The Seybold Report

ISSN 1533-9211

DOI 10.5281/zenodo.7353683

ISOLATED AND CONTINUOUS HAND GESTURE RECOGNITION BASED ON DEEP LEARNING: A REVIEW

BARAA WASFI SALIM*

ITM Dept., Technical College of Administration, Duhok Polytechnic University, Duhok, Iraq.
Corresponding Author Email: baraa.salim@dpu.edu.krd

SUBHI R. M. ZEEBAREE

Energy Eng. Dept., Technical College of Engineering, Duhok Polytechnic University, Duhok, Iraq
Email: subhi.rafeeq@dpu.edu.krd

Abstract

Getting to know sign language is of great research importance as it affects the lives of deaf and mute people and societies in general. The rapid development of deep learning techniques presents new horizons in sign language recognition because it can give more accurate results and deal with large amounts of data. This paper provides an overview of sign language recognition systems that use deep learning as a basis. A review of recent studies in this field and the division of recognition systems into continuous and isolated and the algorithms used in both methods: Recurrent Neural Network (RNN) based method, Convolutional Neural Network (CNN), and Three-Dimensional Convolutional Neural Network (3D-CNN), in addition to the challenges in systems, Identify the signal, problems, and prospects.

Keywords: Deep learning; Recognition of sign language; isolated identification, Continuous identification, CNN, RNN, LSTM.

1. Introduction

Healthy people can use oral language to communicate adequately, while hearing impairments (deaf, aphasia, etc.) need sign language to express their thoughts. Since most healthy people have not learned sign language, it is necessary to strengthen sign language and make it suitable for ordinary society [1]. There are barriers to this kind of communication. Sign language recognition and translation technology facilitate communication between hard of hearing and healthy people [2]. Sign language research should not only allow hearing-impaired people to read sign language, but also allow hearing-impaired people to understand what healthy people are saying [3]. Sign language recognition and translation are the first, and sign language creation research belongs to the second. For people with hearing impairments, this interaction process is especially important, therefore, research into sign language recognition, translation, and sign language generation has important theoretical and practical as well as social significance [4].

In recent years, with the rapid development of videoconferencing, human-computer interaction and virtual reality technology, the research on video-based sign language recognition, translation, and generation has received more attention in the world. It is an important topic in pattern recognition [5]. Since sign language research involves video comprehension, gesture recognition, action recognition, video description creation, and vision generation, its development has implications for video processing, computer vision, human-computer

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

interaction, pattern recognition, natural language processing, etc. Several areas of research are of reference importance [6]. Sign language research mainly includes three aspects: separate sign language recognition, continuous sign language translation, and sign language generation. Sign language recognition and translation aim to convert sign language video into text words or sentences, and sign language generation is the creation of a synthetic video based on natural language or spoken language sentences [7]. Sign language recognition and translation and sign language creation can be two opposite processes.

Traditional methods can solve the problem of sign language recognition within a certain data volume. Still, the algorithms are complex, generalization is low, and the amount of data and types of patterns addressed to it are limited, making it impossible to express the intelligent human understanding of sign language fully. As a result, sign language recognition technology based on deep learning and human vision and perception rules has become unavoidable in this fast big data development era. The deep learning networks and algorithms used vary. In general, improvements are made from three parts: data entry, network architecture, and integration method. This paper identifies isolated and continuous sign language based on deep learning. In addition to the challenges and the direction of the future development of sign language.

2. Sign language recognition systems based on deep learning

As illustrated in Figure 1, the gesture recognition method may be divided into various stages: picture frame capture, hand tracking, feature extraction, and classification. The input for static gesture recognition is single frames of photos, but the input for dynamic sign languages is video, continuous frames of images [8]. Data collection distinguishes deep learning-based techniques from sensor-based approaches. This section focuses on deep learning-based vision-based gesture recognition research methodologies.

Figure (1): Sign language recognition systems

2.1. Image frame acquisition

The data obtained in vision-based gesture recognition is a frame of pictures. Image capture equipment such as a basic camcorder, webcam, stereo camera, thermal camera, or more modern active technologies such as Kinect and LMC are used to acquire data for this system. 3D cameras that capture depth information include stereo, Kinect, and LMC [9]. Sensor-based identification is defined in this study as any data-collecting approach that does not involve cameras.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

2.2. Hand Tracking

Because both tracking and segmentation are hand extractions from the backdrop, hand tracking is a segmentation component. The movement of the hand can be quite quick, and its appearance might vary considerably in only a few frames, making it difficult to track [10].

2.3. Feature extraction

Feature extraction converts important elements of an input data set into feature vector combinations. In the case of gesture recognition, the extracted features must include all important information from the hand gesture input and be represented in a compressed version that acts as the gesture's identity, allowing it to be distinguished from other gestures. In the existing deep learning recognition systems, this stage is part of the basis of the work of the designed neural network [11]. For example, in convolutional neural networks, all the features are entered, and the first layers of the network extract the features.

2.4. Classification

At this stage, the signals are classified into words, letters, or sentences according to the work of each system. The classification process in deep learning is usually done using the final layers of the convolutional neural network in a full-link network [12]. As a final output of this process, it is a value that is later represented in text or sound for the required letter, word, or sentence.

2.5 Suggested Methodology

Sign language recognition systems were reviewed as a first stage and the challenges facing sign language recognition. In this study, a field survey of the literature that dealt with sign language discrimination using deep learning was conducted, and the discrimination systems were separated on the basis of discrimination (isolated and continuous). In this research, 23 research papers were studied for isolated discrimination systems that used deep learning. On the other hand, 13 research papers were studied. It is noted that the number of studies that distinguish continuous sign language is less common. The comparison between systems was made on the basis of the data set used, algorithms and methods used by researchers, the language that was distinguished, the type of system in terms of static or dynamic, and the adoption of accuracy as a basis for the efficiency of the system. After comparing discrimination systems, the problems, obstacles and recommendations for future work were summarized, in addition to the conclusion reached by the study.

3. Issues and Challenges in Sign Language Recognition Research

The main sign language research group is deaf and hard of hearing people (including aphasia, deaf people, etc.). While considering technological innovation, the applicability of Artificial Intelligence technology to sign language recognition today and its practical application in the aphasia community cannot be ignored. Successful sign language research requires understanding the culture, background, and living environment of aphasia and creating a sign language application system that matches the user's region, age, gender, education level, type

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

of sign language used, and language proficiency. The challenges faced by sign language recognition systems can be summarized in the following points [13].

