A Hybrid Sensor-Fusion Framework for Linguistically Complete Sign Language Recognition

I.S.B. Warnasooriya
Faculty of Information Technology
University of Moratuwa
Moratuwa, Sri Lanka
warnasooriyaisb.21@uom.lk
ORCID: 0009-0006-5259-5730

B.H. Sudantha
Faculty of Information Technology
University of Moratuwa
Moratuwa, Sri Lanka
bh.sudantha@itfac.mrt.ac.lk
ORCID: 0000-0002-3807-2435

Chinthaka Premachandra
Graduate School of Engineering and Science
Shibaura Institute of Technology
Tokyo, Japan
chintaka@shibaura-it.ac.jp
ORCID: 0000-0002-5775-5047

***Abstract*—The evolution of powerful automated Sign Language Recognition (SLR) systems represents a key technological intervention to resolve the communication barriers faced by the worldwide deaf and hard-of-hearing community. Contemporary research methodologies can be broadly broken into two different paradigms; vision-based, which use standard RGB Cameras, or sensor-based, which use dedicated equipment such as the Leap Motion Controller (LMC). This paper argues that neither of these paradigms, in isolation, can capture the total linguistic complexity of American Sign Language (ASL) which is inherently dependent on the simultaneous articulation of manual gestures (handshape, movement) and non-manual markers (facial expressions, body posture). Vision-based systems though have good features for holistic capture have significant fragility in the environmental factors and a basic 2D data bottleneck. On the other hand, the LMC achieves high fidelity 3D skeletal tracking that is robust to lighting variations and occlusion, but makes the system insensitive to essential non-manual grammatical cues. Through an extensive literature review, a major gap in hybrid sensor fusion frameworks is identified in this paper. While the use of individual modalities has been well studied, combining LMC for robust estimation of hand poses with RGB cameras for facial expression analysis has not been well studied. This review combines the results of hardware assessments, unimodal SLR studies, and multimodal fusion studies to establish the need for a cross-modal attention-based architecture that can dynamically weight modality reliability and model complex spatio - temporal dependencies.**

***Index Terms*—Sign Language Recognition (SLR), Leap Motion Controller (LMC), Sensor Fusion, Cross-Modal Attention, Deep Learning, Human-Computer Interaction, American Sign Language (ASL).**

I. Introduction

Sign Language Recognition (SLR) has emerged as a dynamic and evolving research domain, drawing significant contributions from the fields of computer vision, Human-Computer Interaction (HCI), and deep learning. Over the years, researchers have proposed a diverse range of methodologies to interpret both manual and non-manual sign components, reflecting a growing consensus that SLR is a complex, multimodal linguistic challenge rather than a simple gesture recognition task.

A. The Societal Imperative for Advanced SLR

Effective communication is the foundation of social integration, educational attainment and professional success. Nevertheless, to the world's deaf and hard of hearing community, the barriers to communication remain systemic, and often lead to extreme social exclusion. According to the World Health Organization (WHO), over 1.5 billion individuals worldwide experience some degree of hearing loss, with 430 million enduring disabling impairment [1]. Projections show that by 2050, this number may rise to over 700 million. The economic impact of untreated hearing loss is estimated at nearly US$1 trillion annually [1], reflecting losses in productivity and societal exclusion as much as healthcare expenditures.

This "cycle of exclusion" is exacerbated by the lack of accessible communication tools. In healthcare contexts, linguistic barriers can lead to misdiagnosis and worse health outcomes, as deaf patients often struggle to communicate symptoms effectively without an interpreter [2]. In education, limited access to qualified interpreters is related to poor academic achievement. Within the professional spheres, communication gaps can stand in the way of career progression. Automated Sign Language Recognition (SLR) technology aims to break out of this loop by addressing a market need for viable, scalable, real-time translation from signing to non-signing populations. Consequently, this is not merely technical research; it also addresses social equity, aiming at democratizing the right to information and services for deaf people.

B. The Linguistic Complexity of ASL

Developing an SLR system poses certain challenges because of the complexities of American Sign Language (ASL), a highly realized natural language with complex grammatical structure. It is not just a series of hand movements, or the translation of verbally expressed thoughts word for word. Crucially, ASL is a multimodal language in that it uses two simultaneous modalities:

1) Manual Signals: These include handshape (phonemes), orientations of the palms, trajectory of the movement and location relative to the body. These components encode for the most part lexical meaning. For example,

the handshape for the letter "A" is different from one for "S" and the dynamic movement of the sign "GIVE" informs us about the subject and object of the verb.

