AIMS Mathematics

AIMS Mathematics, 11(2): 3839–3902.
DOI: 10.3934/math.2026156
Received: 29 September 2025
Revised: 24 December 2025
Accepted: 12 January 2026
Published: 09 February 2026

https://www.aimspress.com/journal/Math

Review

Mathematical and computational perspectives on next-generation neural networks for sign language recognition: A systematic review of advances, challenges, and assistive applications

Yahia Said[1], Mohammad Barr[2,*], Yazan A. Alsariera[3] and Ahmed A. Alsheikhy[4]

[1] Center for Scientific Research and Entrepreneurship, Northern Border University, 73213, Arar, Saudi Arabia

[2] Department of Electrical Engineering, College of Engineering, Northern Border University, Arar 91431, Saudi Arabia

[3] Department of Computer Science, College of Information and Communications Technology, Tafila Technical University, Tafila 66110, Jordan

[4] Department of Computer & Network Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah 21959, Saudi Arabia

* **Correspondence:** Email: mohammed.barr@nbu.edu.sa.

Abstract: Artificial intelligence (AI) and machine learning (ML) have revolutionized assistive technologies, particularly for individuals with hearing and speech impairments. This systematic review critically examines recent innovations in next-generation neural network architectures for sign language recognition (SLR), emphasizing their mathematical and computational foundations. Following PRISMA guidelines, we analyze state-of-the-art models, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM), and hybrid approaches integrating classical machine learning methods such as support vector machines (SVMs). We explore strategies for feature extraction, data augmentation, multimodal fusion, and optimization, highlighting their roles in improving accuracy, robustness, and real-time adaptability. Persistent challenges include dataset scarcity, limited generalizability, and computational trade-offs. From a mathematical perspective, optimization techniques, probabilistic modeling, and explainable AI frameworks are emerging as key enablers for safe and trustworthy SLR systems. This review identifies research gaps and proposes future directions toward responsible, mathematically grounded, and computationally efficient AI-powered assistive technologies.

Keywords: artificial intelligence; machine learning; mathematical models; neural networks; sign language recognition; multimodal fusion; optimization methods; explainable AI; assistive technologies; systematic review
Mathematics Subject Classification: 68T07, 68T45

1. Introduction

Effective sign language recognition (SLR) systems have a lot of potential uses since sign language communication is the main way that millions of individuals engage with the world [1,2]. Some potential applications include the translation of sign language into broadcasts, the development of equipment that responds to orders given in sign language, and the creation of sophisticated systems to aid the disabled in performing everyday tasks. Specifically, AI has become a potentially game-changing tool for researchers, and its use in solving the SLR problem will certainly have an immediate and far-reaching effect [3,4]. A subfield of speech and language processing, SLR focuses on the automatic interpretation of non-verbal cues such as hand gestures that assist the deaf and hard of hearing in communicating with one another.

Building fully functional systems that can understand sign language and respond to commands given in this format has been the goal of numerous exciting and innovative solutions proposed and tested in recent years [5–9]. This is all because hardware and software components have evolved to the point where developing advanced systems with real-time translation capacities appears to be within reach. However, it is crucial to refine the interpretation algorithms until false positives are uncommon before considering any genuinely useful applications [10–13]. There are a lot of obstacles to overcome before creating SLR technologies that can achieve near-perfect accuracy on a big vocabulary [14,15]. Therefore, it is crucial to keep coming up with new approaches and assessing their respective benefits, ultimately arriving at solutions that are more and more dependable.

While the majority of researchers think that deep learning models are the way to go, there is still some debate on the best network architecture, even if numerous other architectures have shown promise. The only method to find the top algorithms and improve them utilizing other teams' discoveries where relevant is to conduct extensive experimental evaluations. Most countries have their own distinct sign languages, thus most of the study is done at the local level with people who are fluent in the signs of the area. Considering this, it's not unexpected that SLR issues have been the focus of numerous scientific articles, and that the suggested solutions' performance levels have been climbing at a rapid pace over the past few years [16,17].

Based on the main data-gathering strategy, the different SLR solutions in the existing literature can be roughly categorized into two main classes. One set of techniques makes use of third-party sensors like data gloves to learn more about the signer's behavior. Many authors have built upon the work of Starner et al. [18] by utilizing wearable sensors in various ways. Most current efforts have focused on vision-based approaches, which use pictures, video, and depth data to deduce the semantic meaning of hand signals; this is in contrast to sensor-based approaches, which have some practical limitations.

Many more methods, some based on filtering principles, have been suggested since Chen et al. [19] introduced a skin-color-based hand gesture detection system, and many more have followed. Regarding the best neural network model for stereo vision SLR applications, the convolutional neural network

(CNN) model [20] was an early front-runner [21–24]. The commercial release of the microsoft kinect device has opened up a whole new realm of understanding [25,26], and scientists are still investigating ways to harness the power of depth vision to create more precise SLR tools. Aside from CNNs, other designs like RNNs [27] and Hidden Markov Models (HMMs) [19] are commonly used. While random forest (RF) and K-nearest neighbor (k-NN) are occasionally selected for the classification task, the SVM model is also commonly utilized for this purpose [28,29].

In this review paper, we will present a discussion of the most important works in the field of assistive techniques for the deaf and hearing impaired with a presentation of existing challenges and future directions for handling those limitations. First, a taxonomy and summary of the literature on automatic SLR is presented. We meticulously reviewed all published articles on Machine Learning and Deep Learning-based automatic sign language recognition from 2014 to 2024. Our analysis revealed that the vast amount of available data necessitates a conceptual classification of existing SLR approaches to better understand and organize the field. Consequently, this work evaluates the relative strengths and weaknesses of various SLR methodologies, focusing on the key features and commonalities shared by the majority of these approaches in relation to specific tasks and functionalities.

Second, we propose a foundational framework for SLR models. This framework is developed based on the limitations and challenges identified in the literature. While debates continue regarding the most promising areas of research, it is widely acknowledged that machine learning and deep learning techniques play a crucial role in advancing sign language recognition. Despite significant progress, even the most advanced models currently fall short of the reliability required for real-world applications. However, there is a general consensus that deeper models hold greater potential for the future of practical SLR systems compared to traditional machine learning methods.

Third, performance and benchmark datasets are examined. We analyze the use of benchmark datasets in the literature and their impact on performance. High-quality sign language datasets are critical for training SLR technologies to produce accurate and reliable predictions. However, the availability of such datasets is limited, and even when they are accessible, they are often insufficient for comprehensive testing. It is standard practice to partition datasets into training, validation, and testing subsets, enabling models to be evaluated using the same data used for optimization. Unfortunately, the lack of standardized datasets makes it challenging to directly compare results across studies, as each employs different datasets, hindering consistent performance evaluation and benchmarking.

Finally, we recognize the potential of current approaches while addressing their limitations, unresolved questions, and associated challenges. Our analysis highlights several key findings. The scarcity of high-quality datasets for less widely spoken sign languages, coupled with the regional variations in sign language alphabets and vocabularies, poses significant barriers to cross-border collaboration. This lack of standardization complicates the development and testing of more advanced applications, which require training on substantially larger vocabularies. While many proposed solutions demonstrate innovative concepts, they often fall short in terms of precision and reliability. Moreover, the complexity of semantic information further complicates its capture through statistical analysis, presenting a critical challenge for the field of continuous SLR.

The structure of this paper is as follows: Section 2 introduces the foundational concepts, including deep learning and machine learning, along with essential background information. Section 3 outlines the methodology employed in this investigation. Section 4 provides an in-depth discussion of the proposed framework and the machine learning and deep learning methodologies used in designing sign language recognition models. Section 5 explores various SLR models, while Section 6 focuses on the languages involved in the recognition process. Section 7 presents a comparative analysis of ML/DL algorithms for sign language recognition and highlights the benchmark SLR datasets used for training and validation. Section 8 addresses challenges, open questions, and future research directions. Finally, Section 9 concludes with the key findings of this study.

2. Background

The use of sophisticated algorithms that can learn from their experiences has been the subject of persistent research and development in recent years intending to automate a wide variety of language tasks [30]. Automating SLR could greatly enrich the lives of many persons who use sign language as their primary means of communication [31]. Automated SLR tools must be precise enough to prevent producing misleading or non-functional responses, otherwise, a plethora of specialized services would be impossible to develop. To set the stage for automated SLR, we give some historical context below regarding several key methods.

2.1. Machine learning

Machine learning encompasses a range of stochastic methods capable of predicting the value of a given parameter when provided with sufficient examples. For instance, using Algorithm 1 as a reference, the learning process typically involves forwarding samples through a mapping function. This category includes a variety of well-known approaches, such as naïve Bayes, random forest, K-nearest neighbor, logistic regression, and Support Vector Machine (SVM) [30–32]. Training is a fundamental step in these approaches and can be either supervised—using labeled data to establish relationships between variables—or unsupervised—where no labels are provided, and the model learns to make predictions based on input features. However, due to their inherent simplicity, these methods often fall short in capturing complex semantic cues, which are crucial for many language-related tasks. Nevertheless, they serve as valuable benchmarks for evaluating success or failure and provide a foundation for developing more sophisticated analytical techniques.

Algorithm 1: Training process

Input: x (data in a d dimension vector)
Output: y (prediction)
Mapping function f : predict labels from input data
Training data: select data, label pairs
Hyperparameters: configure model parameters
Learning algorithm: minimize loss between prediction and target

With the help of machine learning algorithms, SLR has been somewhat successful. Initial research in this area relied on information gleaned from wearable sensors, which translate a user's motions with remarkable precision. Methods like SVM can filter the data in order to get a reasonably accurate identification of the target sign. It has been attempted to interpret continuous segments of sign language speech using dynamic models such as dynamic time warping or relevance vector machines, but most of the aforementioned machine learning methods are used to analyze static content, which consists of individual signs that are isolated in time and space. Due to its superiority for easy SLR problems, fundamental stochastic models saw heavy application throughout the initial phases of the research process. While the number of features studied and the size of the dataset determine the computing power requirements, these statistical models usually use less power than more complicated systems. Despite the fact that more advanced SLR applications sometimes call for more variables and even more modalities, the simplicity of simpler models is still appealing. Therefore, simpler machine learning approaches are still useful since they can be used to compare and contrast the features of more complex methods subsequently suggested.

2.2. Deep learning

Deeper architectures, which use several layers and communicate input in vector format between them, have lately supplanted simpler Machine Learning methods. These structures progressively refine the estimation until positive recognition is obtained. Often referred to as "deep learning" systems or deep neural networks (DNN), these algorithms follow concepts comparable to the aforementioned machine learning methodologies, however with somewhat more intricacy. Recurrent neural networks (RNNs) with a minimum of one recurrent layer and convolutional neural networks (CNNs) with a minimum of one convolutional layer are the two most popular network topologies utilized for various applications.

While the training phase determines the algorithm's effectiveness, these networks can display diverse properties and typically operate better for different kinds of tasks depending on the number and kind of layers. A key consideration is the quality of the training set, since bigger and more targeted datasets typically result in more resilient network training. Typically, one can further refine a model by adjusting a few pertinent hyper-parameters that characterize the training process [33]. Currently, most studies on SLR automation use approaches that combine images with depth data; this produces a mountain of data that frequently necessitates real-time analysis, or at least consideration of the temporal dimension.

Many more complex models are built using RNN or CNN architecture since basic machine learning approaches fail to perform well with bigger and more varied datasets. In certain applications, deep networks can attain an ideal recognition accuracy of over 98% when trained with multi-modal input, such as skeletal data paired with depth images from microsoft kinect. Konstantinidis et al. [34] proved the benefits of deep learning by identifying individual sign language terms using data from many sources; nevertheless, their model's performance varied between datasets. Increasing the number of layers (depth) is sometimes necessary for more complex models used for SLR tasks like real-time translation or continuous voice interpretation. Although deep models seem like a sure bet to power automated SLR, it's unclear if the existing architectures will stay the same or if new models will emerge that are better able to get the semantics of sign language. Deep belief networks with many layers and autoencoder-based networks are two potential models that could see increased application in the future.

3. Review strategy

For the benefit of all researchers, we have evaluated and organized all available scientific information related to SLR in this paper. To better serve anyone looking for the groundwork of this area of study, we supplemented the study's basic data with an unbiased evaluation of its quality and potential for beneficial contributions. Here are some of the primary research questions that we want to address.

- Question 1: What datasets are available for automated sign language recognition research?
- Question 2: What methods are being used in SLR for different languages?
- Question 3: What problems in this scientific area have not yet been fully addressed?

The long-term goal of this work is to clear up any confusion that may arise among academics and provide a foundation for future studies on SLR. We broke it down into three distinct but interconnected stages: planning, carrying out, and presenting. First, determine which research questions are most pertinent. Second, establish ground rules for the evaluation process. Third, formalize the selection threshold. Fourth, evaluate the work's premises and results. Fifth, investigate the experimental setup methodologically. Finally, extract any relevant information that may provide answers to the mentioned questions.

3.1. Evaluation protocol

While conducting the literature review, we adhered to a particular procedure in order to conduct an objective evaluation of the content of the paper. In the first step of this procedure, acceptable variables were identified; in the second step, approaches taken by the authors were identified and analyzed; in the third step, the quantitative output was organized; and in the final step, the criteria for generalization and summary were outlined.

3.2. Inclusion and exclusion

To identify the scientific works included in this review, a clear set of criteria was established. Only studies specifically related to Sign Language Recognition (SLR) were considered, as this aligns with the focus of this article. As shown in Figure 1, the review spans the period from 2014 to 2024, aiming to systematically analyze recent advancements in the field. Table 1 presents a concise and comprehensive overview of the guidelines used for selecting research papers.

Table 1. Inclusion and exclusion criteria.

Inclusion	Exclusion
English language	Other language than English
Related to central questions	duplicated
Publisher after 2014	Out of time range
Full-text available	No access to full-text
Related to SLR tasks	Non relevant

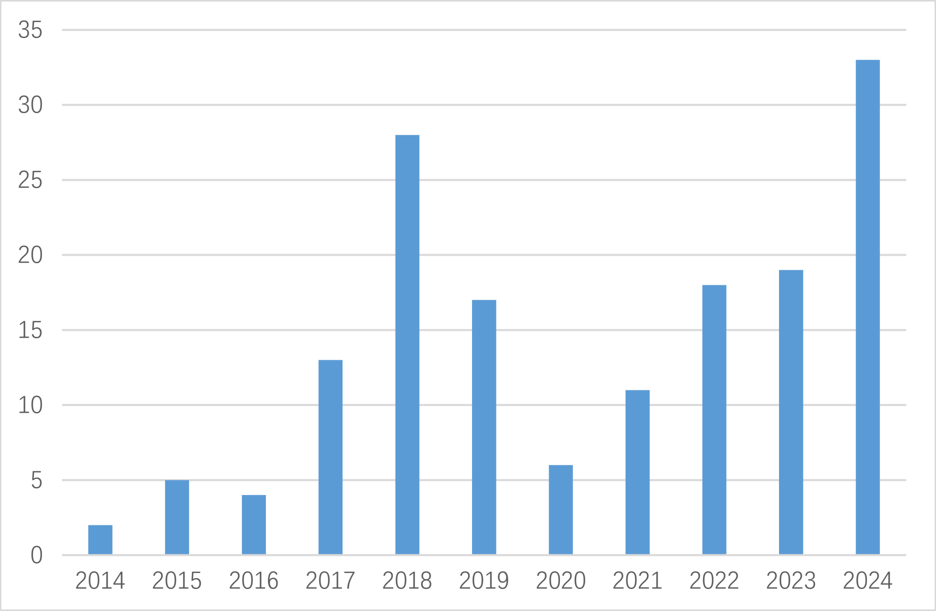

Figure 1. Number of publications on SLR by year.

3.3. Search method

By searching through sources that were accessible to the general public, it took a considerable amount of time and a combination of automated methods and manual labor to locate the most pertinent study material. The automated segment was driven by a number of keywords, which are: sign language, sign language recognition, sign language identification, automatic sign language recognition, hearing impaired, deaf, mute, deep learning, machine learning, artificial intelligence, hand gesture, pose estimation, and sign translation.

Additionally, the collection expands each time the algorithm discovers a new paper that is equally as pertinent as the ones that are already present. We ran a comprehensive search, and some of the resources that we looked through included Scopus and Web of Sciences databases. At this point, our primary objective was to locate as many publications as possible that are related to SLR. Immediately following this stage, we conducted a thorough examination of the full corpus of material that we had obtained by employing the forward/back technique. In order to acquire a more comprehensive understanding of each work, it was helpful to be able to trace the references and follow the primary research lines. Because of this, we were able to ensure that the study did not overlook any significant foundational studies and that the final collection of SLR papers appropriately reflects the most effective research directions. For the purpose of processing the collection, we utilized the Mendeley technique, which enabled us to easily classify the works in accordance with the regional sign languages that they referenced. It is evident from Figure 2 that there are several variants of sign language, the most frequent of which is the American variation. Nevertheless, there are other works that belong to American, Argentinian, Arabic, and other languages. The type of architecture that was offered for the solution was another factor that was used to differentiate across the articles. Among the criteria that were used to differentiate between the articles was the architecture of the solution that was presented. Figure 3 presents an all-encompassing summary of the situation.

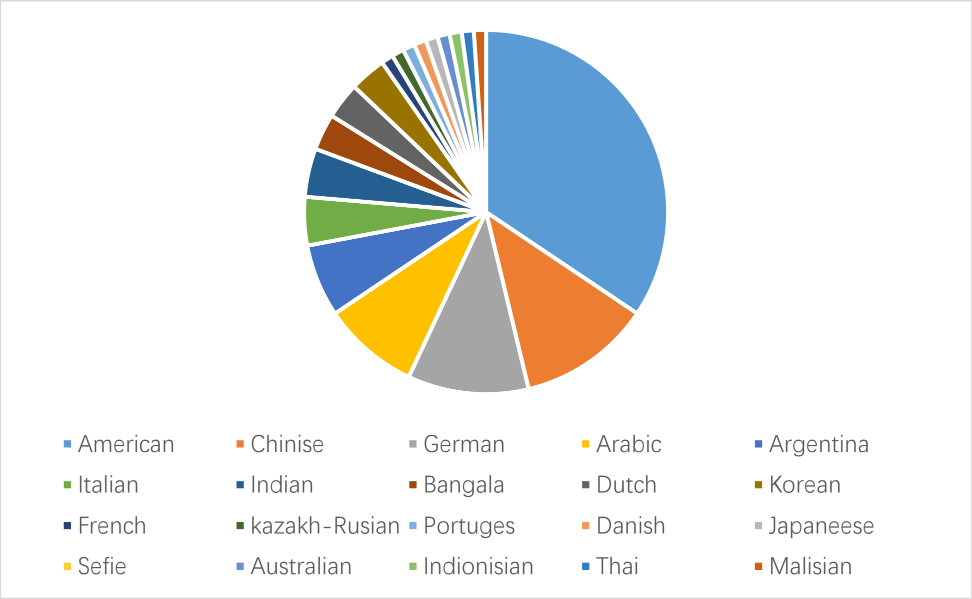

Figure 2. Percentage of publication on SLR based on the spoken language.

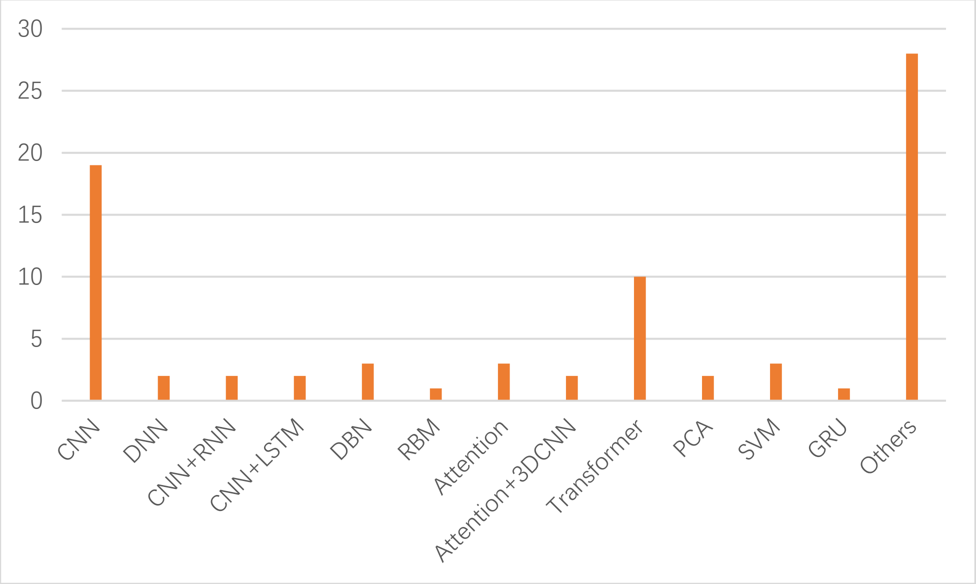

Figure 3. Percentage of publication on SLR based on the used technique.

3.4. Selection method

The initial search yielded 218 papers, of which 11 were promptly excluded as duplicates. Each remaining paper was carefully reviewed based on the information provided on its first page and the criteria outlined in Table 1. This process allowed us to filter out studies that were of low quality, unrelated to the research area, or obtained from unreliable sources. As a result, 63 papers were excluded for failing to meet the inclusion criteria, leaving 144 core and relevant papers for further analysis.

After that, we went over each study in its entirety and rejected the ones that did not particularly deal with SLR or that did not give adequate evidence to support their point of view. Following that, we made it a point to check the internet source articles for authorship and graded the quotations according to the quality and accuracy of the information they included. Finally, in order to choose the research that ought to have been disclosed, we carried out a qualitative evaluation. The selection method resulted in a reduction of the total number of publications that were included in the analysis to 84. Despite lacking scientific merit and relevance to the main research questions, these publications were excluded from the final analysis.

