A Systematic Study of Sign Language Recognition Systems Employing Machine Learning Algorithms

Pranav and Rahul Katarya

Abstract The deaf and mute population struggles a lot in expressing their thoughts and ideas to others; Sign Language is the most expressive means of communication for them, but a majority of the general population is callow of sign language, hence the mute and deaf experience difficulties while communicating to the rest of the world. To overcome this communication barrier, a device that can accurately translate sign language gestures to speech and vice-versa in real-time is needed. There exist solutions for converting verbal or written language to sign language in real-time reliably and accurately, however the same cannot be said about translating sign language to textual and/or vocal format. The currently existing systems either do not support communication in both directions, are not real-time, have low recognition accuracy, or require static surrounding conditions. Some systems require additional hardware components like expensive sensors, which tend to increase the cost. In this survey, we have reviewed numerous existing solutions and have categorized them depending on the method used. We hope that the results obtained from this study may serve as a road map to guide future study in the domain of Sign Language Recognition (SLR).

Keywords Inertial Measurement Unit (IMU) · Neural Networks (NN) · Surface Electromyogram (sEMG)

1 Introduction

As of March 2020, World Health Organization reported that almost 430 million people in the world suffer from severe hearing problems [1]. Also going by the 2011 population census of the government of India [2], there were approximately 1.64 million people who suffer from speech disability, and 1.26 million people suffer from hearing impairment. The deaf and mute population resorts to using sign language

Pranav (✉) · R. Katarya
Big Data Analytics and Web Intelligence Laboratory, Department of Computer Science, Delhi Technological University, New Delhi, India
e-mail: pranav_2k20cse16@dtu.ac.in

S. Majhi et al. (eds.), *Distributed Computing and Optimization Techniques*, Lecture Notes in Electrical Engineering 903, https://doi.org/10.1007/978-981-19-2281-7_11

for communicating, which involves hand gestures, body posture, and facial expressions. However, as the mass population is uninformed of the meaning of signs in sign language, the Deaf-Mute population experiences difficulties in expressing their thoughts. There are three possible solutions to this problem, the first of them being teaching sign language to the masses, which is not possible, as most of the population might not be willing to do so. The other two methods being vision-based SLR systems, and sensors-based SLR systems. Some of the existing solutions in the field are discussed in Sect. 2. The taxonomy of signs and some sign languages are discussed in Sect. 3. Section 4 describes the working of SLR. Sections 5 and 6 provides the conclusion of the paper and aims to provide a roadmap for future development in the field.

2 Some Existing Sign Language Recognition Systems

SLR systems can be broadly classified into two categories: 1) Sensor-based systems; 2) Vision-based systems.

2.1 Sign Language Recognition Systems Based upon Sensors

For data collection, this method uses a data glove with sensors embedded on the fingers, palm, and arm of the signer. Some of the existing sensor-based systems in the field are demonstrated in Table 1.

In [3], they used a Myo armband [23] worn on the signer's forearm, which consists of sEMG sensors to recognize muscle activity. This data was passed on to a custom Artificial Neural Network (ANN) for recognizing some basic words of the Chinese sign language (CSL). In [4], they used a Photoplethysmography (PPG) sensor (to detect contractions of the forearm muscles) with IMU (to detect orientation and

Table 1 Some existing sensor-based SLR systems that use Machine Learning Techniques

Ref.	Sensors	Classifier	Type	No. of signs	Accuracy (%)
[3]	sEMG	ANN	Static	15	88.7
[4]	IMU, PPG	ResNet, GBT	Static	9	98
[5]	IMU	CapsNet	Dynamic only	60*	94
[6]	Flex, IMU	kNN, DTW; CNN	Dynamic only	10	96.6, 98
[7]	sEMG, IMU	DBN	Dynamic	150	95.1
[8]	IMU	LSTM	Dynamic	28	99.89
[9]	EMG, IMU	ANN; SVM	Dynamic	13	93.8, 85.6

