

Comparison of Deep Learning Architectures for Sign Language Recognition: From Pixel-based to Holistic Keypoints on Web-based Applications

Niracha Chaiwong^{1*}

¹School of Information Technology and Innovation Computer Science and Information Technology

Department Bangkok University

*niracha.c@bu.ac.th

Abstract

This research aims to propose and evaluate the performance of deep learning models for the classification of 50 sign language glosses (WLASL-50) to enhance communication efficiency between hearing-impaired individuals and the general public. The comparative study investigates three distinct architectures: 1) a hybrid Convolutional Neural Network and Long Short-Term Memory model (YOLO-LRCN) leveraging raw pixel data, 2) a hand landmark-based model (MediaPipe Hands) integrated with LSTM, and 3) a holistic skeletal landmark model (MediaPipe Holistic) that fuses hand and upper-body pose coordinates with LSTM. Experimental results demonstrate that the holistic-based approach achieves superior performance, significantly increasing the Top-10 Accuracy from 45.98% in the baseline pixel-based model (YOLO-LRCN) to 81.17%. Furthermore, the holistic approach achieved a substantial Top-1 Accuracy of 44.84%. Beyond accuracy enhancements, the coordinate-based architectures exhibit a significant reduction in model complexity, featuring parameters approximately ten times fewer than those of pixel-based models (approx. 850 KB). This efficiency results in high-speed processing, with an average inference time of approximately 95 ms per prediction, and minimal computational resource requirements. The optimized model was deployed as a video-upload-based web application prototype using Streamlit, allowing users to submit recorded sign-language clips for near-instantaneous prediction. This study concludes that the holistic skeletal approach is a promising and effective strategy for developing lightweight sign language translation systems on resource-constrained devices, although further validation with a larger vocabulary, repeated experimental trials, and real users is required prior to practical deployment.

Keywords: Video-based Sign Language Recognition, Automatic Sign Language Processing, Long Short-Term Memory (LSTM) Networks, Skeleton-based Feature Extraction, Comparative Analysis of Pixel-based and Coordinate-based Data

การเปรียบเทียบประสิทธิภาพสถาปัตยกรรมเรียนรู้เชิงลึกสำหรับการจำแนกภาษามือจากการประมวลผลฟิกเชลภาพ สู่พิกัดโครงร่างแบบองค์รวมบนเว็บแอปพลิเคชัน

นิรชา ชัยวงศ์^{1*}

¹ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศและนวัตกรรม มหาวิทยาลัยกรุงเทพ

*niracha.c@bu.ac.th

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอและเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้เชิงลึกในการจำแนกภาษามือจำนวน 50 คำ (WLASL-50) เพื่อเพิ่มประสิทธิภาพในการสื่อสารระหว่างผู้พิการทางการได้ยินและบุคคลทั่วไป โดยดำเนินการทดสอบเปรียบเทียบผ่าน 3 แนวทางหลัก ได้แก่ 1) แบบจำลองโครงข่ายประสาทเทียมแบบคอนโวลูชันร่วมกับหน่วยความจำระยะสั้นยาว (YOLO-LRCN) ซึ่งใช้ข้อมูลฟิกเชลภาพโดยตรง 2) แบบจำลองที่ใช้พิกัดจุดเชื่อมต่อมือ (MediaPipe Hands) ร่วมกับ LSTM และ 3) แบบจำลองที่ใช้พิกัดโครงร่างแบบองค์รวม (MediaPipe Holistic) ซึ่งรวบรวมพิกัดมือและพิกัดร่างกายท่อนบนร่วมกับ LSTM ผลการทดลองพบว่าแบบจำลองที่ใช้พิกัดโครงร่างแบบองค์รวม (Holistic-based) ให้ประสิทธิภาพสูงสุดอย่างมีนัยสำคัญ โดยสามารถยกระดับค่าความแม่นยำสูงสุด 10 อันดับแรก (Top-10 Accuracy) จากร้อยละ 45.98 ในแบบจำลองดั้งเดิม (YOLO-LRCN) เพิ่มขึ้นเป็นร้อยละ 81.17 และให้ค่าความแม่นยำอันดับหนึ่ง (Top-1 Accuracy) สูงถึงร้อยละ 44.84 นอกจากนี้ความแม่นยำที่เพิ่มขึ้น แบบจำลองแบบพิกัดโครงร่างยังมีขนาดพารามิเตอร์ที่น้อยกว่าแบบจำลองฟิกเชลภาพถึง 10 เท่า (ประมาณ 850 KB) ส่งผลให้มีความเร็วในการประมวลผลสูง โดยมีเวลาประมวลผลเฉลี่ยประมาณ 95 มิลลิวินาทีต่อการทำนายหนึ่งครั้ง และใช้ทรัพยากรคำนวณต่ำ ผู้วิจัยได้นำแบบจำลองที่พัฒนาขึ้นไปประยุกต์ใช้ในรูปแบบเว็บแอปพลิเคชันต้นแบบด้วย Streamlit ในลักษณะแอปพลิเคชันเพื่อทำนายผลแบบเกือบทันที ซึ่งผลการศึกษานี้ชี้ให้เห็นว่าการใช้พิกัดโครงร่างแบบองค์รวมเป็นแนวทางที่มีศักยภาพและมีประสิทธิภาพสำหรับการพัฒนาระบบแปลภาษามือแบบเบาบนอุปกรณ์ที่มีทรัพยากรจำกัด แม้ยังต้องมีการทดสอบเพิ่มเติมกับคำศัพท์ที่หลากหลายขึ้น การทดลองซ้ำหลายรอบ และผู้ใช้งานจริงก่อนนำไปใช้งานจริง