4. Sign language recognition based on AI techniques

The vast majority of SLR studies concentrate on the same difficulties, mostly focusing on how to interpret the hand and body gestures that are used to indicate sign language signals. Studies in this area frequently use the same methodology, despite the fact that their techniques are different. This is because the essential goals of these studies are same. The overarching paradigm that the majority of researchers in this subject agree upon is depicted in Figure 4. The first layer of the solution combines both a visual display and wearable sensors of hand signs for SLR data collection. The second layer is the organization layer that filters gesture data and has the ability to decode a sign into the appropriate data format while assigning labels. One example of an additional step that can be required is the integration of data from various video frames or the normalizing of samples. The procedure of feature extraction is started by the system as soon as it receives the sign data. Feature extraction and entry categorization are the two most important tasks that all suggested systems need to accomplish in order to determine which sign is the most likely to be present. The visual features, the hand movement features, the characteristics of the three-dimensional skeleton, the face features, and a great deal of other types of features can all function as primary sources of information. When it comes to the success of the SLR approach, the selection of characteristics that will be used for algorithm training is an essential component. Typically, the data is processed and converted to a vector format before being fed into the model. This is done in order to ensure that the data is accurate. The numerous channels are combined in order to investigate the combined impact that they have on the process of sign identification. In the next layer, the model is trained using an optimization algorithm then evaluated based on specific protocol such the k-folds validation. In the final layer, the model is tested on new data for potential use in real applications if high confidence prediction is generated.

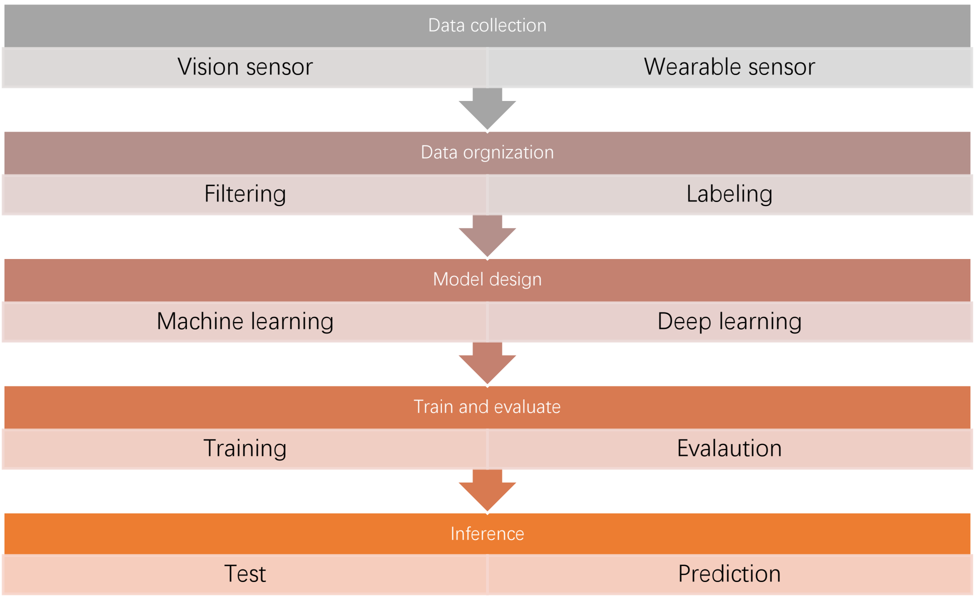

Figure 4. The main paradigm for building an SLR system.

4.1. Data collection

The discipline of interactive computing has experienced considerable growth over the course of the past several years. Since this is the case, it is necessary to develop efficient techniques of human-computer interaction. One method that has the potential to assist in the development of this sector is the recognition of sign language. A receiver can acquire the ability to recognize familiar motions through the use of sign language. Obtaining information regarding the recognition of sign language can be accomplished through the use of hardware-based, vision-based, or hybrid approaches.

Since they don't impose many limitations on users, vision-based methods have recently attracted more attention in the field of sign language recognition systems than sensor-based alternatives. Users' depth and posture estimation data is gathered via vision-based sensors. The topic of depth data and pose estimation is covered later. A few of the more recent SLR investigations depend on visual input. Formats such as depth information and RBG are examples of what is often encountered in this industry, as shown by [17].

According to earlier studies conducted by Rioux-Maldague and Giguere [35], the proliferation of 3D sensors has led to a rise in the utilization of depth data. In their investigation, they utilized a Microsoft Kinect sensor, which captures depth images using a conventional intensity camera and boasts an image resolution of 640 × 480. Also, depth data has been acquired using vision-based methods in recent papers [36–38]. Videos [39–44] or images [45–48] captured with a regular camera or a mobile device can provide depth data. The hand gesture grayscale images utilized by Oyedotun and Khashman [48] have dimensions of 248 × 256 pixels. Zheng et al. [17] states that using depth data helps with human body extraction since it keeps things private and makes it easier.

In addition, depth measurements remain unchanged regardless of changes in lighting, hairstyle, apparel, skin tone, or backdrop [17]. Pose estimation has been utilized to support vision-based techniques in addition to depth data. To classify various hand positions, Rioux-Maldague and Giguere [35] employed depth information in conjunction with regular intensity images. Using OpenNI+NITE framework functions that are publicly available, they were able to track the hands. A 3D hand pose was inferred using inverse kinematics and depth channels, and computationally expensive heat maps for 2D joint positions were generated during pose estimation.

The most recent development in hand form identification was elaborated upon by Koller et al. [37], who detailed how the joints' locations and angles dictate the shape of a hand stance. Since these joint locations and angles can be approximated using depth pictures and pixel-wise hand segmentation, they are now utilized in many experiments. A hand pose estimation system coupled with a classifier trained to recognize hand gestures is used in other research, for example, by Zimmermann and Brox [49]. The limited performance of standard cameras limits vision-based approaches, despite the fact that they do not involve invasive procedures. Another issue is that complex hand characteristics need more processing time to implement, while simple ones can lead to ambiguities [50].

When it comes to sign language recognition, hardware-based methods aim to sidestep computer vision issues. For instance, these difficulties can arise when trying to identify indications in a movie. Methods that rely on physical components often make use of gadgets or wearable sensors. A glove-based technique or user-attached sensors are common in wearable devices used for sign language detection. The sensors, gloves, or rings can decode sign language and render it audible or textual. With regard to wearable sensors and technologies, many works were proposed [50–52] for capturing depth and intensity images using data collected from a SOFTKINECT and a Microsoft Kinect sensor. Direct measuring techniques involving sensors attached to the body or the hands, as well as motion capture devices, are part of a related class of observations [53].

As pointed out by Huang et al. [54], sensor-based techniques are inherently unnatural due to the necessity of wearing cumbersome devices. Real-Sense, a new method they suggest, can detect and follow hand locations in a more organic way. A resurgence of enthusiasm for research into human action and gesture detection has occurred in recent years, spurred on by the enormous success of device-based systems.

The Kinect is the most widely used device-based technique, outpacing both Google Tango and the leap motion controller (LMC) [13,50–55]. Leap motion is a top-notch device that employs computer vision to accomplish a practical interactive function, according to Wang et al. [56]. Learning and practicing sign language is not prevalent in society, which further emphasizes the significance of LMC (as described in [57]).

Alternate approaches use input from specialized gloves, as in [58–60], and also make use of a variety of technical tools, including accelerometers [61] and depth recording devices [62]. One of the simplest sensor setups that allows for cheap and easy motion tracking is the coloration of the fingers on gloves, as seen in [63,64].

According to [65], digitally capture-capable gloves were used to deduce Arabic sign language variant hand signals using a smaller number of sensors. Although there is a significant investment required to design and operate such specialized machinery, the final cost is far lower than that of competing high-tech products. By utilizing a motion controller as their principal input device, the authors of [66] were able to achieve extremely precise three-dimensional object tracking at a rate of 120 frames per second. Their controller was specifically designed to capture hand movements, allowing the researchers to keep track of multiple important hand positions from frame to frame. With the same instrument, [67] achieved pinpoint accuracy in differentiating fifty distinct isolated hand signs.

When gathering data on sign language recognition, hybrid methods have been employed. When it comes to proportional automatic voice or handwriting recognition, hybrid approaches perform as well as, if not better than, other methods. Combining vision-based cameras with other kinds of sensors, including infrared depth sensors, allows hybrid techniques to obtain multi-mode information about the hand forms [68]. Calibration between hardware and vision-based modalities is a key component of this technique, and it can be somewhat tough. The method's speed and suitability for investigating the effects of deep learning approaches are both enhanced by the fact that it does not necessitate retraining.

In order to test how this data directly affects a CNN, Koller et al. [69] used the cleaner hybrid approach, often known as Automated Speech Recognition (ASR). While high-resolution still images or continuous RGB recordings have their uses, depth imaging is superior for estimating distances from a given point. A few algorithms combine the two forms of visual data [46]. Although it is utilized less frequently than the preceding two types, thermal imaging is still an intriguing potential.

Near-infrared (IR) heat sensors can also use radio wave emission and light reflection to create an image. While this data has shown promise in other areas, such as face identification and body contouring, it has not yet made it into stereo vision research [70]. The position of the joints in the hands as they make SLR movements is one example of how skeletal data can be used as input. Motion capture also provides some feedback in the form of monitored information changes between images. In these types of models, the optical sequence is typically defined as a vector that describes the movement of pixels in a series of images. On the other hand, in video materials, the so-called scene sequence can be tracked, which refers to the motion of three-dimensional objects within the scene, relative to the distance from the camera lens [71]. All of the input devices have the potential to be useful in certain situations, but how well they work depends heavily on those circumstances. Deep sensors and RealSense/Kinect recording systems are examples of more sophisticated input sources.

4.2. Data organization

Research for sign language recognition relies heavily on data organization. It may include tasks such as representing signs, filtering and normalizing data, organizing and displaying data, and labelling.

As a visual language, sign language allows people to communicate through the use of both manual and non-manual sign representations that are grammatically structured. Hand form, palm orientation, finger and hand movement and placement, head tilt, mouthing, and other facial expressions are all examples of what might be represented. Eight time-ordered representative frames were utilized by Tang et al. [52]. The two hands, depicted by them, moved closer together before gradually pulling apart.

In an experiment described in [36], the signer's hand was utilized to represent all gestures. To further illustrate the form of the hand sign, a hand segmentation phase was also employed. Just as Koller et al. [37] used a double state to represent 60 different hand shape classes, a single state was used to represent the rubbish class. Zhou et al. [72] conducted an additional study that solely included signers with their right hand. Here, the dominant hand was the right hand and the submissive hand was the left.

The Bengali Sign Language was the concentration of Hossen et al. [73]. The language has 51 letters and was represented in the experiment using 38 signs. These signs were created by merging related sound alphabets into one sign. As mentioned in [69], the Bahasa Indonesian language uses a maximum of five marks to represent a single word. This means that there is a single, consistently performed sign for each word and prefix in signed Indonesian (SIBI).

Huang et al. [54] conducted an additional experiment using 26 indications represented by 26 output units and 66 input units. Attempts to compare hand and body characteristics have also been made in previous investigations. According to research in [15], when it comes to sign language identification, body features are somewhat more accurate representations than hand aspects. Essentially, a 2.27% improvement in sign language identification was achieved by utilizing body features [44]. Joints in the torso are more reliable and stronger than joints in the hands, which explains these findings.

Normalization is the process of normalizing input according to a set of criteria in machine learning and deep learning. The goal is to make the AI tool work better. Data pre-processing is when this technique, which may involve media processing chores or statistical processes, is carried out. Considerations such as input format (e.g., text, image, or video), sample variability, machine learning architecture type, automation tool purpose, etc. determine the specific normalization technique that is best used.

Modern methods for sign language recognition often use normalization because of the positive effect it has on performance, and its inclusion has been supported by empirical evidence [64,74]. Given the diverse range of input modalities and purposes used in SLR investigations, it is not surprising that the discipline employs a wide range of normalization procedures. Changing images to fit them into a standard format that the algorithm can easily understand is a common practice in most of the visual approaches.

Due to the pixel-level encoding of information in machine learning models during feature extraction and network training, this is a common method for accomplishing this. Kratimenos et al. [75] and other studies [64,76] demonstrate some of the basic instances of normalizing methods utilized in SLR, such as image scaling and re-shaping. In order to make the feature map dimensions fit, Garurel et al. [77] additionally use the training mean values and standard deviations to determine the best size for each frame. Another common technique, cropping can remove potential causes of algorithmic misunderstanding, improving the quality of visual input and leading to more reliable sign recognition.

To facilitate sign language communication, input images are usually cropped to exclude everything but the parts showing the hands and face. Cropped photos are normalized in [78] using the average neck length, which eliminates the effect of camera distance for all photographs. According to [79], the positioning of important joints allows for the selection of a benchmark signer and the standardization of input from other signers. Also, contour extraction is employed for this purpose; for instance, in [80], the hand-related regions are extracted while the backdrop is eliminated.

To standardize the quality of different clips and decrease computing demands, frame down sampling is commonly employed by SLR systems that mainly use video as raw input. The procedures of normalizing and filtration were utilized in [35]. All of the image's pixels were adjusted to lie on the [0, 1] interval, and the intensity histogram was levelled out. The produced images were subsequently subjected to four distinct orientation and scale Gabor filters. The main hand outlines were attempted to be obtained by applying bar filters to the depth and intensity images.

In their experiment, Li et al. [68] also employed gabor filters to extract classifiable hand features. Prior to applying Gabor filters, the images were scaled down to 96 × 96 pixels. Another study used component analysis (PCA) filter convolutions trained on input images in their experiment [36]. Koller et al. [37] utilized pre-trained convolutional filters in the CNN model's lower layers and performed a per-pixel normalization on images as part of the preprocessing. In their experiment, Zhou et al. [72] refrained from doing any normalization technique since the retrieved features naturally fell within the interval of $[-1, 1]$. Another experiment by Yang and Zhu [39] set a threshold to filter the minor skin-color area, an approach that enhanced the robustness of the system by using the second layer of their CNN model as a filter.

The experiments conducted by [41,44,45] also show instances of normalization. Balayn et al. [41] normalized Japanese sign language (JSL) motion sentences and used them as inputs and outputs for Seq2Seq models. The classifier was fed normalized hand positions and cropped hand areas, as described by Konstantinidis et al. [44]. In their attempt to examine Chinese Sign Language, [45] obtained a total of 1,260 images of basic signs in Chinese, which were normalized to 256 × 256 optimized background samples. Their model used 16 filters in the first convolutional layer. The filters had a width and height of 7 and a channel width of 3. Similarly, Koller et al. [69] applied a global mean normalization process to images before finetuning their CNN model.

Experiments to format and organize data in various ways have been reported. Tang et al. [52] organized the hidden layers of their models using various planes within which all units shared similar weights. In another experiment by Jiang and Zhang [45], the data were divided into training and test sets, with the training set containing 80% of the total images and the test set containing the remaining 20%. In a different experiment that used a Kinect sign language dataset, Huang et al. [51] formatted and organized their data into 25 vocabularies that were extensively used in daily life. Each word was played by nine signers, and each signer repeated each word three times. Using this approach, each word was organized into 27 samples, yielding a total of 25 × 27 samples.

Eighteen samples were selected for training, and the remaining samples were used for testing. Many studies from this field also include filtering and data augmentation steps, which have the purpose of improving the quality of input and consequently boosting the accuracy of the model. Random sampling or discarding of frames is one of the most straightforward techniques found in literature, where approximately 20% of input is eliminated.

In [81], this technique is complemented by random changes of brightness, saturation, and other image parameters. Some of the data augmentation methods used in [82] include Gaussian Noise, Just Counter, and Future Prediction. The PoseLTSM tool also employs some operations aimed at augmenting the input images, with rotation of the hands around fixed points in the wrists as one of the most original ideas. As with normalization, the choice of filtering and data augmentation techniques is directly related to the properties of the model and the type of input, so it must be made with full understanding of each individual implementation and its objectives.

4.3. Model design

Feature extraction is an essential part model design since it determines the training process and, by extension, how fast the models can learn to differentiate between various signs and words. Features in sign language communication are always based on raw data and relate to the locations of various body parts, such as important places on the face and hands. Statistical processes are used to compute features, which are then given weights that are directly proportional to their discriminatory value [82].

The neural model is able to learn the probability of features' association with particular classes by expressing them as vectors in latent space. In some cases, a specialized tool was utilized to extract features from the various feature engineering techniques that are detailed. The impact on accuracy and scalability of the model is usually taken into account while optimizing the final number of features and their weight distribution [36,72].

In their experiments on sign language recognition, multiple writers used feature extraction algorithms [50,52]. By establishing the network's architecture as (NX, N2, 1000, 1000, 1000, 1000, NTC), Wu and Shao [38] were able to perform high-level feature extraction, with NX representing the observation domain dimension and N2 the number of hidden nodes. For the purpose of identifying hand positions from depth and intensity pictures, Rioux-Maldague and Giguere [35] introduced a new feature extraction approach. In order to downsize the images from 128 × 64 to 64 × 64, they were de-interlaced by keeping every other line in the image. A 1 × 4096 intensity vector was extracted from every 64 × 64 image that was generated.

The recognition procedure was significantly improved by Tang et al. [52] when they retrieved hand traits by taking the two hands into account collectively. To overcome the difficulties of processing many visual modalities, a related experiment in [40] employed PCANet for feature extraction. Li et al. [43] demonstrated feature extraction in action by transforming data from two-handed sensors into vectors of useful information. Doing away with the need to recreate the hand's exact form, orientation, and location is the goal of this method.

The spatial feature extraction performed by Camgoz et al. [38] also made use of 2D CNNs. The feature maps were created by convolving images with weights in the 2D convolution layers. Furthermore, findings from [21] further proved that spatial-temporal information may be extracted using many layers of convolution and subsampling. To train a Gaussian mixture model-hidden Markov model (GMM-HMM), Huang et al. [51] employed these principles to extract characteristics from a movie that included sign language.

In a different study, features like finger length, finger width, and finger angle were fed directly into the DNN, in contrast to Huang et al. [51], who manually supervised the feature extraction procedure. Due to their capacity to address spatial and temporal correlations, 3D-CNNs have been utilized in several experiments instead of 2D CNNs. For example, in order to create a representation of every video clip that was taken into consideration, the authors in [11] utilized a ResNet model that was based on a 3D CNN model.

In a related area, the authors of [83] created a neural network for feature extraction using a multi-layer architecture. Several input features were extracted using a convolution layer in [45]. As a feature extractor for an SVM, the authors in [46] utilized a trained CNN. Konstantinidis et al. [34] conducted an additional study that used video sequences to extract skeletal elements in addition to video content. Skeletal features included the body, hands, and face, whereas video features included the image and optical flow.

For video feature extraction, the VGG-16 network that had been pre-trained on ImageNet was utilized. Features were extracted using a combination of the ImageNet VCG-19 network and conv44 in an analogous study conducted by Konstantinidis et al. [44]. Among the most important characteristics retrieved from the experiment were the 18 2D body joints and the 21 2D hand joints. Humanoid feature extraction and recognition were carried out by Rao and Kishore [42]. Human interpreters rely on these characteristics to reliably remember signs. Some trials have attempted to streamline or do away with feature extraction altogether. To simplify their feature extraction procedures, Yang and Zhu [39] employed a CNN. As a result, the sign language recognition system may receive images directly.

Building a model for sign language recognition using machine learning require the feature selection step. The basic idea is to simplify the data such that just a few important statistical parameters remain, and then feed those into the machine learning network [84]. The goal is to reduce the amount of calculations needed to get an accurate forecast by include just the features that drastically improve the algorithm's ability to recognize different classes.

Since different models use different algorithms, raw data structures and volumes, and the primary tasks anticipated of the machine learning classifier might cause the exact number of picked features to vary from model to model [85]. Researchers rate features according to their significance and pick those that are worthy of inclusion using a variety of approaches. One may classify feature selection methods as either supervised or unsupervised [84]. Some of the features' inherent characteristics are captured by filter methods (e.g., variance threshold, correlation coefficient, or Chi-square test) and evaluated by wrapper methods (e.g., forward feature selection or backward feature elimination) in order to determine their relative importance in a given algorithm [85].

Embedded methods incorporate LASSO regularization or random forest importance, while hybrid approaches combine the best features of both the filter and wrapper approaches. Given the variety of feature selection schemes available, researchers should apply the one that works best with their particular classifier, important tasks, and data [86]. Findings from the experiments reported in [52,81] are examples of feature selection experiments.

There was less need for human feature selection in [39] because a DNN was employed. Feature extraction and autonomous detection are both accomplished by the DNN. Using 215 separate test words to stand in for typical sign language conversations, another example of the feature selection method was given in [72]. Among the 18 features retrieved from the joints of the human body, Konstantinidis et al. [44] chose to focus on just 12 for their experiment. The candidates were chosen because, in most sign language datasets, the signers are seated and their leg skeletal joints are not apparent.

Not only did some trials employ CNN, but PCA was also utilized to help with feature selection in others. The fact that principal PCA is a tried-and-true method for reducing the number of dimensions in a space might inform its application to the processing of image data, which often contains information about spaces with many dimensions. One example is the use of PCA for feature selection and dimensional reduction in [68]. DNN, also known as feature learning, were demonstrated in a separate experiment by Huang et al. [54] to generate and choose features. To put it simply, a DNN can automatically evaluate and produce features from unprocessed input.

Building a coherent model for SLR from the phonetic to the semantic levels is the primary goal of the model design step. From the utilization of the signing space to the synchronization of both manual and non-manual elements like eye gazing and facial emotions, the modelling process encompasses a wide range of techniques. Contrarily, natural language processing, pattern recognition, computer vision, and linguistics are all involved in SLR [87]. The goal of SLR is to create various algorithms and methods that can identify preexisting signals and understand their meaning. Models for classic, deep learning, SLR continuous, and SLR isolated sign language processing are covered in this section.

4.3.1. Machine learning

The field studying how computers can learn to do tasks automatically, without human intervention, is called machine learning. Along with the required data, machine learning algorithms are often given broad guidelines that describe the model. Typically, the data contains instructions for how to execute the specified job by the model. Machine learning algorithms are able to accomplish their goals when they modify the model using the data that is linked to it. Numerous machine learning algorithms are available, some of which are SVM, PCA, and HMM.

To solve classification problems with two groups, supervised machine learning models like SVM is used. When you feed an SVM model with labelled training data, it may classify the new instances into groups. This is just one of many uses for SVM in previous investigations. To learn the retrieved data, Nguyen and Do [46] used multiclass SVM. While the combination of histogram of oriented gradient (HOG) with local binary patterns (LBP) and SVM model had higher validation accuracy, the CNN-SVM model had lower accuracy.