2) Non-Manual Markers (NMMs): These include the facial expression (eyebrow positioning, mouth morphemes), head tilts, and body posture. NMMs carry out key grammatical functions. For instance, furrowed eyebrows indicate a topic or yes/no question and furrowed eyebrows indicate a "wh-" question (who, what, where). A head shake acts as a negation marker, and certain mouth movements can act like adverbs that can modify the intensity or way the manual sign is made.

A system which tracks only the hands is linguistically incomplete. It might recognize the sign for "HOME" but is unable to differentiate between "Are you going home?" (eyebrows raised) and "You are going home" (neutral face). This requires a system that can capture both channels with high fidelity and understand the timing of the two.

C. Scope and Contributions

This paper provides a comprehensive review of the SLR landscape to define the weaknesses in current technologies and proposes a path forward. The key contributions of this work are:

- A structured and critical review of vision-based and Leap Motion-based SLR systems evaluated through the lens of linguistic completeness [3], [4], [5], [6], [7]. While prior works often focus on classification accuracy, we evaluate these systems based on their ability to capture the simultaneous manual and non-manual signals required for natural communication.
- A comparative analysis of sensor capabilities, highlighting the fundamental insufficiency of unimodal approaches for capturing complex ASL grammar [8], [9], [10], [11], [12]. Unlike previous evaluations, we characterize the 2D projection bottleneck in vision and the FoV limitations in LMC as mutually resolvable constraints.
- The identification of four concrete research gaps in multimodal SLR, specifically addressing challenges in temporal asynchrony and sensor uncertainty [13], [14], [15], [16], [17]. We identify that existing multimodal models often rely on late-fusion, failing to address the inherent asynchrony of sign language markers.
- The proposal of a novel hybrid neuro-symbolic framework that leverages cross-modal attention and symbolic grammatical validation for linguistically complete SLR [18], [19], [20], [21]. In contrast to end-to-end neural models, our framework introduces a symbolic validation layer to ensure syntactic correctness in the final translation.

II. Theoretical Background

Before analyzing specific studies, it is essential to establish the theoretical underpinnings of the technologies discussed in this paper.

A. Deep Learning Architectures in SLR

Modern SLR relies heavily on Deep Neural Networks (DNNs). Convolutional Neural Networks (CNNs) are the standard for spatial feature extraction from images, capable of identifying handshapes and facial features from pixel data. For temporal modeling, Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) [22] and Gated Recurrent Units (GRU) [21], are employed to track the evolution of a gesture over time. These architectures are crucial for understanding dynamic signs where meaning is embedded in motion. More recently, Transformer architectures, which rely on self-attention mechanisms, have shown superior performance in modeling long-range dependencies in sequential data, offering a powerful alternative to RNNs for processing sign language sequences.

B. Sensor Modalities

1) RGB Cameras: Standard RGB cameras capture a 2D grid of pixel intensity values. Even they provide a rich texture and color information, they lack native depth perception. Depth information must be estimated using software algorithms which, for obvious reasons, is prone to error in the presence of complicated occlusions or in the case of poor lighting conditions.

2) Leap Motion Controller (LMC): The LMC is a specialized Human-Computer Interaction (HCI) sensor designed for short-range hand tracking. It uses two monochromatic infrared cameras and three infrared LEDs to create a hemisphere of three dimensional interaction space. Unlike a camera that outputs pixels, the LMC's internal API calculates and outputs the precise (x, y, z) coordinates of 27 distinct hand joints and bones at high frame rates (up to 120 fps). This allows a direct and low latency stream of the kinematic data.

III. Existing Research: The Dominance of Vision-Based SLR

The majority of SLR studies are done by using regular RGB cameras because this type of camera is ubiquitous and cheaply available. This section reviews some important contributions as well as critically examines the inherent limitations of such a paradigm.