คำสำคัญ: การจดจำภาษามือแบบวิดีโอ การประมวลผลภาษามืออัตโนมัติ โครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นยาว การสกัดคุณลักษณะจากจุดเชื่อมต่อร่างกาย การเปรียบเทียบข้อมูลฟิกเชลและพิกัด

1. Introduction

Effective communication is a fundamental human right, yet a significant gap persists between the deaf community and non-signers: over 1.5 billion people worldwide experience hearing impairment, projected to exceed 2.5 billion within three decades [1]. Sign language relies on hand movements, facial expressions, and body posture, but technological solutions to bridge this gap remain scarce compared to spoken-language technologies [2]. Automated sign language recognition (SLR) is further complicated by high intra-class variation, subtle distinctions between similar gestures, and variable lighting [3]. Deep learning has provided robust tools for this video-based Human-Computer Interaction problem.

Modern SLR has historically relied on pixel-based approaches, combining 2D CNNs for spatial features with LSTM units for temporal modeling (LRCN) [4, 5]. While capable of learning directly from raw images, these models degrade sharply under variable lighting and cluttered backgrounds, and their heavy compute requirements make them impractical for real-time use on local web browsers or low-resource hardware.

Recent work has therefore shifted from raw pixels to coordinate-based skeletal landmarks for greater robustness and efficiency [6, 7]. This study rigorously compares three progressive architectures for 50-class isolated sign language recognition (WLASL-50) [8]: a pixel-based baseline (YOLO-LRCN), a hand-landmark model (MediaPipe Hands + LSTM), and a holistic skeletal-landmark model (MediaPipe Holistic + LSTM) [6]. Its main contribution is showing how fusing upper-body pose with hand trajectories resolves ambiguity between visually similar signs while substantially reducing model size, enabling a lightweight system deployed as a functional web application via Streamlit.

2. Related Work

2.1 Sign Language Datasets and Ambiguity Challenges

The establishment of comprehensive, signer-independent benchmarks is critical for developing practical sign language recognition systems. Recent literature highlights the development of the Word-Level American Sign Language (WLASL) dataset, which currently stands as the largest public repository, comprising 21,083 video samples across 2,000 lexical signs performed by 119 diverse signers [10]. A primary challenge emphasized in large-scale scenarios is visual ambiguity, where structurally distinct glosses share nearly identical manual gestures. For example, gestures like "dance" and "read" vary subtly only in hand orientations, whereas nouns and verbs from the same lemma often share overlapping spatial trajectories. Consequently, models evaluated on small-scale vocabularies fail to generalize to real-world deployment due to these complex linguistic constraints.