On the other hand, the CNN-SVM model was more likely to prevent overfitting. In order to compare the most popular classifiers, which use a combination of softmax and linear SVM, the demand for real-time performance was assessed in [68]. When compared to other sophisticated classifiers, SVM and softmax achieved superior accuracy. It was also noted that an SVM classifier with a linear kernel outperformed the softmax-based classifier, but it took more time to train.

Similarly, using the same dataset, an experiment by [54] sought to compare the performance of DNN and SVM. The results showed that compared to SVM, DNN achieved a higher recognition rate. As an example, SVM was chosen by the authors of [88] as an appropriate classifier for real-time SLR. SVM and DNN were employed by Chong and Lee [57] in their investigation of American Sign Language. According to the results, when using SVM, the rate of sign language recognition for 26 letters was 80.30%, while using DNN, it was 93.8%. Additionally, it was noted that the recognition rates for a grouping of 26 letters and 10 numbers were slightly lower for SVM (72.79%) and DNN (88.79%). When it came to sign language recognition, the DNN outperformed the SVM.

A large-vocabulary SLR method was also used by Huang et al. [89] with SVM. In order to represent video features as a fixed-dimensional vector, the experiment's SVM approach made it easier to do mean pooling across clipped data. Using SVM for video feature-based categorization was proposed by Huang et al. [89]. It was pointed out that their machine learning method ignores time-related data while mean-pooling, even though SVM are used.

In addition, [90] assessed how well the SVM performed in a hybrid setup. The experiment tested how well a HOG+SVM system could classify data. An SVM classifier was fed canonical hand shapes into the hybrid system, and a HOG feature extractor was used to generate 64-dimensional features. The accuracy improvements achieved by combining HOG and SVM ranged from 14.18% to 18.33% as compared to using SVM.

To extract features or decrease dimensionality, PCA is employed in computer vision. A number of recent studies have employed PCA to reduce the number of dimensions in sign language recognition. An orthogonal linear transformation is the easiest way to explain PCA, which changes the original data’s coordinate system to one with less dimensions. There was a proposal for a PCA-based fingerspelling recognition system in [36].

Using PCA, Koller et al. [37] were able to decrease the dimensionality from 1024 to 200 using feature maps. In another study, PCA was employed to identify data streams with around 492 dimensions that showed a lot of variation [41]. One other way that PCA has helped cut down on overfitting is by using it on Kinect data. Another experiment employed principal component analysis (PCA) to convert a matrix to a vector with 210 dimensions [56].

An improved technique for the mel frequency cepstral coefficient (MFCC), which is helpful for sign language recognition, can be created with the help of these dimensional vectors. There was a comparison of the suggested approach to others, including SAE+PCA, in [91]. Based on the results of the comparison, SAE+PCA outperformed the proposed technique and attained an accuracy rate of

99.05%. A variant of PCA called recursive principal component analysis (RPCA) has also garnered attention in other trials for feature extraction. Using RPCA, the authors of [92] were able to attain a 98% classification rate when investigating the characteristics of SLR systems.

Utilizing statistical processes, HMM is able to discern patterns arising from the intricate interplay of motions within a space-time continuum. While [93] was the first to utilize it in the context of SLR in 1996, [94] achieved good performance with the best settings in 1997 when using it to categorize individual hand motions from visual input. In an effort to expand upon the underlying model's promising performance, variations like factorial HMM [95] or dual HMM [96] were proposed about the same time. According to those researches, the model needs a large amount of training data to produce reliable statistical predictions.

Shortly thereafter, Wilson and Bobick [97] suggested enhancing this method based on parameters, while authors in [98] suggested including parallel computing into this paradigm. In order to address issues related to language, the same idea was expanded upon by [99]. By training the model with 80% of the sample and testing it with 20%, this method proved to be more cost-efficient than any of the previous HMM implementations. It achieved an accuracy of more than 94% for static signs and more than 84% for dynamic signs in continuous speech.

A different subset of these models, input/output HMM, was initially proposed by [100] for use with less homogeneous data. Using the same idea, it is possible to successfully track hand locations during sign language communication; for example, as shown in [101], the output accuracy was over 70% when 16 different signs were distinguished solely by hand movement. In 2009, another paper improved upon the input/output HMM model [102]. The authors established a cut-off point and increased the accuracy to above 90%, but only for cases with fewer than 20 signs to be detected.

After failing to noticeably enhance SLR performance over earlier versions, [103] offered an alternative in 2003, naming their approach Left & Right HMM. Even with limited data, a hybrid of HMM and Gaussian mixture model (GMM) models can improve hand sign recognition as demonstrated in [104], albeit at the cost of reduced system reliability. Data gathered from a number of video cameras was also analysed using HMM by [105]. Although those approaches have their uses, further research is needed to apply them to SLR.

Some academics have attempted to improve their results by combining HMM with other approaches in recent years. An effort in this direction was made in 2011 by [106], who used this method in conjunction with PCA to extract important characteristics from hand signals. Meanwhile, in order to follow the contours of hands during sign language communication, the authors of [107] integrated HMM into an RNN model; nevertheless, they only achieved success when dealing with a small set of known signs.

While Yang et al. [108] did their best to reduce calculation time by creating a variant of HMM, there are some requirements that must be satisfied for this method to work, such as a maximum length for each gesture. Training samples with limited distribution were processed using a combination of the CRF approach and HMM in the study by Belgacem et al. [109]. However, even with a large number of alternatives, the discrimination process remains challenging. HMM are a common solution to the terrestrial alignment problems that plague many continuous processing workloads. Incorporating an EM-based approach into HMMs helped with weak supervision and video processing issues in [37].

In order to enable continuous sign language recognition, Zhou et al. [72] utilized HMM techniques to create a model framework. Thanks to HMM, the final system can handle a bigger vocabulary, model individual signs and their transitions, and train and decode using even the most cutting-edge techniques. The authors of [52] conducted an additional experiment that looked at the GMM-HMM as a starting point. To train the GMM-HMM for recognition, characteristics such as trajectory and hand-shape were retrieved. When utilizing both trajectory and hand-shape information, an average accuracy rate of 90.8% was attained.

4.3.2. Deep learning

Learning representations of data is the primary goal of deep learning, a relatively new area of machine learning [50]. Nevertheless, the intricacy of the models and the input details to the system limit deep learning techniques' capacity to extract data semantics [34,50]. Neural network-based SLR seems to benefit greatly from recent developments in deep learning. In recent experiments, various deep learning techniques have been utilized, such as convolutional neural networks, recurrent neural networks, attention-based approaches, deep belief networks, and autoencoders architecture.

In order to distinguish between images, Convolutional Neural Networks take in input images, amplify certain parts of those images, and then output the results. For the purpose of sign language recognition, Figure 5 depicts the fundamental CNN construction mode. When compared to other deep learning algorithms, CNNs need significantly less pre-processing [37]. There are a lot of tasks where neural networks excel [39], but they need a lot of labelled data to train on [41,45]. There is an additional burden to gather training data for hand shape identification because of the extremely high rate of intra-class ambiguity in this procedure, which is affected by the subject's position.

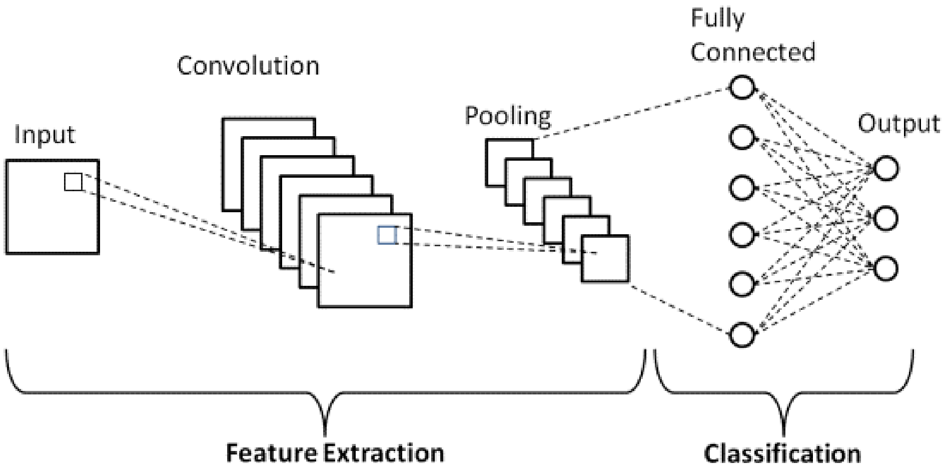

Figure 5. CNN architecture.

Among the many models that make sequential data modeling easier, RNN stands out. Voice recognition, video recognition, language translation, and natural language processing are just a few of the many critical activities that have benefited greatly from this style of approach. To understand how RNN Encoder-Decoders work for sign language recognition, Figure 6 illustrate the basic idea.

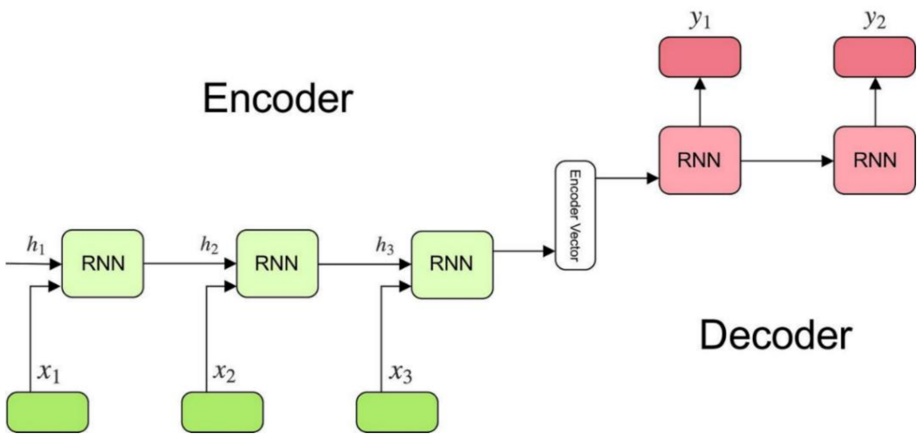

Figure 6. RNN encoder-decoder architecture.

A bidirectional RNN and long short-term memory (LSTM) were employed by Fang et al. [110] to enable universal and non-intrusive sign language word and sentence translation in their experiment. Results from the experiment showed that RNN model could accurately represent American sign language words with all the necessary characteristics.

Long Short-Term Memory (LSTM) is an RNN feature that has been used in a few experiments. In their study, Kavarthapu and Mitra [111] utilized a bidirectional LSTM for encoding and a second LSTM at the embedding layer for decoding. Bidirectional LSTM is a game-changer for sign language recognition since it enables abstract data collecting. It was clear from the outcomes that the bidirectional LSTM worked quite effectively.

Rakun et al. [112] used LSTM for the recognition of Indonesian Sign Language. In this experiment, LSTM was utilized due to its independence from pre-clustered per-frame data and its ability to accept entire sequences as input. According to the results of the experiment, the 2-layer LSTM model outperformed all of the other models and correctly classified root words with 95.4% of accuracy. Inflectional words presented a considerably greater challenge for the LSTM model, which resulted in a significantly lower accuracy of 77% when trained on these words.

In [113], an model made of LSTM cells was utilized for SLR system. Every time step in the design took the feature vector from all the frames as input. In the output layer, a softmax classifier was used. Real-time sign language translation was ensured with the use of LSTM. As a result, the model was able to convert long-form sign language films into full-text English sentences, which greatly improved sign language communication.

A small number of recent studies have also employed LSTM to identify motions in Indonesian sign language. The researchers in [114] employed 2-layer LSTM neural networks to recognize SIBI gestures. With accuracy rates of 91.74% for prefix, 98.94% for root, and 97.71% for suffix datasets, the neural network demonstrated very high performance [115].

Increasing use of hierarchical deep recurrent fusion (HRF) networks has resulted from efforts to tackle the difficulties of sign language translation. Visual semantics can be encrypted with several levels of visual granularity using a hierarchical recurrent architecture that was created by Guo et al. [54]. To decipher a text, the HRF employs skeletal signemes and complimentary RGB visemes. In order to encode the complete visual material, the HRF translated the video into multiple neural languages.

Up next, Guo et al. investigated sign language translation action patterns using adaptive clip summarization (ACS). Instead than using a set interval to acquire key frames or clips, as previous models have done, they suggested an adaptive temporal segmentation technique. To further reduce the duration, a hierarchical adaptive temporal encoding network was then created. Aside from HRF, LSTM was chosen as the fundamental RNN component. Long short-term memory (LSTM) learned the original features' persistent qualities in its top layer. Condensed visemes or signemes' recurring characteristics were taught by the medium layer. Into textual semantics the visual data was transformed by the bottom layer. Earlier, we established that learning the descriptors of sub-visual words like visemes and signemes was the central premise of the proposed approach. Thorough trials demonstrated the exceptional efficacy of the HRF framework, which is based on LSTM.

The classification of learning representations in several sign language experiments has been accomplished using a deep belief network (DBN). Double-layer perceptron DBN are similar to multilayer perceptron (MLPs), but DBNs have a lot more layers. Although DBNs' additional layers are notoriously tough to train, they greatly improve the network’s learning capability. Nevertheless, DBN training has been made easier by recent efforts.

An example of a DBN is the one employed by Rioux-Maldague et al. [35], which consists of three restricted Boltzmann machine (RBM) and one additional translation layer. By utilizing DBNs, Tang et al. [52] were able to accomplish hand posture identification. A higher recognition accuracy of 98.12% was achieved by the DBN compared to the baseline HOG+SVM technique, according to the recognition findings.

A deep belief network's architecture and performance in gesture identification were investigated using an American Sign Language dataset. In the experiment, DBN was tested against two other typical methods for gesture recognition—a convolutional neural network and a stacking denoise auto encoder—and the results showed that the proposed DBN performed significantly better.

There are many situations when using just one deep learning technique can be difficult. Consequently, there have been experiments that incorporated deep learning techniques. For example, it was pointed out in [52] that training DBNs was not an easy task to parallelize across many machines. In order to assess this matter, they compared it using CNNs. While the hybrid DBN method produced a higher recognition accuracy rate (95.17%), CNN still managed to pull ahead with decent results.

A hybrid deep architecture was suggested by Wang et al. [40] to tackle the continuous sign language translation challenge. The hybrid model included a fusion layer (FL), a bidirectional gated recurrent unit (BGRU) module, and a temporal convolution module. Here in the model, temporal convolution is in charge of collecting the quick changes in time, while BGRU holds on to the big changes in context that happen across several time dimensions. To discover the correlations between the corresponding features in the temporal convolution and BGRU outputs, the FL next fuses them. Results from experiments showed that as compared to using just deep learning approaches, this hybrid architecture enhanced accuracy by 6.1% in terms of Word Error Rate (WER).

A CNN and a bidirectional recurrent neural network (Bi-RNN) have both been employed in tandem. By combining both methods, the authors of [43] were able to derive features from each video frame using a 3D convolutional neural network (CNN), and they were able to generate unique features from the sequential behavior in each frame using a bi-RNN. While the Lipnet model had a lower average character error rate, the hybrid approach had a greater average word error rate.

Cui et al. [3] used a deep CNN and a Bi-LSTM together to get features. By feeding video streams into the CNN model, spatiotemporal representations could be learned. The next step was to train Bi-LSTMs to understand more nuanced interactions. Repetition in LSTM computations is achieved by

Bi-LSTMs through the use of forward and backward hidden sequence calculations. Unidirectional RNNs have a limitation that the authors took advantage of by using Bi-LSTMs. they can only determine hidden steps by looking at previous time steps.

Researchers in [89] used attention-based 3D-CNNs to improve sign language vocabulary recognition for huge datasets. In particular, the attention-based approach offers two benefits. Before anything else, the model can pick up spatio-temporal features from unprocessed video data without any prior training. Second, we can pick up hints with the help of attention mechanisms. Here, utilizing continuous sign language data and the ChaLearn14 benchmark [116], attention-based 3D-CNNs were evaluated. Compared to more sophisticated algorithms, the results showed that the method was more accurate.

A study that employed transfer learning to train a CNN model to recognize Indian sign language was proposed [117]. With the use of transfer learning, new classes could be learned even when training sets were small. Using a combination of deep learning-based networks, Oyedotun and Khashman [48] identified hand motions taken from a public database. Stack denoising autoencoder (SDAE) and convolutional neural network (CNN) methods were utilized. Contrasted with SDAE's 92.83% identification rate on test data that wasn't part of the training set, CNN got 91.33%.

Using a combination of CNN and RNN, Bantupalli and Xie [118] conducted an additional experiment that investigated American Sign Language. When it came time to recognize sign language in a video stream, the Inception CNN model was brought into play. Its job was to extract spatial information from the feed. The experiment then proceeded to extract temporal features from video sequences using an RNN model and LSTM. The softmax and pooling layers of the CNN were utilized to generate the outputs. Despite the experiment's success, the scientists speculated that capsule networks, not Inception, would have performed better when it came to sign language recognition.

Muslims who are deaf or hard of hearing face significant barriers that prevent them from obtaining higher degrees. Because of this, they have a much more difficult time studying the Holy Qur'an than the average person, and they have a much harder time understanding its meanings and interpretations. As a result, they are unable to practice Islamic rituals like prayer, which need knowledge of the Holy Qur'an. A novel model [119] for Qur'anic sign language recognition that is based on CNN was proposed. Aiming to assist hearing impaired persons in learning Islamic rituals, the proposed model is designed to recognize the hand gestures that relate to the dashed Qur'anic letters in Arabic sign language.

With the use of deep recurrent neural networks, hand feature representation, and hand semantic segmentation, a new framework is suggested for SLR employing deep learning [120]. A newly-developed semantic segmentation algorithm called DeepLabv3+ [121] is trained to extract hand regions from each frame of the input video using a set of pixel-labeled hand images. After that, in order to rectify the hand scale differences, the extracted hand regions are cropped and scaled to a set size. An alternative to employing pretrained deep convolutional neural networks for feature extraction is a single-layer Convolutional Self-Organizing Map (CSOM). Later on, a deep Bi-directional Long Short-Term Memory (BiLSTM) recurrent neural network is employed for feature vector sequence recognition. The three layers that make up a BiLSTM network are the fully connected and softmax layers. We test the suggested strategy on a difficult Arabic sign language database with 23 individual terms uploaded by three individuals.

Although the examined literature covers a wide variety of sign language recognition methodologies, nearly all of them adhere to a small set of core principles. Specifically, attention-based neural models with transformer architecture are the main focus of the studies [122]. As shown in Figure 7, this computing paradigm trains the model to classify sign language data using encoder and decoder stacks. Not only has this method outperformed previous models, but it has also been successful with many kinds of tasks. The goal here is for the models to pick up on the connection between spatial and temporal signals and use that information to infer the desired sign.

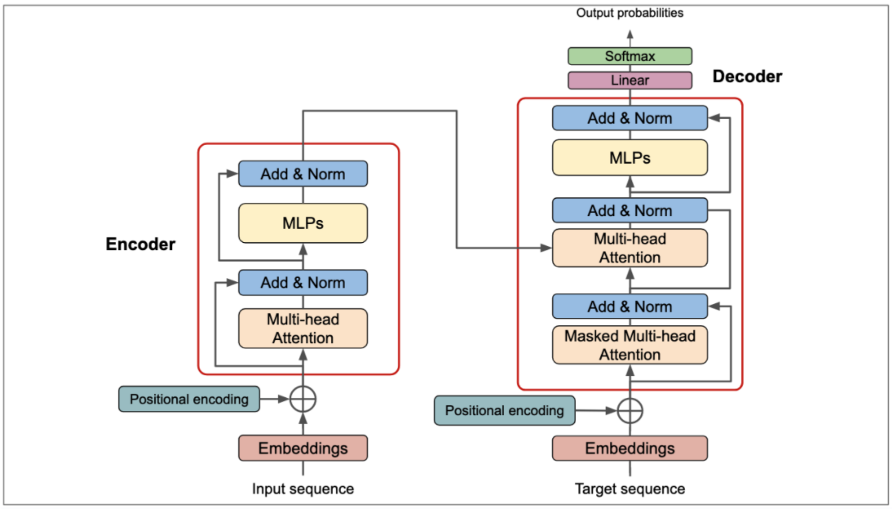

Figure 7. Transformer main architecture.

Tokenization is a process that takes input and output and uses them to create frames, key points, and word embeddings [123]. Adding a temporal ordering step is necessary since transformer models do not provide positional information for the sequences that are being inspected, which is one of their distinctive constraints. This leads us to the next essential component of transformer-based neural models for feature extraction. This process involves selecting the most important features from the input tokens and then using them to train the model [77,123]. There are characteristics that distinguish one gloss from another (intra-cue features) and those that distinguish between signals (inter-cue features) [81,124].

A hybrid approach significantly improves efficiency by using a separate CNN-type neural network to extract information from video input. The categorization step is usually handled by an encoder-decoder stack, which consists of multiple successive layers, or a Bi-directional Long Short Term Memory (Bi-LSTM) module. The number of deployed attention heads and the exact depth of the model can be fine-tuned for optimal performance based on empirical evaluations; these parameters vary with the model's intended usage and other circumstances.

For instance, at the top of the stack, some research suggested a linear projection layer and a softmax attention layer, while others suggest employing just two layers in transformer models, instead of the usual six used for NLP [75,78]. The model is fine-tuned for the particular goal through a validation approach, and its efficiency during training is enhanced by a normalization procedure, which is motivated by maximizing conditional probabilities and minimizing cross-entropy loss. This type of network has been evaluated in several contexts, such as sign language translation, isolated [125] and continuous SLR [126], and other similar tasks. While this methodology was tested with video footage and skeletal data as input modalities, it may theoretically be applied with other modalities as well [80].

The deep learning method's transformer architecture's adaptability is well appreciated in this demanding domain, as it allows for the output to be tailored by choosing the training dataset, features, and training hyperparameters. Several intriguing suggestions, such as gloss-level supervision or the usage of specific posture estimation methods, were made in the examined literature that could further improve encoder models' sign language understanding capabilities. Those enhancements have the potential to finally put an end to some of the persistent problems in the SLR sector [127].