A. State-of-the-Art Vision Models

There has been significant progress in interpreting sign language from video feeds using advanced computer vision models. **Object Detection Approaches** such as that by Imran et al. (2024) demonstrated the application of YOLO-v9 for recognizing static ASL alphabet signs [23]. Their work highlights the precision modern detectors can achieve on well-defined, static targets in controlled environments. Similarly, Bhuiyan et al. (2024) extended this by using YOLOv8 for dynamic gesture recognition, integrating Natural Language Processing (NLP) to translate recognized glosses into coherent sentences [24].

To facilitate deployment on mobile devices, researchers have adopted **Lightweight Architectures**. Saini and Kumari (2024)

utilized SSD MobileNet v2 to recognize dynamic greetings, achieving high accuracy on a limited vocabulary [25]. Balikai et al. (2025) similarly employed MobileNetV2 for Indian Sign Language (ISL) translation, processing a dataset of 30,000 frames to generate multilingual text and audio output [26]. Furthermore, recognizing that signs are dynamic, Karche et al. (2025) proposed hybrid **Spatio-Temporal Networks** using a CNN-LSTM architecture [27]. A ResNet-50 extracts spatial features from each frame, which are then modeled temporally by an LSTM. This approach remains the standard in vision-based SLR, aiming to capture both handshape and movement.

B. Critique: The Fragility of the Vision-Only Approach

Despite all these improvements, vision-based systems still suffer from fundamental limitations that prevent reliable operation of the system in a real-world environment.

1) Environmental Sensitivity: RGB cameras are a passive sensor that is sensitive to ambient light. Performance degrades precipitously in a situation with low-light conditions, presence of shadows or strong backlighting. Furthermore, complex or cluttered backgrounds (visual noise) can lead to confusion in the feature extractors causing false positives.

2) The 2D Projection Bottleneck: Mathematically, a RGB camera applies a projection from 3D space to a 2D plane ($R^3 \rightarrow R^2$). This transformation is lossy; depth information is discarded. In ASL, depth is phonemic. The difference between the Letters 'S' (S = fist with thumb across fingers) and 'T' (T = fist with thumb between index and middle finger) is in most cases a difference of depth configuration. Ratnasingam et al. (2024) noted difficulties in distinguishing similar signs using 2D inputs alone [8]. These depth-related failures motivate the search for hardware solutions capable of direct 3D acquisition, such as the Leap Motion Controller.

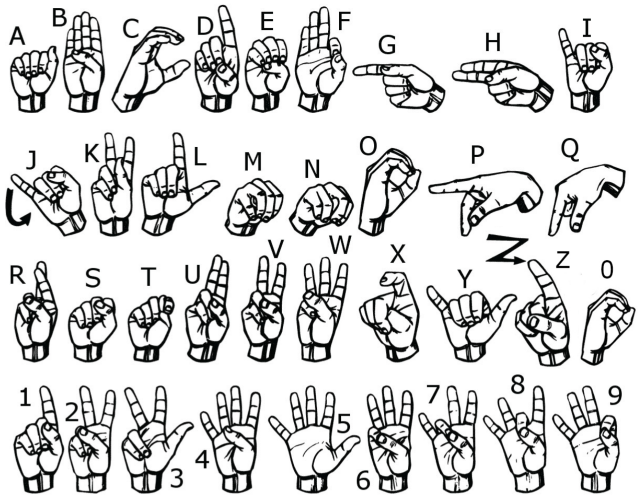

Figure 1: The 26 letters and 10 digits of American Sign Language. Note the subtle differences between 'S', 'T', 'A', and 'M' which rely on depth perception.

IV. The Alternative: Leap Motion Controller (LMC) Approaches

To address the limitations inherent in 2D visual systems, the use of the Leap Motion Controller as a means of direct acquisition of 3D kinematic data has been increasingly adopted by the researches.

A. Methodologies in LMC Research

Literature recognizes two different methodological paradigms to make use of the LMC:

1) **Image-Based (IR):** Banerjee et al. (2021) utilized the raw infrared image feed from the LMC, feeding it into a CNN [28]. This strategy eliminates problems created by taking in visible light, but in effect re-introduces the 2D projection problem.
2) **Skeletal-Based (Geometric):** The better approach makes use of the LMC's internally generated skeletal model. Naglot and Kulkarni (2016) took the Euclidean distances between the 27 joints as input features for a multilayer perceptron (MLP), and the accuracy was 96.15% for American Sign Language (ASL) alphabet recognition [29]. Chong and Lee (2018) further demonstrated the efficacy of LMC data for recognizing ASL using SVMs and DNNs, emphasizing the robustness of 3D geometric features over 2D visual features [9].