2.2 The Shift from Pixel-based to Skeleton-based Features

Traditional deep learning paradigms predominantly utilized visual appearance representations. Hybrid Gated Recurrent Units or LSTMs capture spatial features from raw RGB frames before modeling chronological sequences via sequential layers [4, 5]. Although effective in controlled laboratory environments, pixel-based baselines demonstrate severe vulnerability to environmental artifacts, variations

in lighting, and changes in background context. To overcome these visual noises, recent studies have pioneered skeleton-based feature representations [6]. Multi-stream skeleton-aware ensembles that fuse hand, body-pose, and depth-based modalities have further advanced state-of-the-art performance on isolated sign recognition benchmarks [11]. By converting raw image frames into geometric coordinates of body joints, the model filters out irrelevant visual attributes, allowing the classification architecture to focus exclusively on human motion dynamics [2]. Standard tracking setups historically relied on hand-crafted structural descriptors or partial keypoint extractors [3]. However, the recent advent of advanced local coordinate estimation libraries, such as the MediaPipe framework, enables real-time tracking of isolated hands and global upper body features on standard monocular camera inputs without specialized hardware. This study bridges the existing technological literature by analyzing how the sequential fusion of hand and upper-body tracking layers optimally resolves lexical errors during high-frequency gesture processing.

3. Research Framework and Methodology

3.1 Research Framework

The research follows a structured workflow shown in Figure. 1: (1) Data Processing: selecting target classes, extracting coordinate keypoints, normalizing landmarks, and splitting data into training and test partitions (2) Model Training and Tuning: constructing and training YOLO-LRCN, MediaPipe Hands + LSTM, and MediaPipe Holistic + LSTM with iterative hyperparameter tuning and early stopping to prevent overfitting (3) Performance Evaluation: assessing Top-1/Top-10 accuracy, parameter size, and latency via a multi-class confusion matrix and (4) Comparative Analysis and Deployment: analyzing the transition from pixel-based to holistic coordinate frameworks and deploying the best-performing model as a web application prototype.

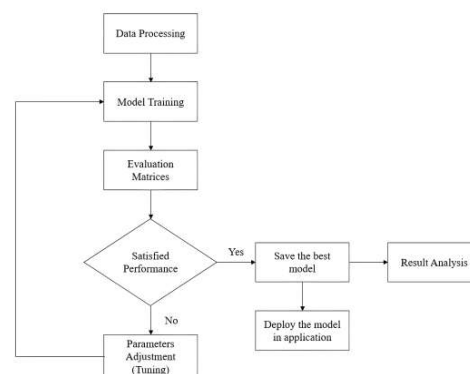


Figure. 1. Overall Experimental Design.

3.1.1 Framework for a Fair Comparative Study

To ensure performance differences resulted solely from architectural variations, a strict experimental framework was enforced. All models were trained and tested on identical data partitions using a fixed random seed split. Hyperparameters remained constant, including the Adam optimizer, categorical cross-entropy loss, learning rate, and batch size. Additionally, uniform training conditions with identical



epoch limits and early stopping criteria ensured all architectures had an equal opportunity to converge perfectly. The complete training configuration was as follows: Adam optimizer with an initial learning rate of 0.001 (reduced by a factor of 0.5 when validation loss plateaued for 7 consecutive epochs) batch size of 32 a maximum of 150 epochs with early stopping (patience = 20 epochs, monitoring validation loss, restoring best weights) LSTM layer widths of 64-128-64 units with a dropout rate of 0.2 and batch normalization applied between recurrent blocks and sparse categorical cross-entropy as the loss function. These settings were held constant across all three architectures to isolate the effect of input representation on performance. All experiments were conducted in a Google Colab environment equipped with an NVIDIA T4 GPU.

3.2 Research Methodology

This study uses 50 high-frequency WLASL classes, selected for everyday relevance and sufficient per-class samples (Table II).

