A Comprehensive Review of Sign Language Recognition: Different Types, Modalities, and Datasets

M. MADHIARASAN[1*], Member, IEEE [1*]*Department of Computer Science and Engineering, Indian Institute of Technology Roorkee,* PARTHA PRATIM ROY[2], Member, IEEE [2]*Department of Computer Science and Engineering, Indian Institute of Technology Roorkee*

***Abstract*—A machine can understand human activities, and the meaning of signs can help overcome the communication barriers between the inaudible and ordinary people. Sign Language Recognition (SLR) is a fascinating research area and a crucial task concerning computer vision and pattern recognition. Recently, SLR usage has increased in many applications, but the environment, background image resolution, modalities, and datasets affect the performance a lot. Many researchers have been striving to carry out generic real-time SLR models. This review paper facilitates a comprehensive overview of SLR and discusses the needs, challenges, and problems associated with SLR. We study related works about manual and non-manual, various modalities, and datasets. Research progress and existing state-of-the-art SLR models over the past decade have been reviewed. Finally, we find the research gap and limitations in this domain and suggest future directions. This review paper will be helpful for readers and researchers to get complete guidance about SLR and the progressive design of the state-of-the-art SLR model.**

***Index Terms*—Artificial Intelligence, Sign Language Recognition, Datasets, and Human-Computer Interaction.**

I. Introduction

According to the WHO (World Health Organization) report, over 466 million people are speech or hearing impaired, and 80% of them are semi-illiterate or illiterate [1]. Non-verbal manner conveys and communicates our views, emotions, and thoughts visually through sign language. Compared to spoken language, sign language grammar is quite different. A sign comprises specific hands, shapes, or signals produced in a particular location on or around the signer's body combined with a specific movement.

Hand gestures, signals, body movements, facial expressions, and lip movements are the visual means of communication used by the hand-talk community and ordinary people to convey the meaning; We recognize this language as a sign language. Sign language recognition (SLR) is challenging and complex, and many research opportunities are available with the present technology of artificial intelligence. A taxonomy of SLR is shown in Figure 1. It comprises datasets, input modality, features, classification, computational resources, and applications. The dataset is further classified into isolated sign dataset and continuous sign dataset. Vision-based modality and sensor-based modality are the general types of input modality. Hand movement, facial expression, and body movement are the major features that concern SLR. Classification is typified into traditional methods (HMM, RNN, etc.), deep learning (CNN), and hybrid method (combination of traditional and deep learning or combination of deep learning and optimization algorithm).

SLR aims to understand the gestures by suitable techniques, which requires identifying the features and classifying the sign as gesture recognition. In the literature, there is no comprehensive review paper addressing the aspect of the modality (vision and sensor), different types (isolated (manual and no manual) and continuous (manual and no manual)), various sign language datasets, and state-of-the-art methods based studies. This review study focuses on SLR-based research work, recent trends, and barrier concerns to sign language. Different sign languages, modalities, and datasets in sign language have been discussed and presented in tabular form to understand better. From databases like IEEE explore digital library, science direct, springer, web of science, and google scholar, we used the keywords sign language recognition to identify significant related works that exist in the past two decades have included for this review work. We excluded papers other than out-of-scope sign language recognition and not written in English. The contributions to this comprehensive SLR review paper are as follows:

- Carried out a review of the past two decades of published related work on isolated manual SLR, isolated non-manual SLR, continuous manual SLR, and continuous non-manual SLR.
- Discussed different sensing approaches for sign language recognition and modality
- This paper presents SLR datasets concerned with isolated and continuous, various sign languages, and the complexity of the datasets discussed.
- Discussed the framework of SLR and provided insightful guidance on SLR

Dr. M. MADHIARASAN was with the Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India – 247667, *Corresponding Author, e-mail: [1*]mmadhiarasan89@gmail.com, mmadhiarasan.cse@sric.iitr.ac.in, [1*] Orcid ID: 0000-0003-2552-0400.

Dr. PARTHA PRATIM ROY was with the Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India – 247667, e-mail:partha@cs.iitr.ac.in, Orcid ID: 0000-0002-5735-5254.

Manuscript received April xx, 2022; revised xx, 2022.

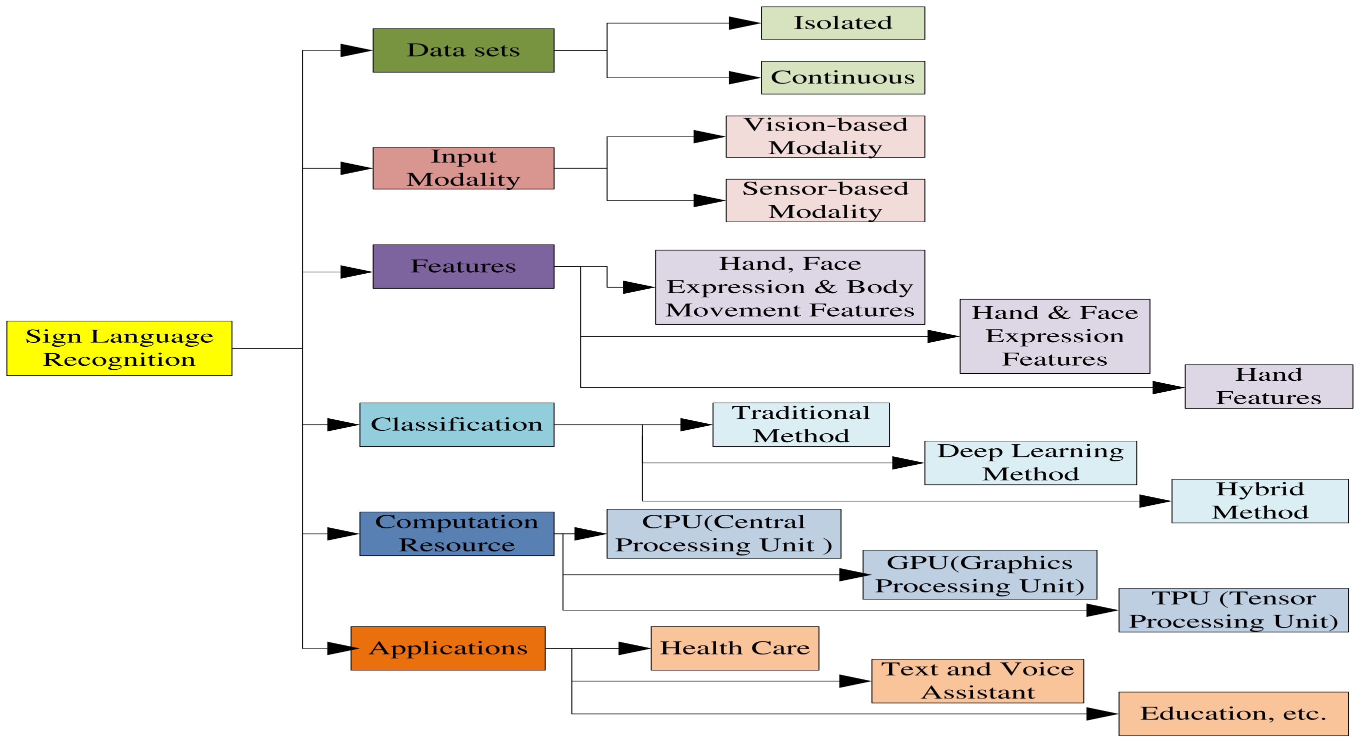

Fig. 1: SLR Taxonomy: the fundamental attributes of SLR like datasets, modality, features, classification, computation resources, and application, along with each attribute's categorization are shown here. Large-scale datasets and modalities affect the recognition performance. The efficient features extraction method and classification model with efficient power computation resources lead to high performance.

- Point out the limitations related to the dataset and current trends available in the SLR and potential application of SLR with human-computer interaction.
- This paper studied the results of the current state-of-the-art SLR model regarding the various benchmark SLR datasets for isolated and continuous SLR.
- This paper analyzes current SLR issues and advises future SLR research direction.

A. Need of SLR

As per WHO statistics, around 5% of the population in the world suffers from a lack of hearing power. According to the prediction of the United Nations, the number of deaf people in 2050 will be 900 million [1]. Hence, SLR receives a lot of attention at present. SLR can eliminate the communication gap between the hand-talk community (deaf and dumb). Also, SLR helps to improve communication in the following ways.

- It reduces the frustration of the hand-talk community.
- The communication barrier overcome by SLR leads to effective communication.

Much research endeavored to develop high-performance SLR. Despite that, it is challenging, and it is one of the recent research fields with enormous research scope.

B. Challenges

SLR comprises numerous gestures and facial expressions, making it complex and challenging. In addition, to the manual components, lip shapes and eyebrow positions distinguish similar signs; e.g., many manual signs seem to be of a similar pose. However, these can be differentiated with the help of facial expression and lip movement. Sign language comprises hand movement, shape, position, orientation, palm posture, finger movement, facial expression, and body movements. These components highly influence the performance of SLR. Some of the barriers and problems of SLR are tabulated in Table I. With the advance of hardware, efficient algorithms can improve the processing speed. The scaling and image orientation problems can be resolved with recent deep learning techniques. The illumination problem can be overcome if the RGB is converted to HSV (Hue Saturation Value) or Ycbcr (Luminance Chrominance). Dynamic and non-uniform background environment problems could be resolved using the skin region and background subtraction method.

C. Procedure involved in SLR

The SLR involves data collection, preprocessing, feature extraction, and classification phase. The block diagram of SLR and its general process is demonstrated in Figure 2. These stages are discussed in the following. Note that, for the sensor-based approach, preprocessing and segmentation are optional.

Data Collection: In SLR, the data acquisition is performed using one of two modes; Vision and Sensor. In a vision-based approach, the input is an image or video [2], [3]. A single

TABLE I: SLR Barrier, and Problem: We discussed how the barriers (dynamics, illumination of lights and environment, scaling, and computation time) of SLR cause a problem.

Barrier	**Problem**
Computation speed and time	Create complexity to the system and take a lot of computation time.
Scaling and image orientation problem	The distance of input data capturing various signers.
Illumination of light	Performance varies with different illumination scenarios because most models use the RGB model. It is highly illumination sensitive.
Dynamic and non-uniform background environment	The noise, improper detection of hand, and face lead to affect the performance and mislead the sign recognition system.

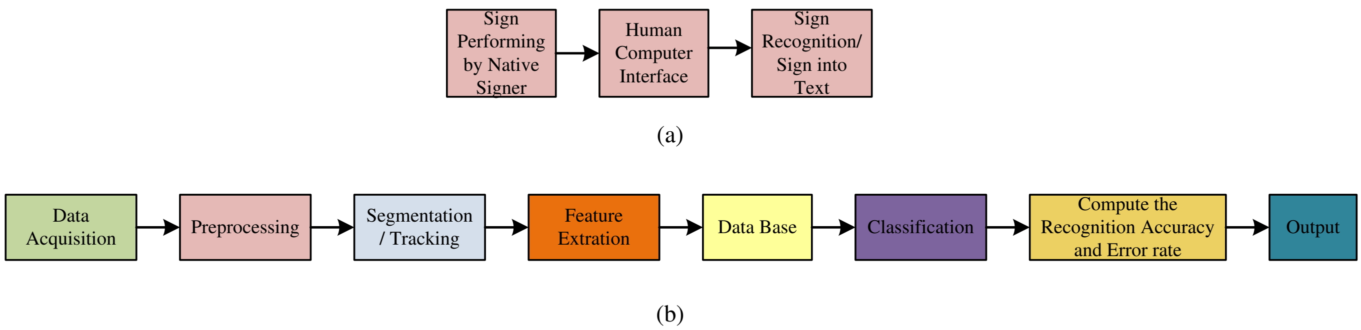

Fig. 2: (a) SLR (b) General process flow of SLR. Figure (a) illustrates the native signer performing sign conversion into text with the help of a human-computer interface, and figure (b) illustrates the procedural stages of SLR. The recognition rate highly depends on the data set, preprocessing, feature extraction, and classifier.

camera is used to collect standard signs while multiple cameras, active and invasive devices, help collect the depth information. Video camera, webcam, or smartphone device [4], [5], [6], [7] captured the continuous motion. The sensor-based approach collects the signal with the use of the sensor [8], [9], [10], [11].

Image Preprocessing: The performance of the SLR system can be improved by preprocessing methods such as dimension reduction, normalization, and noise removal. [12].

Segmentation: The segmentation stage splits the image into various parts or ROI (Region of Interest) [13], Skin Colour Segmentation [14], HTS (Hands Tracking and Segmentation) [15], Entropy Analysis and PIM (Picture Information Measure) [16]. The background requires the hand gesture extraction to be done effectively by segmentation and tracking process.

Tracking: Tracking of hand position and facial expression from the acquired image/video can be performed using camshaft (continuously adaptive mean shift used to track the head position) [17], Adaboost with HOG (Histogram of the gradient) [18], Particle filtering (KPF-Kalman Particle Filter) [19].

Feature Extraction: Transforming preprocessed input data into the feature space is known as feature extraction. Further, it is discussed in detail in section 2.

Data Base: The acquired data (image/video) is stored in the database and classified into two sets, namely training and testing datasets [20]. The classifier learns by training dataset and the performance is evaluated by testing data.

Classification: The classifiers perform the classification by extracting features and classify the sign gesture. The Hidden Markov Model (HMM) [9], [21], Long-Short Term Memory (LSTM) [22] Deep Learning network [23], and hybrid classifier [2], [24] are used as classifiers to recognize sign language.

Evaluation Stage: The performance of a trained classifier is validated with a testing dataset (unseen data during training) [25]. The error incurred during classification gauges sign recognition performance.

Although there are few review papers in the literature [26], [12], however, they lack focus and understanding of SLR. This paper provides a comprehensive SLR preamble, recent research progress, barriers or limitations, research gap, and future research direction and scope. We organized the rest of the review paper as follows. Section 2 presents sign language modality, preprocessing, and the various feature extraction methods in SLR. Carried out a literature review concerning the manual and non-manual aspects of SLR in Section 3; Section 4 discusses and illustrates the classification architecture of SLR. Section 5 presents various types of SLR, datasets concerning SLR, and reviews work related to the modalities, current state-of-the-art models based on SLR. The recent trends, challenges, and limitations are highlighted in Section 6. Sections 7 and 8 pointed out future research discussion and conclusion, respectively.

TABLE II: The importance of vision-based and sensor-based methods are shown here according to the obstacle, cost, merits, and demerits. Much research work focuses on vision-based SLR because of its feasibility in real-time applications.

Method	**Capturing Device**	**Obstacle**	**Efficiency**	**Cost**	**Limitation**	**Advantage**
Vision-based Method	Video Camera	Environment, disturbance, and noise.	Low (depends on the resolution).	Low	Possess challenging concerns for time, speed, and overlapping. More Feature extraction techniques are required.	Fast Speed.
Sensor-based or gloves-based Method	Sensors and gloves	Environment, disturbance, and noise.	Better than vision-based method (depends on sensor performance).	High	Not suitable for real-time application.	Better performance. Require minimal feature extraction.

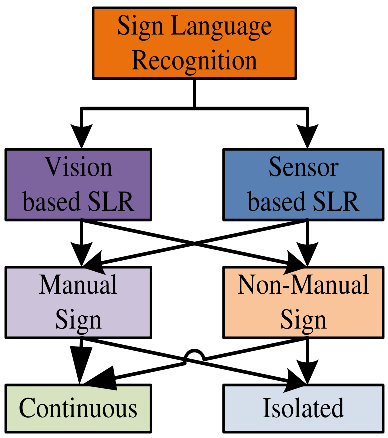

Fig. 3: SLR Types: Vision-based SLR and sensor-based SLR are the SLR types. It is further, classified into manual and non-manual, then isolated and continuous.

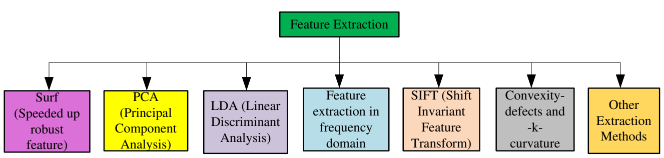

Fig. 4: Various methods are used to extract the significant features.

TABLE III: Various features extraction methods associated with SLR work exist in the literature.

Method	**Year**	**Author**
FPM (Feature Pooling Module)	2020	Sincan and Keles [27]
Histograms of oriented gradients	2016	Chansri and Srinonchat [28]
Euclidean distance	2016	Pansare and Ingle [29]
	2013	Singha and Das [30]
DWT (Discrete Wavelet Transform)	2017	Ahmed et al. [31]
	2016	Prasad et al. [32]
SIFT (Scale Invariant Feature Transform)	2012	Gurjal P [33]
SURF (Speeded Up Robust Feature)	2012	Yao and Li [34]
Fourier Descriptors	2017	Kumar [35]
	2015	Shukla et al. [36]
PCA (Principal Component Analysis)	2021	Gurbuz et al. [37]
	2015	Tripathi and Nandi [4]
Fuzzy neural network	2016	Dour and Sharma [38]

II. Modalities of SLR

SLR is one of the most prominent research areas in computer vision and natural language processing. In concern to the acquisition process, the SLR system is classified as a sensor-based and vision-based approach. Both approaches are next classified as manual and non-manual, and further classified as isolated and continuous. Figure 3 illustrates the SLR types. Much research work focused on isolated manual-based SLR. Only a little research work addressed continuous non-manual SLR.

