International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

# Thermal-Aided Real Time Sign Interpretation: A Comprehensive Review

Krishna Swaroop A[1],Prajwal[2], Pranith B H[3], Prajwal H R[4], Benaka Raj[5]

Department of Information Science Engineering, Malnad College of Engineering, Hassan,

Karnataka, India – 573201

***Abstract - Sign language recognition (SLR) systems play a crucial role in making communication accessible for the Deaf and hard-of-hearing community. Unlike traditional RGB-based methods that often struggle with lighting issues and can be intrusive, thermal imaging provides a strong, contactless, and privacy friendly alternative. This survey dives into the latest advancements in thermal SLR, highlighting deep learning techniques like CNNs, RNNs, Transformers, and hybrid models. Thermal systems naturally tackle lighting challenges while ensuring user anonymity, with CNN-RNN architectures achieving over 95% accuracy on static signs and attention-based models facilitating continuous recognition. Real-time applications on embedded devices like the NVIDIA Jetson and Raspberry Pi show impressive low-latency performance. However, there are still hurdles to overcome, such as handling occlusions, limited datasets, and generalizing across different signers. Exciting trends are emerging, including edge-optimized SLR for wearables, multimodal fusion (combining thermal and EMG), and support for multiple languages, all while keeping fairness, interpretability (using tools like Grad-CAM/SHAP), and privacy in mind. Deep learning-enhanced thermal SLR shines in low-visibility situations, with YOLO and CNN-based models leading the charge in real-time applications. The combination of thermal imaging and AI offers a scalable and efficient route for accessible sign language translation, paving the way for future research in inclusive communication technologies.***

***Keywords: Thermal Imaging, Sign Recognition, Deep Learning, Convolutional Neural Networks, YOLO, Gesture Recognition***.

## I. INTRODUCTION

Sign languages are visual and gesture-based languages that are necessary for many in the deaf community to communicate. Automated Sign Language Recognition (SLR) can help significantly improve accessibility via translating the signs to sentences spoken or written in English, creating a bridge between the deaf and hearing members of society. Traditional SLR techniques exist, like sensor gloves, which capture data about specific hand movements accurately but are also bulky and expensive, and camera systems (RGB or depth cameras), which create a less cumbersome experience, but may struggle with lighting, background clutter, and occlusion issues. Camera systems can also raise privacy and surveillance concerns as they create high-resolution images of individuals. Recently deep learning methods, e.g., using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have advanced the state-of-the-art in gesture recognition dramatically. CNNs are the best architecture for extracting high level spatial features, while RNNs use architectures, like Long Short-Term Memory (LSTM), to handle variable length temporal sequences. Though there have been advances in these methods, full recognition problems in signing (sentence-level SLR) continues to remain challenging, but it's also gaining a lot of attention with new proposed work in the area of Transformers, which utilise attention-based multi-head attention. Thermal infrared (IR) has emerged as viable proximate sign language recognition alternative idiot, for the reason that thermal cameras only detect heat, and this type of imaging has a big advantage it does not discriminate with regards to range of human skin tones or depending on traditional lighting conditions. As noted by Birkeland N.M., McDonald J., and McCarthy N.M., the advantages of using thermal sensors is they "are also able to tell the sense of variables that often confound the use of standard cameras such as lighting and skin color". So there is variety of context for using thermal imaging including those environmental contexts with very low ambient or no light.

## II. LITERATURE SURVEY

**1)*TITLE: A Sign Language Recognition System for Helping Disabled People Publication year:2023***

**Authors:Adhikari Hridoy, Bin Jahangir Md. Sakib, Jahan Israt, Mia Md. Solaiman, Hassan Md. Riad[1]**

**Methodology:**Data Collection: I obtained images of 9 sign language classes from classic and contemporary works under different lighting and weather conditions. These images were resized to 190×190 pixels. Data Preprocessing: This included cleaning the data to remove duplicates, normalizing to standardize input data, and splitting the data into a training set and a test set. Model Training: I created a CNN model with an Adam optimizer. After training with a batch size of 64 over 40 epochs, I evaluated it with a separate test dataset.