3.1. The video medium used

Video is the main carrier of sign language display content. Background video data collected by current sign language research is clean and includes only sign language procedures demonstrated by sign language providers in the camera area with a typical background in a lab environment. While in reality, there are challenges facing recognition systems such as noise and changing background that pose a challenge to isolate the part of the body required for recognition [4]. The complexity of the video scene reduces the recognition accuracy, and slight fluctuations can give different meanings. In addition to the number of frames, shifting from one syllable to another, and finding the point of distinction between one gesture and another, according to the different cast, is critical for the systems and a semantic deficit. That is why these characteristics are considered a challenge to sign language recognition systems in real-time [13].

3.2. Translation of sentences under weak supervision

Sign language is based on the time segmentation relationship and the transfer relationship between gestures. Hash relationship refers to the reasonable stopping point between independent words or phrases in a sentence; the transfer relationship refers to the formation of gestures according to grammatical rules. Due to the lack of professional knowledge of sign language and the high cost of annotations by most commentators, current sign language datasets typically contain only sentence-level labels without providing accurate annotations. This also poses higher requirements for capturing the consistency of Spatio-temporal detail in videos [13, 14].

3.3. Special language restrictions

Similar to the grammar of the spoken language, sign language consists of a series of actions according to the rules of semantic constraints. For example, the moving direction of gestures has the grammatical function of indicating the subject and object; it may express multiple parts of speech or semantics such as nouns and verbs. Changes in the head, hand shape, and posture are the main manifestations of sign language verbs. Semantic constraints on the head include head movements, facial expressions, mouth shape changes, shoulder jerks, eye gaze, and other subtle actions expressed in sign language. Basic semantic elements. As language habits and grammar differ across languages, there are also differences in the language restrictions of sign languages in different countries. For example, the same procedure has different meanings in different languages. In addition, in practical application scenarios, different parts of speech in natural language are often treated differently. It can be seen that sign language research needs to fully consider knowledge Background to the Semantic Constraints of Sign Language to effectively solve the above semantic problems and challenges in Sign Language Linguistics [14, 15].

ISSN 1533-9211

DOI 10.5281/zenodo.7353683

3.4. Multimedia dynamic sequence

Single-modal data has some natural flaws, such as color images being easily affected by light, angle, etc.; Depth data lacks details of a finger, face, etc. Effective multimodal data integration helps compensate for the shortcomings of -input modeling. Currently, the research direction presents more work on discrete sign language recognition and less work-related to continuous sign language translation [13].

3.5. Expand learning vocabulary

Nowadays, sign language research is moving from a small task of synthetic vocabulary to understanding sentences that involve extensive vocabulary in the real world. To apply sign language technology to various real-life scenarios, it is necessary to explore sign language recognition and translation methods for a wide vocabulary. In addition, a good sign language recognition system needs a certain amount of flexibility and power and can provide reasonable approximate semantic inferences for new, unseen words [16].

4. Literature Review

Recognition of sign language can be achieved by using either isolated or continuous approaches.

4.1. Sign language recognition of isolated words based on deep learning

The objective of learning about isolated sign language is the vocabulary of isolated sign language expressed in the video. Compared with continuous sentences, isolated sign language videos have a shorter duration and simple and clear connotations [17]. The main focus is on how to describe the basic features of the word sign language more effectively and reduce the rate of misjudgment. The technological approaches for processing time-series data are grouped into three primary frameworks: convolutional neural networks, three-dimensional convolutional neural networks, and recurrent neural networks. Research in this area is relatively active [18]. This part summarizes the most important literature on identifying isolated sign language.

4.1.1. 2D Convolutional Neural Networks

Convolutional neural networks are one of the most widely used techniques in computer vision of deep learning techniques. Researchers have conducted many kinds of research that included the use of convolutional neural networks to distinguish sign language for isolated words or letters, including depth, skeleton, basic body points, etc., and focuses on the features of Manual mode, integrating features and other related optimization strategies, to achieve better discrimination results. Data entry mainly involves the integration of the data layer before entering the network. This fusion method is highly dependent on the data set, and the pre-processing is more complex, but it also improves the recognition accuracy to a certain extent. Convolutional Neural Networks [32, 22, 28, 34, and 35] are used in many sign language recognition systems. Datasets contain a set of images of sign language words and semantics. CNN's have proven to be accurate in identifying this type of image. [23] Proposed the Bengali

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

Sign Language recognition method based on multimedia data and using DCNN that captures image features at different levels at multiple scales to learn manual image features automatically. In [24], it is proposed to use ROI segmentation pre-processing of input data through an object detection network. Hand area detection is implemented through the YOLO network, and sign language learning is implemented through Convolutional Neural Network (CNN). In [41] a model of convolutional neural networks based on collective research in recognizing gestures related to the HCI domain was implemented. This is followed by implementing the Collective Search Algorithm (CSA) to choose the optimal hyper parameters for training the data set using convolutional neural networks. In [39], the connected component analysis algorithm was used to divide the fingertips from the image of the hand. The segmented finger regions of the hand image are given to a CNN classification algorithm that categorizes the image into different categories. In [26] they proposed a framework based on deep learning, where a CNN and skeleton (body, hand and face) were used where each signed video corresponds to a single word video. In [27], it is proposed to recognize Indian Sign Language gestures using Convolutional Neural Networks (CNN) for continuous sign language video in Selfie mode. A hard of hearing person can operate the SLR mobile application independently. In [29] a system is built using KINECT as a sensor for image capture, image segmentation, feature detection and extraction from ROI, supervised and unsupervised classification of images using CNN and (KNN) and text-to-speech algorithms. In [36] the gesture recognition of static ASL using deep learning and CNN has been proposed. The contribution consists of two solutions to the problem. The first was resized using ASL still bi-cubic images. In addition, good recognition results in boundary hand detection using Robert's edge detection method. In a study [38], CNN was applied to (NUS hand position data set and American fingerspelling data set) to recognize hand gestures.

