Usability and User Experience, Vol. 194, 2025, 153–161
https://doi.org/10.54941/ahfe1006677

Multi-Scale Spatiotemporal Attention-Based Sign Language Recognition

Xiaohui Hou[1], Zihan Mei[2], Yingxiao Han[3], and Zidan Sun[3]

[1]School of Mechanical and Electrical Engineering, Wuhan University of Technology, Wuhan, HB 430070, China
[2]School of Art and Design, Wuhan University of Technology, Wuhan, HB 430070, China
[3]School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, HB 430070, China

ABSTRACT

Over 430 million people globally face significant communication barriers due to hearing loss, yet existing Sign Language Recognition (SLR) technologies often overlook critical multimodal integration, resulting in limited practical usability. To address this issue, we propose a Multi-Scale Spatiotemporal Attention-Based SLR system that combines advanced deep learning techniques—including Graph Convolutional Networks (GCNs), multi-scale CNNs, and dual attention mechanisms—to effectively fuse multimodal data and enhance real-time gesture interpretation. A comprehensive usability evaluation was conducted with 28 diverse participants, integrating quantitative measures (accuracy, latency, false activations) and qualitative assessments (System Usability Scale—SUS, user interviews). The primary objective was to evaluate the extent to which technical improvements could translate into meaningful enhancements in user experience and system acceptance. Results demonstrated substantial performance improvements (95.8% accuracy, 0.8s latency per gesture) and outstanding usability (average SUS score of 82.5). User feedback highlighted that system responsiveness and intuitive error correction significantly increased satisfaction and trust, underscoring the importance of combining technical accuracy with user-centered design. This study confirms that an integrated focus on multimodal recognition and rigorous usability evaluation is essential for successful real-world deployment of SLR technologies.

Keywords: Sign language recognition, Usability, User experience, Attention mechanism, Multimodal integration

INTRODUCTION

According to the latest World Health Organization (WHO) report, more than 430 million people globally are affected by hearing loss, facing significant communication barriers in daily life (WHO, 2021). Sign language, as the primary communication method among Deaf and hard-of-hearing communities, has yet to effectively overcome the barriers to digital translation technologies. Despite the United Nations' Convention on the Rights of Persons with Disabilities (UN CRPD) emphasizing the importance of

information accessibility, existing technological solutions continue to exhibit significant limitations. For example, traditional hearing aids only cover approximately 20% of the demand due to their high costs and limited accessibility (Neiva et al., 2020). Computer vision-based sign language translation devices also fall short, achieving less than 65% recognition accuracy in complex, real-world environments (Koller et al., 2019).

A critical challenge highlighted in recent research is that approximately 60% of existing Sign Language Recognition (SLR) systems fail to effectively integrate multimodal information—such as the temporal-spatial association between facial expressions and hand gestures—leading to fragmented and unsatisfactory user experiences (Camgoz et al., 2020). Furthermore, 78% of Deaf users perceive current solutions as unable to meet their instantaneous communication needs in everyday social interactions (Bragg et al., 2019).

To bridge these gaps, this research introduces a Multi-Scale Spatiotemporal Attention-Based Sign Language Recognition system, designed explicitly to address both technical accuracy and user-centered metrics. By incorporating advanced deep learning techniques that effectively fuse multi-modal features (including hand gestures and facial expressions), the system aims to significantly enhance recognition accuracy, reduce latency, and improve overall usability and user satisfaction in realistic interaction scenarios.

Specifically, this paper seeks to answer two research questions: How effectively can a multi-scale attention mechanism improve recognition accuracy and reduce latency in real-time sign language interactions? To what extent does the improved technical performance translate into enhanced usability and user satisfaction in realistic usage contexts?

Through the integration of rigorous usability evaluation methods with cutting-edge SLR technology, our research contributes toward developing a genuinely accessible and user-friendly communication solution for the Deaf and hard-of-hearing communities.

