Human Sign Language Recognition System

Gitesh pal, Deepak kumar, Anshul Lakhera, Asst. Prof. Tanya Sharma
Department of Information & Technology
Indraprastha Engineering College, Uttar Pradesh

ABSTRACT

Human Sign Language Recognition (HSLR) systems aim to bridge communication gaps between the deaf and hearing communities by accurately interpreting sign language gestures into written or spoken language. Such systems involve advanced computer vision, machine learning, and natural language processing techniques to capture, process, and recognize sign language gestures from human users in real time. The development of HSLR systems presents unique challenges, such as differentiating similar gestures, recognizing subtle finger and hand movements, and managing variations in sign language across regions. This paper proposes a comprehensive approach to HSLR by integrating deep learning techniques, including convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) for sequence processing. The proposed HSLR system could significantly impact accessibility for the deaf community, enhancing inclusivity in various settings such as education, customer service, and public communications.

I. INTRODUCTION

Sign language serves as a vital communication tool for individuals who are deaf, hard of hearing, or mute, allowing them to convey information and interact meaningfully with others. Despite its importance, a significant barrier exists for widespread understanding, as most people lack the specialized knowledge to interpret sign language [1, 3]. Traditionally, human interpreters have been employed to facilitate communication, but this approach can be costly and logistically challenging, limiting the accessibility of sign language in everyday settings [2, 4].

Advances in technology have given rise to various approaches for Sign Language Recognition (SLR), including the use of wearable IoT sensors, data gloves, and vision-based systems [2, 5, 7]. While data gloves and wearable sensors can provide reliable recognition accuracy, they require users to wear additional hardware that may interfere with natural movement, making them impractical for everyday communication [6, 7]. Vision-based SLR, on the other hand, utilizes computer vision and deep learning to interpret gestures without additional devices, offering a non-invasive and convenient solution [4, 6].

Recent developments in artificial intelligence and computer vision have enabled the use of Convolutional Neural Networks (CNNs) for feature extraction in SLR [4, 5]. CNNs excel in image-based recognition tasks, yet their computational demands pose challenges for real-time applications [8, 10]. Addressing these limitations, this study proposes a lightweight, computer vision-based SLR model designed to support real-time communication. Leveraging the American Sign Language (ASL) alphabet and commonly used phrases, this model provides an accessible solution for communication across different user groups, including deaf, mute, and visually impaired individuals [4, 9].

In addition to CNNs, Long Short-Term Memory (LSTM) networks are integrated into our framework to improve the temporal accuracy of gesture recognition. While CNNs effectively extract spatial features from individual frames, LSTMs specialize in handling sequential data, allowing them to capture the dynamic nature of sign language gestures over time [4, 8]. By leveraging an LSTM layer after the CNN, our model can analyze the flow of gestures in a video sequence, distinguishing similar signs based on subtle temporal differences [8]. This combination of CNN and LSTM provides a robust solution that maintains high accuracy even in complex, multi-frame gestures [7, 8].

Our system, built on Mediapipe for feature extraction and a random forest classifier for gesture recognition, is designed for efficiency and ease of deployment [6]. By integrating speech-to-text, text-to-speech, and auto-completion features, the framework enhances usability, enabling seamless communication without the need for interpreters [9, 10]. The following sections will discuss related work, outline the methodology, and evaluate the model's performance, highlighting its potential to improve accessibility in real-time communication scenarios [3, 6, 9].

II. LITERATURE SURVEY

Over the past decade, significant advances have been made in the field of Human Sign Language Recognition (HSLR), particularly with the evolution of wearable sensors, computer vision, and deep learning techniques. Between 2014 and 2018, wearable sensor-based systems were predominant in the research landscape. These systems utilized data gloves and electromyography (EMG) sensors to capture hand gestures. For instance, a wearable data glove with optical sensors was proposed, which demonstrated promising accuracy in recognizing hand gestures for sign language translation [1]. Similarly, another study focused on using EMG sensors to capture muscle activity in the hand for prosthetics and sign language gesture recognition, providing accurate gesture recognition [2]. These approaches, while accurate, required cumbersome hardware setups, limiting their widespread application.

With the advent of deep learning techniques, computer vision-based systems gained prominence for sign language recognition. One study applied deep learning methods to recognize mouth shapes for sign language translation, marking a significant leap in the integration of AI into HSLR systems [3]. Another study further advanced the field by introducing a neural network-based approach to translate sign language into text, integrating multiple neural networks for improved performance [4]. These early efforts laid the foundation for incorporating deep learning into HSLR but highlighted the need for better handling of temporal and continuous sign data.