Note.—Static = Static signs only; Dynamic = Both Static and Dynamic signs unless explicitly specified; * = approximately

motion of hands) on the signer's wrist. The collected data were pre-processed and passed on to a Gradient Boost Tree (GBT) and deep Residual Network (ResNet) for classifying signs for numbers one to nine of the ASL. In [5], they used a deep capsule network (CapsNet) for recognizing gesture data of approximately 40–80 (they used 20 sentences, with each sentence consisting of 2–4 words) words of the ISL, captured using an IMU. In [6], they used multiple flex sensors attached to a glove, in conjunction with three IMUs on the signer's arm. They used two methods for classifying the gestures: 1) k-Nearest neighbors (kNN) with Dynamic time warping (DTW) method, 2) Convolutional neural networks (CNN). They tested their models on 10 different gestures of the Italian sign language. In [7], they used an sEMG sensor along with IMU to capture muscle activity data along with hand motion data to capture gesture features and employed a Deep Belief Net (DBN) deep learning model for recognizing 81 single-handed and 69 two-handed words of the CSL. They achieved recognition accuracy of up to 88% for user-independent tests and 95% for user-dependent tests. In [8], six IMUs were used on fingers and back of the palm of the signer's hand to acquire hand motion and finger movement data for gestures. They used Long-term short memory (LSTM) deep learning algorithm for recognizing some commonly used sign words of the ASL. In [9], they used a Myo armband consisting of EMG sensors and IMU to capture features of sign gestures. The system uses Artificial Neural Networks (ANN), and support vector machines (SVM) for gesture recognition. They tested their system using a set of 13 commonly used ASL sign gestures.

2.2 Sign Language Recognition Systems Based on Vision-Based Techniques

For data collection, this method uses a camera module directed towards the signer to capture images/video of the gestures. Some of the existing solutions utilizing computer vision techniques for SLR are demonstrated in Table 2.

In [10], RGB gesture images were initially cropped down to reduce processing load and then converted to grayscale. They used multiple images of the same gesture for training and testing the performance of the You Look Only Once (YOLO) system, which is a CNN based object detection method. Their systems achieve 100 hundred percent accuracy on English alphabet fingerspelling images of the Indonesian sign language, however, accuracy drops significantly when testing on videos. In [11], they used RGB images along with a depth map (obtained using Kinect sensor [24]) of fingerspelling of English alphabets of the ASL. They used CNN for recognizing the signs and achieved up to 99.79% accuracy. In [12], they used VGGnet and ResNet architectures of CNN for gesture recognition. Their dataset includes 32 static Arabic sign language gestures, with 800 images of each. In [13], they used a Leap motion sensor [25] to acquire the spatial orientation of the signer's hand and fingers. They used a Hidden Markov classifier (HMC) for recognizing the gestures. In [14], they

Table 2 Some existing vision-based SLR systems

Ref.	Classifier	Type	No. of signs	Accuracy (%)
[10]	YOLO	Static	24	73
[11]	CNN	Static	24	99.79
[12]	CNN	Static	32	99
[13]	HMC	Static	24	86
[14]	Multi-class SVM	Dynamic	35	87.6
[15]	I3D CNN	Dynamic	249	64.44
[16]	PCA, SVM	Static	35	95.31
[17]	CNN	Static	71	93.04