Sensor-based approach: Physically attached sensors acquire trajectories of the head, finger, and motion of the signer. Sensor-associated gloves track the signer's hand articulations and recognize the sign. The comparison of SLR methods shown in Table II clarifies vision and sensor-based approaches. In contrast with vision-based SLR, sensor-based SLR provides efficient performance.

Vision-based approach: The gestures captured by multiple cameras (or webcam) are recognized using the vision/image-based approach. From the captured image/video, it extracts palm, finger, and hand movement features. With the help of these extracted features, classification was performed. Poor illumination or lighting environment, noisy background, and blurring present in the image result in misclassification. Although vision-based SLR is suitable for real-time conditions, it must adequately care for preprocessing, feature extraction, and classification.

A. Preprocessing

The computational burden of data processing could be reduced by preprocessing methods. Image reduction and image conversion methods do the size reduction and conversion from color to gray scale. Image reduction methods reduce the burden of data processing. The unwanted object can be removed by the histogram equalization [39]. The noise present in the image are removed using the filter, like median, moving average method, and so on [40]. Gaussian average methods are used to remove the image background component [41]. Filters perform removal of the unwanted components and minimize the size of the data with the help of image edge detection algorithm [42]. The filter process speeds up with the help of fast Fourier transformation because instead of an image, the frequency domain is used [43]. The image is split into possible segments [44]; masking is used in segmentation to improve processing. Elimination of background effect using binarization histogram equalization aid for better image contrast. Normalization methods can effectively handle the variance in the data [45].

B. Feature Extraction

In SLR, relevant feature extraction plays a vital role. It is crucial for sign language, as irrelevant features lead to misclassification [46]. The feature extraction aid in accuracy improvement, and speed [47]. Some of these feature extraction method include SURF (Speeded Up Robust Feature) [34], speed up robust feature (Laplace of Gaussian with box filter) [34], SIFT (shift-invariant feature transform) [33], PCA (Principal Component Analysis) [37], [4], LDA (Linear Discriminant Analysis) [48], Convexity defects and k-curvature [49], time domain to frequency domain [31], [35], Local binary pattern, etc. The feature extraction methods used for SLR-based study is tabulated in Table III. Various feature extraction methods are showed in Figure 4. Feature vector dimension reduction performed by PCA, LDA, etc. aid in reducing the computational burden on the classifiers. The dimensionality pruning, features reduction, and lowering of the dimension keep the significant features of high variance and minimizing remaining features, thus, reduces the training complexity. Fourier descriptors are noise resistance and invariant to scale, orientation, and normalization is easy. The process of transforming the correlated into an uncorrected value is known as principal component analysis. Original data are linearly transformed effectively, and the feature vectors get reduced.

The preprocessing and feature extraction methods aid the classifier. Also, they reduce the computation burden, avoid overfitting issues, and wrong recognition possibilities. SIFT's merits are invariant to lighting, orientation, and scale. However, the performance is not satisfactory [33]. Using Histogram of Oriented Gradients (HOG) [28], the unwanted information is removed, keeping the significant features to ease the image processing. The feature vectors are obtained using the computation of gradient margin and angle. As HOG cell size and the number of bins increase, the extracted feature also increases. Larger subdivisions furnish global information, and small subdivisions given local information that is worthwhile. The demerits of both the methods are that they require more memory. SURF is invariant to image transformation and a faster feature extractor than SIFT. Still, it has the requirement of camera setup in horizontal position for better performance, and the disadvantage is illumination-dependent, not rational. Location and frequency captured using a Discrete Wavelet Transform. Temporal resolution is the critical merit of DWT [31], [32].

III. Literature studies about SLR

Sign language is not generic; it varies according to the region and country [1]. The sign language classification is available in over 300 sign languages worldwide, namely ASL, BSL, ISL, etc. According to Ethnologue 2014 [50] in the United States, ASL is a native language for around 2,50,000-5,00,000 people. Chinese Sign Language is being used in China by approximately 1M to 20M deaf people. Approximately 1,50,000 people in the United Kingdom use British Sign Language (BSL). In Brazil, approximately 3 million signers use Brazilian Sign Language to communicate, like Portuguese Sign Language or French Sign Language. According to Ethnologue 2008 in India, approximately 1.5 million signers use Indo-Pakistani Sign Language.

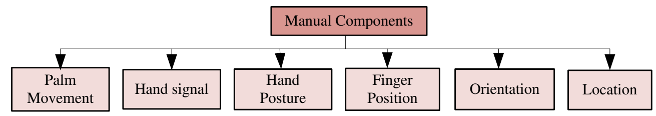

Fig. 5: Manual Components. The important manual features related to sign language are shown here.

SLR is not only meant for deaf and mute people. Ordinary people also communicate information in the noisy area of public places and the library without disturbing others. Manuel (communication by hands) and non-manual (communication by body posture or facial expression) medium are usually used in sign language. People use sometimes finger spelling which is communicated by splitting words into letters, then spelling the letter using fingers). Manual and non-manual SLR are discussed in detail in following subsections.

A. Manual SLR

Hand motion, hand posture, hand shape, and hand location are the manual sign components. Figure 5 shows the manual sign components. With one hand or two hands, the signer usually communicates with others. The manual SLR is classified into isolated and continuous.

1) Isolated Manual SLR: The literature work on isolated manual SLR are as follows:

Classical methods: Ong et al. [51] suggested sequential Pattern Tree-based multi-class classifier for DGS (German Sign Language (Deutsche Gebärdensprache) and Greek Sign Language (GSL) recognition. Their proposed SP-Tree Boosting algorithm-based recognition model performs better than the Hidden Markov Model. Chansri and Srinonchat [28] proposed data fusion incurred ANN-based Thai SLR model. They extracted the hand feature using histograms of oriented gradients, and they did classification using a back-propagation algorithm associated with an ANN. Yin et al. [52] performed hand gesture recognition using a joint algorithm based on BP and template matching method combination. The joint algorithm takes computation time as 0.0134 and an accuracy of 99.8% was achieved for isolated hand gesture recognition. Jane and Sasidhar [53] carried out an ANN classifier with an association of data fusion. They performed three hidden layers of artificial neural network with wavelet denoising and TKEO (TeagerKaiser energy operator) methods for a SEE (Signing Exact English). Based on this approach, the recognition rate is 93.27%. Korean finger language recognition model was developed based on ensemble ANN [11]. The performance was analyzed by varying dataset size (50 to 1500) and classifier (1 to 10). The comparative analysis of eight ANN classifier-associated ensemble models identifies 300 training datasets as an optimal structure to lead to 97.4 % recognition accuracy for Korean finger language recognition.

Almeida et al. [54] extracted seven vision-based features using RGB-D sensor. They recognized Brazilian Sign Language with an average of 80% using the SVM. They did phonological structure-based decomposition and extraction of signs. Hence, they suggested a model suitable for other SLR purposes. Fatmi et al. [9] performed SLR based on ANN and SVM. They have compared their performance with HMM. Comparison with other machine learning techniques to ASL words, higher accuracy achieved by proposing ANN. Lee and Lee [55] developed SVM based on a sign language interpretation device with 98.2 % recognition accuracy. SVM classifier-based sign interpretation device developed system. Wei et al. [10] presented the CSL sign recognition model using a code matching method by including a fuzzy K-mean algorithm. They determined subclass by a fuzzy K-mean algorithm and classification was done with the Code matching method. Li et al. [56] suggested ASL recognition prototype model based on KNN, LDA, and SVM classifiers. They carried out a prototype model based on LDA, KNN, and SVM classifiers using a firmly stretchable strain sensor for ASL 0-9 number sign recognition. The authors reported that the model achieved an average accuracy of 98%.

Yang et al. [57] performed a Chinese Sign Language (CSL) recognition model based on sensor fusion decision tree and Multi-Stream Hidden Markov Models classifier. They developed a wearable sensor associated with the Chinese SLR model with user-dependent and user-independent using Multi-Stream Hidden Markov Models. The searching range improved by optimized tree-structure classification. Dawod and Chakpitak [58] carried out work on real-time recognition model for ASL alphabets and numbers sign recognition. They used RDF (Random Decision Forest) and HCRF (Hidden Conditional Random Field) based classifiers and Microsoft Kinect sensor v2 for the data collection. The HCRF classifier-based recognition model gets the mean accuracy for numbers-based sign recognition as 99.99% and alphabets sign recognition as 99.9%. The RDF-based recognition model achieved mean accuracy for number sign recognition as 96.3% and alphabets sign recognition as 97.7%. Hence, the HCRF based sign recognition model leads to better performance than RDF for both ASL numbers and alphabets recognition. Hrúz et al. [59] presented a Hidden Markov Model-based Czech SLR model with an association of kiosk. Also, they performed SLR, automatic speech recognition, and sign language synthesis.

Mummadi et al. [60] proposed an LSF model based on IMU sensors associated with wearable hand gloves with various classifiers like naïve Bayes, MLP, and RF. Real-time wearable IMU sensor-based glove-associated sign recognition model developed for LSF recognition instead of complimentary filter advanced fusion strategy and the advanced classifier can improve the accuracy rate. Botros et al. [8] presented a comparative analysis of wrist-based gesture recognition using EMG signal. Forearm and wrist level-based gesture are recognized using EMG signal. Gupta and Kumar [61] performed a wearable sensor-based multi-class label incurred SLR model. The LP-based SLR model has a minimal error and computation time than the tree-based, BR (binary relevance), and CC (Classifier Chain) based sign recognition models. Compared to the classic tree classification model, the suggested model performs well with minimal classification errors. Hoang [62] presented a new vision-based captured ASL alphabets sign dataset (HGM-4). With this dataset, using a classifier, developed a contactless SLR system.

Deep learning approaches: Al-Hammadi et al. [46] performed sign dependent and sign independent SLR using three datasets using single and fusion parallel 3DCNN. The proposed model gets a better recognition rate than other considered six existing literature methods. Sincan and Keles [27] performed CNN and LSTM based SLR model for Turkish SLR. The feature extraction improved by FPM (Feature Pooling Module), convergence speeds up using the attention model. Yuan et al. [24] pointed out DCNN (deep convolution neural network and LSTM (long short-term memory) based model for hand gesture recognition. The residual module has overcome the gradient vanishing and overfitting problem. Complex hand gesture long-distance dependency problem addressed by improved deep feature fusion network. Compared to Bayes, KNN, SVM, CNN, LSTM, and CNN-LSTM, the DFFN based model performs well on ASL and CSL datasets.

Aly and Aly [2] designed an Arabic SLR model using BiLSTM (deep Bi-directional Long Short Term Memory recurrent neural network). Convolutional Self-Organizing Map for hand shape feature extraction, and DeepLabv3+ extracts hand regions. The suggested model proved validity on signer-independent real Arabic SLR. The proposed model is suitable for an isolated sign, and continuous sign-based analysis can be a future direction. Rastgoo et al. [3] carried out work on a multi-modal and multi-view hand skeleton-based SLR model. Features fusion and single-view vs. a multi-view projection of hand skeleton-based performance analysis performed. SSD (Single Shot Detector), 2DCNN (2D Convolutional Neural Network), 3DCNN (3D Convolutional Neural Network), and LSTM (long short-term memory) based deep pipe-line architectures were proposed to recognize the hand sign language automatically. Lee et al. [22] designed the k-Nearest-Neighbour method associated with Long-Short Term Memory (LSTM) recurrent neural network-based American SLR model. The leap motion controller is used to gain the sign data. Compared to SVM, RNN, and LSTM models, the proposed model (LSTM with KNN) outperforms 99.44%.

For a clear understanding, the research work related to isolated manual SLR are tabulated in Table IV and graphical representation is shown in Figure 6. The recognition model results in good accuracy for isolated sign recognition, not assured to be generalized for continuous sign recognition with better precision.

2) Continuous Manual SLR: Processing one-dimensional data is simpler compared to handling a high-dimension dataset like video [63]. Continuous SLR with uncontrolled environment-based SLR is quite complex as there is no clear pause after each gesture.

TABLE IV: Isolated Manual SLR Literature Work. Related work with regard to vision and sensor based SLR model concerns the isolated manual sign comprehensively summarized in a tabular form for better understanding.

Year	Author	Method	Pros	Cons	Accuracy or Result	Sensor	Sample and Lexicon Size	Model Type	Dataset
2021	Yuan et al. [24]	CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory)	Wearable gloves based hand gesture recognition model.	Real-time live data-based analysis with the uncontrolled environment not performed to prove the validity.	Recognition accuracy: 99.93% for ASL and 96.1% for CSL.	3-dimensional flex sensor-based data gloves, gyroscope, accelerometer, and bending sensor.	26 ASL alphabets, daily activities for CSL, 6 subjects	America sign language, Chinese sign language.	A total of 5178 hand motions of ASL alphabets and 100 daily life activities sign of CSL with a total of 34452 hand motions.
2021	Botros et al. [8]	Linear Discriminant Analysis (LDA) used as a classifier, PCA used to reduce the dimension.	The reliable wearable device, less affected by noise	Continuous sign based experimentation was lacking limited regard to the dataset.	For multi-finger gestures, accuracy: 91.2%, single finger accuracy: 92.1%, and conventional wrist gestures accuracy: 94.7%.	EMG sensor	17 various single fingers, multi-finger, and wrist gestures, 21 subjects	human-computer interaction: Hand gesture sign.	72 repetitions per participant of 5 single-finger, 6 multi-finger gestures, and 6 wrist gestures are collected.
2021	Lee et al. [22]	K-Nearest-Neighbour method, Long-Short Term Memory (LSTM), and Recurrent Neural Network.	Suitable for real-time environment/ applications.	One hand sign (right hand) only considered for recognition.	Recognition accuracy: 99.44% and validation (5-fold cross-validation) accuracy: 91.8%.	Leap motion controller	26 ASL alphabets, 100 subjects.	American SLR	2600 samples (26 *100 samples).
2020	Gupta and Kumar [61]	Multi class classification.	Computation time is less.	Expensive and continuous sign based on experimentation was lacking.	LP (label power set) model average classification error: 2.73%.	Surface electromyogram and inertial measurement units.	100 signs, 10 subjects.	Indian SLR	20,000 samples.
2020	Hoang [62]	Multi camera-based sign (ASL) data collection.	Different position of hand gesture captured	Gesture recognition not performed.	NA	4 cameras (front, back, right, and left), laptop camera.	26 letter sign, 5 subjects.	America sign language.	4,160 samples.
2020	Al-Hammadi et al. [46]	Single 3DCNN, and PARALLEL 3DCNN.	Better recognition rate, generalize well with three datasets.	Optimal selections of hyperparameter problem, real-time practical implementation with a live sign, was lacking.	Recognition rate for dataset 1 (SSL -SAUDI SIGN LANGUAGE) single 3DCNN signer-dependent mode: 96.69%, signer independent recognition rate: 72.32%, parallel 3DCNN signer-dependent mode: 98.12%, and signer independent recognition rate: 84.38%, for dataset 2 (ArSL-ARABIC SIGN LANGUAGE) signer-dependent mode: 100%, signer independent recognition rate: 34.9%, for dataset 3 (ASL-AMERICAN SIGN LANGUAGE) signer-dependent mode: 76.67% and signer independent recognition rate: 70%.	RGB cameras, Microsoft Kinect, analog camcorder.	Dataset 1 (40 subjects, forty gesture classes(200 gesture)), dataset 2 (3 subjects, 23 gestures, 150 samples), dataset 3 (43 classes, 40 subjects).	SAUDI SIGN LANGUAGE, ARABIC SIGN LANGUAGE, AMERICAN SIGN LANGUAGE.	Dataset 1: signer dependent total 8000 samples, signer independent total 6400 video samples. Dataset 2: signer dependent total 3444 samples, signer independent total 3444 samples. Dataset 2: signer dependent total 280 samples, signer independent total 280 samples.
2020	Aly and Aly [2]	Convolutional SOM, deep Bi-directional LSTM network, and DeepLabv3+.	Signers independent combinational sign recognition model.	Experimentation performed with a limited number of signers.	Accuracy without segmentation: 69.0% and with segmentation accuracy: 89.5%.	Video camera: 25 fps (frames per second).	23 words, 150 sequences, and 3 subjects.	Arabic sign language.	3450 samples.
2020	Rastgoo et al. [3]	Midpoint algorithm, SSD (Single Shot Detector), 2DCNN (2D Convolutional Neural Network), 3DCNN (3D Convolutional Neural Network), and LSTM (Long Short-Term Memory).	Large-scale hand sign language dataset presented	Complex model.	Accuracy: 96.2% for the RKS-PERSIANSIGN dataset, 82.10% for First-Person datasets, and an average estimation error: 12.82 for the NYU dataset.	Video camera (RGB).	100 sign words, 10 subjects, 10 various environment.	Persian signs language.	First-Person (45 hand action, 100 K frames), NYU (36 joints, 81,009 image sequence), and RKS-PERSIANSIGN (100 signs, 10,000 samples) datasets.
2020	Sincan and Keles [27]	CNNs (Convolutional Neural Networks), Feature Pooling Module, unidirectional and bidirectional LSTM (Long Short-Term Memory).	New AUTSL dataset presented	Accuracy is lagging state-of-the-art methods. It is affected by the dynamic background.	AUTSL dataset based recognition model accuracy: 95.46%, Montalbano dataset based recognition model accuracy: 96.11% and for user-independent benchmark dataset accuracy: 62.02%.	Microsoft Kinect v2.	AUTSL and Montalbano datasets, 226 signs, 43 subjects.	Turkish Sign Language (TSL).	AUTSL: 38,336 samples, Montalbano: 14,000 samples.
2019	Dawod and Chakpitak [58]	HCRF (Hidden Conditional Random Field), and RDF (Random Decision Forest).	Higher recognition rates	The authors fail to perform quantitative-based analysis.	HCRF classifier mean accuracy (numbers): 99.99% and alphabets: 99.9%, and the RDF mean accuracy (number):96.3% and (alphabets): 97.7%.	Kinect sensor v2.	A-Z alphabets and 1-20 numbers, 30 subjects (signer).	American Sign Language.	345000 samples.
2019	Fatmi et al. [9]	ANN, SVM, and HMM.	Functioning ability on a PC with Bluetooth Low-Energy (BLE) connections.	The limitation is not considered non-manual, and dictionary size is limited.	1. ANN: 93.79%, 2. HMM: 85.90%, and 3. SVM: 85.56%.	Myo Armband (x2) (sense motion and depth).	13 ASL gestures signs, 3 subject.	American Sign Language.	26,000 instances captured, among 66 % used for training and remaining used for testing.
2018	Jiang et al. [48]	Linear Discriminant Analysis (LDA).	Simpler than forearm device.	The limitation is that calibration is required, and classification accuracy depends on calibration, and time also elapsed.	Accuracy for air gesture: 92.6% and for surface gestures: 88.8%.	Sensing fusion using 4-Channels sEMG Electromyography, one IMU inertial measurement unit.	8 air gestures and 4 surface gestures with 2 distinct force levels, 10 subjects.	Human-computer interaction: Hand gesture Sign (Signing Exact English).	3 trials of datasets.
2018	Mummadi et al. [60]	Naïve Bayes, Feed-forward Neural Network(MLP), Random Forest are used as a classifier. A complementary Filter with a coefficient factor of 0.93 was used to obtain low drift and less noise data.	63 milliseconds taken to recognize the sign (settling times and delays are faster and minimal respectively than local fusion algorithm with IMU motion sensor-based method).	The performance of the model is depending on sensor noise and drift. Distance between sensor and hand increase, accuracy decreased.	Naïve Bayes method accuracy: 89.1%, and FI score: 87%, MLP method accuracy: 92.2%, and FI score: 91.1%, RF method accuracy: 92.95%, and FI score: 91.98%. Mean accuracy: 92% and mean F1 score: 91%.	3D magnetometer, 3D gyroscope (GYRO), and 3D accelerometer (ACC).	22 hand gestures, 57 subjects.	French Sign Language (LSF).	1.25 million samples.
2018	Kim et al. [11]	Ensemble Feed-forward Neural Networks.	Easy to wear, adaptable to portable devices.	Selection of hyperparameter was not addressed, suffered by convergence and computation problem.	E- ANN with 8 Classifiers accuracy: 97.4%.	Armband module (8-channel electromyography) sensors.	7 numbers, 17 vowels and 14 consonants, 17 subjects.	Korean finger language.	1500 training data.
2018	Jane and Sasidhar [53]	Artificial neural network, wavelet denoising techniques, and TKEO (TeagerKaiser energy operator).	Easy to implement.	Two-handed and user-independence signs recognition was not addressed.	Average accuracy: 97.12%.	Myo Armband (gyroscope, accelerometer, magnetometer, and sEMG (surface electromyography)) sensors.	48 lexicon word.	Signing Exact English (SEE-II).	4927 samples, 204 features.
2018	Li et al. [56]	K-Nearest Neighbour, Linear Discriminant Analysis, and Support Vector Machines.	Less interference, stretchable and wearable comfort.	The model was not robust in nature. Hysteresis characteristics and noise present in the sensor lead to misclassification.	LDA accuracy: 97.81%, KNN accuracy: 97.86%, SVM accuracy: 97.89%, and ASL recognition for 0-9 average accuracy: 98%.	Firmly stretchable strain sensor (Custom Glove).	Six subjects. 0-9 ASL sign.	American sign language.	10 trials of each gesture by 6 subjects, 540 data samples.