The proposed deep learning model is experimentally evaluated in all of these works, and the results are usually compared to those of other SLR methods. In most cases, techniques that rely on transformer design achieve much better results than simple sequence to sequence models and other standards. When it comes to tasks like posture estimation, the best version of the system can usually get predictions right up to 85% of the time. For isolated SLR, it's around 70%–75%, and for the more challenging translation task, it can get it right up to 45%. The advantages over alternative approaches were negligible in some circumstances and substantial in others.

The objective isn't the only variable that could impact output quality; other variables include the amount of the vocabulary, the size of the training dataset, the precise setup of the network, etc. [128]. The results of those tests are certainly helpful, but it is still difficult to say for sure what configuration would provide the best results independent of variables like signer identity, regional sign language variations, and environmental factors. So far, data suggests that transformer-type deep neural networks play a role in this area of study; however, it is unclear what that role should be and how to use it to broaden the scope of potential SLR applications [74].

Although there are noticeable advancements in accuracy compared to previous deep learning SLR systems, the methods based on transformer architecture are still far from being suitable for use in daily practice. Accuracy tends to improve with increasing complexity of studied sign language samples, and it becomes more noticeable with increasingly complicated assignments [126]. Additional input modalities and localized sign language variations, as well as more thorough testing, are necessary to conclude whether performance gaps are caused by training samples and selected features or by the fundamental data processing approach [81].

Given these findings, the development of universal autonomous tools that can perform signer-and language-independent continuous SLR is still in its early stages. The results of the evaluation of the encoder models indicate that SLR may benefit from a slightly different architecture than linguistic tasks, so it would be fascinating to witness creative efforts to rethink transformer models and create them specifically for sign language interpretation [129].

This review highlights the most crucial aspects while referencing the most significant studies due to the extensive scientific literature on the topic and the significance of hand gestures for SLR. Over the course of several decades, scientists have studied gesture interpretation, leading to a deluge of reviews covering the topic at different points in time. Gavrila [130] conducted one of the first reviews, looking at various 2D and 3D models for human motion analysis. While Ribeiro and Gonzaga [131] mostly concentrated on real-time methodologies available at the time, Moeslund and Granum [132] offered a thorough summary of twenty years of research including gesture tracking and recognition. Rautaray and Agrawal [133] revised assessment of possibilities and obstacles in this area is one example of a more recent article. While Mohandes et al. [134] investigated sensor-based and direct measurement approaches to sign language identification, Kumar and Bhatia [135] covered a range of feature extraction methods.

We present a concise synopsis of the present status of research in the area of automated hand gesture and sign language identification due to the fact that this area has seen substantial development and numerous evaluations throughout the previous twenty years. Accurate hand form recognition is a highly useful characteristic for automated systems since most sign language characters and words can be conveyed with simple hand gestures. But there are a lot of obstacles to overcome when trying to recognize hand motions, and some of those obstacles may be associated with the fact that signers' hands are different sizes and shapes and have varying skin tones. Furthermore, different people may sign with different styles that highlight different aspects.

The application of sophisticated analytical methods that seek to detect patterns apart from the signer's identity or the physical characteristics of their hands can overcome such challenges [136]. An efficient method for analysing hand motions in SLR is to employ deep learning networks, which can detect latent relationships among numerous variables. Depending on the regional sign languages, certain words or phrases can be expressed using either one-handed or two-handed motions.

Typically, one-handed signals are assigned fundamental meanings like letters or numbers. Therefore, it is possible to accurately identify simple linguistic content from various sources, including still photos or movies, using only hand motion analysis. Some uses may benefit from combining hand gesture analysis with additional methods, such as monitoring head motions [137].

Despite growing interest in full-body tracking and continuous sign language interpretation, this facet of SLR is expected to remain relevant because hand motion is the foundation of all sign language communication systems. The most effective use of pure hand gesture analysis approaches, however, would likely need a combination of methods. As an example, there are a growing number of hybrid models that take into account various aspects of a signer's behavior [137,138].

In the field of sign language recognition, posture estimation algorithms are fundamental tools due to the significant significance that body form plays. Finding the precise position of the whole body from the measurements of a few fixed spots is the main concept. Deep learning algorithms, when trained adequately with carefully selected examples, have shown to be effective in this task, albeit there are other techniques to get the same result. In the case of high-quality input, ideally from multiple sources/modalities, this is especially the case [139].

By comparing the spatial organization of distinct body components in images of varying sizes, a convolutional neural network-based approach was proposed by [140] for establishing the human body's stance. For the final prediction, it was necessary to repeat the pooling and upsampling operations multiple times. Experiments using two separate datasets showed that this model significantly outperformed the baselines by 1.7%–2.4%.

In order to forecast the body's location, another model based on the same neural network type was introduced in [141], which made use of interdependent variables. Using an approach that doesn't require the creation of a graphical representation, this method utilizes a CNN network along with pre-prepared knowledge maps to generate appropriate output. Evaluative results on the MPII set (with a 9% improvement), the LSP dataset (with a 6% improvement), and the FLIC dataset (with a 3% improvement) corroborated this as well.

Using DNN as the foundational tool for estimating the locations and interrelationships of the body's joints, [142] built a cascade architecture model in 2014. The model's performance, which outperformed past solutions on two regularly used datasets by 2% and 17%, proves that framing the problem as a question of regression is a very acceptable paradigm.

To compare different deep learning-based pose estimation methods, the authors of [143] introduced a new dataset for SLR research and established a standard for predicting body positions.

They discovered evidence that transfer learning applies to SLR [144] after studying its potential applications. Continuing from the linear Skinned Multi-Person Linear model (SMPL), a comparable approach for posture estimation was proposed by [145] using RGB images and deep learning. The authors of this study conducted a parameter regression using three-dimensional models of human joints as intermediaries.

In order to ensure that any structural flaws are rectified, the model depends on autoencoders to connect the regressed SMPL to a convolutional neural network. When tested on Surreal and Human 3.6M datasets, the enhanced SMPL demonstrated a noticeable performance gain over the baseline. Aside from physical gestures, facial expressions, and body language are also important components in sign language communication. Although there is a plethora of research on the topic of automatic hand gesture recognition, there is less on the subject of body posture analysis.

To tackle this issue, Jain et al. [146] used a CNN to examine the interrelationships of different body parts. However, using a tree-like data organization and an SVM as their classifier, Yang and Ramanan [28] came up with a different approach. Using a graphical model to depict the spatial configuration of human joints, Chen and Yuille [147] performed another noteworthy study in this field. Charles [148] enhanced this method by enhancing the system's ability to comprehend body positions by extracting temporal information from successive photos; Toshev and Szegedy [142] offered another strategy to evaluating the location of body joints.

In order to determine the best methodology for body posture detection, the authors of this work conducted trials on a new dataset using the latter two methodologies, despite the fact that there are many conflicting principles and ideas. More recently, [149] suggested a method that uses a convolutional network to analyze graphs; in this method, the human body is shown in three dimensions using a network of points and connections. To distinguish between data and put this schematic representation into context, this approach uses an attention mechanism. Experimental testing on a variety of SLR datasets shows that this model can outperform alternative techniques by a small margin (0.7%–3.4% points).

Combining features of convolutional and recurring neural networks, as well as a self-correcting feature that can enhance prior predictions, is a model developed by [150]. This model accounts for noise as it constructs a 3D vector space from local input and uses it to recover partial body positions. By comparing it to other models on a new dataset, the authors confirmed that their creation is the best. Depth Ranking Pose Estimation for 3D pictures, the technique proposed by [151], likewise heavily relies on depth imaging. Combining depth data with two-dimensional photos, this approach uses a CNN network to decide between candidate pairs in the initial phase and then makes 3D posture predictions in the second step. Compared to other 3D posture estimation methods, this one performed far better on a scale of more than 6 mm when tested on the industry-standard Human 3.6M dataset.

With the use of depth information to generate maps, a model called DDP (Deep Depth Pose) was suggested by [152] for approximating body positions. These maps were made in advance and included every joint that was relevant as well as several body positions. This strategy surpassed the benchmarks by over 11%, proving its effectiveness in practice.

There have been numerous efforts to develop a good model using convolutional and recurrent forms of deep learning networks due to the importance of body position estimate in various research disciplines, including SLR. With the advent of 3D imagery and the creation of depth maps, these models' identification capabilities have been substantially enhanced. Cascade or tree-like structures, the imposition of specific constraints, etc., are some of the methods that try to achieve greater advances in precision. Experimental assessments show that newer models are significantly more effective and dependable than older ones, yet no solution, no matter how complicated, will have 100% universal application [74].

Improving the ability to understand people's body positions is an important area of study. In instance, scientists are putting in a lot of time and effort to make sure that joint locations can be pinpointed even when photos have background noise or some body components are obscured. Although significant strides have been made in 3D body position mapping, one source of complexity is the fact that a single 2D pose can correspond to numerous 3D locations.

Labeling 3D joint images is challenging, which adds another layer of complexity and calls for high-tech input devices. In contrast, accurate mapping of spatial interactions between critical body locations is necessary for effective 3D data regression. Among the many features tracked by current models is the exact three-dimensional placement of every joint, as seen from different angles and in relation to different body shapes. These models lay the groundwork for additional SLR research that can be expanded upon with different methodologies.

The capacity of modern systems to recognize poses and forecast shapes has been enhanced by technological advancements in capturing equipment. An encouraging area of study is the integration of several data sources such as thermal imaging or hybrid data with indications based on vision, which can increase the systems' reliability in real-world scenarios. In contrast to image-based approaches, which deduce the positions of the critical points (i.e., limbs and joints) from 2D images, sensor technology directly transfers these positions.

Due to this crucial distinction, the input type and desired outcome should inform the procedures used to complete this activity [153]. Correcting the interpretation of sign language information relies heavily on deducing the stance [137]. This is especially crucial for continuous SLR, since it displays individual indications in a continuous stream and how the subject's body moves can convey the whole meaning of the expression. The selection of features, which can incorporate both two- and three-dimensional data points, as well as the depth and architecture of the classifier, are just a few of the numerous elements that might impact the efficiency of pose estimation methods. In spite of their impressive accuracy, several of the most recent pose estimate algorithms are still too vulnerable to false positives to be considered ready for widespread use just yet [154].

Recent research has shown a trend toward using cutting-edge tech, such as the Microsoft Kinect, to identify body poses using a variety of parameters; this is obviously an area that will be further explored in the coming years as improved sensors and tracking devices become accessible [138,142]. At long last, reliable tools for testing out novel approaches are appearing. More thorough testing is encouraged by the availability of publicly available big SLR datasets, which moves us closer to the commercialization stage of this technique.

5. Sign language recognition models

5.1. Continuous sign language recognition models

Continuous models have been utilized in certain investigations pertaining to sign language recognition and modelling. In order to continuously recognize gestures, for instance, Wu and Shao [50] suggested a novel bimodal dynamic network. Both the spatial locations of the 3D joints and the spoken commands of the gesture tokens were used to build the model. Using an expectation–maximization

(EM)-based method, Koller et al. [37] showed how to recognize sign language continuously. To solve the issue of temporal alignment in continuous video processing tasks, an EM-based algorithm was developed. Continuous sign language recognition also has scalability issues, which Li et al. [53] attempted to solve with their suggested system.

Camgoz et al. [38] created a complete system for continuous sign language recognition and alignment. Explanation: The model relies on explicit subunit modelling. In a similar vein, Wang et al. [40] proposed a connectionist temporal fusion method that might convert video's continuous visual languages into textual sentences. Moreover, Rao and Kishore [42] have performed further research on continuous SLR models. Over the course of several iterations, a system was constructed and tested using 282 words of continuous Indian Sign Language.

Koller et al. [69] also utilized a database that included of continuous German Sign Language signing. Graphics were handled in a continuous fashion in [88]. Because the animations were so tough to manipulate after processing, this method was incredibly difficult to implement. Deep residual networks may learn patterns in continuous films containing motions and signs, as Pigou et al. [155] discovered when studying the challenges of continuous translation. Deep residual networks can reduce preprocessing requirements.

The model shown in [17] can improve upon current methods of sign language recognition by a range of 15% to 38% in relative terms and by 13.3% in absolute terms. In addition, Cui et al. [156] proposed a weakly supervised method that, with the aid of deep neural networks, could constantly recognize sign language. The result was on par with what is accomplished by state-of-the-art methods.

5.2. Isolated sign language recognition models

Most investigations on sign language recognition have relied on single sign samples up until recently. Based on hand movements captured by sensor gloves, these models process a series of pictures or signals [92]. In many cases, sensor gloves stand in for a full sign. As an example, a dataset containing isolated signs from the sign languages of Denmark and New Zealand was utilized by Koller et al. [37]. Each signed video corresponded to one word in another experiment by [34], which suggested an isolated SLR system to extract discriminative characteristics from videos.

Following their assessment of the difficulties associated with continuous translation, Escudeiro et al. [88] adopted a standalone strategy. Basically, each gesture was made independently, which makes it much easier to work with animations. In contrast, Fang et al. [110] found that deep recurrent neural networks were the most effective in a hierarchical model. An structured high-level representation usable for translation was generated from the model by combining the isolated low-level American Sign Language characteristics. The utilization of regions of interest (ROIs) to isolate hand motions and sign language characteristics has the potential to improve recognition accuracy, according to recent advances in sign language research [118].

An isolated SLR system was employed to enable real-time sign language translation in [113]. A time-series neural network module and video pre-processing were components of the standalone gloss recognition system. Latif et al. [157] conducted an additional study that examined video portions using an estimated "gloss-level". Cui et al. [3] adjusted their receptive field to match the predicted duration of a single sign while they were conducting their observations.

An isolated SLR task was the subject of a recent study by Huang et al. [116]. To identify a huge vocabulary, it was suggested to employ an attention-based 3D-CNN. The model's strength lay in the fact that it made use of the 3D-CNN's spatio-temporal feature learning capabilities. The American sign language lexicon video dataset, which includes video sequences of isolated American sign language signs, was utilized by Papadimitriou and Potamianos [90].

5.3. Deliberation on sign language recognition models

Isolated and continuous modes make up SLR, and they each present unique difficulties and need for unique solutions. Specifically, continuous SLR requires significantly more direct monitoring, which is a major difference. Unlike isolated SLR, which concentrates all the important information into a small area of a single image, continuous SLR requires meticulous alignment of the video's portions in chronological order and accurate tagging of each sentence.

That's just one illustration of the computationally intensive complexity of continuous sign language recognition. This is something that needs to be considered when evaluating methodologies and choosing features. Continuous video analysis improves the model's accuracy prospects if sequential labeling is executed properly and the most predictive features are chosen. While clever uses of deep learning systems have helped to automate a lot of related chores and this area in recent years, there is still a long way to go before we see advances that the general public can benefit from.

Graph neural networks applications, for instance, make advantage of the attention mechanism, which is fascinating since it works effectively with various kinds of data and can explain complicated connections in space and time. If this method is the best way to fix the current problems with continuous SLR, more study will reveal it.

6. Sign language recognition based on region and spoken language

Sign language is based on many fundamental ideas. To start with, sign languages are never really global. The majority of countries employ a variety of sign languages. There are a lot of countries where sign language is used, including the US, UK, Arabic world, and China. You can see a summary of the studies that used different sign languages in Table 3. As an example, the most widely used localization, American Sign Language (ASL), adheres to its own set of grammar norms apart from visual English.

In their experiment, Rioux-Maldague and Giguere [35] used their proposed technique to classify ASL according to grammatical norms, demonstrating the application of this localization. In order to train and recognize postures, Tang et al. [52] conducted an experiment that took 36 hand postures derived from American Sign Language into consideration. On the other hand, some systems use non-ASL indicators in an English-ordered fashion. Research centered on Italian Sign Language is one such example.

There was an evaluation of a new bimodal dynamic network for gesture recognition in [50], using a dataset of twenty signs from Italian culture or anthropology. The Italian dataset included 7,754 gestures and 393 labelled sequences. For many people who are hard of hearing, Arabic Sign Language is the best way to communicate. A method that can distinguish connected indicators was developed using Arabic depth and intensity images in [36]. An accuracy of 99.5% was achieved when testing the suggested technique with a dataset acquired from three distinct users. A dataset in Arabic Sign Language was also utilized by the writers in [157,158].

Chinese has been the subject of some sign language experiments. The 510 individual words taken from Chinese Sign Language were used as a vocabulary in [72]. Of these words, 353 had only one sign while the rest had multiple signs. In order to accomplish their experiment's goal, Yang and Zhu [39] utilized the instructional film We Learn Sign Language, which demonstrates an interest in Chinese Sign Language. Jiang and Zhang [45] conducted an additional study that included Chinese Sign Language to aid in the fingerspelling procedure. In addition, tests were conducted using Chinese Sign Language by the authors in [56,92,116].

There were a few of more studies that looked at Argentine Sign Language. As an example, consider the study [34], which used Argentine Sign Language to collect data from 10 participants. Similarly, Konstantinidis et al. [44] investigated bone recognition for hands and bodies using Argentine Sign Language with ten participants. Some studies use a combination of sign languages rather than just one.

To investigate CNN training on 1 million hand images, for instance, Koller et al. [37] used a combination of Danish and New Zealand sign languages. Using publicly available lexicons, the sign languages were culled from two representative videos. While there was minor motion blur in the New Zealand version, it was nonexistent in the Danish data. To further investigate the function of SubUNets in sign language recognition, Camgoz et al. [38] conducted an experiment with Danish, New Zealand, and German sign languages.

7. Training and evaluation

7.1. Training with backpropagation

For the purpose of training artificial neural networks, specifically feed-forward networks, backpropagation is a potent deep learning technique. Iteratively, it minimizes the cost function by modifying biases and weights. To minimize loss, the model updates these parameters at each epoch in response to the error gradient. Gradient descent and stochastic gradient descent are two optimization methods that are commonly used in backpropagation. By calculating the gradient according to the calculus chain rule, the method is able to efficiently traverse the many layers of the neural network in order to minimize the cost function. Equations (1) and (2) provide the fundamental equations that characterize the learning process.

$$\theta^{t+1} = \theta^{t} - \alpha\frac{\partial E}{\partial\theta}, \tag{1}$$

$$E = \frac{1}{2N}\sum_{i=1}^{N}(y_i - y_i'), \tag{2}$$

where α is the learning rate and θ is the weight.

In order to train a translation layer, Rioux-Maldague and Giguere [35] utilized the standard multilayer perceptron (MLP) approach. Every 24 letters were translated into a 24-dimensional softmax vector by the output layer during training using normal backpropagation, which interpreted the activations of different restricted Boltzmann machines (RBMs). The training process included weight decay and early halting, and it was based on 200 epochs of backpropagation. They also used the whole network for a fine-grained backpropagation phase, although they slowed down the learning pace significantly. Wu et al. [50] also used the conventional backpropagation method to fine-tune the relative importance of each modality.

7.2. Loss function

SLR can be formulated as a sequence-to-sequence learning problem. Given an input video sequence $X = \{x1, x2, \dots, xT\}$, where each frame xt represents visual or multimodal observations (e.g., RGB, depth, pose), the goal is to predict a linguistic label sequence $Y = \{y1, y2, \dots, yL\}$, with L≤ TL, corresponding to glosses, words, or characters. A neural model $f_\theta(\cdot)$, parameterized by θ, maps the input sequence to frame-level logits as 3.

$$Z \;=\; f_\theta(X) Z \in \mathbb{R}^{T\times K}, \tag{3}$$

where K is the vocabulary size. The learning objective is to find optimal parameters based on 4.

$$\theta^* \;=\; \arg\min_\theta L(Y, f_\theta(X)), \tag{4}$$

where L is a task-dependent loss function.

Cross-entropy loss (CE)

For isolated sign recognition or frame-level classification, cross-entropy loss is commonly used. Let $\hat{p}_t\,(k)$ denote the predicted probability of class k at time t, and y_t be the ground-truth label. the CE los function is computed as (5).

$$L_{CE} \;=\; -\textstyle\sum_{t=1}^{T} \log\, \hat{p}_t\,(y_t). \tag{5}$$

CE enforces discriminative frame-level learning but requires explicit temporal alignment, limiting its applicability in continuous SLR.

Connectionist temporal classification (CTC)

As previously formalized, CTC removes the need for frame-level annotations by marginalizing over all valid alignments. The CTC loss function is presented in (6).

$$L_{CTC} \;=\; -\log \textstyle\sum_{\pi\in\mathcal{B}^{-1}(Y)} \prod_{t=1}^{T} P(\pi_t | x_t). \tag{6}$$

CTC is central to continuous SLR due to its alignment-free supervision and robustness to variable signing speed.

Attention-based sequence-to-sequence loss

Transformer and encoder-decoder models optimize a conditional log-likelihood over output tokens. The loss fuction is defined as (7).

$$L_{seq2seq} \;=\; -\textstyle\sum_{l=1}^{L} \log\, P(y_l | y_{<l}, X). \tag{7}$$

This loss captures long-range dependencies and linguistic structure but is sensitive to alignment noise and requires large training data.

Hybrid CTC–attention loss

To combine the strengths of CTC and attention mechanisms, a multi-objective loss is often used.

The hybrid CTC loss function is computed as (8).

$$L_{hyrid} \;=\; \lambda L_{CTC} + (1-\lambda)L_{seq2seq}, \tag{8}$$

where λ∈[0,1] balances alignment stability and semantic modeling. This formulation improves convergence, stabilizes training, and enhances recognition accuracy in continuous SLR.

Contrastive loss for representation learning

For multimodal or self-supervised SLR, contrastive learning enforces alignment between representations. The loss function is defined as (9).

$$L_{contrastive} \;=\; -\log\frac{e^{sim(z_i,z_i^{+})/\tau}}{e^{sim(z_i,j)/\tau}}, \tag{9}$$

where τ is a temperature parameter. Contrastive loss enhances modality-invariant and signer-independent representations. Table 2 present a summary comparison between different loss functions for SLR tasks.

Table 2. Comparative table of loss functions in SLR.