Note.—Static = Static signs only; Dynamic = Both Static and Dynamic signs

used a Kinect sensor to capture skeletal images of the signer. However, the Kinect does not capture does not capture finger joints and cannot distinguish signs involving finger articulations and orientation, which greatly reduces the number of signs that can be recognized using it. They used "Multi-class SVM with radial basis function kernel" on a dataset of 35 ISL gestures. In [15], they used a "two-stream model of inflated 3D (I3D) ConvNets". The first stream was fed RGB data and the second was fed optical flow data of sign videos. Using only RGB, or optical flow data, recognition accuracy was significantly reduced compared to when using both of them together. In [16], they used static images of ISL gestures English alphabets and numbers 1 to 9. They used Grayscale conversion followed by segmentation, and noise removal before subjecting the images to the skin thresholding process. The output image of this step is a small-sized image including only the hands of the signer, from which PCA was used for extracting the features of the gesture, which were then fed to an SVM classifier for recognizing the gestures. In [17], they used a dataset of English alphabets, numbers, and 35 common static gestures of the ASL. They used skin-color based segmentation to extract hand region from images, which were then fed to a Keras and CNN classifier for recognizing the signs. The problem with this system is that it takes a lot of time to recognize the gestures, averaging 2.6 s for each sign in the dataset, which is not acceptable for real-time SLR systems.

3 Sign Languages and Taxonomy of Signs

3.1 Types of Signs

SL gestures can be performed using one hand or both hands, they can be static or dynamic, manual (include hand gestures only) or non-manual (include other physical features like mouth movements, facial expressions, and body orientation). Dynamic signs can further be classified as type 0 and type 1, as shown in Fig. 1. The type 0 sign

Fig. 1 Taxonomy of sign language gestures [18]

indicates a two-handed dynamic sign, in which both hands are performing motion, whereas in a type 1 sign is a two-handed dynamic sign with only the primary hand performing motion [18].

3.2 Problem with Universal Sign Language Recognition System

There is no universally accepted sign language. Just like spoken languages, over time as people communicated with each other, sign languages also spread out and took multiple forms. There are more than a hundred different forms of sign language in use today all over the world. [19] Even countries that have a common spoken language do not necessarily share a common sign language, e.g. America, Britain have English as their main language, yet both nations have their own sign language viz. American sign language (ASL), and British sign language (BSL).

Fingerspelling for English alphabets in the ASL is done using a single hand, and signs for most of the alphabets are static in nature, as shown in Fig. 2; while Fingerspelling of English alphabets in the Indian sign language (ISL) mostly involves using both hands, and the signs are static in nature, as shown in Fig. 3. Also, other countries having their own sign language (e.g. Brazilian, Chinese, etc.) makes the existence of a single universal solution for SLR impossible.

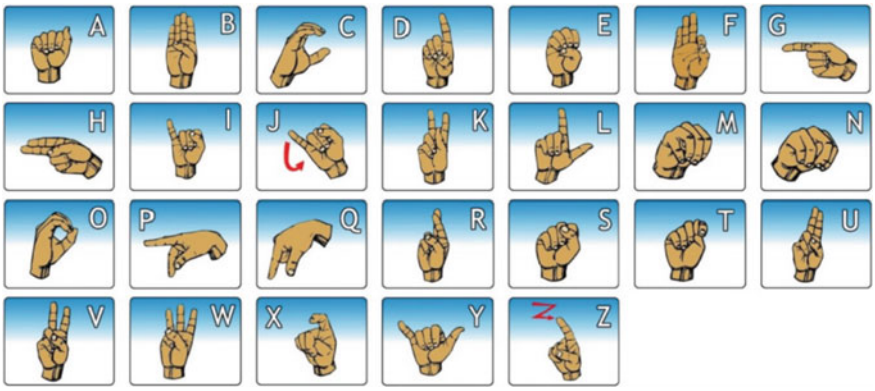

Fig. 2 ASL fingerspelling dataset [20]