Year	Author	Method	Pros	Cons	Accuracy or Result	Sensor	Sample and Lexicon Size	Model Type	Dataset
2018	Yin et al. [52]	Template Matching, BP neural network, and Combined Model.	High recognition rate data glove.	Dynamic gesture recognition based research was not performed, and data glove barrier is there, and only a single background effect considered for the experiment fail to generalize in other background.	1. Template Matching accuracy: 96.7%, 2. Feedforward Neural Network accuracy: 98.4%, combined model accuracy: 99.8%.	Custom Glove (Bend Sensors (x5), FLEX2.2).	6 numbers, 3 letters, 5 subjects.	Hand gesture sign.	1000 different data for each gesture from 5 different signers, a total of 9000 data.
2018	Lee and Lee [55]	Support Vector Machines.	Custom-made devices not required because 3D-printed based device used it can fit different sign irrespective of hand and finger sizes holders.	The crucial factor like background light and other effects fail to be addressed. Sign respect to words and sentences not considered for experiments, and implementation can be on a smaller-sized printed circuit board if the research gap is not addressed.	Without pressure sensor accuracy: 65.7%, with a fusion of pressure sensors accuracy: 98.2%.	Custom Glove (9-Axis IMU, Flex Sens.(x5), Pressure Sens. (x2)), five flex sensors, two pressure sensors, and a three-axis inertial motion sensor.	28 gesture patterns (26 ASL letters, 2 signs), 12 subject.	American sign language.	6,480,000 datasets (12 subjects× 20 times × 10 s × 100 Hz × 27 signs).
2017	Yang et al. [57]	Multi-Stream Hidden Markov Models, and Decision Tree classifier.	Time consumption is reduced, and recognition accuracy is improved by the decision tree based classifier.	The proposed model result still lagging behind the literature state of the methods.	User dependent model accuracy: 94.31%, user-independent model accuracy: 87.02%.	4-Channel sEMG, 3-Axis (gyroscope, accelerometer) sensor.	150 signs (one-handed sub-words sign: 81, two-handed sub-words sign: 69, hand orientations: 3, hand amplitude levels: 3), 8 subjects.	Chinese Sign Language (CSL).	30000 sub word data sample (total 3750 sub-word samples per subject).
2016	Wei et al. [10]	Code matching method, and fuzzy K-mean algorithm.	User's training burden minimized.	Target set size is limited; instead of code matching method, advanced fusion method can improve the recognition accuracy.	The recognition accuracy of two reference subject for one-third gestures of the target set: (82.6 ± 13.2)%, and (79.7 ± 13.4)% and half of the target set: (88 ± 13.7)% and (86.3 ± 13.7)%.	sEMG (surface electromyographic), GYRO (gyroscopes), and ACC (accelerometers).	110 sign words, 5 subjects.	Chinese Sign Language (CSL).	13750 (2750 sign word samples for each subject).
2016	Chansri and Srinonchat [28]	Histograms of oriented gradients, and artificial neural network (BP).	Simple model.	The author can not perform statistical analysis.	Accuracy: 84.05%.	Microsoft Kinect (color and depth).	42 letters, 24 hand gestures.	Thai Sign Language.	420 hand gesture samples.
2014	Almeida et al. [54]	Support Vector Machines (SVM).	Simplified model with generic nature.	Fail to address the aspect of feature selection and recognition rate based on uncertainty presented in the suggested model.	Average accuracy: 80%.	Kinect sensor, nuiCapture Analyze software (RGB-D sensor).	34 specific sign, 1 subject.	Brazilian Sign Language.	170 video samples.
2012	Ong et al. [51]	Sequential Pattern Trees Boosting algorithm.	Run-time complexity is less.	Stability issue.	Accuracy: 55% for the first ranked sign and 87% for within the top 10 signs.	Kinect TM camera.	1. 40 signs, 14 subjects, 2. 982 signs, 1 subject.	DGS (German Sign Language (Deutsche Gebärdensprache), Greek Sign Language (GSL).	2800 samples for DGS Kinect 40 dataset and 4910 samples for GSL 982 Signs dataset.
2009	Hrúz et al. [59]	Hidden Markov Model, and kiosk.	Less difficult concern to usability.	Dataset is tiny, and performance improvement needs more data.	8 states HMM Model recognition rate: 81.63%.	Camera.	50 sign, 2 subject.	CzSL (Czech sign language).	338 samples.

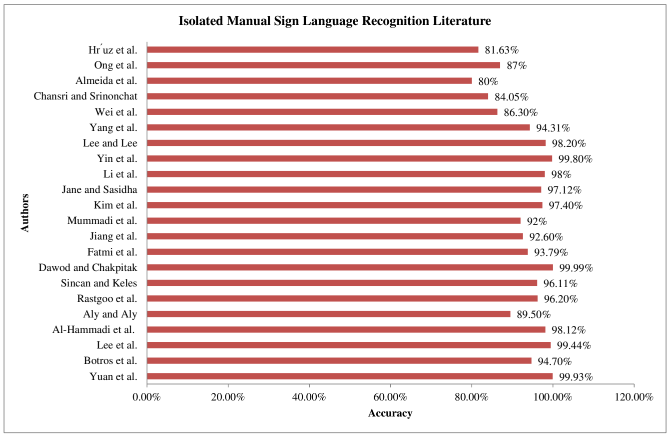

Fig. 6: Isolated Manual SLR Literature. This bar chart representation depicts the mean accuracy and recognition rate achieved in the literature related to the isolated manual SLR concerning different sign languages, models, and modalities.

It makes SLR performance way behind performance of speech recognition. The existing research work on continuous manual SLR are as follows:

Traditional methods: Nayak et al. [64] pointed out the feature extraction approach for continuous sign. Relational distribution are captured from the face and hand present in the images. The parameters are optimized by ICM, so convergence speeds up; they used dynamic time warping for distance computation between two sub-strings. The continuous sign sentence extracts the recurrent features using RD, DWT, and ICM based approaches. Kong and Ranganath [65] performed continuous SLR by merging of CRF (conditional random field) and SVM in a framework of Bayesian network. They performed a semi-Markov CRF decoding scheme-based merge approach for independent continuous SLR. Tripathi and Nandi [4] carried out a gesture recognition model for continuous Indian Sign Language. They extracted meaningful gesture frames using the Key-frame extraction method. The orientation histogram technique extracted each gesture-relevant feature and used the Principal Component Analysis to reduce the feature dimension. They used the distance classifier for classification. According to performance analysis with other considered classifiers, the Correlation and Euclidean distance-based classifier perform with a better recognition rate. Gurbuz et al. [37] developed an ASL model for the RF sensing-based feature fusion approach. They use LDA, SVM, KNN, and RF as classifiers. The random forest classifier-based model for five signs results in 95% recognition accuracy, while 20 signs result in 72 %. They can use the deep learning classifier in the future to improve recognition accuracy. Hassan et al. [5] proposed Modified k-Nearest Neighbor and Hidden Markov Models based on Continuous Arabic SLR. Window-based statistical features and 2D DCT transformation extract the features. The proposed model performance analyzed with sensors, vision-based datasets, and motion tracker dataset leads to a better recognition rate. For sentence recognition (MKNN) Modified k-Nearest Neighbor yields the best recognition rate than the HMM-based Toolkit. For word recognition, RASR performs

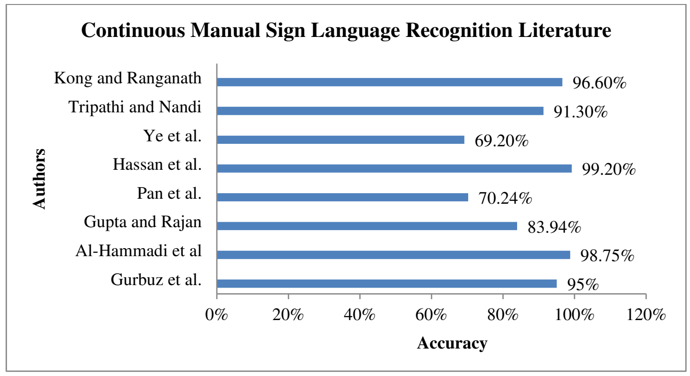

Fig. 7: Continuous Manual SLR Literature. This bar chart presents the mean accuracy and recognition rate achieved in the literature on continuous manual SLR concerning different sign languages, models, and modalities.

better with a higher recognition rate than MKNN GT2K.

CNN, LSTM and Cross model based related work on continuous manual SLR: Ye et al. [66] pointed out a 3D convolutional neural network (3DCNN) with a fully connected recurrent neural network (FC-RNN) to localize the continuous video temporal boundaries and recognize sign actions using an SVM classifier. Designed Convolutional 3D and recurrent neural network-based integrated SLR model for continuous ASL sign recognition. Al-Hammadi et al. [23] presented a single modality-based feature fusion adopted 3DCNN model for dynamic hand gestures recognition. They captured the hand feature using an open pose framework. MLP and auto encoder-based feature extracted 3DCNN model with open pose framework based on hand sign capturing model result in good recognition accuracy for KSU-SSL (King Saud University Saudi Sign Language) dataset using a batch size of 16. Gupta and Rajan [67] examined the performance of three models, namely modified time-LeNet, t-LeNet (time-LeNet), and MC-DCNN based on Indian SLR. continuous Indian SLR models based on MCDCNN, t-Lenet, and modified t- Lenet classifier using sensor-based dataset presents and performance-based investigation carried out. Pan et al. [68] spatial and temporal fused Attention incurred Bi directional long term memory network-based SLR model developed. They detected captured video key action by optimKCC. Multi-Plane Vector Relation (MPVR) is used to get skeletal features. They performed two dataset-based analyses to prove the validity of continuous Chinese SLR concerns sign independent and dependent cases. Papastratis et al. [69] suggested a cross-modal learning-based continuous SLR model, and they have proved validity with three public datasets, namely RWTH-Phoenix-Weather-2014, RWTH- Phoenix-Weather-2014T, and CSL. They achieved the performance improvement of the suggested model by considering additional modalities.
Table V and Figure 7 provide a better understanding of literature work regarding continuous manual SLR.

B. Non-Manual SLR

Facial expressions, head movement, mouth movement, eye movement, eyebrow movement, and body posture are the non-

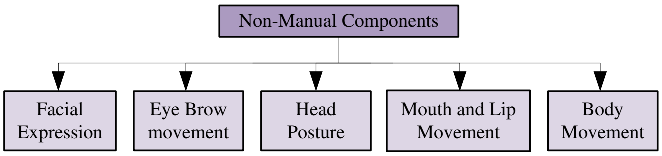

Fig. 8: Non-Manual Components. We showed the important non-manual features related to sign language. These non-manual components are helpful to make sensible recognition of similar types of signs.

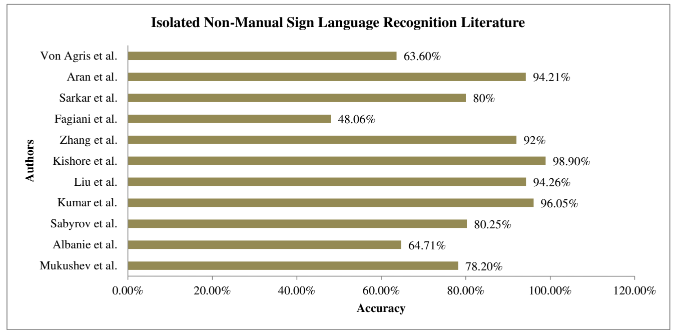

Fig. 9: Isolated Non-Manual SLR Literature. This bar chart represents the mean accuracy and recognition rate achieved in the literature concerning the isolated non-manual SLR regarding the different sign languages, models, and modalities.

manual sign parameters. Non-manual sign components showed in Figure 8. A facial expression considering the lowering and raising of eyebrows expresses grammatical information and emotions. Signers are good listeners and follow eye contact. Similar hand pose signs can be recognized by considering non-manual features. The isolated and continuous are the two types of non-manual SLR models.

1) Isolated Non-Manual SLR: The study of related research works in isolated non-manual-based SLR as follows:

HMM based work: Von Agris et al. [70] designed Hidden Markov Model-based British SLR with manual and non-manual features. Aran et al. [7] performed a Turkish SLR model using a cluster-based Hidden Markov Model. They proved the validity by cross-validation with eight folds (sign independent) and five folds (sign dependent). Sarkar et al. [71] presented an isolated American SLR model using Hidden Markov Model. They improved the segmentation process by a dynamic programming-based approach. Fagiani et al. [72] carried out the Hidden Markov Model-based isolated Italian sign recognition in concern to signer independence. The suggested model gets better accuracy than the support vector machine-based recognition model. Zhang et al. [73] suggested adaptive hidden states incurred Hidden Markov Model for Chinese SLR. The carried-out fusion of trajectories and hand shapes leads to better recognition. Kumar et al. [74] performed an Indian SLR model based on a decision fusion approach with two modalities (facial expression and hand gesture). They used an HMM-based classifier for recognition and used IBCC for decision fusion purposes. They have carried out two modalities (facial expression and hand gesture) associated with IBCC based on HMM classifier decision fusion approach for Indian SLR. Using advanced classifiers and feature extraction algorithms can improve recognition accuracy.