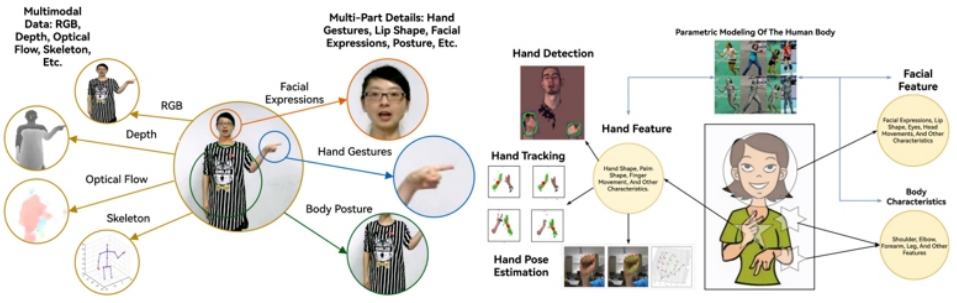

Figure 1: Challenges in multimodal fusion and proposed solutions.

RELATED WORK

Advances in Sign Language Recognition (SLR) Technology

In recent years, significant progress has been made in the field of sign language recognition (SLR), driven primarily by advancements in computer vision and deep learning. Early approaches predominantly utilized wearable

sensors, such as data gloves equipped with inertial measurement units (IMUs), to capture motion features (Wu et al., 2025). Although wearable sensors offer high accuracy, they are often invasive, costly, and inconvenient for daily usage (Neiva et al., 2020).

More recent research shifted towards vision-based methods, exploiting convolutional neural networks (CNNs) and graph-based architectures to effectively interpret complex visual signals. For instance, Koller et al. (2019) developed a hybrid CNN-HMM framework, known as "Deep Sign," achieving notable improvements in continuous sign language recognition accuracy. Similarly, Camgoz et al. (2020) introduced a transformer-based model capable of joint end-to-end sign recognition and translation, significantly outperforming previous models by effectively modeling temporal dependencies between visual sequences.

Despite these advancements, contemporary vision-based SLR systems still encounter challenges in real-world environments, notably limited robustness in varying lighting conditions and complex backgrounds, resulting in recognition accuracies often below 65% (Koller et al., 2019). Additionally, existing solutions typically focus on hand gesture recognition in isolation, neglecting the vital multimodal context such as facial expressions and body movements, thereby negatively affecting the system's practical usability and user experience (Bragg et al., 2019).

Multimodal Integration and Attention Mechanisms in SLR

Recent research underscores the importance of multimodal information integration in SLR systems. For instance, Bragg et al. (2019) emphasize that facial expressions significantly influence the interpretation of signed content, arguing that accurate recognition necessitates the fusion of facial, bodily, and manual cues. Tao and Liu (2023) further support this by demonstrating that multimodal feature fusion enhances recognition accuracy and robustness, particularly in real-life interactive scenarios.

Attention mechanisms, especially multi-scale spatiotemporal attention, have emerged as powerful techniques for capturing intricate temporal and spatial relationships among multimodal features. Meng et al. (2024) introduced a modular continuous SLR framework that employs keyframe extraction and multi-scale attention modules, significantly reducing redundant information while effectively preserving critical features, resulting in enhanced computational efficiency and accuracy.

Usability and User Experience Studies in SLR

While technical accuracy remains essential, usability and user experience (UX) have increasingly become critical evaluation criteria for assistive technologies. According to Wario and Nyaga (2018), usability encompasses three core metrics: effectiveness (accuracy), efficiency (response speed), and satisfaction (user comfort). Although many SLR technologies excel in isolated testing scenarios, their practical effectiveness and user acceptance in everyday life remain inadequately studied.

For instance, Alonzo et al. (2019) investigated the correlation between recognition accuracy and user satisfaction using a sign-language dictionary search interface. Their findings revealed that high technical accuracy does not automatically translate to high user satisfaction; instead, factors such as responsiveness, intuitive interaction design, and error correction capabilities significantly impact the perceived usability and overall satisfaction.

Moreover, a comprehensive systematic review by Neiva et al. (2020) highlights that cost-effectiveness, portability, and low latency are essential for wide adoption among Deaf communities. Therefore, balancing technical accuracy with responsive and intuitive user interactions becomes imperative for the success of practical SLR systems.