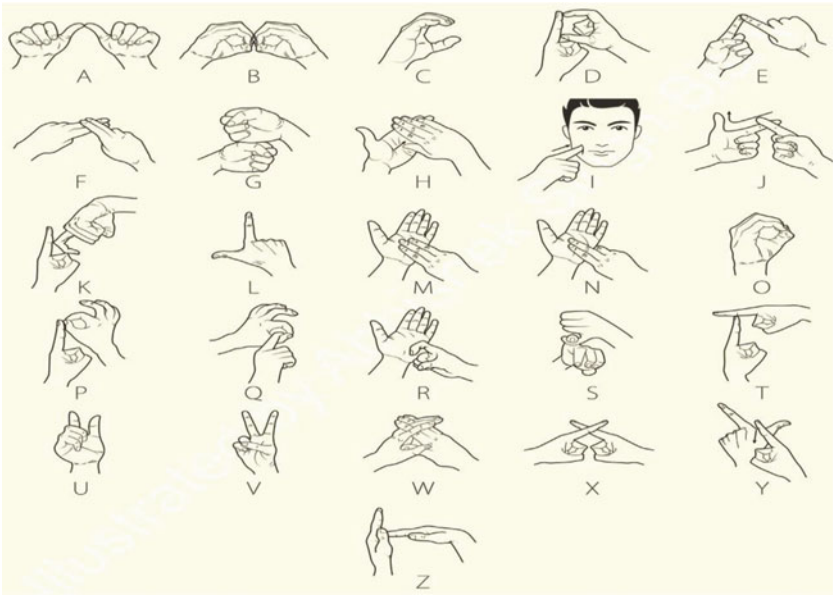

Fig. 3 ISL fingerspelling dataset [21]

4 Sign Language Recognition System Overview

Gestures in sign language are a combination of the following elements:

- Articulation points (finger joints, wrist, elbow, shoulder),
- hand/palm orientation,
- Facial expressions,

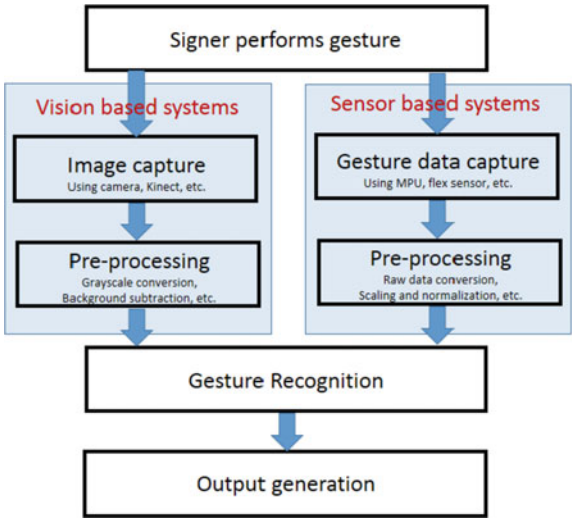

Fig. 4 Overview of the sign language recognition process

- Motion (of fingers, hands, and head).

The process of SLR can be divided into the following stages as shown in Fig. 4:

4.1 Data Acquisition

To detect the motion, position, and orientation of hands, fingers, etc., the available systems can be predominantly categorized into two classes: sensor-based and vision-based.

The vision-based systems make use of a camera unit directed toward the subject and captures images or videos as the person gestures in sign language. The camera can capture RGB or monochrome image sensors, as well as IR image sensors also.

Sensor-based systems majorly make use of some or all of Flex sensors, IMU, Pressure/contact sensors, sEMG sensors, etc. attached to a glove.

4.2 Pre-processing and Features Extraction

The elements of sign language gesture are acquired in this step. In vision-based systems, the captured image is pre-processed to reduce noise and improve image characteristics to make recognition easier. It may include background subtraction, image segmentation, subject isolation, noise removal (data cleaning), face detection and facial expression detection, grayscale conversion, binarization, tracking of hand movements for dynamic gestures, etc.

For the vision-based systems, lighting conditions, camera resolution, frame rate (for dynamic signs), distance from the subject, skin color and background color play a significant role in sign recognition accuracy.

In sensor-based systems, data from the appropriate sensors are collected in raw form and some pre-processing is applied on them to make the data more suitable for use, e.g., orientation data from IMU can vary from a few hundred to thousands [22], however, the desired output data should be in the range from 0 to 360 indicating the orientation angle; also, different sensors have different baud-rate, which must be factored in during the pre-processing stage.