Loss function	Alignment requirement	Suitable SLR task	Advantages	Limitations
Cross-Entropy	Explicit	Isolated / frame-level	Simple, stable	Requires segmentation
CTC	Implicit	Continuous SLR	Alignment-free, robust	Weak language modeling
Seq2Seq (Attention)	None	Learned	Sentence-level	Captures semantics
Hybrid CTC–Attention	Mixed	Continuous translation	Stable + expressive	Higher complexity
Contrastive	N/A	Multimodal / SSL	Improves generalization	Requires careful sampling

7.3. Datasets

In order to evaluate SLR techniques, a selection of the most relevant and accessible datasets that include hand movements are presented. Making sure dictionaries are big enough to support more stringent testing and more complex applications is a top priority. Depending on the selected geographical variety of sign language, there are currently certain high-quality sets that can be utilized for this purpose.

Researchers in the field of UK sign language have access to a variety of datasets, such as RWTH-Boston-1, RWTH-Boston-50, and RWTH-Boston-400, which contain anywhere from ten to four hundred distinct signs. Notable examples of high-quality data corpora for German sign language include DGS Kinect-40, SIGNUM, and RWTHPHOENIX-Weather. There are a lot of real sentences signed by up to nine professional signers in those sets, and the first and ending frames of each sign are labelled with facial and hand feature definitions. The sets also include 35 to 1225 distinct signs.

With more than 30,000 signs performed by six individuals, ASLLVD is the gold standard for ASL research. Like the last set, this one has labels indicating which frames begin and conclude each sentence. There are three high-quality data sets available for studies of polish sign language variation: PSL Kinect 30, PSL ToF 84, and PSL 101. There is a cap of one person working on these datasets, and they only include individual words (with a total of 30 to 101 signs). Indian scholars have access to the Sign Corpus IITA-ROBITA ISL, which was built cooperatively by multiple teams between 2010 and 2017. Sadly, there is just one signer and only 23 signs in the complete set.

Two datasets, ASLLVD and RWTH-PHOENIX-Weather, stand out among the others due to their widespread applicability. In SLR studies, publicly available sign language sets are frequently utilized as benchmarks to assess the efficacy of suggested computing algorithms. This is because these sets are well-suited for sign language interpretation in real-world scenarios. Virtually, all SLR researchers are presently fixated on the problem of limited access to specialized datasets. A further complicating factor is the need for distinct datasets for various linguistic tasks and geographical variants of sign language.

While some studies employed well-known local datasets, others started with video recordings of sign language users and added additional metrics to create new datasets. In order to train a system that can recognize signs independently of signers, a typical dataset contains several instances of the same sign made by different signers. When evaluating the credibility of findings, it is important to remember that certain datasets offered in the literature are much bigger than others.

As shown in Table 3, we relied on the literature reviews and strictly stated criteria to examine the datasets in all the research publications that were reviewed. The databases utilized share numerous commonalities and can be efficiently categorized according to these properties, since all the publications mainly focus on decoding sign language parts of different levels of complexity. Although certain categories may not apply or authors may not have supplied data, the criteria were chosen with the intention of offering a framework for direct comparison between research.

Table 3. Sign language datasets based on the region and spoken language.

Model	Reference	Items	Classes	Subject
European sign language				
NA	[34]	3200+1297	64+50	10+NA
CNN/Stacked LSTM/OpenPose	[44]	32001535 videos	6450	10
CNN-Stacked LSTM	[44]	3200 videos	64	10
3D CNN	[159]	500 videos	10	10
CNN+EM	[37]	1134319 images	60	6, 8, 2009
CNN+BLSTM	[38]	1.2 million images	60	23
NA	[160]	11+200 per class	40	100
Residual network + BiLSTM	[155]	55224, 12599 video-gloss, 22535 video	100, 100, 249	78, 53, 21
NA	[161]	5 hours video	60	18
CNN + B RNN	[3]	6522711874	455	9
Pose estimation	[49]	43986 images	35	20
Temporal CNN	[156]	5672 sentences	9	NA
TCONV + BGRU	[40]	6841 videos	10	40
CNN + BiLSTM	[3]	6841 sentences + 2340 sentences	455	91
DBN	[50]	13858	20	NA
NA	[162]	2000 Videos	10	3
SVM	[88]	NA	57	NA

Continued on next page

Model	Reference	Items	Classes	Subject
American sign language				
NA	[163]	2524	36	NA
NA	[164]	9800	3300	6
NA	[143]	808+479	NA	8
NA	[165]	NA	35	3
DBN	[35]	60000	24	5
DNN	[52]	288 videos	36	8
CNN + HMM	[69]	2137 sentences	40	7
Sparse autoencoder	[68]	120000 images	24	5
3D CNN	[51]	657	25	9
CNN	[47]	60000 images	26	5
PCANet + SVM	[13]	60000 images	24	5
CNN	[90]	3000, 4416 images	24	6, 20
CNN	[166]	78000	26	NA
CNN + SDAE	[48]	2040 gestures	24	NA
Bidirectional DRNN	[110]	7306 images	156	11
NA	[167]	900 images	36	NA
DenseNet	[168]	100000 images	24	NA
CapsuleNet	[169]	34672 images	24	NA
CNN + LSTM	[118]	62400	24	NA
DNN	[57]	NA	36	12
CNN + SVM	[46]	2425 images	5	20
Arabic sign language				
NA	[170]	180	3	10
NA	[134]	900	30	30
NA	[171]	150	150	21
PCANet, SVM	[36]	1400	28	3
NA	[157]	54049	32	40
ResNet 18	[158]	54049	32	40
East Asian countries sign language				
NA	[83]	54000	45	3
CNN	[172]	1074	10	NA
DCNN	[83]	1147 images	37	NA
NA	[173]	9000	90	NA
NA	[54]	78	26	3
3D CNN	[55]	5000 videos	179	50
NA	[92]	100, 16000 sentences	20, 3000	3, 50
3D CNN + attention	[89]	125000 images	500	50
CNN	[39]	NA	40	NA
CNN	[45]	1260 samples	30	NA
3D CNN + attention	[89]	125000, 14000 instances	50020	50
DNN	[117]	30000 images	20	15

Continued on next page

Model	Reference	Items	Classes	Subject
East Asian countries sign language				
CNN	[174]	NA	26, 9	NA
DNN	[42]	282 words	NA	10
LSTM	[112]	1630 words	163	2
3D CNN + BiRNN	[43]	3006 videos, 30 sentences	30	10
3D CNN	[175]	100 images	5	NA
LSTM	[156]	NA	NA	NA
LSTM	[41]	812 sentences	195	1
NA	[144]	14672	419, 105	14
GRU	[144]	14672 videos	524	14
YOLO	[176]	30000 images	25	12

Through this review, we aim to highlight the similarities and differences in the datasets used across various studies. To ensure clarity and address space constraints, training, testing, and assessment datasets are often presented in a combined format. Consequently, the actual structure of a dataset may be more complex than what is depicted in the tables for certain studies. For practical applications of any of these SLR datasets, it is advisable to closely examine each dataset in detail. A glance at Table 4 reveals significant variations in the datasets, particularly in terms of data types.

Table 4. Sign language datasets based on the data type.

Reference	Data type
Alphabetic linguistic content	
[177]	RGB video + depth info
[178] [73]	2 D images
[37] [178] [45] [166] [45] [48] [49] [169] [117] [174] [91] [157] [158] [36] [46] [176]	RGB
[47]	RGBD
[13]	RGBD, Kinect
[92]	RGBD, Kinect, gloves
[90]	RGB video
[36]	RGB+ depth info
[179]	RGB+ depth RGB
[57]	3D models
Words and sentences linguistic content	
[156] [40] [3]	RBG
[34] [43] [44] [159] [143] [144]	RGB video
[41] [112]	RGB, Kinect
[89]	RGB, depth, skeleton
[89]	RGB, Kinect, skeleton point
[52]	RGB, Kinect, 3D skeleton point
[110]	Infrared

Continued on next page

Reference	Data type
Words and sentences linguistic content	
[165] [160] [162]	Video
[164]	Video, Kinect
[171]	RGB, depth, 3D skeleton, facial features
Hand gesture linguistic content	
[38] [3] [167] [168]	RGB
[68]	RGBD
[39] [118]	RGB video
[175] [51]	RGB, Kinect
[35]	Intensity camera, Kinect
[180]	2D and 3D skeleton, depth info
[111]	6D IMU
[155]	RGB, RGBD Kinect

Because studies of sign language use different theoretical frameworks and may investigate seemingly unconnected areas of sign language understanding, this is to be expected. For example, while continuous SLR experiments require sentences or even longer segments of speech, isolated SLR experiments typically use alphanumerical characters or words for recognition of isolated language elements. Knowing the difference between the two approaches and the kinds of datasets appropriate for each is crucial.

When trying to assess a model's generalizability, it's necessary to take into account the dataset’s size and complexity differences. Nevertheless, due to resource constraints and practical considerations, even the most extensive datasets fall well short of being comprehensive. The availability of more statistics documenting several geographical variations of sign language is a positive trend. Since SLR research has broad applicability, it is imperative that we prioritize the development of automated systems that can identify regional variants of hand signals.

Additionally, multi-modal datasets are on the rise, which bodes well for the future of SLR research and provides more room for creative thinking. The lack of diversity in the signers and classes used to compile most datasets casts doubt on their reliability as representations of the real world. Because of this, automated algorithms that use those datasets may not be able to accurately interpret significantly different sign language gesture displays. One of the most important factors influencing the rate of advancement in any area of artificial intelligence research is the accessibility of high-quality datasets for training and testing models.

The studies that were considered show that this is becoming less of an issue, which is encouraging because SLR research is still a young field. When widespread compatibility of the experimental results is sought, there are a number of commonly used datasets that can be regarded as “standards”. However, fresh datasets that are specific to local sign language systems are cropping up, which means that they may be able to be recycled to power more studies in the same area. Although things are looking good, it's important to note that the datasets that are already out there vary substantially in size, structure, quality, and perhaps require the creation of additional datasets to back up certain study paths.

Although a growing number of datasets have been introduced for SLR, dataset scarcity remains a fundamental challenge. This apparent contradiction arises from the distinction between dataset count and effective data coverage. While multiple corpora exist, most suffer from limitations in scale, diversity, and annotation consistency, which significantly restrict model generalizability and real-world applicability.

A dominant source of bias in existing SLR datasets is signer dependency. Many datasets include a small number of signers—often fewer than ten—recorded under controlled laboratory conditions. Models trained on such data tend to overfit to signer-specific characteristics such as hand size, signing speed, posture, and habitual motion patterns. This bias severely limits cross-signer generalization, which is essential for practical SLR systems intended for broad user populations. The lack of demographic diversity in age, gender, and signing style further exacerbates this issue.

Most available datasets focus on limited vocabularies, frequently constrained to isolated signs or predefined gloss sets. While suitable for benchmarking isolated SLR, these datasets fail to capture the linguistic richness of natural sign languages, including co-articulation, grammatical facial expressions, and non-manual markers. Additionally, gloss annotations often abstract away semantic nuance, resulting in models that recognize symbol sequences rather than meaning. This lexical bias restricts the applicability of trained models to real-world continuous signing scenarios.

SLR datasets are commonly recorded in controlled environments with uniform backgrounds, stable lighting, and fixed camera viewpoints. Although this setup simplifies data collection and annotation, it introduces a strong domain bias. Models trained on such data often exhibit significant performance degradation when deployed in unconstrained settings, such as daily communication environments with occlusions, camera motion, or background clutter. This gap highlights the lack of in-the-wild datasets that reflect realistic signing conditions.

Another critical limitation lies in annotation quality. Differences in gloss definitions, temporal segmentation strategies, and labeling conventions across datasets hinder cross-dataset training and evaluation. In continuous SLR, the absence of consistent frame-level annotations further complicates sequence alignment and learning. These inconsistencies introduce noise into the training process and impede the transferability of learned representations.

Many datasets exhibit severe class imbalance, where frequent signs dominate the training distribution while rare signs remain underrepresented. This imbalance biases models toward common patterns and reduces recognition accuracy for less frequent signs. Moreover, most datasets rely exclusively on RGB video, with limited availability of multimodal data such as depth, skeletal pose, or inertial measurements. The lack of multimodal annotations constrains the development of robust, modality-agnostic models.

Collectively, these biases explain why dataset scarcity persists despite the apparent abundance of datasets. The challenge lies not in the number of datasets, but in the absence of large-scale, diverse, consistently annotated corpora that support signer-independent, linguistically grounded, and deployment-ready SLR models. Addressing this issue requires not only larger datasets, but also principled data collection protocols, standardized annotations, and increased emphasis on cross-dataset evaluation.

7.4. Evaluation

Accurate recognition of sign language material is the main focus of most research papers, and the primary criteria used to measure the results is the F1 score which combine the precision and the recall. Training, testing, and validation accuracy are sometimes defined differently by different writers, depending on the phase of an experiment. Another aspect that was monitored in certain experiments was the amount of time needed for training to achieve acceptable accuracy; this was often stated in epochs. Though it was possible to describe processing time and input video length in seconds and/or frames, these metrics were not always deemed important enough to merit direct measurement. For a quick rundown of performance metrics utilized in SLR research, Tables 5 present a detailed summary.

Table 5. Accuracy of state-of-the-art models for SLR.

Model	Accuracy (%)
[34]	99.84
[170]	92
[177]	97.3
[163]	84.68
[83]	94.7
[173]	98
[180]	92.28
[160]	95
[162]	73
[181]	63.56
[165]	89.33
[171]	55.57
[144]	93.28
[50]	70.1
[35]	77
[51]	98.12
[69]	55.7
[68]	99.1
[52]	94.2
[54]	98.9
[47]	80.34
[55]	92.9
[13]	88.7
[117]	97
[144]	93.28
[90]	99.39
[166]	88.7
[89]	91.93
[3]	99.5
[36]	62.8
[37]	87.4
[53]	92.83
[48]	94.5
[110]	73.3
[155]	97.7
[111]	81.7

Continued on next page

Model	Accuracy (%)
[49]	99
[39]	70
[34]	95
[172]	98.05
[167]	90.3
[168]	99.74
[169]	91
[118]	83.78
[57]	84.68
[73]	53
[41]	92.88
[182]	83.72
[91]	99.56
[174]	99.31
[179]	94.7
[183]	77
[112]	90
[42]	98.09
[44]	100
[159]	92.24
[175]	91.93
[3]	88.7
[89]	99.48
[158]	88.1
[45]	98.36
[46]	98.81
[114]	87.31
[125]	92.92
[129]	98.4
[76]	71.9
[80]	77.75
[79]	98.08
[77]	46.96
[78]	74.7
[75]	94.77

While recognition accuracy remains the primary evaluation metric in most SLR studies, practical deployment—particularly in wearable assistive devices, mobile platforms, and interactive systems—requires careful consideration of real-time performance constraints. These constraints include inference latency, memory footprint, and energy consumption, which are directly influenced by architectural design choices.

We consider the following metrics as fundamental for deployment-oriented evaluation:

- Inference latency: Average time required to process a video frame or sequence, directly affecting real-time usability.

- Model size: Memory footprint determined by parameter count and precision, influencing deployability on edge devices.
- Energy consumption: Power usage during inference, critical for battery-powered systems.

Rather than hardware-specific benchmarks, we analyze relative computational behavior across architecture families. 2D CNNs offer low inference latency and moderate model size, making them suitable for isolated sign recognition and short sequences. However, extending them to 3D CNNs significantly increases computational cost and energy consumption due to volumetric convolutions. CNN–RNN hybrids and TCNs provide a favorable balance between temporal modeling and efficiency. TCNs, in particular, benefit from parallelizable convolutions, resulting in lower latency than recurrent models while maintaining competitive accuracy. Pose-based GCNs are computationally efficient, as they operate on sparse skeletal graphs rather than dense pixel grids. This results in small model sizes and low energy consumption, making them highly suitable for real-time and embedded SLR systems, provided reliable pose estimation is available. Transformers deliver strong performance for long-range temporal reasoning but incur high memory and computational costs due to self-attention's quadratic complexity with sequence length. This limits their applicability in real-time scenarios without optimization techniques such as windowed attention or model compression.

Hybrid models combine multiple components and therefore exhibit higher computational demands. Nevertheless, they often achieve superior accuracy per parameter due to complementary inductive biases, making them attractive for medium- to high-budget deployments. Table 6 presents a comparison between different models families based on different metrics.

Table 6. Computational trade-offs of SLR architectures.

Architecture family	Inference latency	Model size	Energy consumption	Real-time suitability	Typical deployment
2D CNN	Low	Small-Medium	Low	High	Mobile/Edge
3D CNN	High	Large	High	Low	Offline
CNN-RNN	Medium	Medium	Medium	Medium	Desktop
TCN	Low-Medium	Medium	Medium	High	Real-time systems
GCN (Pose-based)	Very Low	Small	Very Low	Very High	Wearables/Edge
Transformer	High	Large	High	Low-Medium	Server-scale
Hybrid Models	Medium-High	Large	Medium-High	Medium	Hybrid deployments

This analysis demonstrates that model selection in SLR inherently involves a trade-off between recognition accuracy and computational feasibility. Pose-based and TCN architectures are best suited for real-time, low-power applications, whereas Transformer-based and hybrid models are more appropriate for high-accuracy, offline, or server-assisted scenarios. By explicitly incorporating computational metrics into the discussion, the revised manuscript provides a more comprehensive and deployment-aware perspective on SLR system design.

Proposed sign recognition algorithms are quantitatively evaluated in nearly all of the assessed research papers. The goals of the study dictate which tests are administered, which in turn affects the breadth and depth of the testing. The goal of the testing was to determine the algorithm's performance in differentiating between phrases or words in sign language, typically by comparing it to other benchmark methods.

In general, several approaches worked adequately and identified over 90% of the presented indications; however, there are certain caveats when comparing results across studies due to the varied nature of the tests. The claimed effectiveness was over 97% in a few instances, although these instances were simpler tasks and frequently couldn't be sustained across multiple datasets. Recognition rates of 80% for continuous SLR tasks are strong, particularly when they hold across different datasets.

It was observed that the majority of algorithms showed inconsistent performance when switching between signs, and that a small number of perplexing indicators usually caused the majority of false positives. Due to systemic reasons or the similarity of hand gestures, these frequent mistakes frequently persisted irrespective of the classifier or training technique. This discovery highlights the fact that existing SLR algorithms are still not perfect and should be regularly compared with human-made estimates to prevent misunderstandings, and it also suggests that certain issues with the structure and form of sign language, rather than methodological shortcomings, are preventing the development of more effective tools.

Typically, the proposed models are assessed based on their ability to accurately carry out the main task, which is either sign language translation or recognition. The primary metric for evaluating the model's performance is the whole dataset average accuracy; a greater percentage denotes a more precise method. The accuracy of the model in identifying the “most likely” options was expressed as top-1, top-5, and top-10 in some instances, instead of just one right response.

In order for the neural classifier to perform well in testing and real-world scenarios, it needs to be trained on data that closely matches the samples it will face. It is common practice for human observers to annotate a set of basic sign language symbols, words, or sentences before using them as training data for a sign language decoding system. Following training, the model can be applied to deduce certain sign language elements using the same structure, with different levels of accuracy. While some research focused on finding the best feature combinations, other studies compared several classifiers on the same tasks to see which ones performed better.

Although neural models can only generalize to the signs learnt in the training set, it is possible to attain some degree of accuracy when it comes to individuals expressing the same sign. Consequently, optimizing training parameters is a crucial part of SLR research that can greatly affect the solutions' usefulness. Improving the ability to translate in real-time and understand increasingly complicated portions of continuous sign language speech are goals of more sophisticated systems. Such uses are far more involved than basic character recognition using sign language datasets derived from sentence and word content. Datasets for sign language derived from other language sources are used for low complexity systems. individual words, and they often need to use a combination of indicators to decipher a sequence’s meaning. As a result, scientists are turning to hybrid designs and complex sequence-to-sequence models to help them decipher subtle semantic cues and distinguish between seemingly identical indications.

Hybrid architectures in SLR—such as CNN–RNN, CNN–GCN–Transformer, or multimodal fusion networks—exhibit superior performance due to the synergistic interaction of multiple modeling mechanisms, rather than a single dominant factor.

At the representational level, hybrid models benefit from complementary feature fusion. CNN-based encoders excel at extracting dense spatial and appearance features from RGB inputs, while GCNs capture articulated kinematic structures from skeletal data, and facial encoders model non-manual cues. By jointly learning these representations, hybrid systems reduce information loss inherent in single-modality pipelines. Empirically, this fusion improves robustness to background clutter, signer variability, and viewpoint changes, directly addressing known failure modes of RGB-only models.

Hybrid architectures also enhance temporal reasoning by distributing temporal modeling across different layers. Local temporal dependencies are often captured by 3D CNNs or TCNs operating on short frame windows, while long-range dependencies and linguistic structure are modeled using RNNs or Transformers. This multi-scale temporal decomposition allows hybrid models to align low-level motion dynamics with high-level semantic units, which is particularly critical in continuous and sentence-level SLR. As a result, hybrid systems achieve better sequence alignment and reduced temporal ambiguity compared to single-stage temporal models.

A less explicit but equally important factor is implicit regularization. Hybrid architectures introduce architectural constraints—such as separate modality streams, attention-based fusion, or auxiliary losses (e.g., CTC combined with sequence-to-sequence objectives)—that restrict the hypothesis space. This structured learning acts as a form of regularization, improving convergence stability and reducing overfitting, especially in the presence of limited or biased datasets. Multi-objective optimization further encourages the model to learn representations that are simultaneously temporally coherent and semantically consistent.

Taken together, the performance gains of hybrid approaches arise from the joint effect of richer representations, hierarchical temporal modeling, and regularized optimization, rather than from any single component in isolation. This mechanistic understanding explains why hybrid architectures consistently outperform monolithic CNN-, RNN-, or Transformer-only models across diverse SLR benchmarks and tasks.

8. Summary of principled model and guidelines for practitioners

SLR poses unique challenges due to its reliance on fine-grained spatial cues (hand shape, orientation, facial expression) and complex temporal dynamics (motion trajectories, co-articulation, and long-range linguistic dependencies). Consequently, a wide range of neural network architectures have been explored, each offering distinct advantages and trade-offs. This section provides a detailed comparative analysis of the principal model families, focusing on their representational capabilities, computational characteristics, and applicability to different SLR scenarios.