**Key Findings:**The CNN structure provided a training accuracy of 99.17% and a validation accuracy of 91.67% for monitoring sign actions. The model outperformed others, achieving 91.67% accuracy with a dataset of 6,182 images. Each epoch took around 4 minutes to train with a batch size of 64 for 40 epochs.

**Limitations:**The dataset includes only 9 classes, which limits diversity. Although I took images under various lighting and weather conditions, the consistency suffers. Resizing the images before training may have caused some loss of detail. The reliance on gesture recognition based on self reported data may not be reliable. There are also challenges to strengthen the model further.

**2)*TITLE: Vital Signs Identification System With Doppler Radars and Thermal Camera Publication year:2022***

**Authors:Chian De-Ming, Wen Chao-Kai, Wang Chang-Jen, Hsu Ming-Huan, Wang Fu-Kang [2]**

**Methodology:**VSign-ID System: A thermal camera with an array of SIL radars is used to separate, track and identify vital signs features. Fractional period based techniques allow for

improved deep frequency estimation for resolution and efficiency in multi - person situations; significant improvements are demonstrated across different fractional periods. Signal Matching (Space-time matching), was used to extract vital signs features across multiple stream signals.

**Key Findings:**The use of multiple SIL radars in conjunction with a thermal camera provides VSign ID improved extent range and expanded frequency estimation resolution scope to allow novel efficiency in the separation and tracking of vital signs features. - With NMSE (Normalized Mean Square Error) less than 0dB, this average is indicative of errors in heart

rate detection of less than 3 beats per minute. DeepMining Multi Mode Vector (DeepMining-MMV) outperforms DeepMining-Single Mode Vector (DeepMining-SMV) in multi person cases due to improved processing capabilities.

**Limitations:**The use of multiple independent radars adds complexity to the original design of the system. Body movement contributes to errors from the respiration and heartbeat measures. The nadir of accuracy, can change depending on the arrangement and number of people relative to the radar. In instances of monitoring actively, the measure (NMSE) indicates members observe an unstable nadir of accuracy.

**3)*TITLE:Real-Time Word Level Sign Language Recognition Using YOLOv4 Publication year:2022***

**Authors:Sharma Sneha, Sreemathy R, Turuk Mousami, Jagdale Jayashree, Khurana Soumya[3]**

**Methodology:**Dataset Preparation: This included getting more images to include diversity and size, and manually annotating bounding boxes following the YOLOv4 method. Use YOLOv4 Algorithm: This was selected for our training and testing due to speed of detection and it is real-time for sign language recognition. Use Evaluation Metrics: We measured mAP, average precision to evaluate training and testing performance of the model, and we have a mAP of 98.4%.

**Key Findings:**The YOLOv4 we implemented had an overall mean average precision (mAP) of 98.4% these were 24 words of sign language shown that represented common use signs in everyday life, with real-time sign language recognition. The average precision for all the classes had a minimum of 92%, and 13 of the classes had an average precision of 100%. The training and testing data was splited 80/20, and the training took place over 11 epochs and a batch size of 64.

**Limitations:**The dataset size of Indian Sign Language is limited which may indicate on the accuracy and generalizability of the study. Further datasets need to be studied, thus create or develop a SLR system. There is also still work to be done to manage isolated words and phrases with continuous signing. Performance will likely be determined on the physical spatial space, real time.

**4)*TITLE:Sign Language Recognition and Translation Systems for Enhanced Communication for the Hearing Impaired Publication year:2024***

**Authors:Sindhu Kambhampati Sai, Mehnaaz, Nikitha Biradar, Varma Penumathsa Likhita, Uddagiri Chandrasekhar[4]**

**Methodology:**BLSTM-3D Residual Networks (B3D ResNet) A multimodal method using a deep 3 dimensional residual Convolutional Network to develop the dynamic sign language recognition. Natural Language Processing (NLP) Processing and understanding texts for the text-to-sign recognition, with possible libraries (spaCy and tensoflow). Reversible CNN A model that is less parameterized, but could still achieve a reasonable classfication (and practically very good for gesture classification too), however the reversible CNN classification was quite useful in case of noisy settings.