4.3 Recognition

In this step, the features extracted in the previous section are matched against a pre-defined dataset or passed on to a machine learning classification model (which has been prior trained with reference data) to recognize the signs. The output of this step can be in form of audio from a speaker, or text display on a screen.

5 Discussion

Flex sensors provide data about flexion of the fingers, however, they do not describe the relative position of the fingers, due to this, using only flex sensors only a small number of signs can be differentiated due to the limited resolution/range of the sensors and difficulties in differentiating similar signs such as 'M' and 'N'. PPG sensors [4] suffer from body motion artifacts and hence cannot be employed in a practical device. Another way of recognizing gestures is using an EMG sensor [3, 7, 9], however, an EMG sensor alone cannot differentiate signs having similar hand orientation and different position like 'mother' and 'father' signs in children's sign language have the exact same hand orientation and differ only in position. The inclusion of IMU [4, 6–9] sensors enables for better differentiation of dynamic signs, however, the problem with similar static signs as in the previous case is still present. Pressure/contact sensors provide data about whether the fingers touch each other, and how firmly, which can be used for discerning many similar gestures and improve the accuracy of the system. Also, as only a few [7] of the aforementioned systems have used a large dataset of other than basic signs (usually 26 alphabets, 10 numbers, and some static gestures), it is difficult to make a statement about the scalability of the other sensor-based systems, because as the number of signs increases, it gets more likely for the signs to be similar to each other and interfere in accurately recognizing the signs.

In the vision-based systems also, apart from [15] who used a dataset for 249 signs for their system and [17] whose dataset contains 71 different signs, all other systems have been tested for a significantly fewer number of signs and hence it is difficult

to comment about their performance with a significantly large dataset for the same reasons as in the case of sensor-based systems.

6 Conclusions

We have seen in the previous sections that most of the systems are merely a proof of concept and not a real solution to the problem, as scalability remains a major concern. For the sensor-based systems, the accuracy of the gesture recognition increases as the number and type of sensors used is increased, hence, future research should continue to explore the accuracy of SLR by using flex sensors on the wrist of the signer also, to capture the relative angle of the hand with respect to the arm, along with using flex sensors on all 10 fingers, and IMUs on each hand to capture as much data as possible to make a decision. The inclusion of contact sensors and sEMG sensors will only improve the accuracy of the system.

Also, it can be concluded from the previous sections that vision-based systems cover the areas missed by sensor-based systems, like facial expressions, and sensor-based systems do not need perfect line of sight communication, unlike vision-based systems. This can be extremely useful in discerning signs that are similar to each other. Hence, it will be fascinating to see the two methods combined to make a hybrid SLR system, which should eliminate most of the drawbacks of the current solutions and can make for a real-world solution to sign language recognition, however, the cost constraint of such a system might pose as a roadblock for availability to the masses.

References

1. Who.int. Deafness and hearing loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. Accessed 26 Apr 2021
2. Censusindia.gov.in. Census of India: Disabled Population. https://censusindia.gov.in/census_and_you/disabled_population.aspx. Accessed 26 Apr 2021
3. Zhang Z, Su Z, Yang G (2019) Real-time Chinese Sign Language Recognition based on artificial neural networks. In: 2019 IEEE international conference on robotics and biomimetics (ROBIO). IEEE, pp 1413–1417
4. Zhao T, Liu J, Wang Y, Liu H, Chen Y (2019) Towards low-cost sign language gesture recognition leveraging wearables. IEEE Trans Mob Comput
5. Suri K, Gupta R (2019) Continuous sign language recognition from wearable IMUs using deep capsule networks and game theory. Comput Electr Eng 78:493–503
6. Saggio G, Cavallo P, Ricci M, Errico V, Zea J, Benalcázar ME (2020) Sign language recognition using wearable electronics: implementing k-nearest neighbors with dynamic time warping and convolutional neural network algorithms. Sensors 20(14):3879
7. Yu Y, Chen X, Cao S, Zhang X, Chen X (2019) Exploration of Chinese sign language recognition using wearable sensors based on deep belief net. IEEE J Biomed Health Inform 24(5):1310–1320