2D CNNs process sign language videos on a frame-by-frame basis, excelling at spatial feature extraction such as hand appearance, facial expressions, and local texture cues. Architectures such as ResNet and EfficientNet benefit from large-scale image pretraining, which is particularly advantageous in SLR settings with limited labeled data. However, because 2D CNNs lack intrinsic temporal modeling, they cannot capture motion patterns or temporal dependencies without additional modules. As a result, they are typically combined with temporal pooling, recurrent networks, or attention mechanisms. While efficient and easy to deploy, 2D CNN-based pipelines are best suited for isolated SLR tasks where temporal complexity is limited.

3D CNNs extend spatial convolutions into the temporal dimension, enabling joint modeling of space and time. Models such as I3D and SlowFast capture short-term motion dynamics directly from video clips, making them highly effective for recognizing dynamic signs. Their main advantage lies in end-to-end spatio-temporal feature learning without explicit temporal alignment modules. However, this comes at a high computational and memory cost, and performance is strongly dependent on large-scale pretraining. Additionally, 3D CNNs struggle to model long-range dependencies efficiently, limiting their scalability to continuous SLR tasks.

Two-stream architectures decouple spatial and temporal modeling by processing RGB frames and motion information (typically optical flow) in parallel. This explicit separation allows the model to emphasize dynamic motion cues that are critical for many signs. While such architectures often improve recognition accuracy, they introduce significant computational overhead due to optical flow computation and increased model complexity. Furthermore, optical flow can be sensitive to noise, occlusions, and camera motion, which reduces robustness in real-world scenarios. These models are therefore more appropriate for offline or benchmark-focused studies rather than real-time applications.

CNN–RNN hybrids combine spatial encoders with recurrent networks such as LSTMs or GRUs to model temporal dependencies. This architecture naturally supports variable-length sequences and is widely used in continuous SLR and sign-to-text translation systems. Recurrent models are effective at capturing sequential patterns but exhibit limitations in modeling very long sequences due to vanishing gradients and limited parallelism. Their performance is also highly dependent on the quality of the extracted frame-level features. Despite these drawbacks, CNN–RNN pipelines remain a practical and computationally efficient choice for moderate-length sequences.

Temporal Convolutional Networks use one-dimensional convolutions with dilation to model long-range temporal dependencies in a parallelizable manner. Compared to RNNs, TCNs offer more stable training and better scalability to long sequences. They are particularly effective for temporal segmentation and frame-level labeling in continuous SLR. However, TCNs do not inherently model spatial structure and therefore rely on a separate spatial backbone. Their effectiveness depends on careful design of the receptive field to match the temporal extent of the task.

GCNs and their spatio-temporal variants operate on skeletal representations of the signer, modeling joints as nodes and their relationships as edges. This structured representation enables efficient modeling of kinematic dependencies and reduces sensitivity to background clutter. GCN-based approaches are computationally lightweight and well suited for real-time applications. However, their performance is limited by the accuracy of pose estimation, particularly for hands and fingers, and they lack appearance-based cues such as texture and facial details. As such, they are most effective when combined with visual feature extractors.

Transformer models leverage self-attention to capture long-range temporal dependencies and complex interactions across frames. Unlike RNNs, they enable global temporal reasoning and flexible multimodal fusion, making them particularly suitable for continuous SLR and sign-to-text translation. Transformer-based models often achieve state-of-the-art performance but are computationally expensive and data-intensive. Their quadratic complexity with respect to sequence length necessitates efficient attention variants or temporal downsampling for practical deployment.

Hybrid architectures integrate complementary model families, such as CNNs or GCNs for local feature extraction and Transformers or TCNs for global temporal modeling. The superior performance of hybrid approaches can be attributed to their ability to fuse heterogeneous features, model long-range dependencies, and introduce implicit regularization through architectural modularity. While highly effective, these models are more complex to design and optimize, and they incur higher computational costs.

Lightweight architectures prioritize efficiency through compact backbones, model pruning, and quantization. These approaches trade some accuracy for real-time performance and low power consumption, making them suitable for wearable assistive devices and edge computing scenarios. Their limited capacity, however, restricts their ability to handle highly complex or long-duration sign sequences.

In summary, no single architecture is universally optimal for all SLR tasks. Instead, the choice of model should be guided by task complexity, sequence length, available computational resources, and deployment constraints. While CNN-based models remain effective for isolated SLR, Transformer-based and hybrid architectures are increasingly favored for continuous and semantic-level tasks. Lightweight and pose-based models, on the other hand, offer practical solutions for real-time and resource-constrained applications. This comparative analysis highlights the importance of principled architectural selection in advancing robust and deployable SLR systems. Table 7 represents a detailed comparative analysis of neural network architectures for sign language recognition

Table 7. Detailed comparative analysis of neural network architectures for sign language recognition.

Architecture type	Key advantages	Main limitations	Suitable scenarios
2D CNNs (e.g., ResNet, EfficientNet)	Strong spatial feature extraction; readily available pretrained weights; efficient for frame-based processing	Lack inherent temporal modeling; require additional modules for motion representation; may miss fine temporal cues	Isolated SLR with limited temporal variability; low-data settings leveraging transfer learning
3D CNNs (e.g., I3D, R(2+1)D, SlowFast)	Joint spatio-temporal modeling; high recognition accuracy for dynamic signs; effective with large-scale pretraining	Computationally intensive; memory-heavy; require substantial training data; limited long-range temporal modeling	High-accuracy isolated SLR; offline processing; datasets with rich motion patterns
Two-Stream Networks (RGB + Optical Flow)	Explicit motion modeling; improved performance for highly dynamic signs; separate stream specialization	Optical flow is costly and sensitive to noise; higher complexity; increased inference time	Benchmark-oriented studies; accuracy-focused systems where compute cost is acceptable
RNNs (LSTM/GRU) and CNN–RNN Hybrids	Effective for variable-length sequences; natural fit for sequential decoding; lightweight compared with transformers	Weaker long-range modeling than attention-based models; dependent on frame-level feature quality	Continuous SLR pipelines; low-resource systems requiring moderate temporal reasoning
Temporal Convolutional Networks (TCN)	Parallel temporal modeling; stable training; long receptive fields with dilated convolutions	Requires careful design for long sequences; relies on separate spatial encoder	Framewise labeling, temporal segmentation, and continuous SLR boundary detection
Graph Convolutional Networks (GCN / ST-GCN)	Strong modeling of joint relationships; efficient; robust to background clutter; well-suited for pose-based SLR	Performance tied to pose estimation accuracy; lacks appearance cues; sensitive to hand/finger keypoint errors	Real-time or resource-constrained SLR; pose-centric systems; noisy visual environments

Continued on next page

Architecture type	Key advantages	Main limitations	Suitable scenarios
Transformers (e.g., TimeSformer, ViT-Temporal)	Excellent long-range dependency modeling; flexible multimodal fusion; state-of-the-art performance with adequate data	Data-hungry; high compute/memory cost; challenging for on-device deployment	High-performance continuous SLR; sign-to-text translation; multimodal modeling
Hybrid Models (CNN + Transformer, GCN + Transformer)	Combine strengths of local and global modeling; highly adaptable; superior empirical performance	More complex architecture and tuning; higher computational overhead	Advanced SLR systems prioritizing accuracy; multimodal fusion tasks
Lightweight / Mobile Models (MobileNet + TCN, Tiny-Transformer)	Suitable for on-device and real-time inference; reduced power/runtime cost	Lower accuracy; require distillation/pruning; limited handling of highly complex signs	Wearable and embedded SLR systems; smart glasses; low-latency applications

SLR tasks vary in temporal scope, from isolated signs lasting a few frames to continuous signing over long sequences. The framework categorizes tasks into three sequence-length regimes. Short sequences (isolated signs, 1–2 seconds): These tasks primarily require fine-grained spatial recognition of hand shapes, orientations, and facial cues. 2D CNNs or lightweight 3D CNNs are effective here, as they can capture local spatio-temporal features without the overhead of long-range modeling. Medium sequences (phrases or segmented streams): Medium-length sequences demand modeling of local temporal dependencies. Architectures such as CNN–RNN hybrids or Temporal Convolutional Networks (TCNs) are suitable, as they can capture sequential patterns while remaining computationally tractable. Long sequences (continuous SLR, sentence-level or sign-to-text translation): These tasks require long-range temporal reasoning and semantic alignment. Transformers or hybrid CNN/GCN + Transformer models are recommended due to their ability to model global dependencies and integrate multimodal cues. Table 8 provide a summary on model selection based on sequence length.

Table 8. Model selection based on sequence length.

Sequence length	Dominant challenge	Recommended architectures	Rationale
Short sequences (isolated signs, 1–2 s)	Fine-grained spatial and short-term motion cues	2D CNN + pooling, 3D CNN	Temporal dependencies are limited; local spatio-temporal modeling is sufficient
Medium sequences (phrases, segmented streams)	Temporal ordering and local context	CNN + RNN, TCN	Efficient modeling of sequential patterns with moderate temporal span
Long sequences (continuous SLR, sign-to-text)	Long-range dependency and linguistic structure	Transformer, Hybrid CNN/GCN + Transformer	Self-attention enables global temporal reasoning and language-level modeling

Practical deployment scenarios impose constraints on compute resources, memory, and energy.

The framework incorporates computational considerations to guide architecture selection. Low-budget environments (edge devices, wearable assistive systems): Lightweight GCNs, CNN–TCN pipelines, or pruned CNNs are recommended to ensure real-time performance with minimal energy consumption. Medium-budget environments (desktop or interactive setups): CNN–RNN hybrids, TCNs, or compact Transformers provide a balance between accuracy and efficiency. High-budget environments (server-scale offline analysis): 3D CNNs, full Transformers, or hybrid architectures achieve state-of-the-art accuracy at the cost of higher compute and memory requirements, suitable for offline processing and high-fidelity translation tasks. Table 9 summarizes the model selection based on computation budget.

Table 9. Model selection based on computation budget.

Computational budget	Deployment context	Suitable model families	Trade-off
Low (edge, wearable)	Smart glasses, mobile devices	GCN (pose-based), lightweight CNN + TCN	Lower accuracy but real-time and energy-efficient
Medium (desktop GPU)	Interactive systems, lab setups	CNN–RNN, TCN, compact Transformers	Balanced accuracy and efficiency
High (server-scale)	Offline analysis, translation systems	3D CNNs, full Transformers, hybrid models	Highest accuracy at the cost of compute and memory

By combining sequence length and computational budget, the framework yields clear recommendations for model selection. Real-time, low-power applications: Favor pose-based GCNs or lightweight CNN–TCN pipelines. Moderate accuracy and flexibility: Use CNN–RNN or TCN architectures for medium-length sequences. High-performance continuous SLR: Adopt Transformer-based or hybrid architectures for long sequences requiring semantic-level understanding. Multimodal fusion: Where RGB, skeletal, depth, and facial modalities are available, hybrid models with attention-based fusion provide the most robust performance, despite higher computational cost.

This framework provides a systematic and task-driven approach to model selection, overcoming the limitations of purely descriptive surveys. It allows practitioners to make informed decisions based on task complexity, sequence length, modality availability, and computational constraints, ensuring both practical feasibility and high recognition performance. Based on the proposed principled framework, the selection of neural network architectures for SLR should be guided by task requirements, temporal complexity, and resource constraints. For scenarios where latency and power consumption are critical, such as wearable devices or edge applications, pose-based GCNs or lightweight CNN–TCN pipelines are recommended due to their efficiency and real-time capability. When moderate accuracy and flexibility are desired, CNN–RNN hybrids or TCN-based architectures provide a balanced trade-off between performance and computational cost, making them suitable for interactive or desktop systems. For tasks requiring long-range temporal reasoning and linguistic understanding, such as continuous sign-to-text translation or sentence-level SLR, Transformer-based or hybrid architectures are most appropriate, offering superior modeling of global dependencies. Finally, when multimodal cues—including RGB video, skeletal pose, and facial expressions—are available, hybrid models that integrate these streams provide the most robust recognition performance, albeit at higher computational complexity. These guidelines enable practitioners to select architectures systematically based on operational constraints and task-specific priorities.

9. Challenges and future directions

The most noticeable flaw, after going over a lot of SLR studies, is how disjointed the research is in this area. Although several research teams have come up with promising outcomes utilizing different methodologies, there isn't much overlap between these studies and the use of numerous effective instruments together is taking its sweet time to emerge. One potential roadblock to improving practical results is the absence of a widespread agreement on the most important properties and the best neural network architecture.

Continuous sign language voice recognition is still quite difficult, and even the most advanced automated systems have trouble understanding the subtleties of spoken sign language. This might be due, in part, to the fact that the majority of accessible datasets have very small vocabulary sets and very basic sentences, whereas training models for complex language tasks need much larger libraries with a wide variety of samples. It is still very difficult for automated systems to understand sign language communication. It turns out that the reasons machines still can't reliably decipher sign language sequences aren't as black-and-white as they are initially.

It is challenging to describe any natural language in a mathematical style that computers can be programmed with due to the complicated interplay of many laws and relationships. This clarifies why the latest SLR tools perform admirably with alphabetic characters and basic sentences and phrases, yet.... State-of-the-art sign language recognition accuracy results. has difficulty managing tales and conversations that go on indefinitely. Given the field's social relevance, some of the most prestigious academic institutions in the world devote substantial resources to improving its current state. One may make the case that the next time frame is essential for removing some of the roadblocks to faster advancement.

Although some of the examined models concentrate on RGB images with a higher level of information to enable efficient SLR, most of them use depth imaging. Data presented in a sequential style has also proven valuable, particularly for scene and object tracking as well as skeletal position data. While thermal imaging isn't often used for SLR, it can provide value when paired with more fundamental data types like pictures. At the sign level, we have static signs and dynamic signs, with a subgroup of dynamic signs utilized in continuous SLR. Consistent with recent tendencies, research into complex signs and continuous video is likely to take center stage in the near future. Everything seems to be setting up for this change of emphasis to take place.

An ongoing problem in SLR research is the absence of reliable input databases of high quality. Researchers from smaller nations do not have access to the samples necessary for training and testing models. The only large and diverse datasets accessible are for American Sign Language and a few other varieties, such Chinese or Indian. This is beginning to change, though, as more and more research into SLR is conducted and resources are amassed to support future waves of studies.

Although things are looking up, it's still not easy to test out more complex applications that call for big vocabularies in order to show how well current or future methods work. Meanwhile, more proactive resource sharing and direct collaboration across research teams could significantly alleviate current challenges and set a precedent for more effective networking. An international concerted effort is necessary to find a solution to the problem of sign language recognition, which affects people all over the globe.

However, sign language is highly variable among regions, with each using its own distinct set of hand and facial motions to convey meaning. Therefore, it is very evident that high-quality datasets incorporating all relevant input modalities are required. Until recently, there weren't enough properly labeled sets of hand signs to conduct outside testing of SLR gear, but that's starting to change. In the long run, we hope that better datasets will make it easier to create SLR models that have real-world applications. This requires the labeling of lengthier segments of sign language discourse rather than the current practice of labeling individual pieces.

To put it simply, in order for newly created approaches to become a reality, fresh datasets should reflect the diversity of sign language communication. Modern SLR techniques should have little trouble processing lengthy sequences of signs, as real communications are continuous and unconstrained. Now that deep learning networks are becoming more widely used, this lofty objective may be within reach very soon.

Despite numerous research have tackled this subject, there are still numerous challenges that must be resolved. When attempting to characterize numerous human body components, it can be helpful to integrate attributes. Data might come in various forms, including text, photos, depth and skeleton information, etc., which makes this problem more difficult to solve. Better feature engineering and, by extension, a more precise model, can be achieved by merging portions of this data. These characteristics are most prominent on the hands, face, and trunk of the body. Imperfect models that misinterpret some indications can be the outcome of focusing just on hands. Detecting the position of the hand, estimating its shapes and motions, tracking its movement in real time, and similar activities are all important for success in this area, but they can all be challenging in their own ways.

To illustrate the point, signers' hands can vary greatly in size and shape, and yet, fingers can appear very similar and even obscure one another at times. Light levels and other environmental factors can play a role. When dealing with low-resolution input photos, obstacles in the way, or complex gestures that need analysis, these problems become much worse. Researchers use feature fusion to incorporate face traits into the mix, which helps to relieve some of those problems.

Conversely, there are unique difficulties associated with using sign language, such as the partial blocking of important areas caused by the fast movement of the neck and face. Additionally, the third set of traits, which pertain to the signer's body, can be incorporated to enhance recognition even further. Therefore, more generalizable models that may draw on data from many anatomical regions are preferable and should serve as the basis for further studies.

The field of isolated SLR has shown significant progress in training algorithms to detect individual alphabetic signs or words; however, continuous SLR, which requires interpretation of longer segments of speech, has not been as fruitful. Because of the importance of sign-to-sentence linkages in determining sentence meaning, this task cannot be boiled down to gesture recognition alone. When complex semantic subtleties need to be analyzed, current efforts to build continuous SLR capacity often fail and show only limited effectiveness.

Among the many active areas of SLR research, this will undoubtedly remain a focal point as researchers seek a configuration that may circumvent the obstacles impeding the development of powerful new tools. In light of the present state of the art, we anticipate that future studies will center on more elaborate neural network models with many layers and different compositions of layers used to increase processing power.

Any technology aiming for commercial use and public trust must possess exceptional reliability (>99%) and consistency. Current SLR apps, on the other hand, still indicate a tiny but consistent number of false positives and false negatives, thus this isn't the case. Very few SLR technologies are currently being used in practice since the rate of erroneously detected objects grows with the vocabulary size and task complexity.

Small teams frequently lack the resources to conduct large-scale testing or extensively refine training strategies; therefore, although some proposed solutions are conceptually sound and promising for further development, they are often preliminary and insufficiently validated. The next step is to rally more people behind the cause and collect enough money and materials to optimize accuracy at a high level. The systems need to be evaluated in a range of environments and be able to produce usable findings despite less-than-ideal external conditions, such as input photographs captured in low-light situations.

The ability to meaningfully connect observed hand and body gestures to set units of sign language has long been the focus of scientific inquiry. Although this is reasonable for a preliminary scientific investigation, more focus on the usability aspect is required moving forward. Modern systems are far more efficient and may incorporate as few as a handful of small cameras, in contrast to earlier SLR solutions that necessitated sensors worn on the body and other cumbersome apparatus.

Another area that needs more future attention is user-system interaction, namely how to give users some say over the software that their computers run. Making sure user opinions are acknowledged and having a system in place to quickly find frequent mistakes are both critical. There has been a resurgence of interest in SLR research since the last discovery period, and numerous conflicting theoretical postulations have emerged as a consequence.

Despite widespread agreement that deep learning networks are the best technology to solve this challenging language challenge, a long way still to go before completely automated systems can comprehend live streams of sign language conversation. Some of the known solutions will likely reach a virtually ideal level of maturity in the next decade, and a big breakthrough could happen at any point. More innovative and useful mainstream applications will undoubtedly appear as SLR technologies improve in reliability; these will directly benefit the global population of hearing and speech-impaired people.

While existing studies identify broad challenges in SLR, meaningful progress requires the definition of concrete and measurable benchmarks. A first critical research target is signer-independent continuous SLR at scale, where future models should aim to achieve 75%–80% accuracy (or <25% word error rate) on large-vocabulary datasets (>1,000 glosses) under strictly disjoint train – test signer protocols. Such a benchmark directly addresses the current generalization gap and emphasizes learning signer-invariant linguistic representations rather than appearance-specific cues. Complementarily, a second benchmark should focus on real-time deployability, requiring models to maintain 65%–70% signer-independent accuracy while satisfying practical constraints, including sub-50 ms inference latency per frame, model sizes below 50 MB, and power consumption under 5 W on edge-class hardware. Together, these benchmarks shift the evaluation paradigm from isolated accuracy improvements toward holistic system-level performance, aligning future SLR research with real-world assistive and interactive applications and enabling more rigorous, reproducible comparisons across methods.

Although visual information remains the primary input for most Sign Language Recognition systems, reliance on RGB or depth data alone often limits semantic expressiveness and robustness. Sign languages convey meaning through a combination of manual gestures, non-manual markers (facial expressions and body posture), and linguistic context, creating a semantic gap between observed motion patterns and intended meaning. To address this gap, recent research trends point toward multimodal integration frameworks, where visual features are augmented with skeletal pose, facial landmarks, inertial sensor data, and explicit linguistic representations. Within such frameworks, multimodal Large Language Models (LLMs) offer a unifying abstraction by aligning heterogeneous sensory embeddings with language-level semantics through shared latent spaces and cross-modal attention mechanisms. Visual encoders (e.g., CNNs, GCNs, or Transformers) extract modality-specific features, which are then fused and projected into a language-aware representation space guided by pretrained linguistic knowledge. This integration enables higher-level reasoning, contextual disambiguation, and semantic consistency across long sign sequences. Consequently, multimodal LLM-driven architectures represent a promising direction for moving beyond gesture classification toward semantically grounded sign language understanding and translation, particularly in continuous and large-vocabulary SLR scenarios.

10. Conclusions

This review highlights the critical role of AI-driven pattern recognition systems in advancing SLR and improving communication accessibility for individuals with hearing and speech impairments. The study systematically examined state-of-the-art deep learning and machine learning models, identifying CNNs, RNNs, and hybrid neural architectures as leading approaches for gesture and speech pattern recognition. These models have demonstrated high accuracy in recognizing sign language, but their effectiveness is often constrained by data limitations, processing requirements, and adaptability across different linguistic and cultural contexts.

Despite notable progress, key challenges persist, including data scarcity, generalization issues, and real-time inference limitations. The lack of large, diverse datasets representing multiple sign languages hampers the ability of AI models to generalize across different users and regional variations. Additionally, the high computational cost of deep learning-based models poses a barrier to real-time SLR deployment on resource-constrained devices. Current solutions often require cloud-based computation, which introduces latency issues and limits accessibility in low-connectivity environments.

Author contributions

Yahia Said: Conceptualization, writing—original draft preparation, writing—review and editing, supervision, project administration, funding acquisition; Mohammad Barr: Methodology, formal analysis, data curation, writing—original draft preparation, writing—review and editing, visualization; Yazan A. Alsariera: Conceptualization, software, validation, resources; Ahmed A. Alsheikhy: Methodology, investigation, writing—review and editing, supervision. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare that they have not used any Artificial Intelligence (AI) tools in creating this article.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FPEJ-2026-3030-01".