4.1.2. 3D convolutional network

Although CNN has strong feature extraction ability, it is limited to single-frame image data input. Sign language recognition also needs the help of some methods to remove correlation between frames. 3DCNN can synchronously model the temporal-spatial features of sign language. At the video, level to capture motion information obtained from multiple consecutive frames for more general identification. Various fusion methods, spatiotemporal attention mechanism, and dual-stream 3D-CNN are the three major achievements in sign language recognition research based on 3D convolutional networks. The feature fusion method is different from data fusion, which is a convolutional layer fusion after multimodal data feature extraction. This fusion method is widely used. In [19], the researchers proposed to use an adaptive frame unification strategy to standardize the input frame number. Then the RGB and depth data are sent to the C3D model to extract the spatiotemporal features, respectively. To mitigate the overlap of non-gesture-related factors. In the study [20], a multimodal gesture recognition method is proposed based on the ResC3D network. One of the main ideas is to create a compact and efficient representation of video sequences. Therefore, video optimization techniques, such as average filter, are applied to remove the light and noise contrast in the input video. The weighted frame standardization strategy is used to sample key frames. In paper [25], they developed an assembly method for exploring deep Spatio-temporal features using 3D

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

convolutional neural networks (CNN) with residual architecture (Res-C3D) and building a time series model with structural information based on a long-term memory network (LSTM). In paper [26] they developed an assembly method for exploring deep Spatio-temporal features using 3D convolutional neural networks (CNN) with residual architecture (Res-C3D) and building a time series model with structural information based on a long-term memory network (LSTM). By capturing skeletal information (key points) for the reserved body and arm regions while eliminating other parts, masked Res-C3D is obtained, reducing background effect and other differences, as gestures are primarily derived from arm or hand movements. Moreover, the weights of each voting subclassified as a particular class feature in our ensemble model are obtained adaptively by training rather than fixed weights. In [21], a deep behavior trait extractor was used to deal with the small details of Arabic Sign Language. And the use of a 3D convolutional neural network (CNN) to identify 25 gestures from the Arabic Sign Language Dictionary.

4.1.3. Recurrent Neural Network

Recurrent neural networks (RNNs) are time-series data-processing neural networks that collect long-range contextual semantic information. Therefore, in recent years, the research on sign language recognition based on a recurrent neural network has been intense. A study [30] proposes classifying 60 ASL tags based on data provided by a Leap Motion sensor using a deep learning model called Deep Conv LSTM that integrates convolutional and recursive layers with long-term memory cells. A kinematic model of the right and left forearm/hand/fingers/thumb and the use of a simple data augmentation technique have been proposed to improve the generalization of neural networks. In an introduction paper [31], two deep learning architecture models are proposed. The first model consists of a convolutional neural network (CNN) and a recurrent neural network with long-term memory (RNN-LSTM). The second model includes two parallel convolutional neural networks combined by a merging layer and a recurrent neural network with long-term memory fed by RGB-D. The researchers [33] suggested using the tandem CNN + RNN to recognize the series of gesture images. In [40, 37], it is proposed to use RNN and a pre-trained CNN model (VGG16) to recognize static and dynamic hand gestures.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

The isolated word sign language recognition technology and representative work based on deep learning are shown in Table 1.

Table (1): Comparison of isolated word sign language recognition for previous works.

#	Author	year	Static/ dynamic	Language	Method	Accuracy
1	Yunan et al [19].	2017	Static	China	based on saliency features, RGB-D and SVM	59.43%
2	Qiguang et al [20].	2017	dynamic	China	Based on multimodal data, hand feature enhancement, and CNN	67.71%
3	ElBadawy et al [21].	2017	static	Arabic	3D convolutional neural networks	98%
4	Oyebade et al [22].	2017	Static	American	CNN	92.83 %
5	Hossen et al [23].	2018	static	Bengali	CNN	96.33%
6	Kim et al [24].	2018	static	China	based ROI segmentation, YOLO and CNN	98%
7	Lin et al [25].	2018	static	American	Skeleton and CNN	68 %
8	Konstantinidis et al [26].	2018	static	Argentine	Skeleton and CNN	99.8%
9	Anantha et al [27].	2018	Static	Indian	Deep Convolutional Neural Networks	92.88 %
10	Murat et al [28].	2018	dynamic	American	CNN	98.05
11	Victoria et al [29].	2018	Static	American	FAST and SURF with a KNN	78%
12	Hernandez et al [30].	2020	dynamic	American	RNN and CNN	91%
13	Falah et al [31].	2020	dynamic	American	Deep Convolutional and Recurrent Neural Networks	82.8 %
14	Reshna et al [32].	2020	Static	Indian	CNN	95%
15	Kenneth et al [33].	2020	dynamic	American	CNN+RNN Depth and Skeleton	85.46%
16	Lance et al [34].	2020	dynamic	American	CNN	99.98
17	Aishwarya et al [35].	2020	Static	American	CNN	95%
18	Abdulwahab et al [36].	2020	Static	American	CNN	99.3 %
19	Rajalakshmi et al. [37].	2020	Static/ dynamic	American	CNN and RNN	88%
20	Adithya et al [38].	2020	Static	American	CNN	99.96 %
21	Neethu et al [39].	2020	Static	American	CNN	96.2%
22	Abul et al [40].	2021	Static	American	CNN+SVM	99.82%
23	Thippa et al. [41].	2021	Static	American	CNN-crow search	100%

As shown in Table 1, it is noted that most of the modern systems for the years between 2017 and 2022 have used convolutional neural networks in the cases of static and dynamic discrimination systems. Figure (2) compares static and dynamic systems using CNN and the adoption of a precision scale.