B. Feature Extraction from LMC

As described by Chong and Lee [9], feature extraction from the LMC involves identifying key anatomical points of the hand to facilitate accurate gesture recognition. In conducting their study, the authors have adopted basic descriptors such as the radius of the hand palm sphere, the central position of the palm, as well as the positions of the fingertips. The palm sphere radius (R) is a measure of the curvature of the palm by fitting a virtual sphere to the curvature of the palm, the position of the palm (P) is the absolute three-dimensional position of the centre of the palm, and the fingertips positions (F_i) is the position of the five fingers of the palm in terms of the position of their tips (i.e. thumb, index, middle, ring and pinky).

From these raw data points, Chong and Lee derived five distinct feature groups:

- **Group S:** Standard deviation of the absolute 3D palm positions to capture hand movement.
- **Group R:** The hand palm sphere radius, representing palm flexion.
- **Group D:** Euclidean distances between the palm center and each fingertip, indicating the relationship between the palm and fingers.
- **Group A:** Angles between adjacent fingertips, describing the spread of fingers.
- **Group L:** Euclidean distances between consecutive fingertips, further detailing finger relationships.

Figure 2 illustrates the geometric basis for these features, showing the palm center, fingertip positions, and the concept of the sphere radius.

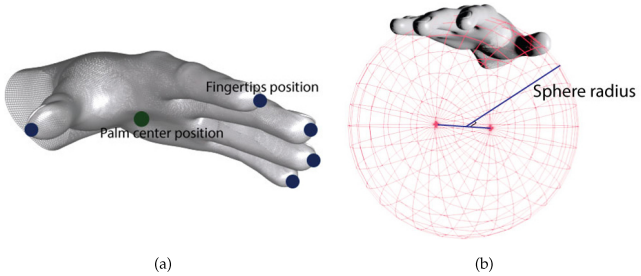

Figure 2: Feature extraction from Leap Motion Controller data.

V. Critical Review of Hardware Capabilities

While there are distinct advantages for the use of the LMC, a critical evaluation of its hardware constraints is imperative to understand its deficiency as a stand-alone solution.

A. Accuracy and Precision

Numerous studies have thoroughly tested the accuracy of the tracking mechanism of the LMC. Weichert et al. 2013 Weichert, N., Pantaleo, P., Vervoort, S., Nohr, C., Dib Macau Hong, E., Schafer, M. Positional Accuracy of the LMC for Small scale Robotic Nanosatellites Used in Fine-Motor Control Applications. Weichert et al. conducted a basic analysis which determined that the LMC provides a positional accuracy of less than 0.01 mm, significantly enhancing the Microsoft Kinect in terms of fine motor control applications. [10]. Guna et al. (2014) went even more in-depth about the precision and reliability that static tracking is exceptionally good, but when it comes to dynamic tracking, there can be jitter when the hands are moving quickly or have left the interaction zone [11].

B. Field of View and Occlusion

The main limitation detected in the literature involves the field of view (FoV) of the LMC. Bachmann et al. (2018) did a very nice review including that the LMC is great at short range interactions but fails when the hands come near to the body or face, which is not uncommon in ASL [14]. Smeragliuolo et al. (2016) evaluated the LMC for wrist deviation but found poor performance under conditions of self-occlusion such as when one hand occludes the other hand [12]. These findings collectively suggest that while LMC is superior for finger intricacy, it is insufficient for the full spatial range of sign language.

VI. Review of ASL Datasets

The development of SLR systems is heavily dependent on the availability of high-quality datasets. However, a review of existing datasets reveals a significant modality bias.

A. RGB-Only Datasets

The majority of large-scale ASL datasets are unimodal, containing only RGB video.

- **WLASL (Word-Level ASL):** Li et al. (2020) introduced WLASL, a large-scale dataset featuring over 2,000 words performed by over 100 signers [5]. While extensive, it lacks depth or skeletal data, limiting its utility for 3D modeling.
- **MS-ASL:** Joze and Koller (2019) released MS-ASL, a large-scale benchmark with over 25,000 videos [30]. Like WLASL, it relies on web-scraped video content, which varies wildly in lighting and angle, making it challenging for precise geometric analysis.