The 50 glosses are: all, bed, before, black, blue, book, bowling, can, candy, chair, clothes, computer, cool, cousin, deaf, dog, drink, family, fine, finish, fish, go, graduate, hat, hearing, help, hot, kiss, language, like, man, many, mother, no, now, orange, shirt, study, table, tall, thanksgiving, thin, walk, what, white, who, woman, wrong, year, yes. Test-set support per class ranged 4–8 samples (mean ≈ 4.5), indicating approximately balanced classes.

Table II Dataset Specifications

Attribute	Description
Total Vocabulary	2,000
Total Videos	21,083
Total Signers	119
Average Samples/Class	10
Classes Used in Study	50
Average Length (seconds)	2.53
Average Frame Rate (fps)	25
Average Frame	63.25

3.2.1 Data Splitting

The final dataset was split at the video level into training and test sets using an 80% / 20% stratified random split (scikit-learn `train_test_split`, `random_state = 42`, `stratify = y`) applied to the original video clips prior to sliding-window augmentation, the resulting windows therefore remain within their originating partition, so no frames from the same source video appear in both the training and test sets. This is a custom random split rather than the official WLASL split, and partitioning was performed at the video level rather than the signer level, so videos from the same signer may appear in both partitions the model is therefore not guaranteed to be fully signer-independent (see Limitations). During training, the held-out test set was also used as the validation set for early-stopping and learning-rate-scheduling decisions

(i.e., no third, fully independent validation partition was maintained) the "Validation" and "Test" rows in Table III correspond to this same held-out set. The summary of dataset sizes is detailed in Table III.

Table III Summary of Dataset Sizes

Dataset	Baseline	After SW, Stride = 5
Training	629	3380
Validation	79	422
Test	79	423

3.3 Baseline Pixel-based Architecture (YOLO-LRCN)

The foundational baseline utilizes a Long-term Recurrent Convolutional Network (LRCN) framework that strictly relies on raw image pixels. To mitigate peripheral background noise and force the spatial model to focus exclusively on the signer, the system incorporates the state-of-the-art YOLOv8n object detection model to identify and isolate body bounding boxes. YOLOv8n is explicitly selected due to its highly optimized architectural efficiency, ultra-lightweight parameter size, and high-speed multi-object localized tracking capabilities, which conform directly to the resource-constrained operational requirements of local computing deployment. The isolated sequence frames are subsequently resized to 128×128 pixels and fed into a spatial feature extraction module. This module is constructed using multiple TimeDistributed 2D-Convolutional layers interleaved with MaxPooling operations to progressively downsample the cropped frames. The resulting high-level spatial maps are flattened into 256-dimensional feature vectors before being processed by a 64-cell LSTM layer for temporal sequence learning. Classification labels are finally decoded via a dense SoftMax layer. The end-to-end framework layout, bounding box localized spatial processing, and recurrent neural architecture of this pixel-based baseline model are systematically illustrated in Figure. 2.

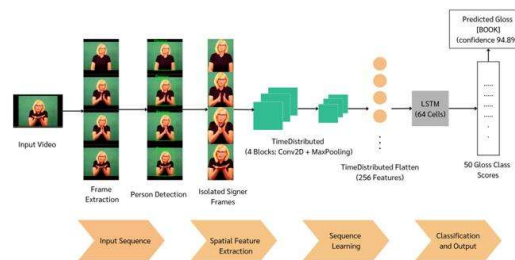


Figure. 2. Schematic diagram of the baseline pixel-based YOLO-LRCN architecture

3.4 Hand Landmark Coordinate Architecture (MediaPipe Hands + LSTM)

The second architecture shifts the input representation from volumetric pixels to geometric coordinates by utilizing a localized hand tracking framework. This approach processes input video streams frame-by-frame to extract the exact coordinate positions of 21 distinct structural hand joints for each hand, yielding a total of 42 spatial keypoints per frame. Each landmark is mathematically translated into a three-dimensional linear vector consisting of localized coordinates (X, Y, Z), which effectively scales down the

structural data matrix while discarding irrelevant visual parameters like lighting variance and clothing styles. These continuous geometric streams are chronologically sequenced and channeled into a standard LSTM layer to learn the specific progression path of isolated signs. The structural framework, coordinate tracking workflow, and sequential learning configuration of this isolated hand-based architecture are systematically illustrated in Figure. 3.