ISSN 1533-9211

DOI 10.5281/zenodo.7353683

Figure (2): Comparison based on method and system type (static or dynamic)

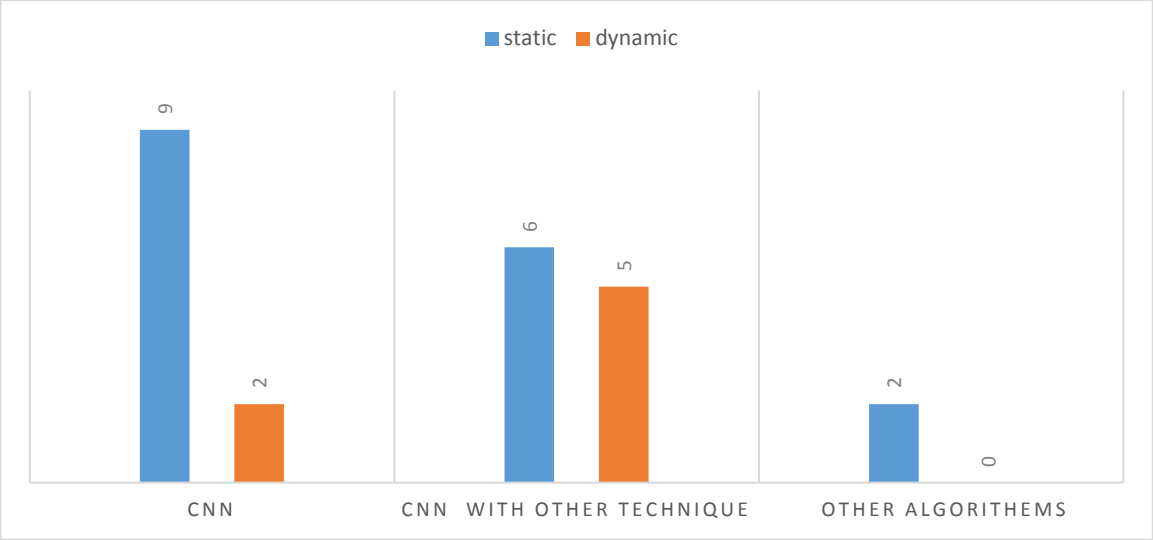
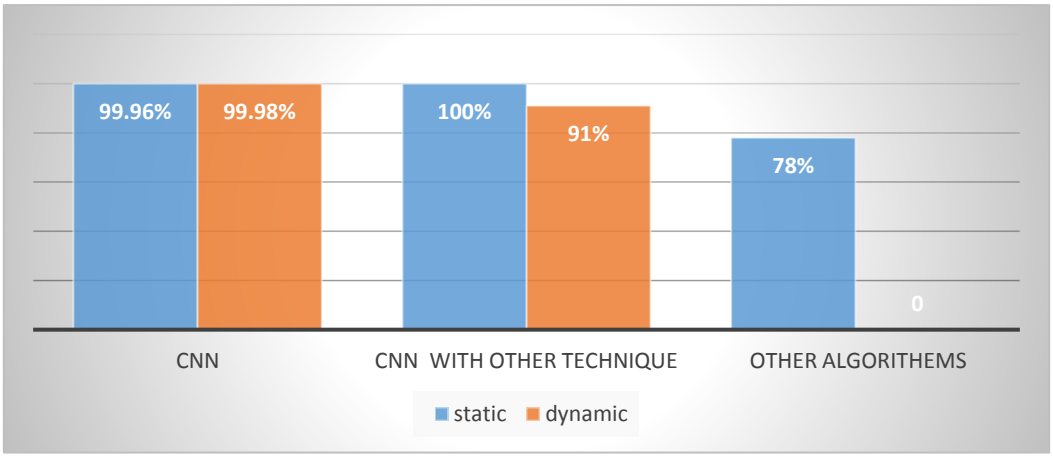


Figure 2 shows that: eleven previous studies depended on CNN technique: 9 for static and 2 for dynamic. Eleven studies depended on CNN together with other techniques: 6 for static and 5 for dynamic. Two studies depended on other techniques without CNN: 2 for static and none for dynamic.

Figure (3): Accuracy-based comparison for depended techniques with (static and dynamic)



When using only CNN technique, the best accuracy ratio achieved with dynamic. While, the best accuracy ration performed with static when using CNN together with other techniques. Finally, when using other techniques without CNN, only the static implementation has been depended by previous studies.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

4.2. Sign language recognition for continuous sentences based on deep learning

Compared with isolated sign language recognition of a word, continuous sentence recognition needs to establish more reliable long-term temporal dependencies. Continuous initial sentence recognition is based on recognizing a single isolated word, which requires relevant algorithms to segment time series. However, due to the complexity of the time-series segmentation process and the high rate of miscalculation, in recent years, scientists have gradually bypassed the time-series segmentation. The time-series alignment algorithm and coding network for speech recognition have been used in previous research in this field. Continuous Sign Language recognition systems depend on different scales, such as: Word Error Rate (WER) and Character Error Rate (CER). However, the accuracy should be depended to reflect the models performance [45, 49].

4.2.1. Long-term bidirectional memory network model

Recognizing sign language for consecutive sentences is more complex and requires stronger timing in the long run. Therefore, semantic information modeling is widely used in the search for sign language recognition of continuous sentences. The researchers [43] proposed using human base points extracted from the face, hand, and body parts as input to a recurrent neural network (RNN) to develop a sign language recognition system. A study [52] on a modified LSTM model proposed a continuous sequence of gestures or a continuous SLR recognizing a series of connected gestures. The researchers [50] attempted to construct two data sets, the first to identify the fingerspelling sequences in this difficult setting.

In contrast to previous work, the video data for the study is challenging due to low frame rates and visual contrast. We train a hand-held signature detector for special purposes to meet the visual challenges using a small subset of our data. Looking at the output of the hand detector, the sequence model decodes the sequence of putative letters with fingers. For the sequencing model, we explore frequent attention-based decoders and CTC-based approaches.

4.2.2. Convolutional Neural Network Model

Compared with the complexity of the BLSTM network model, the continuous sign language recognition based on the three-dimensional convolutional network model avoids the complex modelling of the BLSTM network. It saves complex calculations based on the same time series modelling. The researchers [46] provided the embedding for the CNN in HMM while interpreting the output of the CNN in the Bayesian framework. The hybrid CNN-HMM combines the powerful discriminative capabilities of CNNs with the sequential modelling capabilities of HMMs. They proposed in [49] a hierarchical LSTM decoder (HLSTM) model with visual content and word embedding for SLT. It deals with different resolutions by conveying spatiotemporal transitions between frames, clips, and visual units. It first explores the Spatio-temporal cues of video clips by 3D CNN and packs appropriate insights by online key segment mining with adaptive variable length.