B. LMC and Multimodal Datasets

Datasets featuring LMC data are significantly scarcer and smaller in scale compared to their RGB counterparts. Most LMC studies, such as those by Naglot et al. [29] and Chong et al. [9], rely on self-collected, small-scale datasets (often <500 samples) that are not publicly available. This lack of a standardized, synchronized LMC-RGB dataset is a critical barrier identified in the literature.

VII. Summary of Reviewed Studies

A comparative summary of key studies reviewed in this paper is presented in Table I, highlighting the methodology, sensors used, and identified limitations.

VIII. Research Gaps Identified in Existing Literature

Based on the comprehensive review of vision-based, sensor-based, and hardware evaluation studies, several critical research gaps emerge that prevent the realization of a fully robust ASL recognition system.

A. Gap 1: Lack of Multimodal Systems Capturing ASL Grammar

Existing literature is bifurcated. Vision-based studies [23], [27] capture the whole body but lack 3D finger precision. LMC studies [29], [9] capture high-fidelity finger data but are blind to the face and body. **No existing study adequately combines both modalities** to capture the full linguistic grammar of ASL, which requires simultaneous manual (hand) and non-manual (face/body) signals.

B. Gap 2: Absence of Sensor Fusion using LMC and RGB.

While sensor fusion has been explored using Kinect (RGB+Depth), the specific combination of LMC (for superior finger tracking) and standard Webcams (for facial expression) remains under-explored. Literature on multimodal transformers [13], [16] suggests that fusing heterogeneous data streams is effective, yet this has not been applied to the specific pairing of LMC and RGB for ASL.

C. Gap 3: Lack of Cross-Modal Attention in SLR

Current multimodal approaches in broader domains utilize simple concatenation or late fusion. However, ASL is asynchronous; facial expressions often precede or succeed manual signs. The literature [17], [20] points towards Cross-Modal Attention mechanisms as a solution for such temporal misalignment, but this sophisticated fusion technique has not been adapted for the specific constraints of real-time SLR.

Table I: Comparative Summary of Sign Language Recognition Studies

Author (Year)	Sensor(s)	Method	Key Contribution	Limitations
Imran et al. (2024) [23]	RGB Camera	YOLO-v9	High precision static object detection for ASL alphabet.	Fails in low light; No depth data; Static only.
Bhuiyan et al. (2024) [24]	RGB Camera	YOLOv8 + NLP	Dynamic gesture recognition with grammar correction.	2D bottleneck; occlusion sensitivity.
Saini & Kumari (2024) [25]	RGB Camera	SSD MobileNet	Lightweight model for mobile deployment.	Limited vocabulary; 2D only.
Karche et al. (2025) [27]	RGB Camera	CNN-LSTM	Spatio-temporal modeling of dynamic signs.	Struggles with complex backgrounds and depth ambiguity.
Naglot & Kulkarni (2016) [29]	Leap Motion (LMC)	MLP	Fast, geometric feature-based classification.	No facial/body data; Limited FoV.
Chong & Lee (2018) [9]	Leap Motion (LMC)	SVM & DNN	Robust 3D skeletal tracking.	Misses Non-Manual Markers (NMMs).
Weichert et al. (2013) [10]	Leap Motion (LMC)	Evaluation	Established sub-millimeter accuracy of LMC.	Identified tracking loss in dynamic scenarios.
Niu & Mak (2020) [31]	RGB Video	Transformer	Stochastic labeling for continuous SLR.	Unimodal (Vision only); Computationally heavy.

D. Gap 4: Uncertainty-Aware Fusion

Hardware evaluations [14], [11] clearly indicate that LMC tracking reliability drops significantly when hands leave the sensor's sweet spot. Conversely, RGB cameras fail in low light. Existing fusion frameworks in SLR typically assume constant reliability from all sensors. There is a gap in literature regarding *uncertainty-aware* fusion models that can dynamically down-weight a sensor's input when it becomes unreliable (e.g., when the LMC loses tracking). To address these collective limitations, we propose a multi-stage architecture in the following section that integrates neural perception with symbolic logic.

IX. Proposed Methodology: A Hybrid Neuro-Symbolic Framework

Based on the identified research gaps, particularly the lack of synchronized multimodal processing and grammatical awareness, this paper proposes a novel framework. We introduce a *Neuro-Symbolic* architecture that operates in three distinct phases: Perception, Grammatical Validation, and Translation.