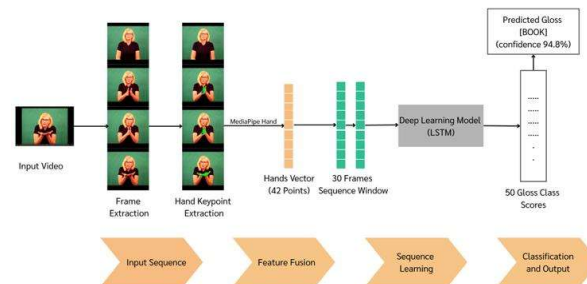


Figure. 3. Schematic diagram of the isolated MediaPipe Hands-LSTM architecture.

3.5 Holistic Skeletal Landmark Architecture (MediaPipe Holistic + LSTM)

The primary proposed paradigm introduces a holistic skeletal landmark framework designed to eliminate lexical ambiguities in high-frequency vocabularies. While isolated hand joint tracking functions efficiently for basic gestures, it demonstrates performance bottlenecks when distinguishing complex signs that require broad physical interactions with body locations, such as the face, shoulders, and chest. To resolve this technical limitation, the holistic model simultaneously fuses the tracking profiles of both hands (42 landmarks) with the coordinates of the upper-body pose structural framework (33 landmarks), capturing a combined total of 75 structural landmarks per visual frame. This multi-modal coordinate topology is normalized to retain structural orientation irrespective of the signer's proximity to the sensor. Concretely, the x and y coordinates of each landmark are normalized by MediaPipe to the range [0, 1] relative to the frame width and height, while the z coordinate is normalized relative to the scale of the wrist landmark, giving a depth value that is physically comparable across frames regardless of the signer's distance from the camera these normalized coordinates are used directly, and are distinct from the Gaussian-noise and random-shift perturbations applied only to the training split as a data augmentation step. The extracted feature vectors are subsequently modeled using an optimized LSTM block with early stopping regularizations, producing a highly lightweight, high-speed architecture optimized for low-latency web deployment. Each video sequence is resampled into fixed-length windows of 30 frames (sliding-window stride of 5 during training-set augmentation); sequences shorter than 30 frames are zero-padded, longer sequences are truncated to the most recent 30 frames, and frames with no detected hand landmarks are represented with zero-filled coordinate vectors. The overall pipeline and data flow of this multi-modal feature fusion and sequential learning process are systematically illustrated in Figure. 4.

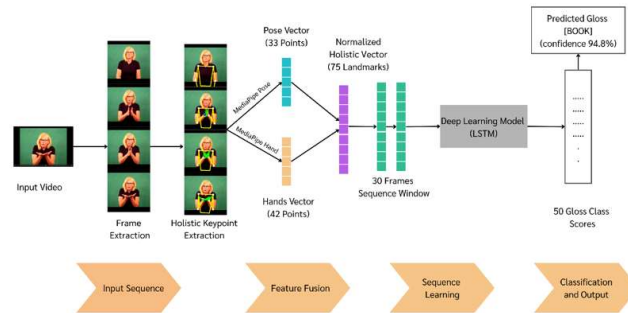


Figure. 4. Schematic diagram of the proposed MediaPipe Holistic-LSTM architecture.

3.6 Web Prototype Implementation

To validate the practical viability of the proposed lightweight MediaPipe Holistic + LSTM model, a functional web browser prototype was implemented. The application architecture was constructed using the Streamlit framework, designed to facilitate native execution on standard commercial computing hardware without local GPU dependence. Rather than requiring complex streaming infrastructure, the interface provides an optimized file-uploader widget allowing users to submit recorded sign language video samples directly.

Upon file submission, the underlying pipeline processes the video container by extracting sequential frames at a standardized rate. The embedded tracking framework extracts the 75 structural landmarks frame-by-frame from the uploaded file, storing the structural configurations inside a temporary temporal buffer matrix. Once the entire frame sequence is processed, the compiled vector array is channeled directly into the optimized 850 KB LSTM network file for instantaneous inference. The final decoded lexical sign label, alongside its corresponding prediction confidence percentage, is then dynamically projected back onto the Streamlit web interface for user validation.