8. Chong TW, Kim BJ (2020) American sign language recognition system using wearable sensors with deep learning approach. J Korea Inst Electron Commun Sci 15(2):291–298
9. Fatmi R, Rashad S, Integlia R (2019) Comparing ANN, SVM, and HMM based machine learning methods for American sign language recognition using wearable motion sensors. In: 2019 IEEE 9th annual computing and communication workshop and conference (CCWC). IEEE, pp 0290–0297
10. Daniels S, Suciati N, Fathichah C (2021) Indonesian sign language recognition using YOLO method. In: IOP Conference Series: Materials Science and Engineering, vol 1077, no 1. IOP Publishing, p 012029
11. Wang CC, Chiu CT, Huang CT, Ding YC, Wang LW (2020) Fast and accurate embedded DCNN for RGB-D based sign language recognition. In: ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1568–1572
12. Saleh Y, Issa G (2020) Arabic sign language recognition through deep neural networks fine-tuning. Int Assoc Online Eng. 71–83
13. Vaitkevičius A, Taroza M, Blažauskas T, Damaševičius R, Maskeliūnas R, Woźniak M (2019) Recognition of American sign language gestures in a virtual reality using leap motion. Appl Sci 9(3):445
14. Gangrade J, Bharti J (2019) Real time sign language recognition using depth sensor. Int J Comput Vis Robot 9(4):329–339
15. Sarhan N, Frintrop S (2020) Transfer learning for videos: from action recognition to sign language recognition. In: 2020 IEEE international conference on image processing (ICIP). IEEE, pp 1811–1815
16. Mali D, Limkar N, Mali S (2019) Indian sign language recognition using SVM classifier. In: Proceedings of international conference on communication and information processing (ICCIP)
17. Tolentino LKS, Juan ROS, Thio-ac AC, Pamahoy MAB, Forteza JRR, Garcia XJO (2019) Static sign language recognition using deep learning. Int J Mach Learn Comput 9(6):821–827
18. Wadhawan A, Kumar P (2019) Sign language recognition systems: a decade systematic literature review. Archives of computational methods in engineering
19. Ethnologue. Sign language. https://www.ethnologue.com/subgroups/sign-language. Accessed 5 June 2021
20. American Sign Language (ASL) Discussion Board. ASL Fingerspelling Alphabet.https://www.fingerspellingalphabet.com/. Accessed 5 June 2021
21. Islrtc.nic.in. Poster of the Manual Alphabet in ISL | Indian Sign Language Research and Training Center (ISLRTC), Government of India. http://www.islrtc.nic.in/poster-manual-alphabet-isl. Accessed 26 Apr 2021
22. Arduino Forum. Understanding MPU-6050 register raw data? Gyro + Accelerometer. https://forum.arduino.cc/index.php?topic=643304.0. Accessed 26 Apr 2021
23. Amazon.com. https://www.amazon.com/Thalmic-Labs-Gesture-Control-Presentations/dp/B00VHWBH02. Accessed 4 June 2021
24. Flipkart.com. MICROSOFT Kinect Sensor for Xbox 360 Motion Controller - MICROSOFT: Flipkart.com. https://www.flipkart.com/microsoft-kinect-sensor-xbox-360-motion-controller/p/itmew4fyxx3gpypn. Accessed 4 June 2021
25. Ultraleap.com. Tracking | Leap Motion Controller | Ultraleap. https://www.ultraleap.com/product/leap-motion-controller/. Accessed 4 June 2021