A Comprehensive Survey Analyzing Research Solutions developed for Speech Impaired Persons

Sai Yerunkar[1], Sahil Dhapare[2], Priyanka Godbole[3]

[1]B.E. Computer, Marathwada Mitra Mandal College of Engineering Pune, Maharashtra, India.
saiyerunkar2018.comp@mmcoe.edu.in
[2]B.E. Computer, Marathwada Mitra Mandal College of Engineering Pune, Maharashtra, India
sahildhapare2018.comp@mmcoe.edu.in
[3]B.E. Computer, Marathwada Mitra Mandal College of Engineering Pune, Maharashtra, India
priyankagodbloe2018.comp@mmcoe.edu.in

Abstract

The ability to express ourselves using gestures and words is a great treasure for mankind. However, there are some unfortunate people deprived of this ability, hence creating a communication gap between them and normal human beings. The underlying reason for this gap is that while deaf and mute persons make use of sign language to communicate among themselves, normal people are either reluctant to learn it or are unable to comprehend the same.Technology is our best asset to bridge this gap. In this paper, we will be discussing the linguistics of sign language, analysing different sign languages, and certain features of their respective datasets and surveying some of the existing research solutions.

Keywords: Sign Language; Survey; Classifiers; Feature Extractor; datasets; linguistics; HMM; AdaBoost; SVM; HoG; ANN

1. Introduction:

Sign language is a way of non-verbal communication adopted by speech and hearing-impaired people. The multitude of people with hearing disabilities exceeds 400 million today. However, the number of non-impaired who can communicate using sign language is very low. In contrast to spoken language, body movements and facial expressions i.e. using hands, arms or even raising eyebrows or using the mouth are used to convey the desired word. There is diversity in the sign languages that are used in every place on earth including their dialects.Every sign language has its grammar, structure and strong context rules and usage. It is also said to be dependent upon the way of life and dialect of a particular place. However, it varies substantially from the spoken language of that respective nation.

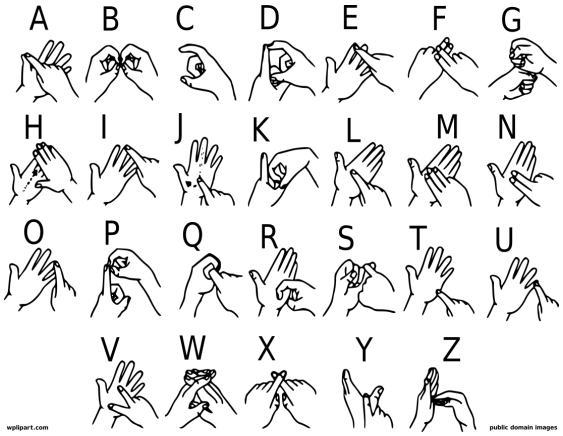

Fig. 1 British Sign Language

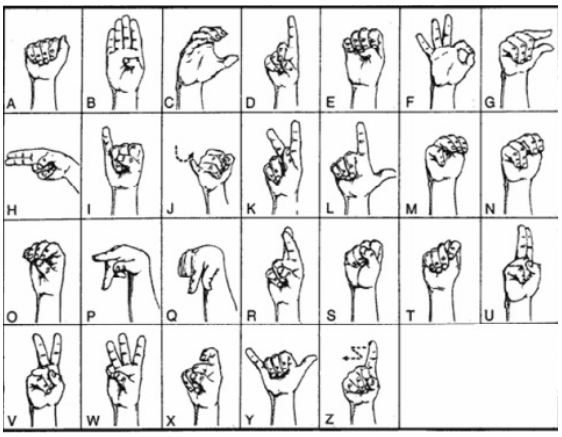

Fig. 2 American Sign Language

This has sparked interest among several researchers in the field to develop a method of communication with the aid of human-computer interaction. Of the several solutions proposed, the most popular one is the sensor-based method. In sensor-based methods, the user is expected to wear wired gloves from which the gestural data can be extracted including the minor details.

Others opt for vision-based methods that acquire data with the help of a camera(s). It includes fingerspelling which is not used in daily life substantially. Gestures that express complete words are used to match the pace of spoken language. This method focuses on the features like colour and texture that account for identifying the gesture. This paper involves segments which are -.

1.1 Motivation:

The motivation behind performing this survey was to enlist and weigh the research that aspires to aid hearing and speech-impaired people. This paper focuses on the hurdles in communication and the insufficiency of technological solutions that are portable and efficient to help them communicate their thoughts in a better manner. This analysis might prove to be a pivotal point for anyone who is trying to develop effective methods to eradicate the barrier of communication between the impaired(hearing and/or speech) and the non-impaired people.