The shift towards leveraging large-scale datasets and pre-trained models emerged as a solution to dataset limitations. One study applied transfer learning using deep convolutional networks to improve sign language recognition accuracy, particularly for languages with limited annotated datasets [5]. The use of transfer learning enabled models to generalize across various datasets, improving performance and reducing training time.

A key development in vision-based systems was the introduction of **Google Mediapipe** in 2020. Mediapipe's real-time hand tracking system allowed for efficient processing of hand landmarks, making it suitable for mobile applications [6]. Mediapipe’s efficiency was further utilized in research, where simpler classifiers, such as Random Forests, were integrated to optimize real-time applications on mobile devices.

Further research into the spatial-temporal features of sign language recognition led to more complex architectures. One study explored the use of 3D convolutional networks (3D-CNNs) with attention mechanisms to improve large-vocabulary sign language recognition, capturing both spatial and temporal aspects of gestures [7]. This approach demonstrated enhanced performance for recognizing a broader range of signs.

In 2022, a study conducted a comprehensive review of Transformer models in sign language recognition, identifying their advantages in handling long-term dependencies and capturing sequential relationships in sign language [8]. Transformers, with their self-attention mechanism, provided a novel and efficient approach to HSLR, offering potential improvements in accuracy and scalability.

Additional multimodal approaches have been explored to improve the robustness and accuracy of recognition systems. One study proposed an RGB-D-based system that combined RGB and depth data, using convolutional neural networks (CNNs) to enhance recognition in varying lighting conditions and hand postures [9]. By incorporating depth information, the system improved the understanding of hand gestures, especially in complex settings.

Lastly, an audio-visual fusion model was introduced that integrated both visual and auditory cues for enhanced recognition of sign language, particularly for sign languages that involve vocal expressions or sound-based cues [10]. This approach leveraged the combination of gesture and audio data, resulting in better comprehension and communication in certain sign language contexts.

Summary Of Literature Survey :

TABLE I. Comparative Analyses Of Literature Survey and Related work

YEAR	APPROACHES USED	MODEL ARCHITECTURE	ACCURACY	SHORTCOMINGS
2015	Data Gloves with Optical Sensors	Sensor-based Glove System	~90%	Intrusive hardware, requires precise sensor placement, limits natural hand movement, lacks scalability
2016	EMG Sensors on Arm	EMG Sensor System	~85%	Invasive setup, high dependency on correct sensor positioning, challenging for real-world use
2017	Vision-Based System	CNN	88% (static signs)	Limited to static gestures, computationally intensive, lacks temporal understanding for dynamic signs
2018	Vision-Based System	CNN + RNN (LSTM)	92%	High computational cost, challenging to deploy on low-power devices, latency in real-time applications
2019	Vision-Based System	CNN-LSTM Hybrid	93% (continuous)	High computational demand, challenges in real-time applications, requires large datasets for training
2020	Vision-Based with Transfer Learning	Pretrained CNN	~90%	Limited by pre-existing datasets, requires large-scale data for fine-tuning, not ideal for continuous signing
2020	Vision-Based Hand Tracking	Lightweight CNN + RF/SVM	~85-90%	Lower accuracy than deeper networks, struggles with complex gestures, can be affected by lighting conditions
2020	RGB-D System (Depth and RGB Cameras)	RGB-D based CNN	~91%	Requires specialized cameras, affected by lighting and depth inconsistencies, hardware limitations
2022	Vision-Based System	Transformer	94%	Computationally expensive, memory-intensive, not optimized for mobile or low-resource devices
2023	Audio-Visual Fusion	Audio-Visual Encoder	89%	Limited to gestures with audio cues, challenging in silent sign recognition, less practical for standard HSLR

III. RESULT

Results of the Studies :

1. **Improved Accuracy for Static and Dynamic Signs**: Vision-based approaches using CNNs and hybrid CNN-LSTM architectures showed strong accuracy in recognizing static and dynamic gestures, with models like **Camgoz et al. (2018)** achieving around 92% for continuous gestures.
2. **Efficient Real-Time Recognition**: Lightweight systems, such as **Google Mediapipe (2020)**, made real-time recognition more feasible by using efficient CNN models combined with simple classifiers like Random Forests. These models demonstrated reasonable accuracy (~85-90%) and were able to function on lower-powered devices.
3. **Enhanced Temporal Understanding**: Models combining CNNs with sequential networks like LSTMs or Transformers (e.g., **Huang et al. (2019)** and **Min et al. (2022)**) significantly improved the ability to recognize sequences in continuous signing, achieving over 90% accuracy.
4. **Robustness with Multimodal Data**: Approaches that incorporated depth or audio data, like **Lee et al. (2020)** and **Kim et al. (2023)**, added robustness to the recognition, with accuracy improvements seen in complex or low-light settings.