Logistic regression and CNN based work: Sabyrov et al. [75] developed K-RSL(Kazakh-Russian Sign Language) interpreted as a human-robot model using Logistic Regression with incurred non-manual components. Mukushev et al. [76] performed Logistic Regression-based SLR using manual and non-manual features. In the captured video, they got key points using OpenPose. Kishore et al. [77] performed Adaptive Kernels Matching algorithm that incurred 3-D Indian SLR model claims improved classification accuracy compared with state-of-the-art methods. Better classification accuracy achieved by 3D motion capture models than Microsoft Kinect and leap motion sensor-based model.Liu et al. [78] pointed out ST-Net (Spatial-Temporal Net) associated with self-boosted intelligent systems for Hong Kong SLR. Compared to a Kinect-based system, the suggested approach performs well with a better recognition rate. Albanie et al. [6] proposed a Spatio-temporal convolutional neural network-based British SLR model. The pretraining has improved by presenting new larger-scale data, namely BSL-1K.
We perform a comprehensive study of the recent developments concerning non-manual SLR. Table VI and Figure 9 show isolated non-manual SLR-related literature work to make a clear understanding.

2) Continuous Non-Manual SLR: Continuous non-manual SLR is highly complex because the issue related to the context sequence has to be handled appropriately for effective performance or enriched accuracy [79]. The temporal boundaries-related problem makes continuous SLR a complex and arduous task. We discuss the related research work as follows:

Classical methods: Farhadi and Forsyth [80] carried out the HMM-based continuous ASL to English subtitles alignment model. With simple HMMs based on a discriminative word model, they perform word spotting. Infantino et al. [81] developed a common-sense engine integrated self-organizing map (SOM) neural network-based SLR model for LIS (Italian sign language). Sarkar et al. [71] performed a HMM-based continuous ASL. They used a dynamic programming-based approach to improve the segmentation. Forster et al. [21] pointed out the German SLR model using Multi-stream HMMs based on combination methods. Compared to system combination and feature combination approaches, synchronous and asynchronous combination-based models achieved better performance. Yang and Lee [82] presented CRF and SVM associated with a continuous ASL using both manual and non-manual features. BoostMap embeddings verified the hand shape, segmenting done by hierarchical CRF, and recognition was performed using SVM. Zhang et al. [83] suggested a Linear SVM based on an automatic ASL by fusing five modalities. The large-scale dataset-based investigation could be future work to improve recognition accuracy.

TABLE V: Continuous Manual SLR Literature Work. Related work concerning the vision and sensor-based SLR model concerns the continuous manual sign, comprehensively summarized here.

Year	Author	Method	Pros	Cons	Accuracy or Result	Sensor	Sample and Lexicon Size	Model Type	Dataset
2021	Gurbuz et al. [37]	Principal Component Analysis (PCA), short term Fourier transform, and RBF (Radial Basis Function) associated SVM (Support Vector Machine).	Contactless sensing, environment independent	The author can not perform a comparative analysis with the existing method.	Recognition accuracy for 20 sign, 150 features: 72.5% and for 5 sign: 95%.	RF sensors and Kinect sensor.	20 signs, 7 subjects.	American SLR.	240 samples.
2020	Al-Hammadi et al. [23]	3D convolutional neural network (3DCNN), auto-encoders, Multi-layer perceptron, and open pose framework.	The expense concerning the training is minimal.	Model results in good accuracy for a small batch.	MLP based feature extraction model achieved recognition accuracy as 98.62% for sign dependent and 87.69% for sign independent and auto encoder based feature extraction model achieved recognition accuracy as 98.75% for sign independent and 84.89% for sign independent.	RGB video camera.	40 dynamic hand gestures, 40 subjects.	Saudi Sign Language.	8000 data samples.
2020	Gupta and Rajan [67]	Modified time-LeNet, MC-DCN (Multi channel deep convolution neural network), t-LeNet (time-LeNet), and SGD (stochastic gradient descent).	Sensor-based on continuous sign recognition model.	Trainable parameters are a large, over-fitting problem.	MCDCNN accuracy: 83.94%, modified t-LeNet accuracy: 81.62%, and t-LeNet accuracy: 79.70%.	Wireless IMUs (inertial measurement units).	11 sentences with 15 words, 10 subjects.	Indian Sign Language.	1100 samples.
2020	Papastratis et al. [69]	Cross-modal learning approach.	Recognition of accuracy improved by latent representations.	Sensitive to Multi modalities.	1. WER: 24.0% (RWTH-Phoe-nix-Weather-2014), 2. WER: 24.3% (RWTH- Phoe-nix-Weather-2014T), 3. WER: 2.4% (CSL).	Stationary color camera, Kinect 2.0.	For Dataset 1: 9 subjects, vocabulary: 1088,dataset 2: 9 subjects, sign:1295, dataset 3: 50 subject, 100 sentences.	German sign language, Chinese Sign Language	RWTH-Phoenix-Weather-2014, RWTH- Phoenix-Weather-2014T, CSL.
2020	Pan et al. [68]	Attention-Based BLSTM, KCC (key-frame-centered clips), and MPVR (Multi-Plane Vector Relation).	Speed up convergence.	Real-time practical implementation was lacking.	Fused Attention incurred BLSTM with optimum KCC and MVPR based sign recognition model accuracy: 70.24% (CSL dataset) and 60.31% (DEVISIGN dataset) for sign dependent and 68.32% (CSL dataset) and 58% (DEVISIGN dataset) for sign independent.	Kinect-1.0, Kinect-2.0.	DEVISIGN dataset: 500 sign words, 8 subjects, CSL dataset: 10 subjects, 200 sign words.	Chinese Sign Language.	CSL dataset: 20000 samples.
2019	Hassan et al. [5]	Modified k-Nearest Neighbor, and Hidden Markov Models (GT^2 K and RASR).	Require less computation time.	Robustness needs to be improved.	For sentence recognition, MKNN based model: 97.78% (DG5-VHand dataset), and for words recognition RASR based model: 99.20% (DG5-VHand dataset).	Camera, DG5-VHand data glove and Polhemus G4 motion tracker.	80 sign, 40 sentences, 1 subject.	ArSL (Arabic sign language).	800 words samples, 400 sentence samples.
2018	Ye et al. [66]	3DRCNN (3D recurrent convolutional neural networks) with SVM classifier.	RGB, motion, and depth channels fusing lead to better accuracy.	Facial information was not considered. Hence, the performance is poor.	Accuracy: 69.2% for Person-dependent and an accuracy: 65.8% for Person-independent.	Kinect 2.0 sensor.	99 sign, 5 sentences, 15 subjects.	American Sign Language.	Sentence videos (100) and Sequence videos (27) are collected.
2015	Tripathi and Nandi [4]	OH (Orientation Histogram), PCA (Principal Component Analysis), and distance-based classifier.	New ISL continuous dataset presented.	Misclassification for a similar type of gesture.	Average recognition rate for Euclidean distance: 90% (18 bins OH), 91.3% (36 bins OH), and Average recognition rate for correlation: 89% (18 bins OH), 89.8% (36 bins OH).	Canon EOS camera.	10 sentences, 5 subjects	Indian Sign Language.	500 video samples.
2014	Kong and Ranganath [65]	Semi-Markov CRF decoding scheme based probabilistic approach.	Complexity related to decoding reduced.	Larger vocabularies' scalability issue.	For an unseen sign from considered signer, accuracy: 96.6% and recall rate: 95.7%, for unseen signers' accuracy: 89.9% and recall rate: 86.6%.	CyberGlove(x2) and 3D Trackers.	74 sentences (107 ASL signs), 8 subjects.	American sign language.	2393 sentences and 10852 sign instances.
2012	Nayak et al. [64]	Relational Distributions, ICM (iterative conditional modes) algorithm, and DTW (Dynamic Time Warping).	Aid for faster training set generation.	They did not address the amplitude variation of the various signer. Comparative qualitative and quantitative analysis are lacking.	NA	Video camera.	155 American Sign Language (ASL), 12 groups.	American Sign Language.	155 video sequences.

TABLE VI: Isolated Non-Manual SLR Literature Work. Related work with regard to vision and sensor based SLR model concern the isolated non-manual sign comprehensively summarized in the tabular form for better understanding.

Year	Author	Method	Pros	Cons	Accuracy or Result	Sensor	Sample and Lexicon Size	Model Type	Dataset
2020	Mukushev et al. [76]	Logistic Regression.	Non-manual components considered as input lead to better accuracy.	Need improvement concern to accuracy.	Accuracy: 78.2%.	LOGITECH-C920 HD PRO WEBCAM.	20 signs, 5 subjects.	K-RSL (Kazakh-Russian sign language).	5200 samples.
2020	Albanie et al. [6]	Spatio-temporal convolutional neural network.	Large-scale dataset (BSL-1K) presented.	Fails to address visual similarity issue.	Accuracy: 46.82% (WLASL) and 64.71% (MSASL) top-1 case.	Camera.	MS-ASL: 1000 words, 222, WSASL:2000 words, 119 subjects BSL-1K: 1064, 40subjects.	BSL (British Sign Language), ASL (American Sign Language).	WSASL: 21K MS-ASL: 25K BSL-1K: 273K.
2019	Sabyrov et al. [75]	Logistic Regression.	Huge dataset not required a humanoid robot sign interpreter.	Experimentation in real-world is lacking.	For 20 sign dataset accuracy: 73% and for 2 class accuracy: 80.25%.	LOGITECH C920HD PRO WEBCAM.	20 words, 3 subjects.	K-RSL n (Kazakh-Russian Sign Language).	2000 videos.
2018	Kumar et al. [74]	HMM (Hidden Markov Model) and IBCC (Independent Bayesian Classification Combination).	The suggested model performed better than BLSTM-NN.	Possibility to classify wrongly fails to present a comparative analysis based on recent methods for the considered applications.	Double hand gestures based recognition rate: 94.27%, and the single-hand sign based recognition rate: 96.05%.	Leap Motion and Kinect sensor.	51 dynamic sign words (31 sign with two hands, 20 sign with a single hand), 10 subjects.	Indian Sign Language (ISL).	4080 data samples include both single and two hands sign.
2018	Liu et al. [78]	ST-Net (SpatialTemporal Net).	Light-weight and robust.	Require more computation time.	Accuracy: 94.26% for Person dependent and 91.19% for Person independent sign.	Microsoft Kinect.	227 words, (86 words – single hand, 33 words), 2 hands separated, and 108 words- 2 hands intersected, 3 subjects based 1802 mouth images.	HKSL (Hong Kong Sign Language).	5221 samples videos.
2018	Kishore et al. [77]	Adaptive Kernels Matching algorithm.	Overcome the issue of spatio-temporal misalignment, Small action changes discovered effectively.	The author fails to perform real-time, live data-based analysis.	Accuracy: 98.9%	Kinect and leap motion sensors.	Five sets of data with 500 signs, 5 subjects each set.	IndianSL (Indian Sign Language).	18000 signs with 36 variations per sign comprises a testing set.
2016	Zhang et al. [73]	Adaptive Hidden Markov Models, enhanced shape context.	Computational cost reduced.	Fail to assure accuracy for the larger dataset.	Accuracy rate for dataset 1: 92% in the top 1, 99% in the top 5, and 100% in the top 10. For dataset 2 accuracy: 86% in the top 1, 96.8% in the top 5, and 98.8% in the top 10.	Microsoft Kinect.	1. 100 signs 1 subject, 2. 500 signs 1 subjects, 99 signs, 5 subjects.	CSL (Chinese sign language).	1. 500 videos, 2. 2500 videos.
2015	Fagiani et al. [72]	Hidden Markov Models.	Easy to implement in real-time case.	Overlapping issue.	Average accuracy rate: 48.06%.	Digital Video Camera.	147 signs, 10 subjects.	LIS (Italian sign language / Lingua Italiana dei segni).	1,470 videos samples (A3LIS-147).
2011	Sarkar et al. [71]	Dynamic programming methods, and Hidden Markov Models.	Adaptable to uncontrolled domain.	Sensitive to dataset.	Recognition rate: 80%.	Video Camera.	147 signs, 10 subjects.	ASL (American sign language).	294 samples.
2009	Aran et al. [7]	Hidden Markov Models with cluster algorithm.	Classify the similar sign accurately.	The small dataset used for validation.	Average accuracy rate: 79.1% (Signer-independent) and 94.21% (Signer-dependent).	Camera.	19 signs, 8 subjects.	TSL (Turkish sign language).	760 samples for each fold (8 folds for Signer-independent and 5folds for Signer-dependent).
2008	Von Agris et al. [70]	Hidden Markov Models.	User-friendliness.	Recognition accuracy decay with low resolutions.	Average recognition rates: 63.6%.	Video camera	263 signs, 4 subjects.	BSL (British sign language).	8100 video samples.

TABLE VII: Continuous Non-Manual SLR Literature Work. Related work concerning the vision and sensor-based SLR model concern the continuous non-manual sign, comprehensively summarized here.

Year	Author	Method	Pros	Cons	Accuracy or Result	Sensor	Sample and Lexicon Size	Model Type	Dataset
2020	Brock et al. [84]	Convolutional Neural Network.	Complex linguistic content handling ability.	Expensive.	Average Word Error Rate (WER): 15.71% for Non-Manual Expression label.	Single-camera video.	1432 signs, 1 subject.	JSL (Japanese Sign Language).	DJSLC corpus a total of 1432 sequential sign.
2020	Zhou et al. [85]	Spatial-temporal multi-cue (STMC) network, spatial multi-cue (SMC) module and a temporal multi-cue (TMC) module, joint optimization strategy.	Able to handle different cues at the same time.	Training time and complexity is high.	STMC WER: 28.6% (CSLunseen), 20.7% (PHOENIX-2014), 21.0% (PHOENIX-2014-T).	Video camera.	PHOENIX-2014: 9 subjects 1295 sign, PHOENIX-2014-T 1115 for sign gloss CSL: 10 subjects, 500 words.	CSL (Chinese sign language), DGS (German Sign Language).	CSL:5000 videos, PHOENIX-2014:6841 samples, PHOENIX-2014-T:8257 samples.
2020	Koller et al. [86]	Hybrid CNN-LSTM-HMMs.	Speedup convergence.	Sensitive to overlapping and computation is difficult.	For RWTHPHOENIX-Weather 2014 dataset, WER: 26.0%.	Video camera.	1. 1066 signs 9 subjects, 2. 1080 sign 9 subjects.	DGS (German Sign Language (Deutsche Gebärdensprache).	One million hand shape images from 23 subjects, Phoenix14T,Phoenix14.
2016	Koller et al. [87]	Iterative Expectation Maximization incurred CNN.	Training is easy.	Convergence took much time.	Recognition accuracy: 62.8%.	Video camera.	1.1080 sign 9 subjects, 2. 455 sign 1 subjects.	DGS (German Sign Language (Deutsche Gebärdensprache).	One million hand shape images Phoenix14, Signum.
2016	Zhang et al. [83]	Linear Support Vector Machine.	Simple concatenation approach.	Unbalancing of data leads to misclassification.	Recognition rate: 36.07% for all lexical items.	Kinect sensor.	99 signs, 5 subjects.	ASL (American sign language).	61 video sequences (segmented in to set of 673 video clips).
2013	Yang and Lee [82]	Conditional random field support vector machine and BoostMap embedding method.	Issues of label bias not occurred.	The author can not perform experimentation with real-time live data.	Recognition rate: 84.1%.	Cameras.	24 signs, alphabets-17, and facial expressions-5.	ASL (American sign language)	98 ASL signed sentences.
2013	Forster et al. [21]	Multi-stream HMMs based on combination methods	Robust model.	An asynchronous combination, not flexible to handle the large-scale dataset.	For SIGNUM dataset WER: 10.7%, and for PHOENIX dataset WER: 41.9%.	Video Camera.	1. 266 signs 1 subjects, 2. 455 signs 1 subjects, 3. 455 signs, 25 subjects.	DGS (German Sign Language (Deutsche Gebärdensprache).	Signum dataset consists 530850 total number of frame and PHOENIX dataset consists 53033 total number of frame.
2011	Sarkar et al. [71]	Hidden Markov Model.	Adaptable to uncontrolled domain.	Sensitive to dataset.	Correct detection: 92%.	Video Camera.	65 signs subjects.	ASL (American sign language).	30 sentences.
2007	Infantino et al. [81]	Self-organizing map (SOM) neural network.	Simple and robust.	Selection of input weights is difficult.	Correctly recognized sentence accuracy: 82.3% for Exp 1(30 videos of sentences with 20 signs) and 82.5% for Exp 2(80 videos of sentences with 40 signs).	Video camera.	40 signs.	LIS (Italian sign language / Lingua Italiana dei segni).	160 video samples.
2006	Farhadi and Forsyth [80]	Hidden Markov Model.	The simple discriminative word model.	The author cannot perform quantitative and qualitative results analysis.	NA	Video camera.	31 words signs.	ASL (American sign language).	80000 frames of film.

TABLE VIII: Sensing approach based SLR. Although the sensor-based approach provides better accuracy than the vision-based approach, the sensor-based approach is not an optimal choice in real-time applications.