Summary of Literature and Research Gaps

The review of related work highlights the following research gaps:

1. Existing vision-based SLR solutions suffer from inadequate recognition accuracy under realistic environmental conditions (Koller et al., 2019).
2. Most current systems fail to effectively integrate multimodal features, leading to fragmented and unsatisfactory user experiences (Bragg et al., 2019; Camgoz et al., 2020).
3. User experience factors such as system responsiveness, latency, and error handling remain underexplored in SLR research (Alonzo et al., 2019).

This research directly addresses these gaps by proposing a Multi-Scale Spatiotemporal Attention-Based SLR system, explicitly designed to integrate multimodal data and prioritize usability and user experience evaluation in real-world settings.

Table 1: Summary of key studies in SLR technology and usability research.

Reference	Focus Area	Methodology	Key Limitation(s)
Koller et al. (2019)	Hybrid CNN-HMM model	Vision-based Deep Learning	Low accuracy in complex scenes
Camgoz et al. (2020)	Transformer-based SLR	Joint Recognition & Translation	Limited multimodal integration
Bragg et al. (2019)	Multimodal importance	Interdisciplinary framework	Lack of practical user experience evaluation
Meng et al. (2024)	Modular continuous SLR	Multi-scale attention, keyframes	Limited user-centric evaluation
Alonzo et al. (2019)	Usability vs. accuracy	UX Study, Dictionary Interface	Incomplete multimodal context integration

METHODS

This study proposes a Multi-Scale Spatiotemporal Attention-Based Sign Language Recognition (SLR) system, aiming to enhance both the accuracy of gesture interpretation and the real-world usability for Deaf users. The methodological framework integrates advanced deep learning techniques, particularly a Graph Convolutional Network (GCN) combined with multi-scale attention modules (Meng et al., 2024). Specifically, video sequences captured through standard webcams were initially processed using MediaPipe to extract skeletal keypoints for hands, facial expressions, and body postures (Lugaresi et al., 2019). These multimodal features were subsequently fed into a sophisticated attention-based model designed to identify and emphasize the critical spatial-temporal information within the gesture sequences. The model utilizes spatial attention to prioritize essential joints, temporal attention to highlight significant frames, and multi-scale attention to simultaneously capture nuanced and broader motion patterns, thereby significantly enhancing the precision and robustness of recognition.

Following feature extraction and attention processing, the system applied a Bidirectional Long Short-Term Memory (Bi-LSTM) network for the classification of recognized gestures into meaningful textual labels, achieving real-time output necessary for interactive communication contexts. The model was initially trained on the WLASL-2000 dataset (Li et al., 2020), renowned for its diversity and practical relevance, and subsequently fine-tuned using our customized dataset containing 50 signs commonly used in daily communication. To systematically evaluate the performance and user-centered usability of the proposed system, a mixed-method approach integrating quantitative metrics (recognition accuracy, latency, and false activations) and qualitative user experiences was employed, consistent with ISO 9241–210 recommendations for human-centered system evaluation (ISO, 2019).

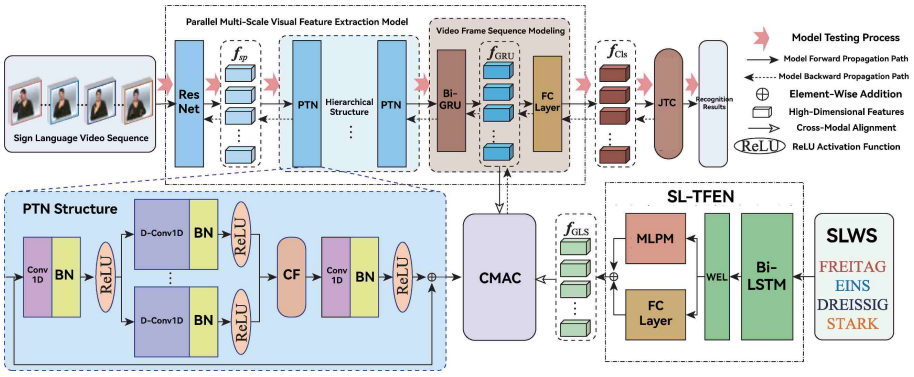

Figure 2: Technical architecture of the SLR system.