2. Literature Survey:

Building an adept system for sign language processing requires an understanding of the signer culture to create systems that consider the user needs and desires, and of various available sign languages to model systems that take into consideration their intricate linguistic characteristics. Here, we will discuss existing research solutions of sign language recognition and processing systems and the background for the same.

2.1 Sign Language Linguistics:

Similar to verbal languages, sign languages are a part of the natural language set, which have their grammar and concordance, and phonological features as well. All of this is integral to organizing elementary units into meaningful semantic units. On performing a linguistic analysis of sign languages, it was revealed that each sign has three important features: 1) the sign or handshape, 2) the region of the action on the body, and 3) the motion or movement. Recent studies of sign languages provide more in-depth and precision-oriented phonological analyses. For example, in some cases, the motion of the sign provides a grammatical function. Particularly, the direction of motion of the verb can help determine the subject and the object. Fingerspelling (spelling out a word using handshapes representing letters) is often used when referring to the names of people or organizations. Its use is subject to great coarticulation, where the change of handshapes depends on the neighbouring letters. Recognition software must have the ability to pinpoint when a hand shape is being used for fingerspelling or other features. Sign languages aren't entirely expressed by handshapes; motion of the head, mouth, and eyebrows, movement of the shoulders, and eye gaze are all crucial. For Example, raised eyebrows are correlated with an open-ended question and furrowed eyebrows with a yes/no question, particularly in ASL. Sign languages also make abundant use of depiction: adding mouth movements to make modifications or the usage of the body to depict an action, dialogue, or cognitive event. Subtle shifts in body positioning and eye gaze are often used to indicate a referent. Sign language recognition systems should appropriately detect such elements. There is a great variety in sign language execution, based on geographical regions, age, gender, ethnicity, proficiency of the language, education level, etc. Unlike spoken languages, sign languages contain a substantial disparity in fluency. Most deaf children are born to parents with unimpaired hearing, who might not know to sign when the kid is born. Hence, most deaf sign language users learn the language later in their youth, typically leading to lower fluency. Accurate detection and modelling of this variety by the processing software is a must. However, this contributes to the bulk and heterogeneity of the necessary training data.

Including all such factors makes it difficult to estimate the vocabulary size for sign languages. For example, the existing ASL-to-English dictionaries contain around 5000 - 10,000 signs[8]. However, the ways signs can be modulated to add nuanced meaning, adjectives and adverbs, different depictions and classifiers are missing.

2.2 Datasets:

Common sign language datasets have a few flaws that limit the capability and generalizability of the models and systems trained using them.

Size: Current, data-oriented machine learning techniques work best in scenarios of abundant data. Accuracy in speech recognition, which in certain respects is comparable to sign recognition, has been made possible by training on a compilation of millions of words. Inversely, the compilation of sign language

gestures is several orders of magnitude smaller, generally containing less than 100,000 articulated signs.

Continuous Signing: Most of the existing datasets for various sign languages contain isolated signs. Such static sign training data is important for certain scenarios (for example: compiling a sign language dictionary), but most real-life cases of sign language communication involve complete sentences, emotional gestures in grammar, longer utterances, etc.

[8]Native Signers: Many datasets allow students or novices to contribute or even use data scraped from unreliable online resources where signer proficiency is unknown. Some of the datasets include professionally trained interpreters, who may be very skilled but are not native, are also used in some of the datasets. Datasets of native signers must always be considered to train models that reflect this core user group.

Signer Variety: The small size of present datasets and excessive dependence on the content from interpreters results in the lack of signer variety. To meticulously reflect the signing population and realistic recognition scenarios, datasets should contain signers of all ages, gender, geography, culture, fluency, etc. It is also critical to have datasets that are signer-independent, which allows people to assess generalizability by training and testing on a diverse range of signers. Such datasets must be generated for all the various sign languages (ASL, ISL, BSL, etc.).