Conflict of interest

The authors declare there is no conflict of interest.

References

1. T. Kim, J. Keane, W. R. Wang, H. Tang, J. Riggle, G. Shakhnarovich, et al., Lexicon-free fingerspelling recognition from video: data, models, and signer adaptation, *Comput. Speech Lang.*, **46** (2017), 209–232. https://doi.org/10.1016/j.csl.2017.05.009
2. M. A. Ahmed, B. B. Zaidan, A. A. Zaidan, M. M. Salih, M. M. B. Lakulu, A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017, *Sensors*, **18** (2018), 2208. https://doi.org/10.3390/s18072208
3. R. P. Cui, H. Liu, C. S. Zhang, A deep neural framework for continuous sign language recognition by iterative training, *IEEE T. Multimedia*, **21** (2019), 1880–1891. https://doi.org/10.1109/TMM.2018.2889563
4. A. A. Hosain, P. S. Santhalingam, P. Pathak, J. Košecká, H. Rangwala, Sign language recognition analysis using multimodal data, *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Washington, DC, USA, 2019, 203–210. https://doi.org/10.1109/DSAA.2019.00035
5. A. C. Duarte, Cross-modal neural sign language translation, In: *Proceedings of the 27th ACM International Conference on Multimedia*, New York: Association for Computing Machinery, 2019, 1650–1654. https://doi.org/10.1145/3343031.3352587
6. M. J. Cheok, Z. Omar, M. H. Jaward, A review of hand gesture and sign language recognition techniques, *Int. J. Mach. Learn. & Cyber.*, **10** (2019), 131–153. https://doi.org/10.1007/s13042-017-0705-5
7. Q. K. Xiao, Y. D. Zhao, W. Huan, Multi-sensor data fusion for sign language recognition based on dynamic Bayesian network and convolutional neural network, *Multimed. Tools Appl.*, **78** (2019), 15335–15352. https://doi.org/10.1007/s11042-018-6939-8
8. E. K. Kumar, P. V. V. Kishore, M. T. K. Kumar, D. A. Kumar, 3D sign language recognition with joint distance and angular coded color topographical descriptor on a 2-stream CNN, *Neurocomputing*, **372** (2020), 40–54. https://doi.org/10.1016/j.neucom.2019.09.059
9. J. Wu, R. Jafari, Wearable computers for sign language recognition, In: *Handbook of large-scale distributed computing in smart healthcare*, Cham: Springer, 2017, 379–401. https://doi.org/10.1007/978-3-319-58280-1_14
10. J. C. Shang, J. Wu, A robust sign language recognition system with multiple Wi-Fi devices, In: *Proceedings of the Workshop on Mobility in the Evolving Internet Architecture*, New York: Association for Computing Machinery, 2017, 19–24. https://doi.org/10.1145/3097620.3097624

11. J. F. Pu, W. G. Zhou, H. Q. Li, Iterative alignment network for continuous sign language recognition, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, 4160–4169. https://doi.org/10.1109/CVPR.2019.00429
12. Q. K. Xiao, M. Y. Qin, Y. T. Yin, Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people, *Neural Networks*, **125** (2020), 41–55. https://doi.org/10.1016/j.neunet.2020.01.030
13. W. Aly, S. Aly, S. Almotairi, User-independent American sign language alphabet recognition based on depth image and PCANet features, *IEEE Access*, **7** (2019), 123138–123150. https://doi.org/10.1109/access.2019.2938829
14. Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, P. Presti, American sign language recognition with the kinect, In: *Proceedings of the 13th international conference on multimodal interfaces*, New York: Association for Computing Machinery, 2011, 279–286. https://doi.org/10.1145/2070481.2070532
15. S. J. Wei, X. Chen, X. D. Yang, S. Cao, X. Zhang, A component-based vocabulary-extensible sign language gesture recognition framework, *Sensors*, **16** (2016), 556. https://doi.org/10.3390/s16040556
16. N. B. Ibrahim, H. Zayed, M. Selim, Advances, challenges and opportunities in continuous sign language recognition, *Journal of Engineering and Applied Sciences*, **15** (2019), 1205–1227. https://doi.org/10.36478/jeasci.2020.1205.1227
17. L. H. Zheng, B. Liang, A. L. Jiang, Recent advances of deep learning for sign language recognition, *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Sydney, NSW, Australia, 2017, 1–7. https://doi.org/10.1109/dicta.2017.8227483
18. T. Starner, J. Weaver, A. Pentland, Real-time american sign language recognition using desk and wearable computer based video, *IEEE T. Pattern Anal.*, **20** (1998), 1371–1375. https://doi.org/10.1109/34.735811
19. F.-S. Chen, C.-M. Fu, C.-L. Huang, Hand gesture recognition using a real-time tracking method and hidden Markov models, *Image Vision Comput.*, **21** (2003), 745–758. https://doi.org/10.1016/s0262-8856(03)00070-2
20. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *P. IEEE*, **86** (1998), 2278–2324. https://doi.org/10.1109/5.726791
21. E. Escobedo, L. Ramirez, G. Camara, Dynamic sign language recognition based on convolutional neural networks and texture maps, *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Rio de Janeiro, Brazil, 2019, 265–272. https://doi.org/10.1109/SIBGRAPI.2019.00043
22. S. Hayani, M. Benaddy, O. El Meslouhi, M. Kardouchi, Arab sign language recognition with convolutional neural networks, *2019 International Conference of Computer Science and Renewable Energies (ICCSRE)*, Agadir, Morocco, 2019, 1–4. https://doi.org/10.1109/iccsre.2019.8807586
23. Y. Q. Liao, P. W. Xiong, W. D. Min, W. Q. Min, J. H. Lu, Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks, *IEEE Access*, **7** (2019), 38044–38054. https://doi.org/10.1109/access.2019.2904749
24. P. Witoonchart, P. Chongstitvatana, Application of structured support vector machine backpropagation to a convolutional neural network for human pose estimation, *Neural Networks*, **92** (2017), 39–46. https://doi.org/10.1016/j.neunet.2017.02.005

25. G. Marin, F. Dominio, P. Zanuttigh, Hand gesture recognition with leap motion and kinect devices, *2014 IEEE International Conference on Image Processing (ICIP)*, Paris, France, 2014, 1565–1569. https://doi.org/10.1109/ICIP.2014.7025313
26. S. Stoll, N. C. Camgoz, S. Hadfield, R. Bowden, Text2Sign: towards sign language production using neural machine translation and generative adversarial networks, *Int. J. Comput. Vis.*, **128** (2020), 891–908. https://doi.org/10.1007/s11263-019-01281-2
27. S. M. He, Research of a sign language translation system based on deep learning, *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, Dublin, Ireland, 2019, 392–396. https://doi.org/10.1109/aiam48774.2019.00083
28. Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, *CVPR 2011*, Colorado Springs, CO, USA, 2011, 1385–1392. https://doi.org/10.1109/CVPR.2011.5995741
29. P. Q. Thang, N. T. Thuy, H. T. Lam, The SVM, SimpSVM and RVM on sign language recognition problem, *2017 Seventh International Conference on Information Science and Technology (ICIST)*, Da Nang, Vietnam, 2017, 398–403. https://doi.org/10.1109/ICIST.2017.7926792
30. S. Badillo, B. Banfai, F. Birzele, I. I. Davydov, L. Hutchinson, T. Kam-Thong, et al., An introduction to machine learning, *Clin. Pharmacol. Ther.*, **107** (2020), 871–885. https://doi.org/10.1002/cpt.1796
31. A. Wadhawan, P. Kumar, Sign language recognition systems: A decade systematic literature review, *Arch. Computat. Methods Eng.*, **28** (2021), 785–813. https://doi.org/10.1007/s11831-019-09384-2
32. Y. S. Abu-Mostafa, M. Magdon-Ismail, H.-T. Lin, *Learning from data*, New York: AMLBook, 2012.
33. Y. Lecun, Y. Bengio, G. Hinton, Deep learning, *Nature*, **521** (2015), 436–444. https://doi.org/10.1038/nature14539
34. D. Konstantinidis, K. Dimitropoulos, P. Daras, A deep learning approach for analyzing video and skeletal features in sign language recognition, *2018 IEEE International Conference on Imaging Systems and Techniques (IST)*, Krakow, Poland, 2018, 1–6. https://doi.org/10.1109/IST.2018.8577085
35. L. Rioux-Maldague, P. Giguère, Sign language fingerspelling classification from depth and color images using a deep belief network, *2014 IEEE Canadian Conference on Computer and Robot Vision*, 2014, 92–97. https://doi.org/10.1109/CRV.2014.20
36. S. Aly, B. Osman, W. Aly, M. Saber, Arabic sign language fingerspelling recognition from depth and intensity images, *2016 12th International Computer Engineering Conference (ICENCO)*, Cairo, Egypt, 2016, 99–104. https://doi.org/10.1109/ICENCO.2016.7856452
37. O. Koller, H. Ney, R. Bowden, Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, 3793–3802. https://doi.org/10.1109/CVPR.2016.412
38. N. C. Camgoz, S. Hadfield, O. Koller, R. Bowden, Subunets: End-to-end hand shape and continuous sign language recognition, In: *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, 3075–3084. https://doi.org/10.1109/iccv.2017.332
39. S. Yang, Q. Zhu, Video-based Chinese sign language recognition using convolutional neural network, *2017 IEEE 9th international conference on communication software and networks (ICCSN)*, Guangzhou, China, 2017, 929–934. https://doi.org/10.1109/iccsn.2017.8230247

40. S. Wang, D. Guo, W. G. Zhou, Z. J. Zha, M. Wang, Connectionist temporal fusion for sign language translation, In: *Proceedings of the 26th ACM international conference on Multimedia*, New York: Association for Computing Machinery, 2018, 1483–1491. https://doi.org/10.1145/3240508.3240671
41. A. Balayn, H. Brock, K. Nakadai, Data-driven development of virtual sign language communication agents, *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Nanjing, China, 2018, 370–377. https://doi.org/10.1109/ROMAN.2018.8525717
42. G. A. Rao, P. V. V. Kishore, Selfie sign language recognition with multiple features on adaboost multilabel multiclass classifier, *Journal of Engineering Science and Technology*, **13** (2018), 2352–2368.
43. M. C. Ariesta, F. Wiryana, Suharjito, A. Zahra, Sentence level Indonesian sign language recognition using 3D convolutional neural network and bidirectional recurrent neural network, *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*, Jakarta, Indonesia, 2018, 16–22. https://doi.org/10.1109/inapr.2018.8627016
44. D. Konstantinidis, K. Dimitropoulos, P. Daras, Sign language recognition based on hand and body skeletal data, *2018-3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Helsinki, Finland, 2018, 1–4. https://doi.org/10.1109/3dtv.2018.8478467
45. X. W. Jiang, Y. D. Zhang, Chinese sign language fingerspelling via six-layer convolutional neural network with leaky rectified linear units for therapy and rehabilitation, *J. Med. Imag. Health In.*, **9** (2019), 2031–2090. https://doi.org/10.1166/jmihi.2019.2804
46. H. B. D. Nguyen, H. N. Do, Deep learning for american sign language fingerspelling recognition system, *2019 26th International Conference on Telecommunications (ICT)*, Hanoi, Vietnam, 2019, 314–318. https://doi.org/10.1109/ict.2019.8798856
47. S. Ameen, S. Vadera, A convolutional neural network to classify American Sign Language fingerspelling from depth and colour images, *Expert Syst.*, **34** (2017), e12197. https://doi.org/10.1111/exsy.12197
48. O. K. Oyedotun, A. Khashman, Deep learning in vision-based static hand gesture recognition, *Neural Comput. & Applic.*, **28** (2017), 3941–3951. https://doi.org/10.1007/s00521-016-2294-8
49. C. Zimmermann, T. Brox, Learning to estimate 3D hand pose from single rgb images, *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, 4913–4921. https://doi.org/10.1109/ICCV.2017.525
50. D. Wu, L. Shao, Multimodal dynamic networks for gesture recognition, In: *Proceedings of the 22nd ACM international conference on Multimedia*, New York: Association for Computing Machinery, 2014, 945–948. https://doi.org/10.1145/2647868.2654969
51. J. Huang, W. G. Zhou, H. Q. Li, W. P. Li, Sign language recognition using 3d convolutional neural networks, *2015 IEEE International Conference on Multimedia and Expo (ICME)*, Turin, Italy, 2015, 1–6. https://doi.org/10.1109/icme.2015.7177428
52. A. Tang, K. Lu, Y. F. Wang, J. Huang, H. Q. Li, A real-time hand posture recognition system using deep neural networks, *ACM T. Intel. Syst. Tec.*, **6** (2015), 1–23. https://doi.org/10.1145/2735952
53. K. H. Li, Z. Y. Zhou, C. H. Lee, Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications, *ACM T. Access. Comput.*, **8** (2016), 1–23. https://doi.org/10.1145/2850421

54. J. Huang, W. G. Zhou, H. Q. Li, W. P. Li, Sign language recognition using real-sense, *2015 IEEE China summit and international conference on signal and information processing (ChinaSIP)*, Chengdu, China, 2015, 166–170. https://doi.org/10.1109/chinasip.2015.7230384
55. D. Guo, W. G. Zhou, A. Y. Li, H. Q. Li, M. Wang, Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation, *IEEE T. Image Process.*, **29** (2020), 1575–1590. https://doi.org/10.1109/tip.2019.2941267
56. N. Wang, Z. Y. Ma, Y. C. Tang, Y. Liu, Y. Li, J. W. Niu, An optimized scheme of mel frequency cepstral coefficient for multi-sensor sign language recognition, In: *Smart Computing and Communication*, Cham: Springer, 2017, 224–235. https://doi.org/10.1007/978-3-319-52015-5_23
57. T.-W. Chong, B.-G. Lee, American sign language recognition using leap motion controller with machine learning approach, *Sensors*, **18** (2018), 3554. https://doi.org/10.3390/s18103554
58. J. Q. Wang, T. Zhang, An ARM-based embedded gesture recognition system using a data glove, *The 26th Chinese Control and Decision Conference (2014 CCDC)*, Changsha, China, 2014, 1580–1584. https://doi.org/10.1109/ccdc.2014.6852419
59. A. Z. Shukor, M. F. Miskon, M. H. Jamaluddin, F. B. Ali, M. F. Asyraf, M. B. B. Bahar, A new data glove approach for Malaysian sign language detection, *Procedia Computer Science*, **76** (2015), 60–67. https://doi.org/10.1016/j.procs.2015.12.276
60. N. B. Ibrahim, M. M. Selim, H. H. Zayed, An automatic Arabic sign language recognition system (ArSLRS), *J. King Saud Univ.-Com.*, **30** (2018), 470–477. https://doi.org/10.1016/j.jksuci.2017.09.007
61. L. Pigou, S. Dieleman, P.-J. Kindermans, B. Schrauwen, Sign language recognition using convolutional neural networks, In: *Computer Vision-ECCV 2014 Workshops*, Cham: Springer, 2015, 572–578. https://doi.org/10.1007/978-3-319-16178-5_40
62. S. G. M. Almeida, F. G. Guimarães, J. A. Ramírez, Feature extraction in Brazilian Sign Language Recognition based on phonological structure and using RGB-D sensors, *Expert Syst. Appl.*, **41** (2014), 7259–7271. https://doi.org/10.1016/j.eswa.2014.05.024
63. B. Hisham, A. Hamouda, Arabic sign language recognition using Ada-Boosting based on a leap motion controller, *Int. J. Inf. Technol.*, **13** (2021), 1221–1234. https://doi.org/10.1007/s41870-020-00518-5
64. U. Farooq, M. S. M. Rahim, N. Sabir, A. Hussain, A. Abid, Advances in machine translation for sign language: approaches, limitations, and challenges, *Neural Comput. & Applic.*, **33** (2021), 14357–14399. https://doi.org/10.1007/s00521-021-06079-3
65. M. I. Sadek, M. N. Mikhael, H. A. Mansour, A new approach for designing a smart glove for Arabic Sign Language Recognition system based on the statistical analysis of the Sign Language, *2017 34th National Radio Science Conference (NRSC)*, Alexandria, Egypt, 2017, 380–388. https://doi.org/10.1109/NRSC.2017.7893499
66. N. Rossol, I. Cheng, A. Basu, A multisensor technique for gesture recognition through intelligent skeletal pose analysis, *IEEE T. Hum.-Mach. Syst.*, **46** (2016), 350–359. https://doi.org/10.1109/thms.2015.2467212
67. L. Quesada, G. López, L. Guerrero, Automatic recognition of the American sign language fingerspelling alphabet to assist people living with speech or hearing impairments, *J. Ambient Intell. Human. Comput.*, **8** (2017), 625–635. https://doi.org/10.1007/s12652-017-0475-7

68. S.-Z. Li, B. Yu, W. Wu, S.-Z. Su, R.-R. Ji, Feature learning based on SAE–PCA network for human gesture recognition in RGBD images, *Neurocomputing*, **151** (2015), 565–573. https://doi.org/10.1016/j.neucom.2014.10.086
69. O. Koller, H. Ney, R. Bowden, Deep learning of mouth shapes for sign language, *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago, Chile, 2015, 477–483. https://doi.org/10.1109/iccvw.2015.69
70. S. Kim, Y. Ban, S. Lee, Tracking and classification of in-air hand gesture based on thermal guided joint filter, *Sensors*, **17** (2017), 166. https://doi.org/10.3390/s17010166
71. T. X. Xu, D. An, Z. H. Wang, S. C. Jiang, C. N. Meng, Y. W. Zhang, et al., 3D joints estimation of the human body in single-frame point cloud, *IEEE Access*, **8** (2020), 178900–178908. https://doi.org/10.1109/access.2020.3027892
72. Y. M. Zhou, G. L. Jiang, Y. R. Lin, A novel finger and hand pose estimation technique for real-time hand gesture recognition, *Pattern Recogn.*, **49** (2016), 102–114. https://doi.org/10.1016/j.patcog.2015.07.014
73. M. A. Hossen, A. Govindaiah, S. Sultana, A. Bhuiyan, Bengali sign language recognition using deep convolutional neural network, *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, Kitakyushu, Japan, 2018, 369–373. https://doi.org/10.1109/iciev.2018.8640962
74. R. Rastgoo, K. Kiani, S. Escalera, M. Sabokrou, Sign language production: A review, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, 2021, 3446–3456. https://doi.org/10.1109/cvprw53098.2021.00384
75. A. Kratimenos, G. Pavlakos, P. Maragos, Independent sign language recognition with 3d body, hands, and face reconstruction, *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, 4270–4274. https://doi.org/10.1109/icassp39728.2021.9414278
76. D. Bansal, P. Ravi, M. So, P. Agrawal, I. Chadha, G. Murugappan, et al., Copycat: Using sign language recognition to help deaf children acquire language skills, *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2021, 1–10. https://doi.org/10.1145/3411763.3451523
77. K. Gajurel, C. C. Zhong, G. H. Wang, A fine-grained visual attention approach for fingerspelling recognition in the wild, *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Melbourne, Australia, 2021, 3266–3271. https://doi.org/10.1109/smc52423.2021.9658982
78. M. De Coster, M. V. Herreweghe, J. Dambre, Sign language recognition with transformer networks, In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Paris: European Language Resources Association, 2020, 6018–6024. https://aclanthology.org/2020.lrec-1.737/
79. L. Meng, R. H. Li, An attention-enhanced multi-scale and dual sign language recognition network based on a graph convolution network, *Sensors*, **21** (2021), 1120. https://doi.org/10.3390/s21041120
80. P. P. Roy, P. Kumar, B.-G. Kim, An efficient sign language recognition (SLR) system using camshift tracker and hidden markov model (HMM), *SN Comput. Sci.*, **2** (2021), 79. https://doi.org/10.1007/s42979-021-00485-z

81. N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, et al., A comprehensive study on deep learning-based methods for sign language recognition, *IEEE T. Multimedia*, **24** (2022), 1750–1762. https://doi.org/10.1109/tmm.2021.3070438

82. B. Saunders, N. C. Camgoz, R. Bowden, Continuous 3D multi-channel sign language production via progressive transformers and mixture density networks, *Int. J. Comput. Vis.*, **129** (2021), 2113–2135. https://doi.org/10.1007/s11263-021-01457-9

83. B. Sara, R. Akmeliawati, A. A. Shafie, M. J. El Salami, Modeling of human upper body for sign language recognition, *The 5th International Conference on Automation, Robotics and Applications*, Wellington, New Zealand, 2011, 104–108. https://doi.org/10.1109/icara.2011.6144865

84. M. Kuhn, K. Johnson, *Applied predictive modeling*, New York: Springer, 2013. https://doi.org/10.1007/978-1-4614-6849-3

85. M. Kuhn, K. Johnson, *Feature engineering and selection: A practical approach for predictive models*, London: Chapman and Hall/CRC, 2019.