To establish a continuous tag recognition system [45], a hierarchical interest network with latent space (LS-HAN) was developed, eliminating temporal segmentation pre-processing. The

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

proposed LS-HAN consists of three components: a dual-stream convolutional neural network (CNN) for generating a representation of video features, a latent space (LS) for semantic gap bridging, and a hierarchical interest network (HAN) for latent space recognition. The proposed system has been extensively tested on two data sets. The experimental results show the effectiveness of the proposed framework. In a study [44], researchers designed a three-dimensional residual convolutional network (3D-ResNet) to extract visual features. Then, a scaling stacked convolutional network with communicative temporal classification (CTC) is applied to learn the mapping between sequential features and the text string. In an attempt to improve the performance of the recursive network [42] it was integrated and fully exploited the representative capacity of deep neural networks with limited data. Then they used the alignment as supervisory information to directly adjust the feature extraction unit. In a study [55], local area images of both hands and face were used, along with skeletal information to capture local information and the positions of both hands relative to the body, respectively, and a multi-stream WSLR framework, where a stream with local area images and a stream with structure information are presented. By expanding the I3D network to improve the accuracy of WSLR recognition. The sequence information and the long-term context sequence of the recursive unit through the time dimension obtained a good recognition effect. Still, the network structure is relatively complex, and the hardware and time requirements are relatively high.

4.2.3. Mixed models

Compared with the above two model architectures of the core network, the continuous sign language recognition based on the hybrid network model is the most widely studied, making full use of the feature extraction ability of convolutional neural networks and the advantages of time-series classification of the recurrent neural network to achieve more accurate recognition? A group of researchers [51] proposed a system at both the word and sentence levels, the use of a multimodal convolutional neural network (CNN) to abstract representations from the inputs of sensory modalities, and the use of bi-directional long-term memory (LSTM) to model temporal dependencies. At the top of the networks. In the study [46], a deep hybrid architecture consisting of a temporal convolution unit (TCOV), a bidirectional recursive module (BGRU), and a fusion layer unit (FL) is presented to address the CSLT problem. TCOV captures a short-term temporal transition on adjacent clip features (local mode). In [48] they review a new approach to using deep learning loss function and temporal classification (CTC) in SLR at the sentence level. The study [54] presents an approach for continuous context-aware sign language recognition using a generative adversarial network architecture, called the Sign Language Recognition Generative Adversarial Network (SLRGAN). The proposed network architecture consists of a constructor that recognizes sign language luminosity by extracting Spatio-temporal features from video sequences. And in [53] a multi-signal Spatio-temporal network (STMC) was presented to solve the vision-based sequential learning problem. The STMC network of the study consists of a spatial multiplexed unit (SMC) and a temporal multiplexed unit (TMC).

The continuous sentence sign language recognition technology and representative work based on deep learning are shown in Table 2.

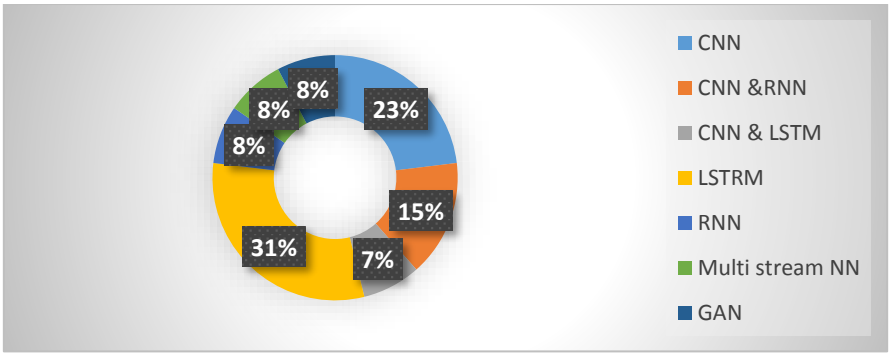
The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

Table (2): Comparison of continuous sign language recognition for previous works

No	Author	year	Static/ dynamic	Language	Method	Accuracy
1	Cui et al [42].	2018	dynamic	American	RNN and CNN	WER:46.9%
2	Ki Ko et al [43].	2018	dynamic	Korean	Recurrent neural network and skeleton	Acc: 89.5%
3	Junfu Pu et al [44].	2018	dynamic	German	3D CNN	WER: 38%
4	Huang et al [45].	2018	dynamic	German	LS-HAN	Acc: 82.7%
5	Wang et al [46].	2018	dynamic	German	3D CNN	WER: 37.9%
6	Koller [47].	2018	dynamic	German	CNN-HMMs	WER: 32.5%
7	Ariesta et al [48].	2018	dynamic	Indonesian	Convolutional Neural Network (CNN) and Bidirectional Recurrent Neural Network (Bi-RNN)	WER: 88.17% CER: 65.33%
8	Guo et al [49]	2018	dynamic	Chinese	HLSTM	WER: 63%
9	Shi et al [50].	2019	dynamic	American	Long and short-term networks	WER:41.9
10	Zhang et al [51].	2019	dynamic	American	BLSTM	Acc: 93.7%
11	Mittal et al [52].	2019	dynamic	American	LSTM	Acc: 72.3
12	Zhou et al [53].	2020	dynamic	Chinese, German	CNN-LSTM-HMM	WER: 28.6 % WER: 20.7%
13	Papastratis et al [54].	2021	dynamic	German, Chinese, Greek	Generative Adversarial Network	Acc: 84.96
14	Maruyama et al [55].	2021	dynamic	American	Skeleton Stream and Multi-stream Neural Networks	Acc: 83.86%

Table (2) illustrates the techniques depended the dynamic sign language recognition addressed the previous studies, such as: RNN, LSTRM. Figure (4) shows the statistical representation of these techniques.

Figure (4): Statistics of used techniques for Continuous Sign Recognition



It can be shown from Figure (4) that LSTM was the most technique used by previous studies for continuous sign language with 31% ratio. While, CNN technique used by 23% of the previous studies. This ratio is decreased to be 07% when using both of (CNN and LSTM) together.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

5. Problems of the Reviewed Studies

Sign language recognition has important research value in computer vision, pattern recognition, human-computer interaction, virtual reality, and other related fields. Although deep learning technology has greatly improved the accuracy and speed of sign language recognition in recent years, it is far from realistic – time and reliability: there is still room for application goals of excellent and accurate sign language recognition and translation. The main challenges are:

(1) Flexibility and specifics of the behaviour of the sign language itself: Sign language is a behavioural sequence consisting of the movements of the upper limb and the hand. The hand is the most flexible limb in the human body [46, 47]. Actions, etc., all affect the semantics of sign language. Some sign languages also include lip coordination and facial expressions. Therefore, recognition accuracy and real-time performance remain the goals of sign language recognition.