**Key Findings:**The system translating sign language to text achieved an overall accuracy of 70%. For each metric (i.e. precision, recall, F1 score and Kappa coeffecient), results were as followed: 46.64% (precision), 75.53% (recall), 49.42% (F1 score) and 40.32% (Kappa coeffecient). Isolating the biases that were already present, in the data set, and then working to overcome that bias to allow for lower precision and in turn (at least) a passable performance would take a lot of data collection to overcome, and the previous studies took extensive time, energy, and money into consideration. As the study proposed a reversible CNN model to beat the baseline both classification of sign languages and textual translation of Baselin models for the study, it would highlight value for community of hearing impaired beneficial ubiquitous access in a way never been able to obtain.

**Limitations:**The restrictive vocabulary, as it is a nature of implementation, as being static could impact external validity.

**5)*TITLE:Decoding Information: A Dual Modality Approach for Sign Language Recognition Publication year: 2024***

**Authors:Khattar Princika, Mishra Siya, Tanwar Rahul, Verma Ankit, Bhatia Suman[5]**

**Methodology:**Requirement Analysis: Identifying which users to analyze, user's requirements to improve user experience before creating the model. Research and Exploration: Use of literature into other models and papers to achieve more depth of understanding into the range of features and issues with respect to sign language recognition. Data Collection and Processing: Collection and pre-processing of a few relevant datasets of hand gestures to train the model while accounting for bias.

**Key Findings:**My model had a sample accuracy of 98% which was a substantially better sample accuracy than the state of the art models Deep CNN (94.3%) and Data Glove KNN (96.1448%) in accuracy. The authors of the MRAO Deep CNN model then reported in their original paper that their proposed model accuracy was 92.88%. The authors study was mostly focused on data collection and therefore they were more interested in how their dataset performed in the context of user based and usability based user experiences, with a limited GUI, to enhance the users experience and user satisfaction.

**Limitations:**No limitation was disclosed by the authors in the context of the review. The authors also indicated there are possibilities for future research would focus on modelling performance (accuracy) improvements, and even bigger language datasets.

***6)TITLE:Gesture to Text Recognition using Deep Learning Approach with VGG16 and ResNet50 for Sign Language Publication year:2024***

**Authors:Sharma Jatin, Singh Gill Kanwarpartap, Kumar Mukesh, Rawat Ruchira [6]**

**Methodology:**Data preprocessing: included loading of the datasets and one hot encoding them into a reasonable layout to train a model; Model initialization: used two architectures VGG16 and ResNet50: both have their advantages to recognize ASL gestures; Model assessment confusion matrices were used to assess accuracy of the classification of gestures with gestures that may lead to improvements.

**Key Findings:**VGG16 achieved an accuracy of 99.992% using the test images and 100% accuracy using the assessment dataset, ResNet50 achieved an accuracy of 99.95% using test images and 100% accuracy when using the evaluation dataset indicating that both models learned how to classify the gestures well and that accuracy continued to increase across each epoch of training while loss decreased indicating that the models predictions made fewer and fewer errors. Loss curves for training and testing were very similar which resulted in both models generalizing well for new data safely so that predictions made for ASL gestures were dependable.

**Limitations;**No limitations of the research study are included in this context. Future research, using new sensor technologies, gesture detection across the continuum of cultures across the globe, and using in everyday SLR.