Sensing Approach	**Devices**	**Year**	**Authors**
Sensor	3-dimensional flex sensor-based data gloves, gyroscope, accelerometer, and bending sensor	2021	Yuan et al. [24]
	EMG sensor	2021	Botros et al. [8]
	Surface electro-myogram and inertial measurement units	2020	Gupta and Kumar [61]
	4-Channels sEMG-Electromyography, one IMU inertial measurement unit	2018	Jiang et al. [48]
	3D magnetometer, 3D gyroscope (GYRO), and 3D accelerometer (ACC)	2018	Mummadi et al. [60]
	Armband module (8-channel electromyography) sensors.	2018	Kim et al. [11]
	Myo Armband (gyroscope, accelerometer, magnetometer, and sEMG (surface electromyography)) sensors	2018	Jane and Sasidhar [53]
	Firmly stretchable strain sensor (Custom Glove)	2018	Li et al. [56]
	Wireless IMUs (inertial measurement units)	2020	Gupta and Rajan [67]
	CyberGlove (x2) and 3D Trackers	2014	Kong and Ranganath [65]
Vision	Microsoft Kinect v2	2020	Sincan and Keles [27]
	Video camera	2020	Aly and Aly [2]
	Video camera	2020	Rastgoo et al. [3]
	RGB video camera	2020	Al-Hammadi et al. [23]
	Laptop camera	2020	Hoang [62]
	Single camera video	2020	Brock et al. [84]
	LOGITECH C920 HD PRO WEBCAM	2019	Sabyrov et al. [75]
	Kinect and leap motion sensors	2018	Kishore et al. [77]
	Leap Motion and Kinect sensor	2018	Kumar et al. [74]
	KinectTMcamera	2012	Ong et al. [51]
	Video camera	2007	Infantino et al. [81]

TABLE IX: Various Datasets in SLR. (P-Publicly available, PP-Publicly available with a password, CA-Contact Author, OR-On Request, SA-Send request along with release agreement)

Language Level	Language	Dataset	Data Type	Subjects	Classes	Samples	Country	Data Size	Link	Data Avail-ability
Isolated	Indian	IIITA -ROBITA	Videos	-	23	605	India	284 MB	https://robita.iiita.ac.in/dataset.php	SA
		INCLUDE [88]	Videos	7	266	4287	India	56.8 GB	https://zenodo.org/record/4010759#.YdqY9lRfg5k	P
	American	Boston ASL LVD	Videos, multiple angles	6	3300+	9800	United States	1-2 GB	http://www.bu.edu/asllrp/av/dai-asllvd.html	P
		ASLLVD	Videos (multiple angles)	6	3,300	9,800	United States	1-2 GB of each	http://www.bu.edu/asllrp/av/dai-asllvd.html	P
		Purdue RVL-SLLL	Videos	14	39	546	United States	-	https://engineering.purdue.edu/RVL/Database/ASL/asl-database-front.htm	OR
		ASLLVD-Skeleton	Skeleton	-	3,300	9,800	United States	1-2 GB of each	https://www.cin.ufpe.br/~cca5/asllvd-skeleton/index.html	P
		RWTH-BOSTON-50	Videos (multiple angles)	3	50	483	United States	295 MB	https://www-i6.informatik.rwth-aachen.de/aslr/database-rwth-boston-50.php	P
		WLASL	Videos	100	2,000	21,083	United States	64 GB	https://dxli94.github.io/WLASL/	P
		MS-ASL [89]	Videos	222	1,000	25,513	United States	1.9 MB	https://www.microsoft.com/en-us/download/details.aspx?id=100121	P
	Argen -tinian	LSA64	Videos	10	64	3,200	Argen -tina	1.9 GB	http://facundoq.github.io/datasets/lsa64/	P
	Chinese	Isolated SLR500	Videos & Depth from Kinect	50	500	125,000	China	-	http://home.ustc.edu.cn/~pjh/openresources/cslr-dataset-2015/index.html	PP
		NMFs-CSL	RGB videos	10	1,067	32,010	China	-	http://home.ustc.edu.cn/~alexhu/Sources/index.html	SA
	German	DGS Kinect 40	Videos (multiple angles)	15	40	3,000	Germany	39.7 MB	https://www.cvssp.org/data/KinectSign/webpages/downloads.html	CA
	Greek	GSL isol.	Videos & Depth from Real Sense	7	310	40,785	Greece	155 KB	https://vcl.iti.gr/dataset/gsl/	P
	Polish	PSL Kinect	Videos & Depth from Kinect	1	30	300	Poland	1.2 GB	http://vision.kia.prz.edu.pl/dynamickinect.php	P
		PSL ToF	Videos & Depth from ToF camera	1	84	1,680	Poland	33 GB	http://vision.kia.prz.edu.pl/dynamictof.php	P
	Spanish	LSE-Sign	Videos	-	2,400	2,400	Spain	-	http://lse-sign.bcbl.eu/web-busqueda/	CA
	Turkish	BUHMAP-DB	Videos	11	8	440	Turkey	505 MB	https://www.cmpe.boun.edu.tr/pilab/pilabfiles/databases/buhmap/	P

Language Level	Language	Dataset	Data Type	Subjects	Classes	Samples	Country	Data Size	Link	Data Availability
		Bosphoru Sign22k [90]	Videos	6	744	22,542	Turkey	-	https://www.cmpe.boun.edu.tr/pilab/BosphorusSign/home_en.html	P
		AUTSL	Videos	43	226	38,336	Turkey	-	https://chalearnlap.cvc.uab.cat/dataset/40/description/	PP
Contin-uous	Indian	ISL-CSLTR	Videos	7	100	700	India	8.29 GB	https://data.mendeley.com/datasets/kcmpdxky7p/1	P
		BVCSL3D	Videos (RGB, depth and skeletal)	10	200	20,000	India	256 B	https://ars.els-cdn.com/content/image/1-s2.0-S1045926X18301927-mmc1.xml	OR
	American	RWTH-BOSTON-104	Videos (multiple angles)	3	104	201	United States	685 MB	https://www-i6.informatik.rwth-aachen.de/aslr/database-rwth-boston-104.php	P
		RWTH-BOSTON-400	Videos	5	400	843	United States	-	http://www-i6.informatik.rwth-aachen.de/~dreuw/database.php	OR
	Chinese	Continuous SLR100	Videos & Depth from Kinect	50	100	25,000	China	-	http://home.ustc.edu.cn/~pjh/openresources/cslr-dataset-2015/index.html	PP
		DEVISI-GN-D	Videos	8	500	6,000	China	-	http://vipl.ict.ac.cn/homepage/ksl/data.html#page2	SA
		DEVIS-IGN-G	Videos	8	36	432	China	-	http://vipl.ict.ac.cn/homepage/ksl/data.html#page2	SA
		DEVIS-IGN-L	Videos	8	2000	24,000	China	-	http://vipl.ict.ac.cn/homepage/ksl/data.html#page2	SA
	German	RWTH-PHOENIX-Weather 2014	Videos	9	1,081	6,841	Germany	52 GB	https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX/	P
		SIGNUM	Videos	25	450	33,210	Germany	920 GB	https://www.phonetik.uni-muenchen.de/forschung/Bas/SIGNUM/	P
		RWTH-PHOENIX-Weather 2014 T	Videos	-	1,066	8,257	Germany	39 GB	https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX-2014-T/	P
	Greek	GSL SD	Videos & Depth from Real Sense	7	310	10,290	Greece	-	https://vcl.iti.gr/dataset/gsl/	CA
		GSL SI	Videos & Depth from Real Sense	7	310	10,290	Greece	-	https://vcl.iti.gr/dataset/gsl/	CA
	Korea	KETI [91]	Videos	14	524 (419 words and 105 sentences)	14,672	South Korea	-	https://arxiv.org/pdf/1811.11436.pdf	CA

TABLE X: SLR Datasets Vs. Modality. Highlight the related work in the literature regarding the various datasets and modalities. Dynamics and multi-modality like RGB, depth, and skeleton lead to better recognition rate in SLR.

Year	Authors	Datasets	Modality
2020	Elboushaki et al. [92]	isoGD, SBU, NATOPS, SKIG	RGB, Depth, Dynamic
2020	Rastgoo et al. [3]	RKS-PERSIANSIGN, NYU	RGB, Dynamic
2019	Köpüklü et al. [93]	EgoGesture, NVIDIA benchmarks	
2019	Lim et al. [94]	RWTH-BOSTON-50, ASLLVD	
2019	Chen et al. [95]	DHG-14/28 Dataset, SHREC'17 Track Dataset	
2019	Ferreira et al. [96]	Real video samples	RGB, Depth, Static
2019	Gomez-Donoso et al. [97]	STB	
2018	Spurr et al. [98]	NYU, STB, MSRA, ICVL	
2018	Kazakos et al. [99]	NYU	
2019	Li et al. [100]	B2RGB-SH, STB	RGB, Static
2018	Mueller et al. [101]	EgoDexter, Dexter, STB	
2017	Victor [102]	Egohands	
2018	Baek et al. [103]	BigHand2.2M, MSRA, ICVL, NYU	Depth, Static
2018	Moon et al. [104]	MSRA, ICVL, NYU	
2018	Ge et al. [105]	MSRA, ICVL, NYU	
2017	Ge and et. al [106]	MSRA, NYU	
2017	Dibra et al. [107]	ICVL, NYU	
2016	Sinha et al. [108]	NYU	
2017	Zimmermann and Brox [109]	Dexter, STB	3D, RGB
2018	Marin-Jimenez et al. [110]	UBC3V, ITOP	3D, Depth
2017	Deng et al. [111]	NYU	
2016	Oberweger et al. [112]	MSRA	
2015	Oberweger et al. [113]	NYU	
2018	Rastgoo et al. [114]	Massey 2012, ASL Fingerspelling A, SL Surrey	2D, Depth, RGB
2016	Duan et al. [115]	RGBD-HuDaAct, isoGD	
2020	Chen et al. [116]	NYU, ICVL, MSRA	2D, Depth
2019	Dadashzadeh et al. [117]	OUHANDS	
2018	Wang et al. [118]	Human3.6M	
2017	Yuan et al. [119]	BigHand2.2M, MSRA, ICVL, NYU	
2017	Guo et al. [120]	ITOP, MSRA, ICVL, NYU	
2017	Fang and Lei [121]	ICVL, NYU	
2017	Madadi et al. [122]	MSRA, NYU	
2016	Wang et al. [123]	isoGD	
2016	Haque et al. [124]	EVAL, ITOP	
2015	Tagliasacchi et al. [125]	Real video samples	
2020	Rastgoo et al. [126]	isoGD	2D, RGB
2016	Wei et al. [127]	MPII, FLIC, LSP	
2016	Newell et al. [128]	MPII, FLIC	
2015	Koller et al. [129]	RWTH-PHOENIX-Weather	
2014	Toshev and Szegedy [130]	LSP, FLIC	

TABLE XI: Study of current state-of-the-art SLR model results with various datasets.

State-of-the-art SLR Model	Author	Year	Results	Datasets
CoT4 CNN	Ravi et al. [131]	2019	Recognition rate -89.69%	BVCSL3D dataset
3 D CNN with score level fusion	Gökçe et al. [132]	2020	Accuracy -94.94 %	Bosphorus Sign22K
STMC	Zhou et al. [85]	2020	WER- 2.1 for Continuous SLR 100 dataset Split I case and WER - 28.6 for Split II case WER-20.7 for RWTH-PHOENIX-Weather 2014 WER - 21.0 for RWTH-PHOENIX-Weather 2014 T	1. Continuous SLR 100 2. RWTH-PHOENIX-Weather 2014 3. RWTH-PHOENIX- Weather 2014-T
TK-3d convNet	Li et al. [133]	2020	Recognition accuracy - 77.55 % for WLASL 100 and 68.75 % for WLASL 200 Recognition accuracy - 83.91 % for MSASL 100 and 81.14 % for MSASL 200	1. WLASL 100 2. WLASL 300 3. MSASL 100 4. MSASL 200
SLRT	Camgoz et al. [134]	2020	BLEU 4 scores -21.80	RWTH-PHOENIX- Weather 2014-T
TSPNet –Joint	Li et al. [135]	2020	BLEU 4 -13.41	RWTH-PHOENIX- Weather 2014-T
Multi-stream Conv Architecture	Zheng et al. [136]	2021	BLEU 4 score – 10.89 (RoI) and 10.73 (stream)	RWTH-PHOENIX- Weather 2014-T
DF- WiSLR (SVM augmented)	Ahmed et al. [137]	2021	Dynamic sign accuracy - 98.5 % and Static sign accuracy - 99.9 %	Wi-Fi CSI dataset (49 gesture (static and dynamic)
SAN	Slimane and Bouguessa [138]	2021	WER – 29.78 %	RWTH-PHOENIX-Weather 2014
Inflated deep CNN	Töngi [139]	2021	Accuracy -0.75	SIGNUM
GLEN	Hu et al. [140]	2021	Accuracy -69.9 % for NMFs-CSL and accuracy - 96.8 % for Isolated SLR 500 dataset	1. NMFs-CSL 2. Isolated SLR 500
VTN-PF	De Coster et al. [141]	2021	Accuracy 92.92%	AUTSL
SAM SLR	Jiang et al. [142]	2021	Accuracy -98.42 % for RGB and 98.53% for RGB-D	AUTSL
SLRGAN	Papastratis et al. [143]	2021	Deaf-to-Deaf SLRGAN WER - 36.05 for GSL SD and WER -2.26 for GSL SI SLRGAN WER - 2.98 for GSL SI SLRGAN WER - 37.11 for GSL SD WER - 23.4 % for RWTH-PHOENIX-Weather 2014-T WER - 2.1 % Continuous SLR 100	1. GSL SD 2. GSL SI 3. RWTH-PHOENIX- Weather 2014-T 4. Continuous SLR 100
VMC	Min et al. [144]	2021	WER - 1.6 % for Continuous SLR 100 WER -22.3 % for RWTH-PHOENIX-Weather 2014	1. Continuous SLR 100 2. RWTH-PHOENIX-Weather 2014
SMA-SLR- v2	Jiang et al. [145]	2021	Accuracy -98.53 % for AUSTL Accuracy - 59.39 % WLASL2000 dataset per instance case and 56.63% per class Accuracy - 99 % for Isolated SLR 500	1. AUSTL 2. WLASL 200 3. Isolated SLR 500
SLR-Net- J+B	Meng and Li [146]	2021	Accuracy -98.08 % for Isolated SLR -500 and 64.57 % for DEVISIGN-L	1. Isolated SLR -500 2. DEVISIGN-L
SVM with RBFK (sEMG and acc)	Pereira-Montiel et al. [147]	2022	Accuracy -96.66%	Colombian sign language (3 subjects, 360 signs, 12 words)
SPOTER	Boháček and Hrúz [148]	2022	Accuracy – 100% for LSA64 Accuracy -63.18% for WLASL 100 and 43.78 % for WLASL 300	1. LSA64 2. WLASL

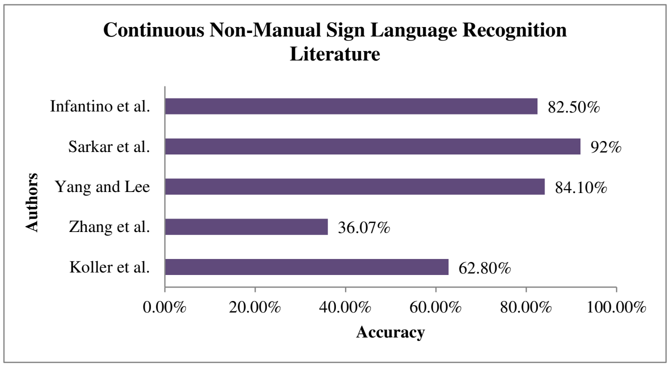

Fig. 10: Continuous Non-Manual SLR Literature. This bar chart representation shows that the literature's mean accuracy and recognition rate concerning the continuous non-manual SLR concerning different sign languages, models, and modalities.

CNN and Hybrid methods: Koller et al. [87] designed a continuous German SLR model using Iterative Expectation Maximization, incurred CNN. They trained the classifier with over a million hand shape sign data. Brock et al. [84] performed Continuous Japanese SLR using CNN. They used frame-wise binary Random Forest for segmentation. The improvement of reliability, accuracy, and robustness for large-scale datasets could be a future research direction. Zhou et al. [85] carried out a continuous SLR model using STMC (Spatial-Temporal Multi-Cue Network) to overcome the vision-based sequence learning problem. Koller et al. [86] proposed a Hybrid CNN-LSTM-HMMs Continuous German SLR model. They performed sign language learning by sequential parallelism and validated it with three public sign language datasets.

The continuous non-manual SLR-related research works are presented in tabular form in Table VII and in graphical chart in Figure 10 for better understating. The research on a continuous sign with a signer independent generic model is important because it has carried very little research on continuous SLR in the past decade.

IV. Classification Architectures

The classification is the brain of the SLR model. It aims to classify the sign accurately with minimum error. Researchers used various classifiers, e.g., traditional machine learning-based approach, deep learning-based approach, and hybrid approach.