86. A. J. Ferreira, M. A. Figueiredo, Efficient feature selection filters for high-dimensional data, *Pattern Recogn. Lett.*, **33** (2012), 1794–1804. https://doi.org/10.1016/j.patrec.2012.05.019

87. K. Yin, A. Moryossef, J. Hochgesang, Y. Goldberg, M. Alikhani, Including signed languages in natural language processing, In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, 2021, 7347–7360. https://doi.org/10.18653/v1/2021.acl-long.570

88. P. Escudeiro, N. Escudeiro, R. Reis, J. Lopes, M. Norberto, A. B. Baltasar, et al., Virtual sign–a real time bidirectional translator of portuguese sign language, *Procedia Computer Science*, **67** (2015), 252–262. https://doi.org/10.1016/j.procs.2015.09.269

89. J. Huang, W. G. Zhou, H. Q. Li, W. P. Li, Attention-based 3D-CNNs for large-vocabulary sign language recognition, *IEEE T. Circ. Syst. Vid.*, **29** (2019), 2822–2832. https://doi.org/10.1109/TCSVT.2018.2870740

90. K. Papadimitriou, G. Potamianos, Fingerspelled alphabet sign recognition in upper-body videos, *2019 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, 2019, 1–5. https://doi.org/10.23919/eusipco.2019.8902541

91. M. Ma, X. D. Xu, J. Wu, M. Guo, Design and analyze the structure based on deep belief network for gesture recognition, *2018 Tenth international conference on advanced computational intelligence (ICACI)*, Xiamen, China, 2018, 40–44. https://doi.org/10.1109/icaci.2018.8377544

92. S. M. Kamal, Y. D. Chen, S. Z. Li, X. D. Shi, J. B. Zheng, Technical approaches to Chinese sign language processing: A review, *IEEE Access*, **7** (2019), 96926–96935. https://doi.org/10.1109/access.2019.2929174

93. R.-H. Liang, M. Ouhyoung, A sign language recognition system using hidden markov model and context sensitive search, In: *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, New York: Association for Computing Machinery, 1996, 59–66. https://doi.org/10.1145/3304181.3304194

94. K. Grobel, M. Assan, Isolated sign language recognition using hidden Markov models, *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, Orlando, FL, USA, 1997, 162–167. https://doi.org/10.1109/icsmc.1997.625742

95. Z. Ghahramani, M. Jordan, Factorial hidden Markov models, In: *Advances in Neural Information Processing Systems*, **8** (1995).
96. L. Rimella, N. Whiteley, Hidden Markov neural networks, 2020, arXiv:2004.06963v3. https://doi.org/10.48550/arxiv.2004.06963
97. A. D. Wilson, A. F. Bobick, Parametric hidden markov models for gesture recognition, *IEEE T. Pattern Anal.*, **21** (1999), 884–900. https://doi.org/10.1109/34.790429
98. M. Brand, N. Oliver, A. Pentland, Coupled hidden markov models for complex action recognition, *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, PR, USA, 1997, 994–999. https://doi.org/10.1109/CVPR.1997.609450
99. E.-J. Ong, O. Koller, N. Pugeault, R. Bowden, Sign spotting using hierarchical sequential patterns with temporal intervals, *2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus*, Columbus, OH, USA, 2014, 1931–1938. https://doi.org/10.1109/CVPR.2014.248
100. Y. Bengio, P. Frasconi, Input-output HMMs for sequence processing, *IEEE T. Neural Networ.*, **7** (1996), 1231–1249. https://doi.org/10.1109/72.536317
101. A. Just, O. Bernier, S. Marcel, HMM and IOHMM for the recognition of mono-and bi-manual 3D hand gestures, In: *Proceedings of the British Machine Vision Conference*, BMVA Press, 2004, 1–10. https://doi.org/10.5244/c.18.28
102. C. Keskin L. Akarun, STARS: Sign tracking and recognition system using input-output HMMs, *Pattern Recogn. Lett.*, **30** (2009), 1086–1095. https://doi.org/10.1016/j.patrec.2009.03.016
103. N. J. Liu, B. C. Lovell, Gesture classification using hidden markov models and viterbi path counting, *VIIth Digital Image Computing: Techniques and Applications*, Sydney, Australia, 2003, 273–282.
104. M. Elmezain, A. Al-Hamadi, J. Appenrodt, B. Michaelis, A hidden markov model-based continuous gesture recognition system for hand motion trajectory, *2008 19th International Conference on Pattern Recognition*, Tampa, FL, USA, 2008, 1–4. https://doi.org/10.1109/icpr.2008.4761080
105. J. Appenrodt, A. Al-Hamadi, B. Michaelis, Data gathering for gesture recognition systems based on single color-, stereo color-and thermal cameras, *International Journal of Signal Processing, Image Processing and Pattern Recognition*, **3** (2010), 37–50. https://doi.org/10.1007/978-3-642-10509-8_10
106. M. M. Zaki, S. I. Shaheen, Sign language recognition using a combination of new vision based features, *Pattern Recogn. Lett.*, **32** (2011), 572–577. https://doi.org/10.1016/j.patrec.2010.11.013
107. P. V. Barros, N. T. Júnior, J. M. Bisneto, B. J. Fernandes, B. L. Bezerra, S. M. Fernandes, An effective dynamic gesture recognition system based on the feature vector reduction for SURF and LCS, In: *Artificial Neural Networks and Machine Learning–ICANN 2013*, Berlin: Springer, 2013, 412–419. https://doi.org/10.1007/978-3-642-40728-4_52
108. W. W. Yang, J. X. Tao, Z. F. Ye, Continuous sign language recognition using level building based on fast hidden Markov model, *Pattern Recogn. Lett.*, **78** (2016), 28–35. https://doi.org/10.1016/j.patrec.2016.03.030
109. S. Belgacem, C. Chatelain, T. Paquet, Gesture sequence recognition with one shot learned CRF/HMM hybrid model, *Image Vision Comput.*, **61** (2017), 12–21. https://doi.org/10.1016/j.imavis.2017.02.003

110. B. Y. Fang, J. Co, M. Zhang, Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation, In: *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, New York: Association for Computing Machinery, 2017, 1–13. https://doi.org/10.1145/3131672.3131693
111. D. C. Kavarthapu, K. Mitra, Hand gesture sequence recognition using inertial motion units (IMUs), *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, Nanjing, China, 2017, 953–957. https://doi.org/10.1109/acpr.2017.159
112. E. Rakun, A. M. Arymurthy, L. Y. Stefanus, A. F. Wicaksono, I. W. W. Wisesa, Recognition of sign language system for Indonesian language using long short-term memory neural networks, *Adv. Sci. Lett.*, **24** (2018), 999–1004. https://doi.org/10.1166/asl.2018.10675
113. S. S. Kumar, T. Wangyal, V. Saboo, R. Srinath, Time series neural networks for real time sign language translation, *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, USA, 2018, 243–248. https://doi.org/10.1109/ICMLA.2018.00043
114. D. M. Adimas, E. Rakun, D. Hardianto, Recognizing indonesian sign language gestures using features generated by elliptical model tracking and angular projection, *2019 2nd International Conference on Intelligent Autonomous Systems (ICoIAS)*, Singapore, 2019, 25–31. https://doi.org/10.1109/icoias.2019.00011
115. S. Kumar, R. Rani, S. K. Pippal, U. Chaudhari, Real time Indian sign language recognition using transfer learning with VGG16, *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, **22** (2024), 1459–1468. https://doi.org/10.12928/telkomnika.v22i6.26498
116. S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, et al., Chalearn looking at people challenge 2014: Dataset and results, *Computer Vision-ECCV 2014 Workshops*, Cham: Springer, 2015, 459–473. https://doi.org/10.1007/978-3-319-16178-5_32
117. J. Joy, K. Balakrishnan, M. Sreeraj, SignQuiz: A quiz based tool for learning fingerspelled signs in indian sign language using ASLR, *IEEE Access*, **7** (2019), 28363–28371. https://doi.org/10.1109/ACCESS.2019.2901863
118. K. Bantupalli, Y. Xie, American sign language recognition using deep learning and computer vision, *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 2018, 4896–4899. https://doi.org/10.1109/BigData.2018.8622141
119. H. A. AbdElghfar, A. M. Ahmed, A. A. Alani, H. M. AbdElaal, B. Bouallegue, M. M. Khattab, et al., A model for qur'anic sign language recognition based on deep learning algorithms, *J. Sensors*, **2023** (2023), 9926245. https://doi.org/10.1155/2023/9926245
120. S. Aly, W. Aly, DeepArSLR: A novel signer-independent deep learning framework for isolated arabic sign language gestures recognition, *IEEE Access*, **8** (2020), 83199–83212. https://doi.org/10.1109/ACCESS.2020.2990699
121. L.-Ch. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, 2017, arXiv:1706.05587. https://doi.org/10.48550/arXiv.1706.05587
122. K. Yin, J. Read, Attention is all you sign: sign language translation with transformers, In: *Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts*, 4 (2020).

123. N. C. Camgoz, O. Koller, S. Hadfield, R. Bowden, Sign language transformers: Joint end-to-end sign language recognition and translation, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, 10020–10030. https://doi.org/10.1109/CVPR42600.2020.01004

124. K. Yin, J. Read, Better sign language translation with STMC-transformer, 2020, arXiv:2004.00588. https://doi.org/10.48550/arXiv.2004.00588

125. M. De Coster, M. V. Herreweghe, J. Dambre, Isolated sign recognition from rgb video using pose flow and self-attention, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, 2021, 3436–3445. https://doi.org/10.1109/cvprw53098.2021.00383

126. I. Papastratis, K. Dimitropoulos, P. Daras, Continuous sign language recognition through a context-aware generative adversarial network, *Sensors*, **21** (2021), 2437. https://doi.org/10.3390/s21072437

127. T. Jiang, N. C. Camgoz, R. Bowden, Skeletor: Skeletal transformers for robust body-pose estimation, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, 2021, 3389–3397. https://doi.org/10.1109/cvprw53098.2021.00378

128. N. C. Camgöz, B. Saunders, G. Rochette, M. Giovanelli, G. Inches, R. Nachtrab-Ribback, et al., Content4all open research sign language translation datasets, *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, Jodhpur, India, 2021, 1–5. https://doi.org/10.1109/fg52635.2021.9667087

129. A. Moryossef, I. Tsochantaridis, J. Dinn, N. C. Camgoz, R. Bowden, T. Jiang, et al., Evaluating the immediate applicability of pose estimation for sign language recognition, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, 2021, 3429–3435. https://doi.org/10.1109/cvprw53098.2021.00382

130. D. M. Gavrila, The visual analysis of human movement: A survey, *Comput. Vis. Image Und.*, **73** (1999), 82–98. https://doi.org/10.1006/cviu.1998.0716

131. H. L. Ribeiro, A. Gonzaga, Hand image segmentation in video sequence by GMM: A comparative analysis, *2006 19th Brazilian Symposium on Computer Graphics and Image Processing*, Amazonas, Brazil, 2006, 357–364. https://doi.org/10.1109/sibgrapi.2006.23

132. T. B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, *Comput. Vis. Image Und.*, **81** (2001), 231–268. https://doi.org/10.1006/cviu.2000.0897

133. S. S. Rautaray, A. Agrawal, Vision based hand gesture recognition for human computer interaction: A survey, *Artif. Intell. Rev.*, **43** (2015), 1–54. https://doi.org/10.1007/s10462-012-9356-9

134. M. Mohandes, M. Deriche, J. Liu, Image-based and sensor-based approaches to Arabic sign language recognition, *IEEE T. Hum.-Mach. Syst.*, **44** (2014), 551–557. https://doi.org/10.1109/THMS.2014.2318280

135. G. Kumar, P. K. Bhatia, A detailed review of feature extraction in image processing systems, *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, Rohtak, India, 2014, 5–12. https://doi.org/10.1109/acct.2014.74

136. S. Y. Jiang, B. Sun, L. C. Wang, Y. Bai, K. P. Li, Y. Fu, Skeleton aware multi-modal sign language recognition, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, 2021, 3408–3418. https://doi.org/10.1109/cvprw53098.2021.00380

137. H. Zhou, W. G. Zhou, Y. Zhou, H. Q. Li, Spatial-temporal multi-cue network for continuous sign language recognition, *Proceedings of the AAAI conference on artificial intelligence*, **34** (2020), 13009–13016. https://doi.org/10.1609/aaai.v34i07.7001
138. R. Rastgoo, K. Kiani, S. Escalera, Sign language recognition: A deep survey, *Expert Syst. Appl.*, **164** (2021), 113794. https://doi.org/10.1016/j.eswa.2020.113794
139. A. Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling, S. Narayanan, Real-time sign language detection using human pose estimation, In: *Computer Vision–ECCV 2020 Workshops*, Cham: Springer, 2020, 237–248. https://doi.org/10.1007/978-3-030-66096-3_17
140. R. Verma, S. Mittal, S. Pawar, M. Sharma, S. Goel, V. H. C. de Albuquerque, Automatic rigging of 3D models with stacked hourglass networks and descriptors, *AIP Conf. Proc.*, **2919** (2024), 050006. https://doi.org/10.1063/5.0184393
141. S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, 2*016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, 4724–4732. https://doi.org/10.1109/CVPR.2016.511
142. A. Toshev, C. Szegedy, DeepPose: Human pose estimation via deep neural networks, *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, 1653–1660. https://doi.org/10.1109/CVPR.2014.214
143. S. Gattupalli, A. Ghaderi, V. Athitsos, Evaluation of deep learning based pose estimation for sign language recognition, In: *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, New York: Association for Computing Machinery, 2016, 1–7. http://dx.doi.org/10.1145/2910674.2910716
144. S.-K. Ko, C. J. Kim, H. Jung, C. Cho, Neural sign language translation based on human keypoint estimation, *Appl. Sci.*, **9** (2019), 2683. https://doi.org/10.3390/app9132683
145. M. Madadi, H. Bertiche, S. Escalera, SMPLR: Deep learning based SMPL reverse for 3D human pose and shape recovery, *Pattern Recogn.*, **106** (2020), 107472. https://doi.org/10.1016/j.patcog.2020.107472
146. A. Jain, J. Tompson, Y. LeCun, C. Bregler, Modeep: A deep learning framework using motion features for human pose estimation, In: *Computer Vision--ACCV 2014*, Cham: Springer, 2015, 302–315. https://doi.org/10.1007/978-3-319-16808-1_21
147. X. J. Chen, A. L. Yuille, Articulated pose estimation by a graphical model with image dependent pairwise relations, 2014, arXiv:1407.3399. https://doi.org/10.48550/arXiv.1407.3399
148. J. Charles, T. Pfister, D. Magee, D. Hogg, A. Zisserman, Personalizing human video pose estimation, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, 3063–3072. https://doi.org/10.1109/CVPR.2016.334
149. Y. R. Bin, Z.-M. Chen, X.-S. Wei, X. Y. Chen, C. X. Gao, N. Sang, Structure-aware human pose estimation with graph convolutional networks, *Pattern Recogn.*, **106** (2020), 107410. https://doi.org/10.1016/j.patcog.2020.107410
150. A. Haque, B. Peng, Z. L. Luo, A. Alahi, S. Yeung, F.-F. Li, Towards viewpoint invariant 3d human pose estimation, *Computer Vision–ECCV 2016*, **14** (2016), 160–177. https://doi.org/10.1007/978-3-319-46448-0_10
151. M. Wang, X. P. Chen, W. T. Liu, C. Qian, L. Lin, L. Z. Ma, Drpose3D: Depth ranking in 3D human pose estimation, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, 978–984. https://doi.org/10.24963/ijcai.2018/136

152. M. J. Marin-Jimenez, F. J. Romero-Ramirez, R. Munoz-Salinas, R. Medina-Carnicer, 3D human pose estimation from depth maps using a deep combination of poses, *J. Vis. Commun. Image R.*, **55** (2018), 627–639. https://doi.org/10.1016/j.jvcir.2018.07.010

153. Q. Dang, J. Q. Yin, B. Wang, W. Q. Zheng, Deep learning based 2D human pose estimation: A survey, *Tsinghua Sci. Technol.*, **24** (2019), 663–676. https://doi.org/10.26599/TST.2018.9010100

154. X. P. Ji, Q. Fang, J. T. Dong, Q. Shuai, W. Jiang, X. W. Zhou, A survey on monocular 3D human pose estimation, *Virtual Reality & Intelligent Hardware*, **2** (2020), 471–500. https://doi.org/10.1016/j.vrih.2020.04.005

155. L. Pigou, M. V. Herreweghe, J. Dambre, Gesture and sign language recognition with temporal residual networks, *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, Italy, 2017, 3086–3093. https://doi.org/10.1109/iccvw.2017.365

156. R. P. Cui, H. Liu, C. S. Zhang, Recurrent convolutional neural networks for continuous sign language recognition by staged optimization, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, 1610–1618. https://doi.org/10.1109/CVPR.2017.175

157. G. Latif, N. Mohammad, J. Alghazo, R. AlKhalaf, R. AlKhalaf, ArASL: Arabic alphabets sign language dataset, *Data Brief*, **23** (2019), 103777. https://doi.org/10.1016/j.dib.2019.103777

158. A. I. Shahin, S. Almotairi, Automated Arabic sign language recognition system based on deep transfer learning, *Int. J. Comput. Sci. Net.*, **19** (2019), 144–152.

159. Suharjito, H. Gunawan, N. Thiracitta, A. Nugroho, Sign language recognition using modified convolutional neural network model, *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*, Jakarta, Indonesia, 2018, 1–5. https://doi.org/10.1109/inapr.2018.8627014

160. B. Mocialov, G. Turner, K. Lohan, H. Hastie, Towards continuous sign language recognition with deep learning, *Proceedings of the workshop on the creating meaning with robot assistants*, 2017, 5525834.

161. V. Belissen, Sign language video analysis for automatic recognition and detection, *14th IEEE International Conference on Automatic Face and Gesture Recognition*, Lille, France, 2019, 1–5.

162. A. Sabyrov, M. Mukushev, V. Kimmelman, Towards real-time sign language interpreting robot: evaluation of non-manual components on recognition accuracy, *CVPR Workshops*, 2019, 75–82.

163. A. T. Magar, P. Parajuli, American sign language recognition using convolution neural network, *Undergraduate Research Conference (URC 2020)*, 2020, 65–80.

164. Q. F. Xue, X. P. Li, D. Wang, W. G. Zhang, Deep forest-based monocular visual sign language recognition, *Appl. Sci.*, **9** (2019), 1945. https://doi.org/10.3390/app9091945

165. K. M. Lim, A. W. Tan, C. P. Lee, S. C. Tan, Isolated sign language recognition using convolutional neural network hand modelling and hand energy image, *Multimed. Tools Appl.*, **78** (2019), 19917–19944. https://doi.org/10.1007/s11042-019-7263-7

166. M. E. M. Cayamcela, W. S. Lim, Fine-tuning a pre-trained convolutional neural network model to translate American sign language in real-time, *2019 International Conference on Computing, Networking and Communications (ICNC)*, Honolulu, HI, USA, 2019, 100–104. https://doi.org/10.1109/ICCNC.2019.8685536

167. M. Taskiran, M. Killioglu, N. Kahraman, A real-time system for recognition of American sign language by using deep learning, *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, Athens, Greece, 2018, 1–5. https://doi.org/10.1109/TSP.2018.8441304

168. R. Daroya, D. Peralta, P. Naval, Alphabet sign language image classification using deep learning, *TENCON 2018-2018 IEEE Region 10 Conference*, Jeju, Korea (South), 2018, 0646–0650. https://doi.org/10.1109/tencon.2018.8650241

169. M. A. Jalal, R. L. Chen, R. K. Moore, L. Mihaylova, American sign language posture understanding with deep neural networks, *2018 21st International Conference on Information Fusion (FUSION)*, Cambridge, UK, 2018, 573–579. https://doi.org/10.23919/icif.2018.8455725

170. T. Aujeszky, M. Eid, A gesture recogintion architecture for Arabic sign language communication system, *Multimed. Tools Appl.*, **75** (2016), 8493–8511. https://doi.org/10.1007/s11042-015-2767-2

171. M. Elpeltagy, M. Abdelwahab, M. E. Hussein, A. Shoukry, A. Shoala, M. Galal, Multi-modality-based Arabic sign language recognition, *IET Comput. Vis.*, **12** (2018), 1031–1039. https://doi.org/10.1049/iet-cvi.2017.0598

172. S. Islam, S. S. S. Mousumi, A. S. A. Rabby, S. A. Hossain, S. Abujar, A potent model to recognize bangla sign language digits using convolutional neural network, *Procedia Computer Science*, **143** (2018), 611–618. https://doi.org/10.1016/j.procs.2018.10.438

173. C. S. Mao, S. L. Huang, X. X. Li, Z. F. Ye, Chinese sign language recognition with sequence to sequence learning, *Communications in Computer and Information Science*, (2017), 180–191. https://doi.org/10.1007/978-981-10-7299-4_15

174. T. D. Sajanraj, M. Beena, Indian sign language numeral recognition using region of interest convolutional neural network, *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, India, 2018, 636–640. https://doi.org/10.1109/icicct.2018.8473141

175. N. Soodtoetong, E. Gedkhaw, The efficiency of sign language recognition using 3D convolutional neural networks, *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Chiang Rai, Thailand, 2018, 70–73. https://doi.org/10.1109/ECTICon.2018.8619984

176. P. Nakjai, P. Maneerat, T. Katanyukul, Thai finger spelling localization and classification under complex background using a YOLO-based deep learning, In: *Proceedings of the 11th International Conference on Computer Modeling and Simulation*, New York: Association for Computing Machinery, 2019, 230–233. https://doi.org/10.1145/3307363.3307403

177. P. F. Sun, F. Chen, G. J. Wang, J. S. Ren, J. W. Dong, A robust static sign language recognition system based on hand key points estimation, *Intelligent Systems Design and Applications: 17th International Conference*, (2018), 548–557. https://doi.org/10.1007/978-3-319-76348-4_53

178. R. Alzohairi, R. Alghonaim, W. Alshehri, S. Aloqeely, Image based Arabic sign language recognition system, *Int. J. Adv. Comput. Sc.*, **9** (2018), 090327. https://doi.org/10.14569/ijacsa.2018.090327

179. R. Rastgoo, K. Kiani, S. Escalera, Multi-modal deep hand sign language recognition in still images using restricted Boltzmann machine, *Entropy*, **20** (2018), 809. https://doi.org/10.3390/e20110809

180. G. Devineau, F. Moutarde, W. Xi, J. Yang, Deep learning for hand gesture recognition on skeletal data, *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*, Xi'an, China, 2018, 106–113. https://doi.org/10.1109/FG.2018.00025

181. R. Alzohairi, R. Alghonaim, W. Alshehri, S. Aloqeely, Image based Arabic sign language recognition system, *Int. J. Adv. Comput. Sc.*, **9** (2018), 090327. https://doi.org/10.14569/IJACSA.2018.090327

182. G. A. Rao, K. Syamala, P. V. V. Kishore, A. S. C. S. Sastry, Deep convolutional neural networks for sign language recognition, *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)*, Vijayawada, India, 2018, 194–197. https://doi.org/10.1109/SPACES.2018.8316344

183. S. Shahriar, A. Siddiquee, T. Islam, A. Ghosh, R. Chakraborty, A. I. Khan, et al., Real-time american sign language recognition using skin segmentation and image category classification with convolutional neural network and deep learning, *TENCON 2018-2018 IEEE Region 10 Conference*, Jeju, Korea (South), 2018, 1168–1171. https://doi.org/10.1109/tencon.2018.8650524