Dataset	Vocabulary	Signers	Signer-independent	Videos	Continuous	Real-life
Purdue RVL-SLLL ASL [65]	104	14	no	2,576	yes	no
RWTH Boston 104 [124]	104	3	no	201	yes	no
Video-Based CSL [54]	178	50	no	25,000	yes	no
Signum [118]	465	(24 train, 1 test) - 25	yes	15,075	yes	no
MS-ASL [62]	1,000	(165 train, 37 dev, 20 test) - 222	yes	25,513	no	yes
RWTH Phoenix [43]	1,081	9	no	6,841	yes	yes
RWTH Phoenix SI5 [74]	1,081	(8 train, 1 test) - 9	yes	4,667	yes	yes
Design [22]	2,000	8	no	24,000	no	no

Table 1. Popular collections of sign language videos are commonly used as datasets for sign language recognition.

	Speech	Sign Language
Modality	aural-oral	visual-gestural
Articulators	vocal tract	manual, non-manual
Seriality	high	low
Simultaneity	Low	High
Iconicity	low	high
Task	recognition, generation, translation	recognition, generation, translation
Typical articulated-compilation size	5 million words	<100,000 signs
Typical annotated-compilation size	1 billion words	<100,000 signs
Typical compilation-vocabulary size	300,000 words	1,500 signs
What is being modelled	1,500 tri-phonemes	1,500 whole signs
Typical compilation-number of speakers	1,000	10

Table 2. A comparative study of speech vs. sign language datasets.
The enormity of existing corpora for sign language is lesser compared to that of spoken and written languages as for sign language there aren't any parallel written corpora.

2.3 Survey of Existing Research Solutions:

Of all the research papers we analysed, some research papers emphasized the use of artificial neural networks and/or deep learning or transfer learning for recognition. Some papers focus on hand gesture recognition using edge detection or edge orientation histograms. Some papers have concentrated on a particular sign language only. Almost all proposed research solutions mentioned in this survey pursue vision-based methods. These papers provide some promising solutions to bridge the gap between non-impaired and impaired people. Following are extracts of the analysed papers:

In [6], the proposed system makes use of one colour camera to track hands in real-time and interprets American Sign Language (ASL) by utilizing Hidden Markov Models.

The system works in the following stages:

1. Usage of Hidden Markov Models in Gesture Recognition
2. Hidden Markov modelling
3. Tracking Hands in Video
4. Feature Extraction and Hand Ambiguity
5. Training an HMM network
6. Experimentation

With the increase in the volume of the training set and context modelling, expected error rates are lower and generalization to a generalized, user-independent ASL recognition system should be feasible [6]. To get closer to this, the following changes appear to be vital:

- Measuring the position of the hand with respect to each shoulder or a fixed point on the body.
- Added follow-up data for fingers and palms.
- Use a two-camera vision system to help disambiguate the hands in 2D and/or track the hands in 3D.
- Collect appropriate domain or task-oriented data and perform context modelling on grammar/phrase level [6].
- Integrate different features like precise face tracking and facial gestures into the feature set.

In [4], the paper proposes a solution following the steps: selecting and extracting the Region Of Interest ie. RoI, process RoI to elicit and classify the features.

Skin masking is done to focus on the RoI i.e. on hand gestures using OpenCV and open library keeping in mind the fingerspelling method of sign language. For the feature extraction, the ORB(Oriented FAST and Rotated Brief) technique is used.It is open-source, effective, and has no cost issues. FAST is used to identify the key points, Harris corner measures to identify the most prominent N points and BRIEF aids in providing descriptors. This approach however has a smaller dataset with 25 sets of commonly used signs, having a random number of images per set. For clustering, the K-Means approach is adopted. The different classifiers used in this system for comparison are Naive Bayes, SVM, Logistic Regression and KNN. This system considers the following parameters for observation: accuracy, precision, f1 score and recall. An experimental study shows that Naive Bayes has the lowest scores 55.4828, 55.4828, 55.4828, 55.4828 and SVM gives the highest scores as follows: 90.5432, 90.5432, 90.5432, 90.5432. However, this system is not implemented as a stand-alone application and can be improved further with a variety of data sets for different requirements and by using CNN to test its performance on other architectures.

In [2], the paper puts forth a method of applying AdaBoost and Haar Like classifiers for recognizing ASL sign gestures and translating them into text and speech. This system focused on ASL consisting of 24 static postures and 2 dynamic gestures, to compose the words letter-by-letter [2]. The key contributor to the high success rate (98.7%) of this system was the large dataset in the training process. It consisted of hand signs in multiple scales and varying illumination in the complex backgrounds for each hand posture. The key drawback of this system was that the alphabets ‘J’ and ‘Z’, which are originally dynamic gestures and use the movement of the hand, are modified into different static signs. Additionally, signs of E, M, T and S are also changed to avoid ambiguity.