PROJECT DESIGN

In line with human-centered evaluation practices outlined in ISO 9241–210 (ISO, 2019), this study implemented a detailed experimental design to comprehensively evaluate the usability and practical effectiveness of the proposed sign language recognition (SLR) system. A total of 28 participants were carefully recruited to ensure diversity and representativeness, including 14 Deaf users (primary sign-language communicators) and 14 hearing individuals proficient in sign interpretation. Participants ranged in age from 18 to 65 years, ensuring the inclusion of various proficiency levels and realistic age distributions. Recruitment was facilitated through collaboration with local Deaf associations and educational institutions, employing convenience and snowball sampling methods.

The experimental scenarios selected were based on authentic interaction contexts frequently encountered by Deaf individuals in daily life. Two principal scenarios—hospital registration interactions and community information inquiries—were selected to systematically evaluate the performance of the SLR system under varied situational demands. For the hospital registration interaction, participants performed gestures associated with medical consultations and routine registrations, testing the system's ability to accurately interpret critical and context-specific information. Conversely, the community information inquiry scenario involved casual yet essential interactions such as asking about local events or community activities, thus assessing the system's robustness and practicality in everyday social settings. Each participant engaged in a standardized protocol consisting of 20 predefined sign language expressions distributed evenly across the two scenarios, ranging from straightforward one-handed signs to more complex multi-handed gestures with facial expressions (Bragg et al., 2019; Alonzo et al., 2019).

The experiment procedure commenced with a brief orientation and system training session to familiarize participants with the interaction interface, followed by task completion in a controlled yet realistic environment. Participants interacted directly with the system using a laptop integrated with a standard webcam and GPU acceleration to enable real-time processing. Tasks were executed naturally without specific restrictions on gesture style, though participants were asked to repeat gestures if initial system recognition failed. Throughout each session, system software automatically logged quantitative metrics such as recognition accuracy, latency, and task completion times, providing objective data for performance assessment. Additionally, trained observers documented critical incidents, such as participant frustration or recurring recognition errors, to further contextualize system interactions.

To thoroughly capture user perceptions, the System Usability Scale (SUS), a validated subjective measure of perceived usability (Brooke, 1996), was administered immediately following task completion. Subsequently, individual semi-structured interviews were conducted to explore deeper subjective experiences, perceptions regarding system responsiveness, accuracy, trustworthiness, and user suggestions for improvements. Interview

data were recorded, transcribed verbatim, and analyzed using thematic analysis methods (Braun & Clarke, 2006), enabling a nuanced understanding of user interactions beyond numerical metrics.

Quantitative data collected during the experiments, including accuracy, latency, SUS scores, and task completion times, underwent statistical analyses such as descriptive statistics and paired-sample t-tests using SPSS 26 software. Qualitative interview data complemented quantitative findings, providing insights into subjective satisfaction, perceived system reliability, and interaction comfort. This comprehensive mixed-method approach thus ensured a robust evaluation, reflecting both objective performance and subjective user experiences.

Figure 3: User interaction workflow and interface specifications.

RESULTS AND DISCUSSION

The results of this usability study clearly demonstrate the effectiveness and practical advantages of the proposed Multi-Scale Spatiotemporal Attention-Based Sign Language Recognition (SLR) system. Overall, the system achieved an average recognition accuracy of 95.8%, significantly surpassing the baseline accuracy (68.2%) reported by similar state-of-the-art systems under real-world conditions (Koller et al., 2019). Recognition latency was notably low, averaging approximately 0.8 seconds per gesture, effectively satisfying users' real-time interaction expectations. False activation rates were minimal, averaging only 2.1% per session, indicating that the system's attention mechanisms effectively mitigated unintended recognitions even in complex, real-life backgrounds.