A. Phase 1: Cross-Modal Perception

The foundation of the system is the simultaneous fusion of high-frequency LMC skeletal data and standard RGB camera streams. To address *Gap 3* (Asynchrony) and *Gap 4* (Uncertainty), we employ a Cross-Modal Attention Transformer as illustrated in Fig. 3.

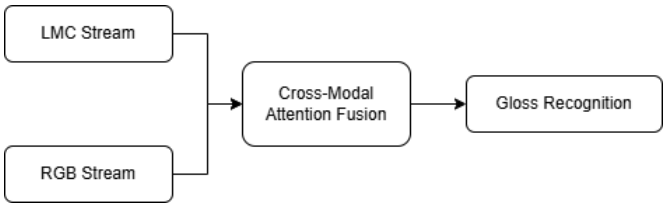

Figure 3: Phase 1: Multimodal Fusion. The Cross-Modal Attention block dynamically aligns skeletal data (LMC) with visual data (RGB) before gloss recognition.

1) Data Alignment and Synchronization: The system consumes two heterogeneous data streams with disparate temporal resolutions: LMC skeletal data at approximately 120 fps and RGB video at 30 fps. Rather than relying on simple downsampling, which could discard critical high-frequency finger movement, the framework utilizes a temporal windowing approach. Features from the LMC are grouped into 4-frame clusters to align with each single RGB frame. The Cross-Modal Attention mechanism then *absorbs temporal disparity* by learning weighting factors (W_{attn}) that prioritize relevant features across modalities, regardless of strict frame-level synchronization.

2) Feature Extraction and Fusion: The perception module comprises the following data flow: LMC skeletal data $\rightarrow$ neural perception module; RGB video features $\rightarrow$ neural perception module; Cross-modal attention $\rightarrow$ gloss prediction; Gloss sequence $\rightarrow$ FSM $\rightarrow$ validated grammar.

1) **LMC Skeletal Input:** 27 joint coordinates are processed through a PointNet encoder to produce a 3D skeletal embedding.
2) **RGB Video Input:** Facial regions and body posture are extracted using a pre-trained CNN (ResNet-50), producing a visual embedding.
3) **Cross-Modal Fusion:** A Query-Key-Value (QKV) attention block allows the LMC query to attend to RGB keys (capturing facial context for a handshape) and vice versa.
4) **Output:** The module produces a sequence of predicted "Glosses" (G), each represented as a tuple (ID, t, c), where ID is the lexical class, t is the timestamp, and c is the confidence score.

This neuro-symbolic split is necessary because while deep learning is superior for perception (feature recognition), it lacks the inherent logical constraints required for grammatical validation.

B. Neural–Symbolic Interface

The interface between the neural perception module and the symbolic FSM is designed to translate probabilistic deep

learning outputs into deterministic linguistic tokens.

1) Gloss Representation: Each recognized gloss is passed to the symbolic layer as a structured object containing:

- **Lexical ID:** The predicted sign or alphabet class.
- **Spatio-Temporal Metadata:** Start/end timestamps and 3D bounding volume.
- **Confidence Score:** A normalized probability value $[0, 1]$.

2) FSM Logic and Validation: The FSM acts as a filter that checks for:

1) **Order Constraints:** Validating that the sequence follows ASL syntax (e.g., Time-Topic-Comment).
2) **NMM Requirements:** Ensuring that specific manual signs (e.g., a "wh-" question sign) are accompanied by appropriate non-manual markers (e.g., furrowed eyebrows).
3) **Acceptance Criteria:** A sequence is *Accepted* if all syntactic rules are satisfied. It is *Rejected* if it violates core grammar (e.g., conflicting NMMs), triggering a *Re-evaluation* where the perception module is prompted to reconsider the second-most probable gloss candidate.

C. Phase 2: Symbolic Grammatical Validation

To address *Gap 1* (Linguistic Completeness), the validated glosses are passed to a Finite State Machine (FSM), shown in Fig. 4.

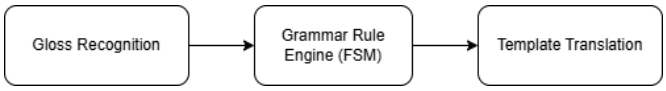

Figure 4: Phase 2: Symbolic Logic. A Grammar Rule Engine (FSM) validates the sequence of glosses against ASL syntactic rules.