4. Experimental Results and Discussion

4.1 Quantitative Performance Evaluation

The experimental results demonstrate a critical architectural paradigm shift between pixel-based appearance processing and holistic coordinate optimization. As detailed in Table IV, the baseline hybrid configuration (YOLO-LRCN) processing raw visual pixels exhibits severe performance bottlenecks, achieving a restrictive Top-1 Accuracy of 18.42% and a Top-10 Accuracy of 45.98%. This limitation stems directly from the spatial dependencies being obscured by background ambient illumination and environmental artifacts inherent in monocular video frames. Conversely, transitioning to geometric vector streams yields substantial accuracy gains. The hand landmark paradigm (MediaPipe Hands + LSTM) isolates biological movement patterns, elevating Top-1 and Top-10 scores to 32.15% and 68.30% respectively.

Table IV Performance and Architectural Complexity Comparison on WLASL-50 Dataset

Architecture Paradigm	Model Framework	Top-1 Accuracy (%)	Top-10 Accuracy (%)	Parameter Size	Inference Latency
1) Pixel-based (Baseline)	YOLO-LRCN	18.42%	45.98%	~8.50 MB	High Latency / GPU-Bound
2) Coordinate-based (Hands)	MediaPipe Hands + LSTM	32.15%	68.30%	~420 KB	Low Latency
3) Holistic Skeleton (Proposed)	MediaPipe Holistic + LSTM	44.84%	81.17%	~850 KB	~95 ms/prediction (measured, CPU, n=50 runs)

Note: only the Holistic model's latency was empirically benchmarked (Section 4.2); Pixel-based and Hands-only latencies are qualitative estimates based on architectural complexity. Benchmarking all three under identical hardware is left for future work.

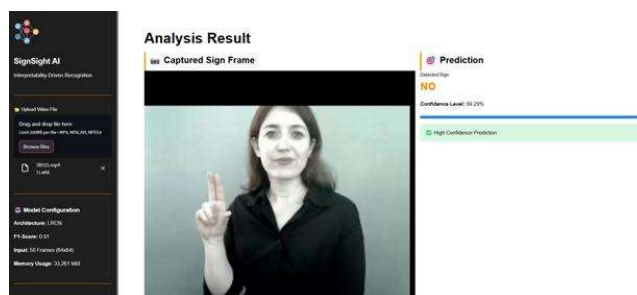
The proposed multi-modal paradigm (MediaPipe Holistic + LSTM) achieves the highest overall performance, demonstrating a substantial Top-1 Accuracy of 44.84% and an exceptional Top-10 Accuracy of 81.17%. The integration of upper-body pose metrics alongside isolated hand tracking effectively resolves deep structural ambiguities among visually identical gestures that share overlapping trajectories. This comparison functions as a partial ablation study: because the Hands-only and Holistic models are identical in every respect except the presence of upper-body pose coordinates, the accuracy gap between them (Top-1: 32.15% to 44.84%) directly isolates the contribution of pose information. Linguistically, this is consistent with location, one of the core phonological parameters in sign language structure [12], where the position of the hands relative to the shoulders, chest, and torso frequently disambiguates signs that are otherwise kinematically similar when only hand trajectories are considered [11]. The pixel-based baseline, by contrast, must implicitly relearn this spatial relationship from raw appearance, which is more susceptible to lighting and background interference.

4.2 Model Complexity and Practical Deployment Analysis

Beyond accuracy, structural efficiency matters for browser deployment. The proposed Holistic framework needs only ~850 KB versus ~8.5 MB for the pixel-based network a roughly ten-fold reduction. This figure covers only the trained LSTM classifier weights; the MediaPipe landmark extractors are fixed, pre-trained modules identical across both coordinate-based models and are excluded, consistent with standard practice of reporting only the newly trained classification head.

This lightweight footprint enables low-latency inference: over 50 repeated runs, the Holistic model averaged 95 ms per prediction (SD = 30 ms; min = 64 ms) on a Colab T4 GPU fast enough to feel instantaneous, though the current prototype is upload-based. At only ~850 KB, the classifier does not require a dedicated GPU at inference time (the T4 GPU was used only for training/benchmarking), enabling accurate recognition without high-end graphics hardware.

