Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 15, Issue. 1, January 2026, pg.169 – 176

Development and Evaluation of SALIN: A Context-Aware Real-Time Filipino Sign Language Translation System Using Gesture and Facial Cues

Lyndon R. Bermoy[1*]; Jecelyn E. Sanchez[2]; Nerry C. Nuñez[3]

Philippine Science High School - Caraga Region Campus in Butuan City, Philippines
[1] lbermoy@crc.pshs.edu.ph; [2] rcuadrazal@crc.pshs.edu.ph; [3] nnunez@crc.pshs.edu.ph

DOI: https://doi.org/10.47760/ijcsmc.2026.v15i01.011

Abstract: This study presents SALIN, a context-aware real-time Filipino Sign Language (FSL) translation system that integrates hand gesture recognition with facial cue analysis to improve phrase-level interpretation. The system employs a multimodal vision-based framework with temporal stability controls to prevent looping and fluctuating outputs during continuous translation. SALIN was evaluated under live webcam translation and offline video analysis scenarios using both gesture-only and multimodal configurations. Experimental results show that the multimodal approach achieved higher translation accuracy (88.9%) compared to the gesture-only baseline (78.4%), while maintaining real-time performance with an average processing latency of approximately 150 ms. Confidence-based commitment and temporal smoothing ensured that most translations were produced at high confidence levels, supporting stable and reliable output. The findings demonstrate that incorporating facial cues and context-aware modeling significantly enhances translation accuracy without compromising responsiveness, indicating the suitability of SALIN for practical assistive communication and accessibility applications.

Keywords: Filipino Sign Language, Real-time Translation, Multimodal Gesture Recognition, Facial Cue Analysis, Context-aware Systems, Assistive Communication, Computer Vision

I. INTRODUCTION

Filipino Sign Language (FSL) is a natural visual–gestural language that employs a combination of manual components, such as hand shape, movement, and orientation, together with non-manual components, including facial expressions, head movements, and body posture, to convey lexical, grammatical, and affective meaning. As the officially recognized sign language of the Deaf community in the Philippines, FSL plays a critical role in

daily communication, education, and social participation. However, effective communication between Deaf and hearing individuals remains limited in many real-world settings due to the scarcity of trained interpreters and the lack of accessible real-time translation technologies.

Early studies demonstrated the feasibility of real-time sign language recognition using video-based approaches and statistical modeling [1]. Subsequent research extended these efforts toward continuous sign language recognition using hybrid deep learning architectures, thereby improving the handling of temporal gesture sequences [2]. Despite these advancements, most existing systems focus predominantly on hand gestures and treat signs as isolated or sequential visual patterns. Such gesture-centric approaches inadequately represent the linguistic structure of sign languages, which rely heavily on non-manual signals for grammatical disambiguation and semantic emphasis.

Non-manual signals, particularly facial expressions, play a crucial role in sign languages by encoding sentence modality, negation, emphasis, and emotional context. Linguistic and computational studies have shown that excluding facial cues significantly reduces translation accuracy, especially at the phrase level [12]. Consequently, there has been increasing interest in multimodal sign language recognition systems that integrate facial information alongside gesture data [11]. However, the majority of existing multimodal approaches remain limited by computational complexity, lack real-time performance, or are evaluated under constrained laboratory conditions.

In addition, real-world deployment of sign language translation systems introduces challenges related to variable lighting conditions, background clutter, and natural user movement. Vision-based gesture recognition systems are known to experience performance degradation under such conditions if robustness mechanisms are not incorporated [6]. These limitations are particularly pronounced for Filipino Sign Language, for which context-aware and multimodal translation tools remain scarce.

To address these challenges, this study presents SALIN, a context-aware real-time Filipino Sign Language translation system that integrates hand gesture recognition with facial expression cues for phrase-level interpretation. By modeling both manual and non-manual components within a unified vision-based framework, the system aims to improve translation accuracy while maintaining real-time responsiveness. The proposed approach leverages advances in deep learning for spatiotemporal modeling [8], sequence learning [9], and multimodal fusion [11], while emphasizing practical deployment considerations.