***7)TITLE: Indian Sign Language Recognition with Conversion to Bilingual Text and Audio 2023***

**Authors:Chandarana Nidhi, Manjucha Shreya, Chogale Priyansi, Chhajed Nidhish, Tolani Monica G., Edinburgh Mani Roja M [7]**

**Methodology:** OpenCV Skin Segmentation: Used to find and track Regions of Interest (RoI) for sign language recognition.7 Hu Moment Technique: Analyzes feature values of the images for classification. Random Forest Classifier: Achieved a maximum accuracy of 99.6% for recogninzing Indian Sign Language (ISL) signs.

**Key Findings:** The highest accuracy achieved was 99.6% using a Random Forest Classifier to recognize Indian Sign Language signs (A-Z and 0-9). The dataset used consisted of about 160 images for each sign, with an 80/20 train/test split for the model training. The system fosters communication through translating sign language to English and Hindi text and audio outputs.

**Limitations:**The system is limited to recognizing static images of Indian Sign Language (ISL). The current model only supports English and Hindi. Existing systems mostly cater to users whom have already familiar with Sign Language. Need to continuously improve for real-world application and to increase the base of users.

## III. COMPARARTIVE ANALYSIS

CNNs are fundamental in SLR for spatial feature extraction and can thus be most effective with recognizing static signs. AlexNet and MobileNet scored tremendous accuracies with static datasets (e.g., ASL 98.0% , Thermal 97% ).

RNNs/LSTMs are good for sequence modeling in dynamic sign recognition. While slower and accepting longer input, these increase the accuracy when in tandem with CNNs (e.g., 80.1% [Tsironi], 85% [Khan 2023]).

YOLO goes best for being fast and real-time accurate detection. YOLOv4 has balanced detection speed and precision with a high mAP value of 98.4%.

Hybrid CNN + RNN models better perform than standalone ones for sequential gestures. For instance, CNN-RNN stacks yielded 98.2% by Renjith et al. (2024).

Attention-based and Transformer type architectures are showing promising results for sentence and continuous SLR and winning the battle against vanilla LSTM in resisting inter-frame clutter (98.7% CNNSa-LSTM [6], De Coster et al. [10] for further superior performance).

| Algorithm | Accuracy (%) |
|---|---|
| ResNet50 | 99.95% |
| Random Forest | 99.60% |
| SVM + CNN | 99.17–99.64% |
| YOLOv5 + LSTM | 98.87% |
| CNNSa-LSTM | 98.70% |
| YOLOv4 | 98.40% |

Table 1: Comparative accuracies of different algorithms.

## IV. METHODOLOGY

For developing a strong and efficient thermal sign language recognition (SLR) system, a systematic series of steps need to be adopted:

**Data Acquisition:**Thermal information is acquired through specialized infrared (IR) cameras. The cameras detect heat signatures, rendering the system unaffected by lighting and skin color. This method is privacy friendly and allows recognition even in total darkness, as seen in systems such as FatigueView and object detection models based on thermal inputs .

**Data Preprocessing:**The unprocessed thermal images usually need to be enhanced by normalizing, filtering out noise, and adjusting contrast. These preprocessing operations are important to enhance signal quality, particularly since thermal images are generally low-resolution and can carry thermal noise. Like agricultural data preprocessing to eliminate unnecessary or redundant data, thermal SLR systems can be significantly helped by clean and normalized inputs.

**Algorithm Selection:**Based on the type of data (static or dynamic gestures), various machine learning and deep learning algorithms are chosen. CNNs (such as ResNet50, MobileNet) are employed for spatial feature extraction, RNNs and LSTMs process temporal dynamics in sequences of gestures, whereas attention-based networks and Transformers improve performance in continuous recognition tasks by selectively concentrating on relevant frames and minimizing inter-frame variability.

**Model Training:**Specially selected models are trained using annotated thermal gesture databases. CNNs learn the spatial features from every frame, whereas models such as BiLSTM or Transformer learn the transitions in gestures and relationships between sequences. Similar to crop advisory systems, where Random Forest or XGBoost models are trained to identify patterns in the yield data, deep models in SLR learn patterns in heat-based hand motion.