To demonstrate the functional viability of the proposed system, the compiled MediaPipe Holistic + LSTM model was integrated into a unified web application interface as illustrated in Figure. 6. The prototype interface, constructed via the Streamlit framework, features an intuitive file-uploader module where users can seamlessly submit pre-recorded sign language video files. Upon file validation, the system visually renders the video player component alongside a centralized execution log. Once the feature extraction and buffer sequencing are complete, the classification output representing the predicted sign language gloss is prominently displayed at the top of the interface layout, accompanied by an explicit prediction confidence score percentage to verify tracking accuracy.



4.4 Error Analysis and Confusion Matrix

To investigate how the Holistic model resolves the visual ambiguity that limits the Hands-only model, per-class confusion matrices and classification reports were generated for both models on the held-out test set (n = 223 sequences). Table V summarizes the resulting Top-1/Top-5/Top-10 accuracy and macro/weighted precision, recall, and F1-score for this evaluation run Figure. 10 and Figure. 11 show the corresponding confusion matrices.

Table V Performance and Architectural Complexity Comparison on WLASL-50 Dataset

Model	Top-1 Acc.	Top-5 Acc.	Top-10 Acc.	Macro Precision	Macro Recall	Macro F1	Weighted F1
Hands-only (MediaPipe Hands + LSTM) [diagnostic run]	26.91%	59.19%	72.65%	0.21	0.26	0.22	0.22
Holistic (MediaPipe Holistic + LSTM, proposed) [diagnostic run]	44.84%	69.51%	81.17%	0.50	0.45	0.43	0.43

Note: headline results use Table IV throughout. Figures. 10-11 and Table V come from a separate diagnostic run for error analysis only; its Holistic figures match Table IV exactly, while its Hands-only figures (26.91%/72.65%) differ slightly, consistent with the single-run variance noted in the Limitations.

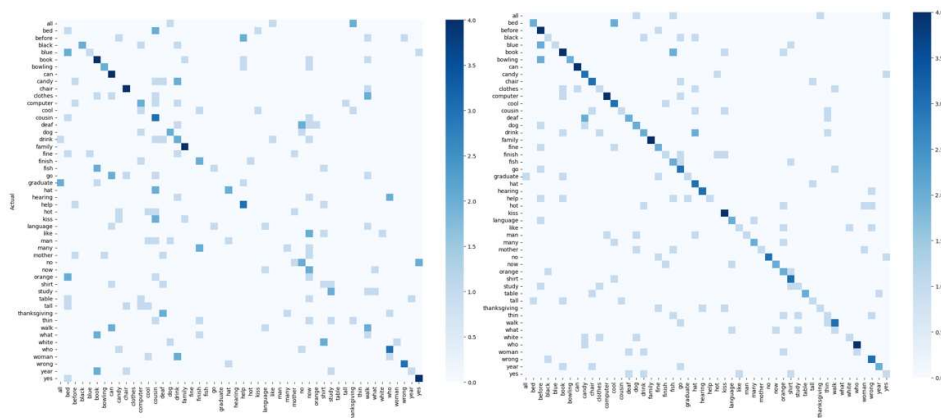


Figure.10. Confusion matrix, Hands-only model (MediaPipe Hands + LSTM), test set (Left).

Figure.11. Confusion matrix, Holistic model (MediaPipe Holistic + LSTM, proposed), test set (Right).

For example, in the Hands-only matrix (Figure. 10), "before" is frequently confused with "help", the two share a similar hand trajectory while the Holistic matrix (Figure. 11) classifies the same samples correctly, consistent with pose information resolving the ambiguity. Off-diagonal mass is also visibly denser and more dispersed in Figure. 10 than in Figure. 11, where predictions concentrate along the diagonal.