ANN like back propagation, multi-layer, and recurrent neural networks are employed as classifiers, but handling large data is difficult. It requires enormous data for training to learn to challenge problems using a machine learning-based approach. The complication associated with HMM: 1. Likelihood of observation, 2. Best hidden state sequence decoding, 3. HMM, parameter framing. The parameter need for the 2 DCNN is excessively more, which makes the design process complex; this is the major drawback of 2 DCNN. In 3DCNN, the Spatio-temporal data has directly represented

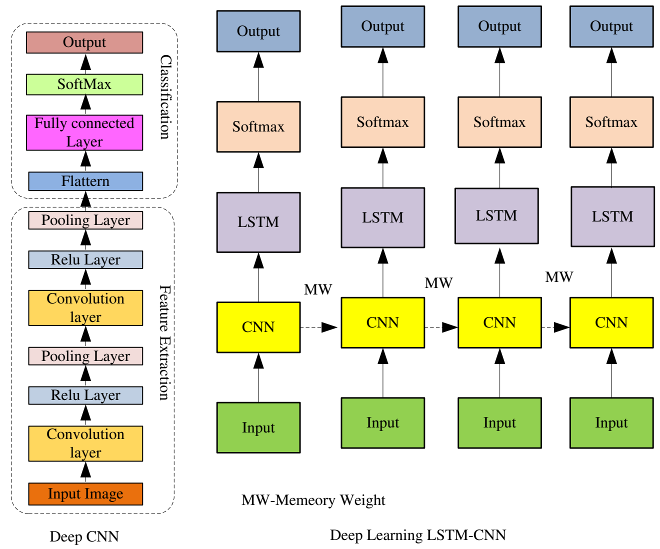

Fig. 11: Deep Learning Architectures (a) Deep CNN, (b) Deep Learning LSTM-CNN Architecture. The Deep CNN (3D) model and hybrid (LSTM+CNN) are recent classifier models. Hybrid deep learning-based classifier enhances recognition rate in SLR.

hierarchically, which is one of the unique features of 3 DCNN. Concerns to the long-term temporal dependence sign capturing 3 DCNN cannot assure robustness. LSTM eliminates the long-term dependence problem. The hybrid-based approach is adopted as a classifier to improve the accuracy.

A. Traditional Architectures

The Artificial Neural Network (ANN), Hidden Markov Model (HMM), and Recurrent Neural Network (RNN) are the most widely used classifiers in SLR due to their sequential data processing ability. Fatmi et al. [9], Lee et al. [22], Von Agris et al. [70] carried the general ANN, RNN, and HMM-based SLR work.

B. Deep Learning Architectures

Deep learning makes massive growth in SLR recently. The spatial and temporal features are easy to handle by the deep learning models. LSTM can handle long-term dependence. Figure 11 highlights the deep learning-based SLR architectures, namely Deep CNN and LSTM-CNN Architecture.

C. Evaluation Metrics

Computation of word error rate, accuracy, and recognition rate evaluates SLR models' performance. The formulations used for evaluation are as follows:

$$Accuracy = (TruePositive + TrueNegative)/(Total) \tag{1}$$

$$\begin{aligned} WER(WordErrorRate) = (&Numberofsubstitutions+ \\ &Numberofdeletions+ \\ &Numberofinsertions)/ \\ &(Totalnumberofwordsinreference) \end{aligned} \tag{2}$$

$$RecognitionRate = (Numberofcorrectlyidentified\ images/TotalNumberofimages) * 100 \quad (3)$$

The cross-validation scheme, namely leave-one-subject-out (LOSO) and k-fold cross-validation, is used to validate the SLR model's effectiveness. The Area Under the Curve (AUC) and ROC (Receiver Operating Characteristic) curve show the trade-off between true positive rate and false positive rate: it is used to measure the classifier performance. The Bilingual Evaluation Understudy (BLEU) score is used to measure the effectiveness of the translation.

V. Different Types of Sensing Approach

According to the acquisition, it classifies SLR into two types, vision and sensor-based approaches. Many research works conducted both vision-based and sensor-based SLR to help the hand-talk community. Table VIII depicts sensing approach-based SLR existing work in literature.

Vision-based sensing devices: Types of cameras used for the vision-based approach are as follows:
Invasive device (body marker method): Examples: LED light, writ band, colored gloves.
Active devices: Kinect sensor, Leap motion sensor.
Stereo camera (depth camera): Capture depth information.
Single-camera: Smartphone, video camera, webcam, thermal camera, etc.
LMC (Leap Motion Controller): The LMC comprises three infrared LEDs and two cameras. It possesses the ability to track 850 nanometers' wavelength of light. The range is 60 cm (2 feet). Hand movement detection converts them into a suitable form of computer (commands) with the leap motion controller. Images are in a gray scalar format, and it acquired raw images using leap motion service software. Demerit: Accuracy is minimal.
Kinect Sensor: The skeleton (depth) image and movement creation done from three-dimensional image data. Multi-array microphone depth sensor and RGB camera are the comprised components in the Kinect sensor. Demerits: it requires more space (6 to 10 feet) distance between the sensor and signer.

Sensor-based sensing devices: The inexpensive and wearable sensor devices such as ACC (Accelerometer), Gyro (gyroscope), and sEMG (surface electromyogram) make sensor-based SLR a prominent tool for SLR.
Data Gloves (sensor-based): Analog form of signal converted into digital format by an ADC converter. It detects hand gestures and signs with the help of various sensors. It comprises an accelerometer and flex sensor (bend signal detection). A gyroscope gets orientation and angular and gains acceleration information with the help of an accelerometer. Finger bending information got by flex sensor. IMU (Inertial measurement units) used for hand movement estimation.
EMG (Electromyography): attach or insert the electrodes into the human muscle. With the help of the inserted electrodes, it recorded muscle movement as an electrical form. sEMG (surface electromyogram) is used for finger movement capturing and distinguishing.

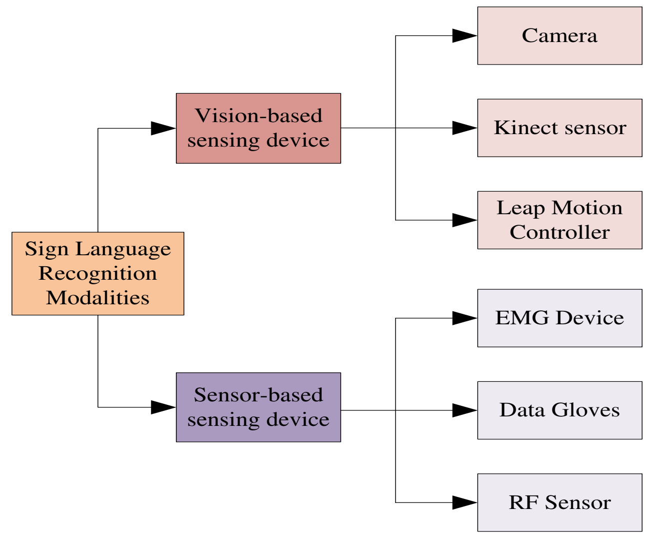

Fig. 12: SLR Modalities. It widely used Kinect and Leap Motion Controller sensor for vision-based SLR. EMG devices produce better results for sensor-based SLR, and it is an ongoing research area.

The RF sensor possesses salient features that make it likable to acquire the sign in the dark environment, contact-less.
Radar and Wi-Fi: The motion of hand movements is collected from the multiple antennas. It extracts features based on Doppler shift and the difference between the magnitude. Advantage orientation and position are flexible.
It has captured the interaction between the environment background and the signer performing the sign using an ambient sensor. Example: 1. Temperature Sensor, 2. Radar Sensor, 3. Pressure Sensor, 4. Sound Sensor. In Figure 12, some SLR modalities are shown.

A. *Various Datasets in SLR*

There is a limitation on the datasets because most of the public datasets lack quality and quantity. The datasets are collected from native signers and ordinary people. Imitation of data was acquired to augment the dataset. We summarize the datasets available to SLR in Table IX. The benchmark dataset details, including the URL link, are detailed here.

1) Datasets Vs. Modality: Datasets and modalities affect largely the performance of SLR models. Many researchers have implemented various SLR models using various methods and datasets. Table X shows Datasets Vs. Modality-based study concerns SLR. The RGB, depth, dynamic-based modality facilitates better performance. Hence, modality-based fusion leads to enhanced performance.

2) The Complexities of SLR Datasets: The acquisition of sign language is performed using the camera and sensor-based sensing devices like Kinect and leap motion controllers, armbands, gloves, electrodes (EMGs), etc. All datasets comprise its specific capturing format, modalities, mapping, environment, and illumination specific to the region or country sign language among 300 sign languages. Some of the complexities in these datasets are as follows:

- Acquire the SLR with various linguistic components based on multiple sensing devices is an arduous task, and the tedious process requires much time and effort.
- Redundant and blur frames in data collection are a significant complexity that affects the SLR model recognition rate.
- Complex background and various lighting are not considered in most datasets and have constraints; therefore, they cannot accurately recognize in a real-time application.
- Most of the datasets were collected with a few native signers, only with a few repetitions. Therefore, it may not guarantee the signer independent recognition performance.
- The signer wearing a long sleeve, occlusion, and object interaction during data collection makes it challenging to preprocessing and recognition.
- The problem of handling real-time applications because most datasets are acquired using the constant background and illumination in a controlled environment.
- Recognition of unseen sign words or sentences is difficult because the limited number of vocabularies and sentences present in the data make them incompatible in the real-time use case.

3) The Solutions to Overcome the SLR Datasets Complexities: The solutions to overcome the SLR datasets complexities are as follows:

- During sign data acquisition, the sample collection must consider various environments, lighting based on multiple times, performing the same words with different signers leads to sign independent SLR model and improves the generic ability.
- The distance between the signers and the recording device should be feasible to overcome the blurring data issue.
- Hand shape-based modality alone is not good enough to recognize the sign; thus, a non-manual feature-based dataset is required to perceive the grammar of the sign language. The isolated and continuous words/sentences include many signers and more repetitions based on a dataset with a more significant number of cues and corpus to improve accuracy, robustness, and generalization.
- Versatile and massive corpus SLR dataset to address all sign components using multi-modal sensing with a complex and more extensive isolated, continuous sign without constraints based on capturing. Thus, it serves as a benchmark for SLR research to validate SLR model validity.

B. Study of current state-of-the-art models for sign language recognition

This paper further explores the state of the models presented in the sign language recognition as follows Ravi et al. [131] performed Indian sign language recognition using RGB-D data using CNN models. They used four-stream inputs for training and tested performance on two streams (RGB spatial and temporal). They got a recognition accuracy rate of 89.69 % for the BVCSL3D dataset. Gökçe et al. [132] carried out an isolated Turkish sign language recognition using 3 D residual CNN with score level fusion and got top 1 accuracy of 94.94 % for the Bosphorus Sign22K dataset. Li et al. [133] presented isolated sign language recognition using TK-3d convNet (transferring cross-domain knowledge-based 3D convolution network). Recognition accuracy of 77.55% for WLASL 100 and 68.75% for WLASL 200, 83.91% for MSASL 100 and 81.14% for MSASL 200 achieved based on the TK-3d convNet SLR model. Camgoz et al. [134] SLRT (Sign language recognition and translation using transformer) applicability verified with RWTH-PHOENIX- Weather 2014-T dataset achieved 21.80 as BLEU 4 score. Li et al. [135] suggested TSPNet – Temporal sematic pyramid network association of hierarchical feature learning based on continuous sign language recognition and result in BLEU 4 of 13.41 for RWTH-PHOENIX- Weather 2014-T dataset.

Zheng et al. [136] suggested a non-independent multi-stream convolutional and RoIs based multi-region convolutional architecture for sign language translation and obtained BLEU 4 scores – 10.89 (RoI) and 10.73 (stream) for RWTH-PHOENIX- Weather 2014-T. Ahmed et al. [137] presented Wi-Fi CSI (channel state information) dataset and developed sign language recognition using device-free Wi-Fi. SVM augmented-based model results with an accuracy of 98.5 % for Dynamic sign and 99.9 % for Static sign. Zhou et al. [85] designed a continuous sign language recognition based on STMC (spatial-temporal multi-cue network). They got the WER (word error rate) of 2.1, 28.6, 20.7, and 21.0 for Continuous SLR 100 dataset Split I case, Split II case, RWTH-PHOENIX-Weather 2014, and RWTH-PHOENIX-Weather 2014 T datasets, respectively. Slimane and Bouguessa [138] performed self-attention network (SAN-sign attention network)-based continuous sign language recognition. They used 2 D CNN with self-attention considered both hand and full-frame as inputs and combined to get final word glosses on evaluation on RWTH-PHOENIX-Weather 2014 dataset achieved WER of 29.78 %. Töngi [139] suggested the inflated deep CNN based on isolated SLR. They used the MSASL dataset to transfer the ASL knowledge to recognize GSL (German Sign Language) on the SIGNUM dataset and achieved an accuracy of 0.75 for high target data. Hu et al. [140] pointed out non-manual feature-aware GLEN (Global local enhancement network) based on the SLR model. They achieved a top 1 accuracy of 69.9% for NMFs-CSL datasets and 96.8% for isolated SLR 500 datasets. De Coster et al. [141] proposed Pose flow and hand cropping associated to video transformer network-based isolated sign language recognition. The VTN-PF (Video Transformer Network with hand cropping and pose) model evaluation on the AUTSL dataset got an accuracy of 92.92 %.

Jiang et al. [142] devised a SAM SLR (Skeleton Aware multimodal framework Sign language recognition) concerning isolated sign language recognition. The skeleton-aware multi-modal (SSTCN–Separable spatial-temporal convolution network) results in better accuracy on the AUTSL dataset, with a top 1 accuracy of 98.42% for RGB and 98.53% for RGB RGB-D. Papastratis et al. [143] performed a generative adversarial network with transformer-based continuous sign language recognition. They used four datasets to validate the performance of SLRGAN (sign language recognition generative adversarial network). SLRGAN Deaf-to-Deaf SLRGAN achieves WER of 36.05 for GSL SD, WER of 2.26 for GSL SI, and SLRGAN WER of 2.98 for GSL SI, WER of 37.11 for GSL SD, WER of 23.4% for RWTH-PHOENIX-Weather 2014-T, and WER of 2.1% Continuous SLR 100. Min et al. [144] conducted VMC (visual alignment constraint) associated Resnet 18 backbone based on continuous sign language recognition model validated on Continuous SLR 100, and RWTH-PHOENIX-Weather 2014 datasets obtained a WER of 1.6% and 22.3%, respectively.

Jiang et al. [145] designed SMA-SLR- v2 (Skeleton aware multimodal framework with global ensemble model) based on isolated sign language recognition. They achieved the top 1 accuracy of 98.53% for AUSTL (RGBD all), the top 1 accuracy of 59.39% for the WLASL2000 dataset per instance case, and 56.63% per class, and the top 1 accuracy of 99% accuracy for isolated SLR 500 dataset. Meng and Li [146] presented a GCN (graph convolution network)-based SLR network (dual sign language recognition model). The fusion of the two-stream models is SLR-Net-J+B, which results in the top 1 accuracy of 98.08% for the isolated SLR -500 dataset and 64.57% for the DEVISIGN-L dataset. Pereira-Montiel et al. [147] devised Colombian sign language automatic recognition using SVM (support vector machine) with RBFK (radial basis function kernel) with four channels of sEMG (surface electromyography) and three-axis acc (accelerometer). Achieve accuracy 96.66% for 12-word recognition. Boháček and Hrúz [148] performed isolated sign language recognition using SPOTER (Sign pose base transformer) validated with LSA64 and WLASL datasets, resulting in a 100% accuracy for LSA64 and 63.18% and 43.78% accuracy for WLASL 100 and WLASL 300, respectively. The current state-of-the-art SLR model is summarized in a Table XI for better understanding. We hope this review paper sets a baseline for futuristic and advanced research in the SLR domain.

VI. Discussion

Sign language possesses dynamic gestures, trajectory property, and multi-dimensional feature vectors. These factors make it challenging to recognize sign language. Still, many researchers are attempting to develop a generalized, reliable, and robust SLR model. Multi-dimensional features are a novel approach that leads to a better recognition rate. This review paper aims to provide an easy understanding and helpful guidance to the research community. To perform research to develop an effective SLR model to assist the hand-talking community is one of the prominent domains in computer vision, pattern recognition, and natural language processing.

A. *Limitation of Current Datasets and their sizes*

The ambiguities and lack of training dataset make the SLR vulnerable. Therefore, the standardized and large-scale datasets with manual and non-manual features are important. The limitation of the current datasets and their sizes are as follows:
Barrier concerning the recording/ collection/ measuring equipment:

- Poor camera quality affect the clarity of the sign in the vision-based system when the resolution is reduced leads to decreased accuracy.
- Improper camera setup is another barrier because it leads to loss of important sign information when a sign is dynamic or static, performing the signer.
- If a multi-camera set-up is used to acquire the signer data, the lack of synchronization lead to information loss, leading to poor performance.
- Device dependability should be reliable, cost-effective, and easy to maintenance.

The environment, background, and illumination profoundly affect the dataset preparation.

- When the background setup comprises noise, it creates misclassification and reduces the recognition rate, so it should be properly dealt with to overcome this barrier.
- Improper light and illumination reduce the clarity and also affect accuracy.
- The distance between the camera and the signer should be maintained at a nominal and workable range. Very much closer and farther, much long distance between the signer and the camera affect the performance.