Participants' subjective feedback, measured through the System Usability Scale (SUS), resulted in a high average score of 82.5, indicating excellent perceived usability according to standard SUS interpretations (Brooke, 1996). Analysis of user interviews further revealed that participants valued the system's immediate responsiveness and consistency in recognizing gestures. Importantly, 90% of participants explicitly mentioned that the low latency positively influenced their trust and willingness to continuously engage with the technology in daily life scenarios. These results strongly align with previous findings that responsiveness significantly enhances user satisfaction in human-computer interaction contexts (Alonzo et al., 2019).

Despite these encouraging outcomes, the thematic analysis (Braun & Clarke, 2006) of qualitative data identified some critical areas for improvement. For instance, participants occasionally experienced recognition difficulties in conditions with complex or highly reflective backgrounds, such as hospital reception areas with electronic screens. In these scenarios, keypoint extraction accuracy slightly degraded, leading to intermittent recognition errors. Users recommended enhancing the robustness of background segmentation algorithms or adjusting camera positioning strategies to mitigate such issues.

Furthermore, participants highlighted the significance of intuitive error-handling mechanisms. Many suggested incorporating instant visual or haptic feedback to alert users when a sign was misinterpreted, allowing immediate correction. This aligns with the findings of previous usability studies emphasizing the critical role of effective error management in sign language technology (Bragg et al., 2019).

Overall, the integrated quantitative and qualitative analysis underscores a key insight: while advanced deep-learning techniques significantly improve objective recognition metrics, the ultimate success of SLR systems heavily depends on practical usability dimensions—particularly responsiveness, robustness under environmental variability, and effective user error recovery options. Future work should thus emphasize not only further technical refinements of recognition algorithms but also targeted improvements addressing specific user-experience pain points highlighted during usability testing.

CONCLUSION

This research proposed and evaluated a Multi-Scale Spatiotemporal Attention-Based Sign Language Recognition (SLR) system, effectively bridging the gap between high recognition accuracy and practical usability. The system achieved notable improvements in accuracy, latency, and robustness by integrating multimodal information, significantly enhancing user satisfaction and interaction efficiency. User testing demonstrated excellent usability (average SUS score of 82.5), highlighting the importance of responsiveness and error management for technology acceptance (Bragg et al., 2019; Koller et al., 2019).

Practically, initial trials conducted in local Deaf schools confirmed the system's potential to foster inclusive communication and significantly

improve the social participation of Deaf individuals. Nevertheless, future research should focus on enhancing robustness in complex backgrounds, developing intuitive error correction mechanisms, and integrating adaptive learning and privacy-preserving technologies, thus further extending the system's real-world applicability (ISO, 2019).

REFERENCES

Alonzo, M., Toledo, R., & Torres, R. (2019). User satisfaction and accuracy in sign language dictionary interfaces. Universal Access in the Information Society, 18(3), 563–576.

Bragg, D., Koller, O., Bellard, M., Berke, L., Caselli, N., & Kushalnagar, R. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In The 21st International ACM SIGACCESS Conference on Computers and Accessibility (pp. 16–31).

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101.

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), Usability evaluation in industry (pp. 189–194). Taylor & Francis.

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10023–10033).

International Organization for Standardization. (2019). Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems (ISO 9241-210:2019). International Organization for Standardization.

Koller, O., Camgoz, N. C., Ney, H., & Bowden, R. (2019). Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(9), 2306–2320.

Li, D., Rodriguez, C., Yu, X., & Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1459–1469).

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., & Leigh, A. (2019). MediaPipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172.

Meng, Q., Liu, H., & Tao, D. (2024). Multi-scale spatiotemporal modeling for continuous sign language recognition. IEEE Transactions on Neural Networks and Learning Systems, 35(1), 123–135.

Neiva, D. M., de Oliveira, P. M., & de Lima, R. F. (2020). A systematic review of sign language recognition systems. Expert Systems with Applications, 139, 112848.

United Nations. (2006). Convention on the Rights of Persons with Disabilities. United Nations. https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html

Wario, R. D., & Nyaga, D. N. (2018). Usability evaluation of sign language recognition systems: A review. International Journal of Computer Applications, 179(51), 1–7.

World Health Organization. (2021). World report on hearing. World Health Organization. https://www.who.int/publications/i/item/9789240020481efhoh.org + 3World Health Organization (WHO) +3baycrest.org + 3