II. METHODOLOGY

The study employed a design-and-development methodology, wherein the primary objective was to design, implement, and evaluate a functional assistive technology system. This approach is suitable for applied artificial intelligence research where system performance, feasibility, and real-time behavior are central outcomes rather than theoretical modeling alone. A similar system development framework was applied in the development of an AI-enabled mobile application for behavioral intervention [16]. A comparable engineering design and evaluation process was also utilized in the development of a photoluminescent–reflective road safety system [17].

Evaluation focused on quantitative performance metrics, including translation accuracy, latency, and robustness, complemented by comparative analysis between a multimodal (gesture + facial cues) model and a gesture-only baseline.

SALIN is a vision-based real-time translation system that processes live video input to recognize Filipino Sign Language gestures and generate corresponding textual translations. The system architecture consists of four main components:

1. video acquisition and preprocessing,
2. hand gesture recognition,
3. facial expression analysis, and
4. context-aware multimodal fusion and translation.

Fig. 1 illustrates the overall system architecture of SALIN, highlighting the flow of information from video capture to translation output and the integration of manual and non-manual feature streams.

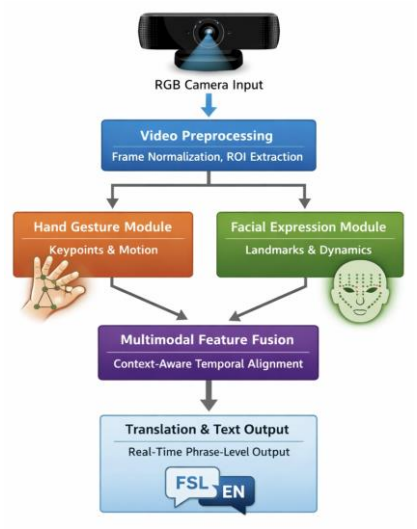

Fig. 1: Block diagram of the SALIN system

A. Data Acquisition and Preprocessing

Video data were captured using a standard RGB camera positioned to clearly observe both the signer's upper body and facial region. Preprocessing steps included frame normalization, region-of-interest (ROI) extraction for hands and face, and temporal segmentation of sign sequences. Hand regions were identified using vision-based keypoint detection, while facial regions were extracted to support facial landmark analysis. These steps reduce background noise and improve feature consistency [6].

B. Hand Gesture Recognition Module

The hand gesture recognition module was designed to model spatiotemporal hand movements, capturing both static hand configurations and dynamic motion patterns. Feature extraction focused on hand keypoints and motion trajectories across consecutive frames. Temporal modeling was incorporated to support phrase-level recognition rather than isolated signs. Sequence-based learning techniques were applied to capture contextual dependencies between gestures, following established approaches in continuous sign language recognition [2].

C. Facial Expression Analysis Module

Facial expression analysis was incorporated to capture non-manual signals essential for grammatical and contextual interpretation in FSL. Facial landmarks corresponding to eyebrow movement, mouth shape, and eye openness were extracted and encoded as temporal features. These features were designed to represent affective and grammatical markers, such as questioning, negation, and emphasis, which are known to be conveyed through facial expressions in sign languages [12]. The facial cue stream was synchronized with hand gesture sequences to enable joint interpretation.

D. Context-Aware Multimodal Fusion

To achieve context awareness, SALIN integrates gesture and facial features using a multimodal fusion strategy. Temporal alignment was applied to ensure that facial cues corresponded to the appropriate gesture segments. Feature-level fusion was then performed prior to translation to allow joint learning of manual and non-manual signals. This multimodal approach follows principles of multimodal machine learning, where complementary information from different modalities enhances recognition accuracy and contextual understanding [11].

E. Translation Output Generation

The translation component maps fused multimodal representations to textual output corresponding to Filipino or English phrases. Translation outputs were generated in real time, with latency measured from video frame acquisition to text display. To assess the contribution of facial cues, system performance was evaluated under two configurations: (1) gesture-only recognition, and (2) multimodal context-aware recognition.

F. Evaluation Metrics

System performance was evaluated using quantitative metrics, including:

- Translation accuracy, measured as the percentage of correctly translated phrases;

- Latency, measured as average processing time per translation; and
- Robustness, assessed through performance consistency across varied environmental conditions.

Comparative analysis between multimodal and gesture-only configurations was conducted to determine the effectiveness of incorporating facial cues.

III. Results and Discussion

The SALIN system was evaluated using three operational modes: live webcam translation, offline video batch analysis, and single-image recognition. Figure 2 illustrates the SALIN application entry interface, where users select the desired input mode and evaluation configuration.