B. *Limitation of Current Trends*

The limitations of the current trends in SLR are as follows:

The barrier regarding the different signers affects the accuracy:

- Break off between the letter/ sign and speed up sign performing: The speedy, continuous, and frequent sign performed by the signer creates challenges for segmentation and feature extraction.
- Blockage of overlapping, occlusion of hand-face, hand-hand.
- Wearing a dress with long sleeves and wearing colored gloves also affects the sign recognition process.
- High variation concerns the interpersonal: Sign varies between signers and instants.

The barrier concerning the video domain:

- The problem of handling the video data in the limited GPU memory is not tractable. Most CNN techniques are only image-based, videos that have an additional temporal dimension. A simple resizing process may cause a loss of crucial temporal information to perform the fine-tuning and classification process on each frame independently.

The barrier concerning network design in machine learning:

- The recognition and classification ability prevailed by the location, illumination, and so on.
- Higher batch size causes a fall in local convergence instead of global convergence. Smaller batch sizes lead to larger iterations and a rise in training expenses.
- Selection of the loss functions during training cause expenses.
- Selection of optimal hyperparameters.

The active research domain is AI-based realistic modeling SLR translation and production of Avatar modeling (manual and non-manual). Developing AI Sign language learning and translation applications (web-based or smartphone) is one of the current trends. Although the advent of deep learning networks improves SLR accuracy, the limitations mentioned above still need to be addressed in the SLR domain.

C. Other Potential Applications of SLR with Human-Computer Interaction

Some potential applications of SLR with human-computer interaction are as follows:

- Virtual reality: With the help of the electronics equipment, the user experiences artificial simulation of real world.
- Smart home: Home attributes to monitor, access, and control using artificial intelligence and electronics devices. It includes a security and alarm alert system.
- Health care: Intended to assist the patients in a better quality of life and good health care service.
- Social safety: To ensure safe and social engagement and to minimize social threats.
- Telehealth: Remotely accessing clinical contacts and care services to enhance patients' health care.
- Virtual shopping: To provide hassle-free, more comfortable shopping with virtual stores.
- Digital signature: To transfer the information as an electronics sign.
- Gaming and playing: To facilitate more entertaining, and gaming experience to users.
- Text and voice assistance: To provide better communication using technology and ease of user comfort.
- Education: To facilitate enhanced learning skills using advanced techniques.

VII. Future Direction and Research Scope

Compared to the recent developmental achievement in automatic speech recognition, SLR is still lagging with a vast gap and remains at an early development stage. According to the literature study, a good number of research exists in SLR. Much research is struggling to achieve a high performance SLR model by exploring advanced techniques like deep learning, machine learning, optimization, and advanced hardware and sensor experimentation. Finally. We need a thorough exploration to solve the following issues in SLR.

- Distinctiveness/contract of sign handling problem.
- Multiple sensors/camera fusion problems.
- Multi-modalities data handling issues.
- Computation problem.
- Consistency issues.
- Difficult to handle a large vocabulary.
- Requirement of standard datasets.

Future Directions

Future directions for SLR are as follows:

- SLR model design needs a better understanding of optimal hyperparameter estimation strategy.
- Building uncontrolled surrounding/environment-based SLR models is a thrust area because researchers develop most of the existing models in the literature with respect to the lab environment-based datasets. Hence, it is demanding.
- Designing a user-friendly, realistic, and robust sign language model is one of the high-scope domains of SLR.
- Design a high-precision sign language capturing device (sensor and camera).
- Devise a novel training strategy to reduce computational training difficulty.
- The lightweight CNN model for SLR is another research scope.
- Develop an SLR with the association multi-modal based leverage to improve recognition accuracy.
- Devising a generic automatic SLR model.

This review paper is presented to provide a complete guide to the research and allow the reader to know about the existing SLR works. It demonstrates challenging problems, research gaps, future research direction, and dataset resources. Therefore, the readers and researchers can move forward towards developing novel models and products to assist the hands-talk community and contribute to social benefits.

VIII. Conclusion

There are several review papers on hand gestures and SLR. Still, existing review papers do not comprehensively discuss facial expression, modality, and dataset-based sign language, lacking in-depth discussion. With this motivation, this review paper studied different types of SLR, various sensing approaches, modalities, and various SLR datasets and listed out the issues of SLR and the future direction of SLR. However, further complete guidance will provide a more precise understanding and acquire knowledge and awareness of the problem's complexity, state-of-the-art models, and challenges in SLR.

This comprehensive review paper will help to guide upcoming researchers about the SLR introduction, needs, applications, and processes involved in SLR. It discussed various manual and non-manual SLR models. Also, it reviewed isolated and continuous of each type (manual and non-manual) and provided easy understanding to the reader with the help table and diagram. The manual and non-manual type-based SLR present effectively and then examine the works related to the various modalities and datasets. Finally, we reviewed recent research progress, challenges, and barriers of existing SLR models, organized in an informative and valuable manner concerning the various types, modalities, and dataset. The improvisation of accuracy concerns vision-based SLR, one of the ongoing and hot research topics. The sensor-based approach is highly suitable for laboratory-based experimentation. But not an appropriate choice for practical real-time applications. The vision-based SLR model's accuracy is less than the sensor-based approach and very much less than the speech recognition model. A robust and sophisticated method is essential for extracting manual and non-manual features and overcoming the barriers. Therefore, a lot of scopes are available to the SLR domain. We hope this review paves insight for readers and researchers to propose a state-of-the-art method that facilitates better communication and improves the hand-talk community's human life.

Declaration

Funding- No Funding was received for this work.
Competing interests- The authors declare no conflict of interest.
Availability of data and materials- Not Applicable.
Code availability- Not Applicable.

Acknowledgment

This work of Dr. M. Madhiarasan was supported by the MHRD (Grant No. OH-31-24-200-328) project.

References

[1] W. H. Organization. (2018) Deafness and hearing loss. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss
[2] S. Aly and W. Aly, "Deeparslr: A novel signer-independent deep learning framework for isolated arabic sign language gestures recognition," *IEEE Access*, vol. 8, pp. 83 199–83 212, 2020.
[3] R. Rastgoo, K. Kiani, and S. Escalera, "Hand sign language recognition using multi-view hand skeleton," *Expert Systems with Applications*, vol. 150, p. 113336, 2020.
[4] K. Tripathi and N. B. G. Nandi, "Continuous indian sign language gesture recognition and sentence formation," *Procedia Computer Science*, vol. 54, pp. 523–531, 2015.
[5] M. Hassan, K. Assaleh, and T. Shanableh, "Multiple proposals for continuous arabic sign language recognition," *Sensing and Imaging*, vol. 20, no. 1, p. 4, 2019.
[6] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman, "Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues," in *European Conference on Computer Vision*. Springer, 2020, pp. 35–53.
[7] O. Aran, I. Ari, L. Akarun, B. Sankur, A. Benoit, A. Caplier, P. Campr, A. H. Carrillo *et al.*, "Signtutor:

An interactive system for sign language tutoring," *IEEE MultiMedia*, vol. 16, no. 1, pp. 81–93, 2009.

[8] F. S. Botros, A. Phinyomark, and E. J. Scheme, "Electromyography-based gesture recognition: Is it time to change focus from the forearm to the wrist?" *environments*, vol. 14, p. 15, 2020.

[9] R. Fatmi, S. Rashad, and R. Integlia, "Comparing ann, svm, and hmm based machine learning methods for american sign language recognition using wearable motion sensors," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2019, pp. 0290–0297.

[10] S. Wei, X. Chen, X. Yang, S. Cao, and X. Zhang, "A component-based vocabulary-extensible sign language gesture recognition framework," *Sensors*, vol. 16, no. 4, p. 556, 2016.

[11] S. Kim, J. Kim, S. Ahn, and Y. Kim, "Finger language recognition based on ensemble artificial neural network learning using armband emg sensors," *Technology and Health Care*, vol. 26, no. S1, pp. 249–258, 2018.

[12] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 1, pp. 131–153, 2019.

[13] S. Kim, Y. Ji, and K.-B. Lee, "An effective sign language learning with object detection based roi segmentation," in *2018 Second IEEE International Conference on Robotic Computing (IRC)*. IEEE, 2018, pp. 330–333.

[14] M. P. Paulraj, S. Yaacob, M. S. bin Zanar Azalan, and R. Palaniappan, "A phoneme based sign language recognition system using skin color segmentation," in *2010 6th International Colloquium on Signal Processing & its Applications*. IEEE, 2010, pp. 1–5.

[15] A. S. Ghotkar and G. K. Kharate, "Vision based real time hand gesture recognition techniques for human computer interaction," *Int J Comput Appl*, vol. 70, no. 16, pp. 1–8, 2013.

[16] J.-H. Shin, J.-S. Lee, S.-K. Kil, D.-F. Shen, J.-G. Ryu, E.-H. Lee, H.-K. Min, and S.-H. Hong, "Hand region extraction and gesture recognition using entropy analysis," *IJCSNS International Journal of Computer Science and Network Security*, vol. 6, no. 2A, pp. 216–222, 2006.

[17] R. Akmeliawati, F. Dadgostar, S. Demidenko, N. Gamage, Y. C. Kuang, C. Messom, M. Ooi, A. Sarrafzadeh, and G. SenGupta, "Towards real-time sign language analysis via markerless gesture tracking," in *2009 IEEE Instrumentation and Measurement Technology Conference*. IEEE, 2009, pp. 1200–1204.

[18] X. Wang, M. Xia, H. Cai, Y. Gao, and C. Cattani, "Hidden-markov-models-based dynamic hand gesture recognition," *Mathematical Problems in Engineering*, vol. 2012, 2012.

[19] P. Li, T. Zhang, and A. E. Pece, "Visual contour tracking based on particle filters," *Image and Vision Computing*, vol. 21, no. 1, pp. 111–123, 2003.

[20] R. A. Kadhim and M. Khamees, "A real-time american sign language recognition system using convolutional neural network for real datasets," *TEM Journal*, vol. 9, no. 3, p. 937, 2020.

[21] J. Forster, C. Oberdörfer, O. Koller, and H. Ney, "Modality combination techniques for continuous sign language recognition," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2013, pp. 89–99.

[22] C. K. Lee, K. K. Ng, C.-H. Chen, H. C. Lau, S. Chung, and T. Tsoi, "American sign language recognition and training method with recurrent neural network," *Expert Systems with Applications*, vol. 167, p. 114403, 2021.

[23] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, T. S. Alrayes, H. Mathkour, and M. A. Mekhtiche, "Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation," *IEEE Access*, vol. 8, pp. 192 527–192 542, 2020.

[24] G. Yuan, X. Liu, Q. Yan, S. Qiao, Z. Wang, and L. Yuan, "Hand gesture recognition using deep feature fusion network based on wearable sensors," *IEEE Sensors Journal*, vol. 21, no. 1, pp. 539–547, 2020.

[25] P. Vamplew and A. Adams, "Recognition of sign language gestures using neural networks," in *European Conference on Disabilities, Virtual Reality and Associated Technologies*, 1996.

[26] K. Kudrinko, E. Flavin, X. Zhu, and Q. Li, "Wearable sensor-based sign language recognition: A comprehensive review," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 82–97, 2020.

[27] O. M. Sincan and H. Y. Keles, "Autsl: A large scale multi-modal turkish sign language dataset and baseline methods," *IEEE Access*, vol. 8, pp. 181 340–181 355, 2020.

[28] C. Chansri and J. Srinonchat, "Hand gesture recognition for thai sign language in complex background using fusion of depth and color video," *Procedia Computer Science*, vol. 86, pp. 257–260, 2016.

[29] J. R. Pansare and M. Ingle, "Vision-based approach for american sign language recognition using edge orientation histogram," in *2016 International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 2016, pp. 86–90.

[30] J. Singha and K. Das, "Indian sign language recognition using eigen value weighted euclidean distance based classification technique," *arXiv preprint arXiv:1303.0634*, 2013.

[31] W. Ahmed, K. Chanda, and S. Mitra, "Vision based hand gesture recognition using dynamic time warping for indian sign language," in *2016 International Conference on Information Science (ICIS)*. IEEE, 2016, pp. 120–125.

[32] M. Prasad, P. Kishore, E. K. Kumar, and D. A. Kumar, "Indian sign language recognition system using new fusion based edge operator." *Journal of Theoretical & Applied Information Technology*, vol. 88, no. 3, 2016.

[33] K. K. Gurjal P, "Real time hand gesture recognition using sift." *Int J Electron Electr Eng*, vol. 2, no. 3,

2012.
[34] Y. Yao and C.-T. Li, “Hand posture recognition using surf with adaptive boosting,” in *British Machine Vision Conference*, 2012.
[35] N. Kumar, “Sign language recognition for hearing impaired people based on hands symbols classification,” in *2017 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 2017, pp. 244–249.
[36] P. Shukla, A. Garg, K. Sharma, and A. Mittal, “A dtw and fourier descriptor based approach for indian sign language recognition,” in *2015 Third International Conference on Image Information Processing (ICIIP)*. IEEE, 2015, pp. 113–118.
[37] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. Crawford, M. M. Rahman, E. Kurtoglu, R. Aksu, T. Macks, and R. Mdrafi, “American sign language recognition using rf sensing,” *IEEE Sensors Journal*, 2021.
[38] G. Dour and S. Sharma, “Recognition of alphabets of indian sign language by sugeno type fuzzy neural network,” *Pattern Recognit Lett*, vol. 30, pp. 737–742, 2016.
[39] A. Sethi, S. Hemanth, K. Kumar, N. Bhaskara Rao, and R. Krishnan, “Signpro-an application suite for deaf and dumb,” *IJCSET*, vol. 2, no. 5, pp. 1203–1206, 2012.
[40] H. Lahiani, M. Elleuch, and M. Kherallah, “Real time hand gesture recognition system for android devices,” in *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*. IEEE, 2015, pp. 591–596.
[41] J. R. Pansare, S. H. Gawande, and M. Ingle, “Real-time static hand gesture recognition for american sign language (asl) in complex background,” *J Signal Inf Process*, vol. 3, pp. 364–367, 2012.
[42] R. Lionnie, I. K. Timotius, and I. Setyawan, “Performance comparison of several pre-processing methods in a hand gesture recognition system based on nearest neighbor for different background conditions,” *Journal of ICT Research and Applications*, vol. 6, no. 3, pp. 183–194, 2012.
[43] Z. Zhang, X. Qin, X. Wu, F. Wang, and Z. Yuan, “Recognition of chinese sign language based on dynamic features extracted by fast fourier transform,” in *Pacific Rim Conference on Multimedia*. Springer, 2016, pp. 508–517.
[44] A. Zorins and P. Grabusts, “Review of data preprocessing methods for sign language recognition systems based on artificial neural networks.” *Information Technology & Management Science (Sciendo)*, vol. 19, no. 1, 2016.
[45] A. Tsagaris and S. Manitsaris, “Colour space comparison for skin detection in finger gesture recognition,” *International Journal of Advances in Engineering & Technology*, vol. 6, no. 4, p. 1431, 2013.
[46] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, and M. A. Mekhtiche, “Hand gesture recognition for sign language using 3dcnn,” *IEEE Access*, vol. 8, pp. 79 491–79 509, 2020.
[47] S. Khalid, T. Khalil, and S. Nasreen, “A survey of feature selection and feature extraction techniques in machine learning,” in *2014 science and information conference*. IEEE, 2014, pp. 372–378.
[48] S. Jiang, B. Lv, W. Guo, C. Zhang, H. Wang, X. Sheng, and P. B. Shull, “Feasibility of wrist-worn, real-time hand, and surface gesture recognition via semg and imu sensing,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3376–3385, 2018.
[49] M. Tariq, A. Iqbal, A. Zahid, Z. Iqbal, and J. Akhtar, “Sign language localization: Learning to eliminate language dialects,” in *2012 15th International Multitopic Conference (INMIC)*. IEEE, 2012, pp. 17–22.
[50] ethnologue. (2014) ethnologue. [Online]. Available: https://www.ethnologue.com/
[51] E.-J. Ong, H. Cooper, N. Pugeault, and R. Bowden, “Sign language recognition using sequential pattern trees,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2200–2207.
[52] S. Yin, J. Yang, Y. Qu, W. Liu, Y. Guo, H. Liu, and D. Wei, “Research on gesture recognition technology of data glove based on joint algorithm,” in *2018 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018)*. Atlantis Press, 2018, pp. 41–50.
[53] S. P. Y. Jane and S. Sasidhar, “Sign language interpreter: Classification of forearm emg and imu signals for signing exact english,” in *2018 IEEE 14th International Conference on Control and Automation (ICCA)*. IEEE, 2018, pp. 947–952.
[54] S. G. M. Almeida, F. G. Guimarães, and J. A. Ramírez, “Feature extraction in brazilian sign language recognition based on phonological structure and using rgb-d sensors,” *Expert Systems with Applications*, vol. 41, no. 16, pp. 7259–7271, 2014.
[55] B. G. Lee and S. M. Lee, “Smart wearable hand device for sign language interpretation system with sensors fusion,” *IEEE Sensors Journal*, vol. 18, no. 3, pp. 1224–1232, 2018.
[56] L. Li, S. Jiang, P. B. Shull, and G. Gu, “Skingest: artificial skin for gesture recognition via filmy stretchable strain sensors,” *Advanced Robotics*, vol. 32, no. 21, pp. 1112–1121, 2018.
[57] X. Yang, X. Chen, X. Cao, S. Wei, and X. Zhang, “Chinese sign language recognition based on an optimized tree-structure framework,” *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 4, pp. 994–1004, 2017.
[58] A. Y. Dawod and N. Chakpitak, “Novel technique for isolated sign language based on fingerspelling recognition,” in *2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*. IEEE, 2019, pp. 1–8.
[59] M. Hrúz, P. Campr, and A. Karpov, “Input and output modalities used in a sign-language-enabled information kiosk,” *SCREEN*, vol. 1, no. C3, p. C2, 2009.
[60] C. K. Mummadi, F. P. P. Leo, K. D. Verma,