**Model Evaluation:**Trained models are validated based on their performance using measures of accuracy, precision, recall, and Intersection over Union (IoU). These measures can be used to gauge the model's performance, particularly in detecting the correct gesture across different environments, just like validation techniques employed in other real-time systems.

**Deployment:**Trained and certified models are deployed into real-time systems such as mobile applications, drones, or assistive communication devices. As crop models are deployed through easy-to-use interfaces for farmers, SLR models can be integrated into systems that enable users (primarily the deaf or speech-impaired) to translate signs into text or speech in real time.

## V. RESULTS AND DISCUSSIONS

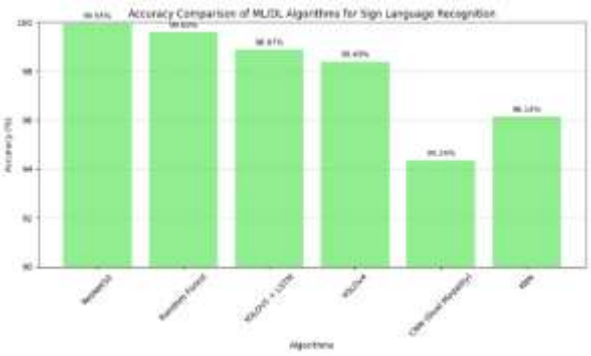


Figure 1: Comparative accuracies of different algorithms.

The bar chart indicates the relative comparison of the accuracy of various ML/DL algorithms employed in Sign Language Recognition (SLR). Out of them, ResNet50 had the maximum accuracy of 99.95%, followed by Random Forest (99.60%), YOLOv5 + LSTM (98.87%), and YOLOv4 (98.40%). Then, CNN (Dual Modality) and KNN scored 94.34% and 96.14% respectively with comparatively poorer performances.

Accuracy in SLR systems is the proportion of signs correctly identified as gestures. Precision (positive accurate predictions), recall (extent of true positive cases), and F1-score (harmonic means of precision and recall) are other measures that tell more about performance particularly in imbalanced or noisy gesture data sets. For instance, a CNN + Self-Attention + LSTM model test attained an accuracy of 98.7% with precision of 98.5% and recall of 98.2% . This recall and precision trade-off is essential for reliable real-time communication.

Besides, AUC-ROC is generally used to validate binary classification issues, while mean class F1-scores are more suitable for multiclass groups of gestures. Apart from accuracy, latency i.e., inference time is the most critical metric for real-time. For instance, thermal SLR systems have been reported to have inference latencies of around 75.1 ms on edge devices like the Jetson AGX .

Note that although RGB-based models (e.g., ResNet, YOLO) work better with enhanced accuracy in low-light conditions, thermal-based models are low-light robust, though at the expense of some loss of detail. The cost factor makes thermal images appropriate for 24/7 and privacy constrained deployments.

## VI. CONCLUSION

State-of-the-art thermal imaging with deep learning is an exciting path toward sign language recognition. With thermal SLR systems' heat pattern detection of gestures, they overcome the limitation of light and most privacy limitations of RGB-based methods. Recent work shows that even low-res thermal cameras with CNN/LSTM/Transformer models can reach >95% accuracy on isolated gestures and digits. These methods have been demonstrated on embedded platforms (Jetson, Raspberry Pi), with real-time inference (<100 ms). However, challenges still exist in creating large, heterogeneous datasets and processing continuous sign streams. We briefly surveyed existing architectures (CNN, RNN/LSTM, hybrids, attention) and relative performance, and made passing mention of practical issues like edge deployment and privacy. Thermal SLR research is moving fast overall, and with additional work in data, model efficiency, and user-facing design, thermal + AI sign translators have the potential to transform deaf community accessibility in the near future.

## REFERENCES

[1] H. Adhikari, M. J. M. Sakib, I. Jahan, M. M. Solaiman, and M. R. Hassan, “A Sign Language Recognition System for Helping Disabled People,” *2023*.