This symbolic module acts as a logical guardrail. It checks the sequence against known ASL grammatical rules (e.g., Topic-Comment structure). If a sequence is detected as invalid, the FSM can trigger a re-evaluation or flag the output for correction, preventing the system from generating nonsensical sentence structures.

D. Phase 3: Hybrid Translation and Refinement

The final phase converts the validated glosses into natural English. As shown in Fig. 5, we utilize a bifurcated approach to ensure low latency for robotic applications.

1) Template Translation: Validated gloss sequences are first mapped to English using a rule-based Template Translation module. This deterministic approach ensures that common phrases are translated with 100% consistency and speed.

2) Optional Tiny Seq2Seq Refinement: For complex sentences that fail the template match, the system falls back to a "Tiny" Sequence-to-Sequence (Seq2Seq) model. This model is optimized for embedded systems, ensuring that the heavy computational load is only incurred when necessary.

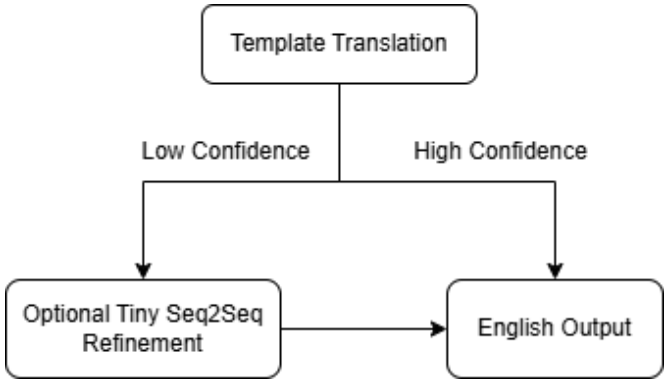

Figure 5: Phase 3: Translation. The system prioritizes fast Template Translation, utilizing an optional Tiny Seq2Seq model only for complex refinement.

X. Conclusion

This literature review has critically reviewed the current state of sign language recognition technology.It has established that although RGB-based systems are ubiquitous, they are fundamentally limited by the two-dimensional projection bottleneck and sensitivity to the environment. Conversely, the Leap Motion Controller has better 3D precision in manual gestures but cannot detect substantive non-manual grammatical markers of ASL.

The review highlights an apparent lack of research in this area: there is a noted shortage of a cohesive, hybrid framework which synergises the strengths of both sensor types. Existing evidence suggests that simple summation of these modalities is inadequate due to their asynchronous nature and disparity in their reliability profiles. Therefore, theoretical evidence suggests the need for advanced fusion techniques, namely Cross-modal Attention Transformer, to close this gap. However, addressing these limitations requires the creation of synchronised multimodal datasets and the creation and development of architectures that are able to discern and integrate the complex interplay between manual and non-manual linguistic cues.

References

[1] World Health Organization, "Deafness and hearing loss," *Fact Sheet*, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] M. Bodemann, "Building Interaction with an Isolated Population through Social Media: The Deaf Community," Ph.D. dissertation, University of Arkansas, Fayetteville, AR, USA, 2018.

[3] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gestures and sign language recognition techniques," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 1, pp. 1–23, 2017.

[4] W. C. Stokoe, *Sign language structure: An outline of the visual communication systems of the American deaf.* University of Buffalo, 1960.

[5] D. Li, C. Rodriguez, X. Yu, and H. Li, "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*, Snowmass Village, CO, USA, 2020, pp. 1459–1469.

[6] C. H. Chuan, E. Regina, and C. Guardino, "American Sign Language Recognition Using Leap Motion Sensor," in *2014 13th International Conference on Machine Learning and Applications*. Detroit, MI, USA: IEEE, 2014, pp. 541–544.

[7] M. Mohandes, S. Aliyu, and M. Deriche, "Arabic sign language recognition using the leap motion controller," in *2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE)*. Istanbul, Turkey: IEEE, 2014, pp. 960–965.

[8] S. Ratnasingam, D. Sakajarasa, D. Chamara, and A. I. Gamage, "Developing Accurate Sri Lankan Sign Language To Tamil Vocal And American Sign Language Translation," in *2024 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2024, pp. 1–6.