S. Kasireddy, P. M. Scholl, J. Kempfle, and K. V. Laerhoven, “Real-time and embedded detection of hand gestures with an imu-based glove,” in *Informatics*, vol. 5, no. 2. Multidisciplinary Digital Publishing Institute, 2018, p. 28.
[61] R. Gupta and A. Kumar, “Indian sign language recognition using wearable sensors and multi-label classification,” *Computers & Electrical Engineering*, p. 106898, 2020.
[62] V. T. Hoang, “Hgm-4: A new multi-cameras dataset for hand gesture recognition,” *Data in Brief*, vol. 30, p. 105676, 2020.
[63] R. Elakkiya, “Machine learning based sign language recognition: A review and its research frontier,” *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–20, 2020.
[64] S. Nayak, K. Duncan, S. Sarkar, and B. Loeding, “Finding recurrent patterns from continuous sign language sentences for automated extraction of signs,” *Journal of Machine Learning Research*, vol. 13, pp. 2589—-2615, 2012.
[65] W. Kong and S. Ranganath, “Towards subject independent continuous sign language recognition: A segment and merge approach,” *Pattern Recognition*, vol. 47, no. 3, pp. 1294–1308, 2014.
[66] Y. Ye, Y. Tian, M. Huenerfauth, and J. Liu, “Recognizing american sign language gestures from within continuous videos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2064–2073.
[67] R. Gupta and S. Rajan, “Comparative analysis of convolution neural network models for continuous indian sign language classification,” *Procedia Computer Science*, vol. 171, pp. 1542–1550, 2020.
[68] W. Pan, X. Zhang, and Z. Ye, “Attention-based sign language recognition network utilizing keyframe sampling and skeletal features,” *IEEE Access*, vol. 8, pp. 215 592–215 602, 2020.
[69] I. Papastratis, K. Dimitropoulos, D. Konstantinidis, and P. Daras, “Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space,” *IEEE Access*, vol. 8, pp. 91 170–91 180, 2020.
[70] U. Von Agris, J. Zieren, U. Canzler, B. Bauer, and K.-F. Kraiss, “Recent developments in visual sign language recognition,” *Universal Access in the Information Society*, vol. 6, no. 4, pp. 323–362, 2008.
[71] S. Sarkar, B. Loeding, R. Yang, S. Nayak, and A. Parashar, “Segmentation-robust representations, matching, and modeling for sign language,” in *CVPR 2011 WORKSHOPS*. IEEE, 2011, pp. 13–19.
[72] M. Fagiani, E. Principi, S. Squartini, and F. Piazza, “Signer independent isolated italian sign recognition based on hidden markov models,” *Pattern Analysis and Applications*, vol. 18, no. 2, pp. 385–402, 2015.
[73] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, “Chinese sign language recognition with adaptive hmm,” in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.
[74] P. Kumar, P. P. Roy, and D. P. Dogra, “Independent bayesian classifier combination based sign language recognition using facial expression,” *Information Sciences*, vol. 428, pp. 30–48, 2018.
[75] A. Sabyrov, M. Mukushev, and V. Kimmelman, “Towards real-time sign language interpreting robot: Evaluation of non-manual components on recognition accuracy.” in *CVPR Workshops*, 2019.
[76] M. Mukushev, A. Sabyrov, A. Imashev, K. Koishybay, V. Kimmelman, and A. Sandygulova, “Evaluation of manual and non-manual components for sign language recognition,” in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 6073–6078.
[77] P. Kishore, D. A. Kumar, A. C. S. Sastry, and E. K. Kumar, “Motionlets matching with adaptive kernels for 3-d indian sign language recognition,” *IEEE Sensors Journal*, vol. 18, no. 8, pp. 3327–3337, 2018.
[78] Z. Liu, X. Qi, and L. Pang, “Self-boosted gesture interactive system with st-net,” in *Proceedings of the 26th ACM International Conference on Multimedia*, 2018, pp. 145–153.
[79] R. B. Wilbur, “Phonological and prosodic layering of nonmanuals in american sign language,” in *The signs of language revisited*. Psychology Press, 2013, pp. 196–220.
[80] A. Farhadi and D. Forsyth, “Aligning asl for statistical translation using a discriminative word model,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, vol. 2. IEEE, 2006, pp. 1471–1476.
[81] I. Infantino, R. Rizzo, and S. Gaglio, “A framework for sign language sentence recognition by commonsense context,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 5, pp. 1034–1039, 2007.
[82] H.-D. Yang and S.-W. Lee, “Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine,” *Pattern Recognition Letters*, vol. 34, no. 16, pp. 2051–2056, 2013.
[83] C. Zhang, Y. Tian, and M. Huenerfauth, “Multimodality american sign language recognition,” in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2881–2885.
[84] H. Brock, I. Farag, and K. Nakadai, “Recognition of non-manual content in continuous japanese sign language,” *Sensors*, vol. 20, no. 19, p. 5621, 2020.
[85] H. Zhou, W. Zhou, Y. Zhou, and H. Li, “Spatial-temporal multi-cue network for continuous sign language recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 009–13 016.
[86] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, “Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos,” *IEEE Transactions on Pattern Analy-

sis and Machine Intelligence, vol. 42, no. 9, pp. 2306–2320, 2020.

[87] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3793–3802.

[88] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, "Include: A large scale dataset for indian sign language recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1366–1375.

[89] H. R. V. Joze and O. Koller, "Ms-asl: A large-scale data set and benchmark for understanding american sign language," *arXiv preprint arXiv:1812.01053*, 2018.

[90] O. Özdemir, A. A. Kındıroğlu, N. Cihan Camgoz, and L. Akarun, "BosphorusSign22k Sign Language Recognition Dataset," in *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, 2020.

[91] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho, "Neural sign language translation based on human keypoint estimation," *Applied Sciences*, vol. 9, no. 13, p. 2683, 2019.

[92] A. Elboushaki, R. Hannane, K. Afdel, and L. Koutti, "Multid-cnn: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in rgb-d image sequences," *Expert Systems with Applications*, vol. 139, p. 112829, 2020.

[93] O. Köpüklü, A. Gunduz, N. Kose, and G. Rigoll, "Real-time hand gesture detection and classification using convolutional neural networks," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[94] K. M. Lim, A. W. C. Tan, C. P. Lee, and S. C. Tan, "Isolated sign language recognition using convolutional neural network hand modelling and hand energy image," *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 19 917–19 944, 2019.

[95] Y. Chen, L. Zhao, X. Peng, J. Yuan, and D. N. Metaxas, "Construct dynamic graphs for hand gesture recognition via spatial-temporal attention," *arXiv preprint arXiv:1907.08871*, 2019.

[96] P. M. Ferreira, J. S. Cardoso, and A. Rebelo, "On the role of multimodal learning in the recognition of sign language," *Multimedia Tools and Applications*, vol. 78, no. 8, pp. 10 035–10 056, 2019.

[97] F. Gomez-Donoso, S. Orts-Escolano, and M. Cazorla, "Accurate and efficient 3d hand pose regression for robot hand teleoperation using a monocular rgb camera," *Expert Systems with Applications*, vol. 136, pp. 327–337, 2019.

[98] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 89–98.

[99] E. Kazakos, C. Nikou, and I. A. Kakadiaris, "On the fusion of rgb and depth information for hand pose estimation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 868–872.

[100] Y. Li, Z. Xue, Y. Wang, L. Ge, Z. Ren, and J. Rodriguez, "End-to-end 3d hand pose estimation from stereo cameras." in *BMVC*, vol. 1, 2019, p. 2.

[101] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "Ganerated hands for real-time 3d hand tracking from monocular rgb," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 49–59.

[102] D. Victor, "Handtrack: A library for prototyping real-time hand trackinginterfaces using convolutional neural networks," *GitHub repository*, 2017.

[103] S. Baek, K. I. Kim, and T.-K. Kim, "Augmented skeleton space transfer for depth-based hand pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8330–8339.

[104] G. Moon, J. Y. Chang, and K. M. Lee, "V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5079–5088.

[105] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Robust 3d hand pose estimation from single depth images using multi-view cnns," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4422–4436, 2018.

[106] L. Ge and et. al, "3d convolutional neural networks for efficient and robust hand pose estimation from single depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1991–2000.

[107] E. Dibra, T. Wolf, C. Oztireli, and M. Gross, "How to refine 3d hand pose estimation from unlabelled depth data?" in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 135–144.

[108] A. Sinha, C. Choi, and K. Ramani, "Deephand: Robust hand pose estimation by completing a matrix imputed with deep features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4150–4158.

[109] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4903–4911.

[110] M. J. Marin-Jimenez, F. J. Romero-Ramirez, R. Munoz-Salinas, and R. Medina-Carnicer, "3d human pose estimation from depth maps using a deep combination of poses," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 627–639, 2018.

[111] X. Deng, S. Yang, Y. Zhang, P. Tan, L. Chang, and H. Wang, "Hand3d: Hand pose estimation using 3d neural network," *arXiv preprint arXiv:1704.02224*, 2017.

[112] M. Oberweger, G. Riegler, P. Wohlhart, and V. Lepetit, "Efficiently creating 3d training data for fine hand pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp.

4957–4965.

[113] M. Oberweger, P. Wohlhart, and V. Lepetit, "Hands deep in deep learning for hand pose estimation," *arXiv preprint arXiv:1502.06807*, 2015.

[114] R. Rastgoo, K. Kiani, and S. Escalera, "Multi-modal deep hand sign language recognition in still images using restricted boltzmann machine," *Entropy*, vol. 20, no. 11, p. 809, 2018.

[115] J. Duan, S. Zhou, J. Wan, X. Guo, and S. Z. Li, "Multi-modality fusion based on consensus-voting and 3d convolution for isolated gesture recognition," *arXiv preprint arXiv:1611.06689*, 2016.

[116] X. Chen, G. Wang, H. Guo, and C. Zhang, "Pose guided structured region ensemble network for cascaded hand pose estimation," *Neurocomputing*, 2020.

[117] A. Dadashzadeh, A. T. Targhi, M. Tahmasbi, and M. Mirmehdi, "Hgr-net: a fusion network for hand gesture segmentation and recognition," *IET Computer Vision*, vol. 13, no. 8, pp. 700–707, 2019.

[118] M. Wang, X. Chen, W. Liu, C. Qian, L. Lin, and L. Ma, "Drpose3d: Depth ranking in 3d human pose estimation," *arXiv preprint arXiv:1805.08973*, 2018.

[119] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim, "Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4866–4874.

[120] H. Guo, G. Wang, X. Chen, and C. Zhang, "Towards good practices for deep 3d hand pose estimation," *arXiv preprint arXiv:1707.07248*, 2017.

[121] X. Fang and X. Lei, "Hand pose estimation on hybrid cnn-ae model," in *2017 IEEE International Conference on Information and Automation (ICIA)*. IEEE, 2017, pp. 1018–1022.

[122] M. Madadi, S. Escalera, X. Baró, and J. Gonzalez, "End-to-end global to local cnn learning for hand pose recovery in depth data," *arXiv preprint arXiv:1705.09606*, 2017.

[123] P. Wang, W. Li, S. Liu, Z. Gao, C. Tang, and P. Ogunbona, "Large-scale isolated gesture recognition using convolutional neural networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 7–12.

[124] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei, "Towards viewpoint invariant 3d human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 160–177.

[125] A. Tagliasacchi, M. Schröder, A. Tkach, S. Bouaziz, M. Botsch, and M. Pauly, "Robust articulated-icp for real-time hand tracking," in *Computer Graphics Forum*, vol. 34, no. 5. Wiley Online Library, 2015, pp. 101–114.

[126] R. Rastgoo, K. Kiani, and S. Escalera, "Video-based isolated hand sign language recognition using a deep cascaded model," *Multimedia Tools and Applications*, vol. 79, pp. 22 965–22 987, 2020.

[127] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[128] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[129] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, 2015.

[130] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660.

[131] S. Ravi, M. Suman, P. Kishore, K. Kumar, A. Kumar *et al.*, "Multi modal spatio temporal co-trained cnns with single modal testing on rgb–d based sign language gesture recognition," *Journal of Computer Languages*, vol. 52, pp. 88–102, 2019.

[132] Ç. Gökçe, O. Özdemir, A. A. Kındıroğlu, and L. Akarun, "Score-level multi cue fusion for sign language recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 294–309.

[133] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li, "Transferring cross-domain knowledge for video sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6205–6214.

[134] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 023–10 033.

[135] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li, "Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation," *arXiv preprint arXiv:2010.05468*, 2020.

[136] J. Zheng, Y. Chen, C. Wu, X. Shi, and S. M. Kamal, "Enhancing neural sign language translation by highlighting the facial expression information," *Neurocomputing*, vol. 464, pp. 462–472, 2021.

[137] H. F. T. Ahmed, H. Ahmad, K. Narasingamurthi, H. Harkat, and S. K. Phang, "Df-wislr: Device-free wi-fi-based sign language recognition," *Pervasive and Mobile Computing*, vol. 69, p. 101289, 2020.

[138] F. B. Slimane and M. Bouguessa, "Context matters: Self-attention for sign language recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 7884–7891.

[139] R. Töngi, "Application of transfer learning to sign language recognition using an inflated 3d deep convolutional neural network," *arXiv preprint arXiv:2103.05111*, 2021.

[140] H. Hu, W. Zhou, J. Pu, and H. Li, "Global-local enhancement network for nmf-aware sign language recognition," *ACM transactions on multimedia computing, communications, and applications (TOMM)*, vol. 17,

no. 3, pp. 1–19, 2021.

[141] M. De Coster, M. Van Herreweghe, and J. Dambre, "Isolated sign recognition from rgb video using pose flow and self-attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3441–3450.

[142] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3413–3423.

[143] I. Papastratis, K. Dimitropoulos, and P. Daras, "Continuous sign language recognition through a context-aware generative adversarial network," *Sensors*, vol. 21, no. 7, p. 2437, 2021.

[144] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," *arXiv preprint arXiv:2104.02330*, 2021.

[145] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Sign language recognition via skeleton-aware multi-model ensemble," *arXiv preprint arXiv:2110.06161*, 2021.

[146] L. Meng and R. Li, "An attention-enhanced multi-scale and dual sign language recognition network based on a graph convolution network," *Sensors*, vol. 21, no. 4, p. 1120, 2021.

[147] E. Pereira-Montiel, E. Pérez-Giraldo, J. Mazo, D. Orrego-Metaute, E. Delgado-Trejos, D. Cuesta-Frau, and J. Murillo-Escobar, "Automatic sign language recognition based on accelerometry and surface electromyography signals: A study for colombian sign language," *Biomedical Signal Processing and Control*, vol. 71, p. 103201, 2022.

[148] M. Boháček and M. Hrúz, "Sign pose-based transformer for word-level sign language recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 182–191.

Authors Biography

Dr. M. MADHIARASAN (Member, IEEE) has completed his Bachelor of Engineering degree in Electrical and Electronics Engineering in the year 2010 from Jaya Engineering College, Thiruninravur, under Anna University, Tamil Nadu, India, Master of Engineering degree in Electrical Drives and Embedded Control (Electrical Engineering) in the year 2013 from Anna University, Regional Centre, Coimbatore, under Anna University, Tamil Nadu, India and Ph.D. (Electrical Engineering) in the year 2018 from Anna University, Tamil Nadu, India.
He has worked as an Assistant Professor and R & D In-charge in the Department of Electrical and Electronics Engineering, Bharat Institute of Engineering and Technology, Hyderabad, India, from 2018 to 2020. He is presently working as a Post-Doctoral Fellow at the Department of Computer Science and Engineering, Indian Institute of Technology, Roorkee (IITR).
His research areas include Renewable Energy Systems, Power Electronics and Control, Computer Vision, Human-Computer Interface, Pattern Recognition, Artificial Intelligence, Neural Networks, Optimization, Machine Learning, Deep Learning, Soft Computing, Internet of Things and Modeling and Simulation. He has been a Technical Program Committee Member, International Scientific Committee Member of many international conferences. He acts as an editorial board member and reviewer for many peer-reviewed international journals (Springer, Elsevier, IEEE Access, etc.) and a guest editor of Energy Engineering.

Dr. PARTHA PRATIM ROY (Member, IEEE) was a Postdoctoral Research Fellow with the RFAI Laboratory, France, in 2012, and the Synchromedia Laboratory, Canada, in 2013. He was with the Advanced Technology Group, Samsung Research Institute, Noida, India, from 2013 to 2014. He is currently an Associate Professor with the Department of Computer Science and Engineering, IIT Roorkee, India. He has authored or coauthored more than 200 articles in international journals and conferences. His research interests are pattern recognition, bio-signal analysis, EEG-based pattern analysis, and multilingual text recognition. He is an Associate Editor of the IET Image Processing, IET Biometrics, IEICE Transactions on Information and Systems and Springer Nature Computer Science.