(2) Sign language behaviour is affected by background interference, lighting, observation angle, and operator uniformity [58]. Operators usually stand fully in the current data set and only move their upper limbs and hands. Still, in real applications, there are many situations, such as complex background, obstruction of multiple people, changes in lighting conditions, whole-body movement of the operator, non-standard sign language, bring More difficulty in identification.

(3) The long-term temporal relationship in successive sentences and the freedom of transitional frameworks between isolated words: Sign language recognition in successive sentences need to remove long-term temporal dependencies to create semantic structures, and at the same time needs to adapt to the complexity of spatial information and isolated words freedom and random transition frames between [59, 60, 61].

High accuracy, scalability, robustness, real-time, and user autonomy remain important challenges for future sign language recognition research. Meanwhile, applying the existing findings to real life and achieving cross-platform deployment is also an urgent need.

6. Recommendations

Although gesture recognition research has been carried out for decades, and made great progress, a large number of research results have emerged, but the gesture recognition technology is not mature. Judging from the published public gesture database, compared to the face recognition database, the total data volume of the gesture database is relatively small, and the background environment is single. Based on the results of published research on gestures, the rate of gesture recognition is relatively high, but the conclusion is often that the test is carried out in an environment with slight changes in lighting and a slight background. Because there are large individual differences in gestures, for example, the waving speed of the elderly is often slower than the waving speed of the young, so how to overcome individual differences also presents a difficulty in recognizing gestures. Static gesture recognition methods and dynamic gesture recognition methods have achieved good results through experimental validation, but there is still much room for improvement.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

1) Training data sets used in research for convolutional neural networks are general data sets. Although the data set has been expanded, the number of such large sample sets is still relatively small. When designing a more complex deep learning network with more layers, datasets of this size cannot meet the requirements, so expanding the gesture database is very useful for gesture recognition research.
2) Convolutional neural networks have achieved great success in computer recognition and vision. They are used in two phases feature extraction and classification. Other classifiers that use the ensemble learning principle to create more accurate systems and make use of convolutional neural networks can be tested in the feature extraction phase.

7. Conclusions and Future Work

Since deep learning is mainly data-driven, the sign language recognition technology based on deep learning is bound to evolve further with the accumulation and in-depth mining of large amounts of data.

1) Isolated single word sign language recognition is mainly deepened to improve accuracy and speed at the same time. In contrast, continuous sentence recognition is based more on the principles and methods of speech recognition, natural language understanding, machine translation and other fields, combined with video Himself innovation characteristics.
2) The establishment of good data sets and evaluation criteria is still urgent, and complete and standardized data nomenclature can greatly improve technical methods. Although there are some relevant standard sign language datasets. Hence, the sign language video datasets in real scenarios are still very scarce, and the production of dataset labels is an expensive and time-consuming process.
3) With deep learning theoretical research, the fundamental innovation in video understanding and analysis will bring breakthroughs in sign language recognition algorithms.
4) Compared with traditional methods, deep learning technology is closely related to network speed, GPU performance, and other hardware performance. Replacing computer hardware and network upgrades will shorten the time for experiments, testing, and evaluation and speed up the process of converting results to enhance theoretical research and algorithm innovation better.

As future work suggestions:

(1) In the future, sign language recognition will be further developed through the cross-integration of different fields. It is expected that more scientists will join the research on sign language recognition so that the research results of sign language recognition can truly serve the public sector, and improve the level of intelligent informatics for the whole society.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

(2) In the future, sign language recognition will bring breakthroughs with the advancement of sign language properties, standardized data sets, recognition algorithms, and efficient computing power.

(3) For syntax recognition, the hybrid network model will be the main network algorithm in the future. For isolated sign language recognition, a 3D CNN requires a lot of computations and highly dependent on the memory and performance of the GPU used. There is still much room for the development of RNNs.

References:

1) Corina, D.P. (2015). Sign Language: Psychological and Neural Aspects. 10.1016/B978-0-08-097086-8.52019-4.

2) Rastgoo, Razieh & Kiani, Kourosh & Escalera, Sergio & Athitsos, Vassilis & Sabokrou, Mohammad. (2022). All You Need In Sign Language Production.

3) Kuenburg, Alexa & Fellinger, Paul & Fellinger, Johannes. (2015). Health Care Access among Deaf People. Journal of Deaf Studies and Deaf Education. 21. Env042. 10.1093/deafed/env042.

4) Adeyanju, Ibrahim & Bello, Oluwaseyi & Adegboye, Mutiu. (2021). Machine learning methods for sign language recognition: A critical review and analysis. Intelligent Systems with Applications. 12. 10.1016/j.iswa.2021.200056.

5) Sarma, Debajit & Bhuyan, M. (2021). Methods, Databases and Recent Advancement of Vision-Based Hand Gesture Recognition for HCI Systems: A Review. SN Computer Science. 2. 10.1007/s42979-021-00827-x.

6) Chakraborty, Biplab & Sarma, Debajit & Bhuyan, Manas & MacDorman, Karl. (2017). A Review of Constraints on Vision-based Gesture Recognition for Human-Computer Interaction. IET Computer Vision. 12. 10.1049/iet-cvi.2017.0052.

7) Bragg, Danielle & Verhoef, Tessa & Vogler, Christian & Morris, Meredith & Koller, Oscar & Bellard, Mary & Berke, Larwan & Boudreault, Patrick & Braffort, Annelies & Caselli, Naomi & Huenerfauth, Matt & Kacorri, Hernisa. (2019). Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. 16-31. 10.1145/3308561.3353774.

8) Mazhar, Osama & Ramdani, Sofiane & Cherubini, Andrea. (2021). A Deep Learning Framework for Recognizing Both Static and Dynamic Gestures. Sensors. 21. 2227. 10.3390/s21062227.

9) Pisharady, Pramod & Saerbeck, Martin. (2015). recent methods and databases in vision-based hand gesture recognition: A review. Computer Vision and Image Understanding. 141. 152-165. 10.1016/j.cviu.2015.08.004.

10) Joksimoski, Boban & Zdravevski, Eftim & Lameski, Petre & Pires, Ivan & Melero, F. & Martinez, Tomas & Garcia, Nuno & Mihajlov, Martin & Chorbev, Ivan & Trajkovik, Vladimir. (2022). Technological Solutions for Sign Language Recognition: A Scoping Review of Research Trends, Challenges, and Opportunities. IEEE Access. 10. 1-1. 10.1109/ACCESS.2022.3161440.