Key Challenges Of These Studies:

1. **High Computational Costs**: Deep learning models, especially CNN-LSTM hybrids and Transformers, are computationally intensive, making them difficult to deploy on mobile devices or in real-time scenarios without significant resources.
2. **Hardware Dependency and Intrusiveness**: Wearable sensor-based systems and RGB-D approaches require specialized hardware, which is often intrusive or inconvenient for users. Systems with data gloves or EMG sensors are uncomfortable for daily use, and depth sensors have limitations in certain lighting conditions.
3. **Limited Generalizability Across Languages**: Most models are trained on specific datasets for individual sign languages (e.g., American Sign Language), which restricts their application across different sign languages or dialects without additional retraining.

IV. CONCLUSION

The advancements in Human Sign Language Recognition (HSLR) research have successfully moved from hardware-intensive, intrusive systems toward more accessible, vision-based, and multimodal solutions. Early sensor-based approaches, while accurate, were limited by their lack of comfort and practicality for daily use. With the rise of deep learning, vision-based models have improved accuracy and made non-intrusive recognition possible, handling both static and dynamic gestures to some extent.

Future Scope Of Human Sign Language recognition System:

1. **Development of Lightweight Models**: Future research can focus on designing efficient architectures, such as MobileNet or TinyML-based models, that maintain high accuracy while being computationally lightweight. This would enable HSLR systems to run smoothly on mobile devices and embedded systems.
2. **Improved Temporal and Sequential Modeling**: Advanced techniques in sequential modeling, such as Transformers, could further enhance the accuracy of continuous signing recognition. Optimizing these models for real-time use with reduced latency would enable more seamless video-based communication.

V. REFERENCES

[1] **Li, R., & Zhu, Z. (2015).** *Hand gesture recognition using wearable data glove with optical sensors.* Journal of Robotics, 2015. [doi:10.1155/2015/965967]

[2] **Zhou, H., Chen, T., & Tong, K. Y. (2016).** *EMG-based hand gesture recognition with wearable sensor for active prosthesis.* Computers in Biology and Medicine, 76, 70-80. [doi:10.1016/j.compbiomed.2016.07.011]

[3] **Koller, O., Ney, H., & Bowden, R. (2017).** *Deep learning of mouth shapes for sign language.* In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017. [doi:10.1109/ICCV.2017.224]

[4] **Camgoz, N. C., Hadfield, S., Koller, O., & Bowden, R. (2018).** *Neural sign language translation.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7784-7793. [doi:10.1109/CVPR.2018.00813]

[5] **Zhao, S., Tian, L., & Dai, Z. (2020).** *Transfer learning for sign language recognition with deep convolutional networks.* IEEE Access, 8, 84744-84752. [doi:10.1109/ACCESS.2020.2992010]

[6] **Google Mediapipe (2020).** *Mediapipe hands: On-device real-time hand tracking.* Retrieved from https://google.github.io/mediapipe/solutions/hands

[7] **Huang, J., Zhou, W., & Li, H. (2019).** *Attention-based 3D-CNNs for large-vocabulary sign language recognition.* IEEE Transactions on Circuits and Systems for Video Technology, 29(9), 2822-2832. [doi:10.1109/TCSVT.2018.2869642]

[8] **Min, S., Seo, S., Kim, H., & Lee, J. (2022).** *Transformers in sign language recognition: A comprehensive review and benchmark.* ACM Computing Surveys, 54(6), 1-36. [doi:10.1145/3453443]

[9] **Lee, J., Kim, J., & Song, S. (2020).** *RGB-D hand gesture recognition using depth CNN.* Pattern Recognition Letters, 133, 233-239. [doi:10.1016/j.patrec.2019.12.010]

[10] **Kim, D., Park, J., & Choi, S. (2023).** *Audio-visual fusion model for sign language recognition.* IEEE Transactions on Multimedia, 25, 456-466. [doi:10.1109/TMM.2023.3103647]