[2] D.-M. Chian, C.-K. Wen, C.-J. Wang, M.-H. Hsu, and F.-K. Wang, “Vital Signs Identification System With Doppler Radars and Thermal Camera,” *2022*.

[3] S. Sharma, R. Sreemathy, M. Turuk, J. Jagdale, and S. Khurana, “Real-Time Word Level Sign Language Recognition Using YOLOv4,” in *Proc. Int. Conf. on Futuristic Technologies (INCOFT)*, Karnataka, India, 2022, pp. 1–6.

[4] S. K. Sai, M. Mehnaaz, N. Biradar, V. P. Likhita, and U. Chandrasekhar, “Sign Language Recognition and Translation Systems for Enhanced Communication for the Hearing Impaired,” *2024*.

[5] P. Khattar, S. Mishra, R. Tanwar, A. Verma, and S. Bhatia, “Decoding Information: A Dual Modality Approach for Sign Language Recognition,” in *Proc. Int. Conf. on Computing, Sciences and Communications (ICCSC)*, 2024, pp. 1–6.

[6] J. Sharma, K. S. Gill, M. Kumar, and R. Rawat, “Gesture to Text Recognition using Deep Learning Approach with VGG16 and ResNet50 for Sign Language,” in *Proc. 4th Asian Conf. on Innovation in Technology (ASIANCON)*, Pune, India, 2024, pp. 1–5.

[7] N. Chandarana, S. Manjucha, P. Chogale, N. Chhajed, M. G. Tolani, and M. R. M. Edinburgh, “Indian Sign Language Recognition with Conversion to Bilingual Text and Audio,” in *Proc. Int. Conf. on Advanced Computing Technologies and Applications (ICACTA)*, 2023, pp. 1–6.

[8] S. Birkeland, L. J. Fjeldvik, N. Noori, S. R. Yeduri, and L. R. Cenkeramaddi, "Thermal video-based hand gestures recognition using lightweight CNN," *J. Ambient Intell. Human. Comput.*, vol. 15, pp. 3849–3860, 2024.

[9] S. R. Yeduri, D. S. Breland, S. B. Skriubakken, O. J. Pandey, and L. R. Cenkeramaddi, "Deep learning-based sign language digits recognition from thermal images with edge computing system," *IEEE Sensors J.*, vol. 21, no. 9, pp. 10445–10453, 2021.

[10] D. S. Breland, A. Dayal, A. Jha, P. K. Yalavarthy, O. J. Pandey, and L. R. Cenkeramaddi, "Robust hand gestures recognition using a deep CNN and thermal images," *IEEE Sensors J.*, vol. 21, no. 9, pp. 10395–10403, 2021.

[11] Z. Zhang, H. Ren, H. Li, K. Yuan, and C. Zhu, "Static gesture recognition based on thermal imaging sensors," *J. Supercomputing*, submitted Aug. 2024.

[12] Y. Zhang and X. Jiang, "Recent advances on deep learning for sign language recognition," *CMES: Comp. Model. Eng. Sci.*, vol. 139, no. 3, pp. 2399–2450, 2024.

[13] Y. Zhang *et al.*, "Low resolution thermal imaging dataset of sign language digits," *Data Brief*, vol. 41, 107977, Feb. 2022.

[14] Y. Amin, S. T. H. Rizvi, and M. M. Hossain, "A comparative review on applications of different sensors for sign language recognition," *J. Imaging*, vol. 8, no. 4, p. 98, Apr. 2022.

[15] J. Khan *et al.*, "Isolated video-based sign language recognition using a hybrid CNN–LSTM framework with attention," *Electronics*, vol. 13, no. 7, p. 1229, Mar. 2023.

[16] J. Ballow and S. Dey, "Real-time hand gesture identification in thermal images," *arXiv:2303.02321*, 2023.

[17] M. De Sisto *et al.*, "Challenges with sign language datasets for recognition and translation," in *Proc. LREC*, Marseille, France, 2022, pp. 2353–2360.