11) Sahoo, Jaya & Prakash, Allam & Pławiak, Paweł & Samantray, Saunak. (2022). Real-Time Hand Gesture Recognition Using Fine-Tuned Convolutional Neural Network. Sensors. 22. 706. 10.3390/s22030706.

12) Yamashita, Rikiya & Nishio, Mizuho & Do, Richard & Togashi, Kaori. (2018). Convolutional neural networks: an overview and application in radiology. Insights into Imaging. 9. 10.1007/s13244-018-0639-9.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

13) Ibrahim, Nada & Zayed, Hala & Selim, Mazen. (2019). Advances, Challenges, and Opportunities in Continuous Sign Language Recognition. Journal of Engineering and Applied Sciences. 15. 1205-1227. 10.36478/jeasci.2020.1205.1227.

14) Adeyanju, Ibrahim & Bello, Oluwaseyi & Adegboye, Mutiu. (2021). Machine learning methods for sign language recognition: A critical review and analysis. Intelligent Systems with Applications. 12. 10.1016/j.iswa.2021.200056.

15) Savant, Rakesh & Nasriwala, Jitendra. (2019). Indian Sign Language Recognition System: Approaches and Challenges.

16) De Sisto, Mirella & Vandeghinste, Vincent & Egea-Gómez, Santiago & Coster, Mathieu & Shterionov, Dimitar & Saggion, Horacio. (2022). Challenges with Sign Language Datasets for Sign Language Recognition and Translation.

17) Wadhawan, Ankita & Kumar, Parteek. (2020). Deep learning-based sign language recognition system for static signs. Neural Computing and Applications. 32. 1-12. 10.1007/s00521-019-04691-y.

18) AL-Qurishi, Muhammad & Khalid, Thariq & Souissi, Riad. (2021). Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3110912.

19) Li, Yunan & Miao, Qiguang & Tian, Kuan & Fan, Yingying & Xu, Xin & Li, Rui & Song, Jianfeng. (2017). Large-scale Gesture Recognition with a Fusion of RGB-D Data Based on Saliency Theory and C3D model. IEEE Transactions on Circuits and Systems for Video Technology. PP. 1-1. 10.1109/TCSVT.2017.2749509.

20) Miao, Qiguang & Li, Yunan & Ouyang, Wanli & Ma, Zhenxin & Xu, Xin & Shi, Weikang & Cao, Xiaochun. (2017). Multimodal Gesture Recognition Based on the ResC3D Network. 3047-3055. 10.1109/ICCVW.2017.360.

21) Elbadawy, Menna & Samir Roshdy, Ahmed & Shedeed, Howida & Tolba, Mohamed. (2017). Arabic sign language recognition with 3D convolutional neural networks. 66-71. 10.1109/INTELCIS.2017.8260028.

22) Oyedotun, Oyebade & Khashman, Adnan. (2017). Deep learning in vision-based static hand gesture recognition. Neural Computing and Applications. 28. 10.1007/s00521-016-2294-8.

23) Hossen, M.A & Govindaiah, Arun & Sultana, Sadia & Bhuiyan, Alauddin. (2018). Bengali Sign Language Recognition Using Deep Convolutional Neural Network. 369-373. 10.1109/ICIEV.2018.8640962.

24) Kim, Sunmok & Ji, Yangho & Lee, Ki-Baek. (2018). An Effective Sign Language Learning with Object Detection Based ROI Segmentation. 330-333. 10.1109/IRC.2018.00069.

25) Lin, Chi & Wan, Jun & Liang, Yanyan & Li, Stan. (2018). Large-Scale Isolated Gesture Recognition Using a Refined Fused Model Based on Masked Res-C3D Network and Skeleton LSTM. 52-58. 10.1109/FG.2018.00018.

26) Konstantinidis, Dimitrios & Dimitropoulos, Kosmas & Daras, Petros. (2018). A Deep Learning Approach for Analyzing Video and Skeletal Features in Sign Language Recognition. 10.1109/IST.2018.8577085.

27) Gondu, Anantha & Syamala, K. & Kishore, PVV & Sastry, A.. (2018). Deep convolutional neural networks for sign language recognition. 194-197. 10.1109/SPACES.2018.8316344.

28) Taskiran, Murat & Kıllıoğlu, Mehmet & Kahraman, Nihan. (2018). A Real-Time System For Recognition of American Sign Language by Using Deep Learning. 10.1109/TSP.2018.8441304.

29) Adewale, Victoria & Olamiti, Adejoke. (2018). Conversion of Sign Language To Text And Speech Using Machine Learning Techniques. JOURNAL OF RESEARCH AND REVIEW IN SCIENCE. 5. 58-65. 10.36108/jrrslasu/8102/50 (0170).

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

30) Hernandez, Vincent & Suzuki, Tomoya & Venture, Gentiane. (2020). Convolutional and recurrent neural network for human activity recognition: Application on American Sign Language. PLOS ONE. 15. E0228869. 10.1371/journal.pone.0228869.

31) Obaid, Falah & Babadi, Amin & Yoosofan, Ahmad. (2020). Hand Gesture Recognition in Video Sequences Using Deep Convolutional and Recurrent Neural Networks. Applied Computer Systems. 25. 57-61. 10.2478/acss-2020-0007.

32) Reshna, S. & Sajeena, A. & Jayaraju, Madhavan. (2020). Recognition of static hand gestures of Indian sign language using CNN. AIP Conference Proceedings. 2222. 030012. 10.1063/5.0004485.

33) Lai, Kenneth & Yanushkevich, Svetlana. (2020). CNN+RNN Depth and Skeleton based Dynamic Hand Gesture Recognition.

34) Fernandes, Lance & Dalvi, Prathamesh & Junnarkar, Akash & Bansode, Manisha. (2020). Convolutional Neural Network based Bidirectional Sign Language Translation System. 769-775. 10.1109/ICSSIT48917.2020.9214272.

35) Sharma, Aishwarya & Panda, Siba & Verma, Saurav. (2020). Sign Language to Speech Translation. 1-8. 10.1109/ICCCNT49239.2020.9225422.