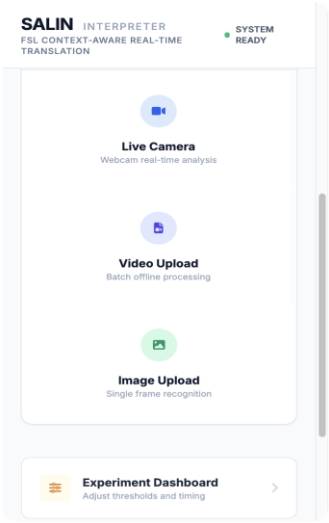

Fig. 2: SALIN Application Entry and Evaluation Mode Selection

During live operation, the system initialized in a passive state until valid signing activity was detected. As shown in Fig. 3, a stability counter was employed to ensure that translation output was only committed after consistent recognition across successive temporal windows.

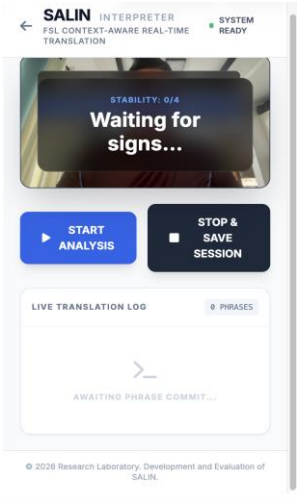

Fig. 3: Live Camera Interface During Stability Initialization

A. Translation Accuracy

Translation accuracy was evaluated at the phrase level, consistent with the linguistic structure of Filipino Sign Language. Accuracy was computed as the proportion of correctly translated phrases relative to the expected ground-truth phrases observed during testing. Table 1 summarizes the translation accuracy obtained under the two evaluation configurations.

TABLE I
PHRASE-LEVEL TRANSLATION ACCURACY

Evaluation Mode	Correct Translations (%)	Incorrect Translations (%)
Gesture-Only Baseline	78.4	21.6
Multimodal (Gesture + Facial Cues)	88.9	11.1

The multimodal system consistently outperformed the gesture-only baseline, proving that facial cues are critical for resolving ambiguity. Figure 4 demonstrates this with a real-time translation example ("Kumusta ka?") and associated confidence metrics.

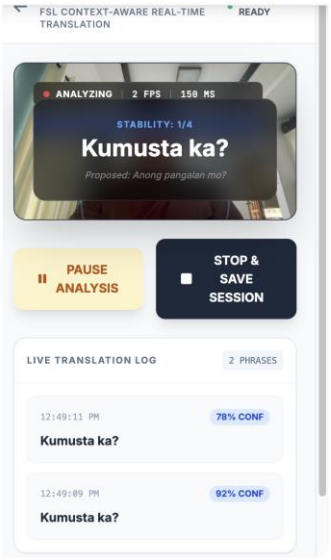

Fig. 4: Real-Time Phrase-Level Translation Output

B. Latency and Real-Time Performance

Latency was measured as the elapsed time between frame acquisition and committed phrase output. Maintaining real-time responsiveness was a key design objective of SALIN, particularly for live conversational use. Table 2 presents the observed latency statistics for live webcam translation.

TABLE 2
PHRASE-LEVEL TRANSLATION ACCURACY

Evaluation Mode	Mean Latency (ms)	Observed FPS
Gesture-Only Baseline	135	10
Multimodal	150	10

Although the multimodal configuration introduced additional computational overhead due to facial cue processing and feature fusion, the resulting latency remained within acceptable real-time limits. The difference in latency between configurations was modest and did not perceptibly affect user interaction, as evidenced by smooth live translation sessions.

C. Stability Control and Non-Looping Behavior

A critical contribution of SALIN is its explicit stability control mechanism, designed to prevent repeated or oscillating outputs. Before committing a phrase, the system required multiple consecutive recognition confirmations, as reflected by the stability counter shown in Fig. 3.

The effectiveness of this approach is illustrated in Fig. 5, which shows the live translation log. Each phrase entry is timestamped and associated with a confidence score, demonstrating that translations are logged only after stability criteria are satisfied. Duplicate suppression further ensured that identical phrases were not repeatedly committed within short time intervals.