5. Conclusion, Limitations, and Future Work

This study presented a rigorous comparison of pixel-based and holistic coordinate frameworks for word-level sign language recognition on WLASL-50. The pixel-based baseline (YOLO-LRCN) suffered from lighting and background sensitivity geometric coordinates isolated motion dynamics from this interference. The proposed paradigm fusing 42 hand and 33 upper-body pose landmarks (75 total) via MediaPipe Holistic + LSTM achieved the best results (44.84% Top-1, 81.17% Top-10) while cutting model complexity roughly ten-fold to ~850 KB, verified via a video-upload-based Streamlit prototype averaging ~95 ms per prediction. Limitations: (1) evaluation is restricted to a 50-word WLASL subset, untested on the full vocabulary or other sign languages (2) the system recognizes isolated words only, not continuous sentences or non-manual grammar (3) the prototype has not been evaluated with real deaf or hard-of-hearing users (4) this study covers only American Sign Language applicability to Thai Sign Language would require a dedicated Thai dataset (5) the train/test split is a custom, video-level (not signer-level or official WLASL) split, so the same signer's videos may appear in both sets, and no separate validation partition was maintained (6) results come from a single training run per architecture, so cross-seed variance is unestablished and (7) reported latency reflects only the LSTM classifier on one hardware configuration end-to-end latency across devices remains future work.

Future work will (1) scale the vocabulary beyond 200 glosses, (2) optimize the pipeline toward continuous, live-webcam inference beyond the current upload-based prototype, and (3) repeat training across multiple random seeds ($n = 5$) to report mean $\pm$ standard deviation of results.



6. References

- [1] WHO. (2019). Deafness and hearing loss. World Health Organization. <https://www.who.int/health-topics/hearing-loss#tab=tab>.
- [2] Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign language recognition: A deep survey. *Expert Systems with Applications*, 164, 113793. <https://doi.org/10.1016/j.eswa.2020.113793>
- [3] Cheok, M. J., Omar, Z., & Jaward, M. H. (2019). A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 10(1), 131-153. <https://doi.org/10.1007/s13042-017-0705-5>
- [4] Basnin, N., Nahar, L., & Hossain, M. S. (2020). An integrated CNN-LSTM model for Bangla lexical sign language recognition. In *Proceedings of International Conference on Trends in Computational and Cognitive Engineering* (pp. 415-425). Springer. https://doi.org/10.1007/978-981-33-4673-4_57
- [5] Yang, S., & Zhu, Q. (2017). Continuous Chinese sign language recognition with CNN-LSTM. In *Proceedings of International Conference on Digital Image Processing* (pp. 1042-1048). SPIE. <https://doi.org/10.1117/12.2281671>
- [6] Sabeenian, R. S., Paramasivam, S., & Dinesh, P. M. (2020). Sign language recognition using deep learning and computer vision. *Journal of Advanced Research in Dynamical and Control Systems*, 12(2), 2354-2361. <https://doi.org/10.5373/JARDCS/V12I2/20201354>
- [7] Liu, Y., Zhang, X., & Wang, L. (2025). SCOPE: Sign language contextual processing with embedding from LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 8912-8920). AAAI Press.
- [8] Arooj, M., Riaz, N., & Fatima, S. (2025). Sign language recognition in the deep learning era: A study of performance and deployment. *International Journal of Science and Research Archive*, 14(1), 512-523. <https://doi.org/10.30574/ijrsra.2025.14.1.0112>
- [9] Rastgoo, R., Kiani, K., & Escalera, S. (2022). Real-time isolated hand sign language recognition using deep networks and SVD. *Journal of Ambient Intelligence and Humanized Computing*, 13(4), 1951-1970. <https://doi.org/10.1007/s12652-021-02920-8>
- [10] Li, D., Rodriguez Opazo, C., Yu, X., & Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1459-1469). IEEE. <https://doi.org/10.48550/arXiv.1910.11006>
- [11] Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021). Sign language recognition via skeleton-aware multi-model ensemble. *arXiv preprint arXiv:2110.06161*. <https://doi.org/10.48550/arXiv.2110.06161>
- [12] Anderson, C. (2018). *Essentials of Linguistics* (Ch. 2.10, Classifying Signs). eCampusOntario. <https://ecampusontario.pressbooks.pub/essentialsoflinguistics/chapter/2-10-classifying-signs/>