36) A.Al, Abdulwahhab & Raheem, Firas. (2020). Hand Gesture Recognition of Static Letters American Sign Language (ASL) Using Deep Learning. 38, Part A. 926-937.

37) Kolla, Bhanu. (2020). Accurate Hand Gesture Recognition using CNN and RNN Approaches. International Journal of Advanced Trends in Computer Science and Engineering. 9. 10.30534/ijatcse/2020/114932020.

38) Adithya, V. & Rajesh, Reghunadhan. (2020). A Deep Convolutional Neural Network Approach for Static Hand Gesture Recognition. Procedia Computer Science. 171. 2353-2361. 10.1016/j.procs.2020.04.255.

39) P S, Neethu & Ramadass, Suguna & Sathish, Divya. (2020). An efficient method for human hand gesture detection and recognition using deep learning convolutional neural networks. Soft Computing. 24. 10.1007/s00500-020-04860-5.

40) Barbhuiya, Abul & Karsh, Ram & Jain, Rahul. (2021). CNN based feature extraction and classification for sign language. Multimedia Tools and Applications. 80. 1-19. 10.1007/s11042-020-09829-y.

41) Gadekallu, Thippa & Alazab, Mamoun & Kaluri, Rajesh & Reddy, Praveen & Bhattacharya, Sweta & Lakshman, Kuruva & M, Parimala. (2021). Hand gesture classification using a novel CNN-crow search algorithm. Complex & Intelligent Systems. 7. 10.1007/s40747-021-00324-x.

42) Cui, Runpeng & Liu, Hu & Zhang, Changshui. (2019). A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. IEEE Transactions on Multimedia. PP. 1-1. 10.1109/TMM.2018.2889563.

43) Ko, Sang-Ki & Son, Jae & Jung, Hyedong. (2018). Sign language recognition with recurrent neural network using human keypoint detection. RACS '18: Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems. 326-328. 10.1145/3264746.3264805.

44) Pu, Junfu & Zhou, Wengang & Li, Houqiang. (2018). Dilated Convolutional Network with Iterative Optimization for Continuous Sign Language Recognition. 885-891. 10.24963/ijcai.2018/123.

45) Huang, Jie & Zhou, Wengang & Zhang, Qilin & Li, Houqiang & Li, Weiping. (2018). Video-based Sign Language Recognition without Temporal Segmentation.

46) Wang, Shuo & Guo, Dan & Zhou, Wen-gang & Zha, Zheng-Jun & Wang, Meng. (2018). Connectionist Temporal Fusion for Sign Language Translation. 1483-1491. 10.1145/3240508.3240671.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7353683

47) Koller, Oscar & Zargaran, Sepehr & Ney, Hermann & Bowden, Richard. (2018). Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. International Journal of Computer Vision. 126. 10.1007/s11263-018-1121-3.

48) Ariesta, Meita & Wiryana, Fanny & Suharjito, Suharjito & Zahra, Amalia. (2018). Sentence Level Indonesian Sign Language Recognition Using 3D Convolutional Neural Network and Bidirectional Recurrent Neural Network. 16-22. 10.1109/INAPR.2018.8627016.

49) Guo, D., Zhou, W., Li, H., & Wang, M. (2018). Hierarchical LSTM for Sign Language Translation. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).

50) Shi, Bowen & Rio, Aurora & Keane, Jonathan & Michaux, Jonathan & Brentari, Diane & Shakhnarovich, Greg & Livescu, Karen. (2018). American Sign Language Fingerspelling Recognition in the Wild. 145-152. 10.1109/SLT.2018.8639639.

51) Zhang, Qian & Wang, Dong & Zhao, Run & Yu, Yinggang. (2019). MyoSign: enabling end-to-end sign language recognition with wearables. 650-660. 10.1145/3301275.3302296.

52) Mittal, Anshul & Kumar, Pradeep & Roy, Partha & Balasubramanian, Raman & Chaudhuri, Bidyut. (2019). A Modified LSTM Model for Continuous Sign Language Recognition Using Leap Motion. IEEE Sensors Journal. 19. 7056-7063. 10.1109/JSEN.2019.2909837.

53) Zhou, Hao & Zhou, Wengang & Zhou, Yun & Li, Houqiang. (2020). Spatial-Temporal Multi-Cue Network for Continuous Sign Language Recognition. Proceedings of the AAAI Conference on Artificial Intelligence. 34. 13009-13016. 10.1609/aaai. v34i07.7001.

54) Papastratis, Ilias & Dimitropoulos, Kosmas & Daras, Petros. (2021). Continuous Sign Language Recognition through a Context-Aware Generative Adversarial Network. Sensors. 21. 2437. 10.3390/s21072437.

55) Maruyama, Mizuki & Ghose, Shuvozit & Inoue, Katsufumi & Roy, Partha & Iwamura, Masakazu & Yoshioka, Michifumi. (2021). Word-level Sign Language Recognition with Multi-stream Neural Networks Focusing on Local Regions.

56) Sandler W. THE PHONOLOGICAL ORGANIZATION OF SIGN LANGUAGES. Lang Linguist Compass. 2012 Mar 1; 6 (3):162-182. doi: 10.1002/lnc3.326. Epub 2012 Mar 2. PMID: 23539295; PMCID: PMC3608481.

57) Dhulipala, Shiva & Adedoyin, Festus & Bruno, Alessandro. (2022). Sign and Human Action Detection Using Deep Learning. Journal of Imaging. 8. 10.3390/jimaging8070192.

58) Stoll C, Dye MWG. Sign language experience redistributes attentional resources to the inferior visual field. Cognition. 2019 Oct; 191:103957. doi: 10.1016/j.cognition.2019.04.026. Epub 2019 Jun 27. PMID: 31255921.

59) Dawod, Ahmad. (2020). Novel Technique for Isolated Sign Language Based on Fingerspelling Recognition. 10.1109/SKIMA47702.2019.8982452.

60) Middi, Venkata Sai Rishita & Raju, Middi. (2020). Sign Language Recognizer Using HMMs. 10.1007/978-981-15-7078-0_71.

61) O. Mercanoglu Sincan, A. O. Tur and H. Yalim Keles, "Isolated Sign Language Recognition with Multi-scale Features using LSTM," 2019 27th Signal Processing and Communications Applications Conference (SIU), 2019, pp. 1-4, doi: 10.1109/SIU.2019.8806467.