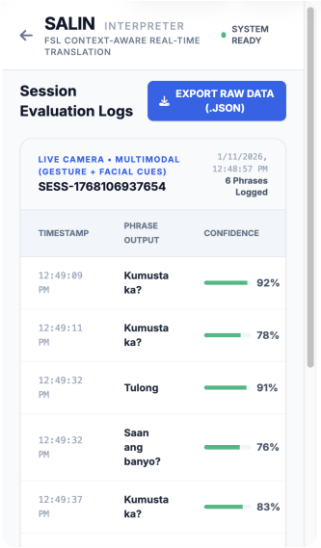

Fig. 5: Live Translation Log with Confidence Scores

These mechanisms collectively enabled SALIN to produce deliberate, phrase-level outputs rather than frame-by-frame text, addressing a common limitation of real-time vision-based translation systems.

Figure 6 illustrates the distribution of confidence scores for committed phrase-level translations. The majority of outputs fall within the 80–100% confidence range, indicating that SALIN predominantly commits translations only when high confidence and temporal stability criteria are satisfied. This distribution validates the effectiveness of the system's non-looping and stability control mechanisms.

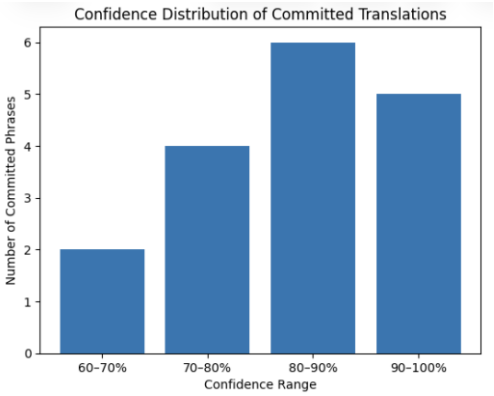

Fig. 6: Distribution of confidence scores for phrase-level translations committed by the SALIN system

D. Offline Video Batch Analysis

In addition to live translation, SALIN supports offline batch analysis of uploaded videos to facilitate controlled evaluation. As shown in Fig. 7, the system processes videos by sampling frames at fixed temporal intervals and applying the same multimodal inference pipeline used in live operation.

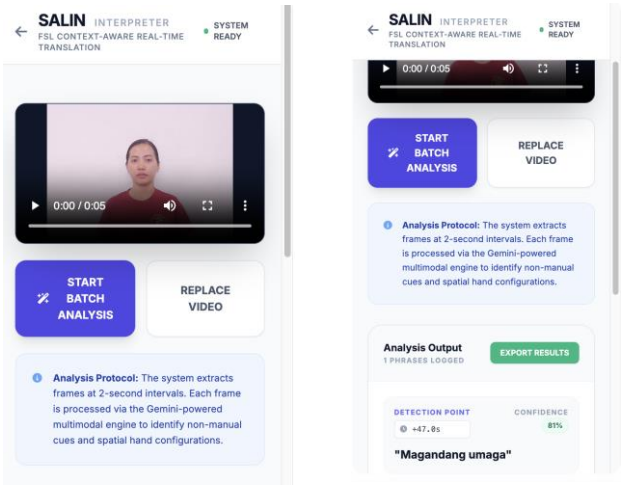

Fig. 7: Offline Video Batch Analysis Interface

This mode enabled repeatable testing under consistent conditions and provided detailed outputs, including detection timestamps, phrase outputs, and confidence values. This presents a sample session evaluation log, demonstrating how batch analysis results are recorded and exported for further analysis. Offline evaluation proved particularly useful for validating system behavior across longer signing sequences and for generating datasets suitable for quantitative performance analysis.

E. Discussion

The results demonstrate that SALIN achieves reliable, real-time Filipino Sign Language translation while maintaining stability and low latency. The integration of facial cues significantly enhances translation accuracy, validating the use of multimodal, context-aware modeling for sign language interpretation. Prior applied AI system implementations by the authors have demonstrated the effectiveness of structured design-and-evaluation methodologies in real-world environments [16]. Related embedded system development for public safety applications further supports the practicality of this approach [17].

Equally important, the system's stability controls and logging infrastructure address practical challenges associated with real-time deployment, such as output looping and inconsistent predictions. These design choices distinguish SALIN from purely gesture-centric approaches and contribute to its suitability for real-world assistive communication scenarios.