In [3], the proposed system works in the following order: it starts with data acquisition followed by pre-processing and segmentation followed by feature extraction and classification. The database is generated

by capturing videos of 10 signers (using ISL) signing the numbers from 0-9. Along with this, 100 images per number are also included, making a total of 1000 images. For data pre-processing, recognizing skin shading was the primary focus. The most popular method for extracting the hand gesture portion of an image using chrominance values that are available in the MATLAB environment, YCbCr, was used [3]. For the next two steps, different feature extractors and classifiers were used in combination to check accuracies like Shape Descriptor & SVM, SIFT & SVM, HOG & SVM, Combined and HOG & ANN. Shape Descriptor & SVM had the lowest accuracy of 15%, SIFT & SVM had 24%, HOG & SVM was the 2nd most accurate with 96% accuracy, Combined had an accuracy of 93%, HOG and ANN proved to be the best with an accuracy of 99%. MATLAB R2013a was used to implement the entire framework [3].

In [1], the system proposes a desktop application that captures hand gesture images which are fed for pre-processing, and the extracted feature values from the input images are then used as input to the classifier. The motive behind why the authors chose to use the Haar-like features and AdaBoost algorithm is:

- Haar-like feature constructively differentiates between the dark and bright areas within the captured image on a kernel.
- Pixel-based systems are slower as compared to Haar-like feature-based systems.
- Additionally, the Haar-like features are comparatively robust to noise in the background of the image and various lighting conditions as it calculates the grey-level difference between the light and dark rectangles [2].
- The AdaBoost algorithm effectively enhances the system's learning accuracy.
- It can pliantly select the best features at each stage and combine a series of weak classifiers into a strong one [2].

The limitations of this system are that it does not employ any existing dataset and is only tested on two signs: the "palm" gesture and the "fist" gesture [2]. It also does not consider the robustness of real-life cases such as a different human hand in different colours and sizes.

This system put forward by [7] makes use of OpenCV and CNN to capture images and convert the ISL gestures to text, which are then converted to an mp3 audio file using the gTTS library. The image capturing of hand gestures is done using a green coloured glove worn by the user.Although ISL itself involves two-handed gestures, all the existing datasets for ISL consist of single-handed gestures [7]. The authors have themselves created a dataset consisting of ISL A-Z alphabets having 1750 images each, which were further used for training the model [7]. The procedure involves recognizing each alphabet and concatenating them to form sentences, which are further transformed into speech.
Although this system provides the benefit of scalability and no hardware requirements, apart from a laptop, it also has the drawbacks of noise in the background of the captured images and lighting in the room and does not take into consideration the signer variety.

3. CONCLUSION:

In this paper, we provide a survey of different research solutions proposed for sign language recognition, generation, and translation, providing background on hearing and speech-impaired culture and sign language linguistics that is often overlooked, an aggregation of vital difficulties, and an appeal for the researchers. In doing so, this paper serves to direct readers both outside and within the computer science domain to this problem, highlights the various opportunities for a collaborative approach with the deaf-mute community, and helps the researchers prioritize which hurdles to tackle next.

4. REFERENCES :

[1] Ruchi. M. Gurav & Premanand K. Kadbe "Vision Based Hand Gesture Recognition with Haar Classifier and AdaBoost Algorithm" International Journal of Latest Trends in Engineering and Technology Vol.5 Issue 2 March 2015.

[2] Vi N.T. Truong, Chuan-Kai Yang & Quoc-Viet Tran "A Translator for American Sign Language to Text and Speech" IEEE 5th Global Conference on Consumer Electronics, 2017.

[3] Miss. Juhi Ekbote & Mrs. Mahasweta Joshi “Indian Sign Language Recognition Using ANN And SVM Classifiers” International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS), 2017.

[4] K.Revanth & N. Sri Madhava Raja “Comprehensive SVM based Indian Sign Language Recognition” Proceedings of International Conference on Systems Computation Automation and Networking, 2019.

[5] Kusurnika Krori Dutta, Satheesh Kumar Raju, Anil Kumar & Sunny Arokia Swarny “Double Handed Indian Sign Language to Speech and Text” Third International Conference on Image Information Processing, 2015

[6] Thad Starner & Alex Pentland “Real-Time American Sign Language Recognition from Video using Hidden Markov Models” Kluwer Academic Publishers, 1997.

[7] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans & Benjamin Schrauwen “Sign Language Recognition Using Convolutional Neural Networks” Springer International Publishing Switzerland, 2015.

[8]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. 10.1145/3308561.33537