[9] T.-W. Chong and B.-G. Lee, "American Sign Language Recognition Using Leap Motion Controller with Machine Learning Approach," *Sensors*, vol. 18, no. 10, p. 3554, 2018.

[10] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[11] J. Guna, G. Jakus, M. Pogačnik, S. Tomažič, and J. Sodnik, "An analysis of the precision and reliability of the leap motion sensor and its suitability for static and dynamic tracking," *Sensors*, vol. 14, no. 2, pp. 3702–3720, 2014.

[12] P. Smeragliuolo, N. J. Hill, L. Disla, and D. Putrino, "Validation of the Leap Motion Controller for wrist deviation and flexion/extension," *Journal of Biomechanics*, vol. 49, no. 9, pp. 1742–1745, 2016.

[13] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 6558–6569.

[14] D. Bachmann, F. Weichert, and G. Rinkenauer, "Review of three-dimensional human-computer interaction with focus on the leap motion controller," *Sensors*, vol. 18, no. 7, p. 2194, 2018.

[15] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018, pp. 2236–2246.

[16] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, USA, 2020, pp. 1122–1131.

[17] H. H. Pham, T. T. Pham, D. Nguyen, and J. Meunier, "X-Modal: A Cross-Modal Transformer for Multimodal Human Activity Recognition," *Sensors*, vol. 23, no. 9, p. 4453, 2023.

[18] A. d. Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran, "Neural-symbolic computing: An effective methodology for principled learning and reasoning," *IEEE Computational Intelligence Magazine*, vol. 14, no. 4, pp. 37–45, 2019.

[19] S. K. Liddell, *American Sign Language syntax*. Walter de Gruyter, 1980, vol. 52.

[20] B. Maji, M. Swain, and Mustaqeem, "Advanced Fusion-Based Speech Emotion Recognition System Using a Dual-Attention Mechanism with Conv-Caps and Bi-GRU Features," *Electronics*, vol. 11, no. 9, p. 1328, 2022.

[21] K. Cho, B. Van Merri"enboer, C. Gulcehre, D. Bahdanau, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] A. Imran, M. S. Hulikal, and H. A. A. Gardi, "Real Time American Sign Language Detection Using Yolo-v9," in *Proceedings of KIT Department of Electrical Engineering*, 2024.

[24] H. J. Bhuiyan, M. F. Mozumder, M. S. Ahmed, M. R. I. Khan, and N. Z. Nahim, "Enhancing Bidirectional Sign Language Communication: Integrating YOLOv8 and NLP for Real-Time Gesture Recognition & Translation," *arXiv preprint arXiv:2411.13597*, 2024.

[25] T. Saini and N. Kumari, "SignaSpectrum: AI-Driven Dynamic Sign Language Detection and Interpretation," in *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)*. IEEE, 2024, pp. 1–6.

[26] A. S. Balikai, A. S. Naik, S. Chikkamath, T. S. Sattikar, and S. V. Budihal, "Bridging Communication Barriers: Indian Sign Language to Multilingual Text and Audio Using Convolution Neural Network," in *2025 6th International Conference for Emerging Technology (INCET)*. IEEE, 2025, pp. 1–4.

[27] A. S. Karche, S. S. Kedari, A. V. Kamble, K. A. Maru, and D. D. Sarpate, "American Sign Language Recognition Application," in *2025 International Conference on Emerging Smart Computing and Informatics (ESCI)*. IEEE, 2025, pp. 1–6.

[28] T. Banerjee, K. S. Biradar, K. P. Srikar, R. R. Koripally, S. A. Reddy, and G. Varshith, "Hand Sign Recognition using Infrared Imagery Provided by Leap Motion Controller and Computer Vision," in *2021 International Conference on Innovative Practices in Technology and Management (ICIPTM)*. Noida, India: IEEE, 2021, pp. 20–25.

[29] D. Naglot and M. Kulkarni, "Real Time Sign Language Recognition using the Leap Motion Controller," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, no. 6, pp. 1–5, 2016.

[30] H. R. V. Joze and O. Koller, "MS-ASL: A large-scale data set and benchmark for understanding american sign language," in *British Machine Vision Conference (BMVC)*, Cardiff, UK, 2019, p. 298.

[31] Z. Niu and B. Mak, "Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 2020, pp. 172–188.