camgoz, N. C., & Bowden, R. (2020). Progressive Transformers for end-to-end sign language production. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12324–12333.

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.