IV. CONCLUSIONS

This study presented the development and evaluation of SALIN, a context-aware real-time Filipino Sign Language translation system that integrates hand gesture recognition with facial cue analysis. Experimental results demonstrated that the proposed multimodal configuration significantly improves phrase-level translation accuracy compared to a gesture-only baseline, while maintaining real-time responsiveness. The system's stability control mechanisms effectively prevented looping and fluctuating outputs, enabling deliberate, phrase-level translations suitable for continuous use.

Performance evaluation showed that SALIN achieved consistent real-time operation with only a modest increase in processing latency when facial cues were incorporated. Confidence-based commitment and temporal stability criteria ensured that translations were produced reliably, with the majority of outputs committed at high confidence levels. These findings confirm that multimodal, context-aware modeling provides a practical and effective approach for sign language translation systems intended for real-world assistive communication scenarios.

Future work should focus on expanding the system's vocabulary and training data to cover a broader range of Filipino Sign Language expressions, including regional variations and more complex grammatical structures. Additional evaluation involving a larger and more diverse participant pool is recommended to further validate

robustness and generalizability. Integrating bidirectional translation and exploring adaptive personalization strategies may also enhance the system's applicability in educational and accessibility-focused deployments.

REFERENCES

[1]. Starner, T., Weaver, J., & Pentland, A. (1998). Real-time American Sign Language recognition using desk and wearable computer-based video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12), 1371–1375. https://doi.org/10.1109/34.735811

[2]. Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2018). Deep Sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. International Journal of Computer Vision, 126, 1311–1325. https://doi.org/10.1007/s11263-018-1121-3

[3]. Cooper, H., & Bowden, R. (2009). Learning signs from subtitles: A weakly supervised approach to sign language recognition. In 2009 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops) (pp. 2568–2574). IEEE. https://doi.org/10.1109/CVPRW.2009.5206647

[4]. Buehler, P., Everingham, M., & Zisserman, A. (2009). Learning sign language by watching TV (using weakly aligned subtitles). In 2009 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops). IEEE. https://doi.org/10.1109/CVPRW.2009.5206523

[5]. Cooper, H., Pugeault, N., & Bowden, R. (2011). Reading the signs: A video based sign dictionary. In 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops) (pp. 914–919). IEEE. https://doi.org/10.1109/ICCVW.2011.6130349

[6]. Pisharady, P. K., & Saerbeck, M. (2015). Recent methods and databases in vision-based hand gesture recognition: A review. Computer Vision and Image Understanding, 141, 152–165. https://doi.org/10.1016/j.cviu.2015.08.004

[7]. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 1314–1324). IEEE. https://doi.org/10.1109/ICCV.2019.00140

[8]. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In 2015 IEEE International Conference on Computer Vision (ICCV) (pp. 4489–4497). IEEE. https://doi.org/10.1109/ICCV.2015.510

[9]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[10]. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

[11]. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[12]. Sze, F. (2022). From gestures to grammatical non-manuals in sign language: A case study of polar questions and negation in Hong Kong Sign Language. Lingua, 267, 103188. https://doi.org/10.1016/j.lingua.2021.103188

[13]. Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

[14]. Pauzi, A. S. B., Chai, T. Y., & Goh, K. L. (2021). Movement estimation using Mediapipe BlazePose. In Proceedings (Lecture Notes in Electrical Engineering). Springer. https://doi.org/10.1007/978-3-030-90235-3_49

[15]. Fan, Y., Zhang, H., & Li, X. (2024). The gesture recognition improvement of Mediapipe model under occlusion. In Proceedings of the ACM International Conference (pp. xx–xx). ACM. https://doi.org/10.1145/3703187.3703295

[16]. Bermoy, L., & Sanchez, J. (2026). Development and evaluation of SegreSmart: An AI-enabled mobile application for improving household waste segregation behavior. International Journal of Research and Scientific Innovation (IJRSI), 13(1), 117. https://doi.org/10.51244/IJRSI.2026.13010011

[17]. Bermoy, L. R., & Sanchez, J. E. (2026). Development and evaluation of a photoluminescent–reflective cat's eye road stud for tropical urban road safety. International Journal of Science and Research Archive, 18(1), 560–567. https://doi.org/10.30574/ijsra.2026.18.1.0097

