EasyChair Preprint

№ 4855

Camera-based Sign Language Recognition and Simultaneous Speech Generation: A Survey

Ayushi Patani, Varun Gawande, Jash Gujarathi and Vedant Puranik

January 3, 2021

Camera-based Sign Language Recognition and Simultaneous Speech Generation: A Survey

Ayushi N. Patani, Varun S. Gawande, Jash V. Gujarathi, Vedant K. Puranik, Prof. Tushar A .Rane
Department of Information Technology
Society for Computer Technology and Research's Pune Institute of Computer Technology
Affiliated to Savitribai Phule Pune University (formerly known as Pune University)
Survey no 27, Dhankawadi, Pune - 411043 , Maharashtra, India

***Abstract*—People regularly face problems for interpreting deaf-mute people, who primarily use sign language for communication amongst themselves and others. Despite efforts being conducted by different governments worldwide such as the provision of a sign language expert for interpreting and communicating all news to the impaired by the New Zealand media, active participation of the impaired is still at a very rudimentary stage. Further, only a few people today are proficient in communicating via sign language and hence the majority of the population at large is still devoid of any understanding of the matter. This could be problematic for deaf-mute people especially during situations of distress like pain, fraud, or other emergency situations like fire, kidnapping, etc. All these problems could be minimized substantially if this language barrier is effectively bridged. This paper aims to contemplate a number of research papers that deliver regarding the same topic. We look to understand the various ways of machine-based sign-based sign language. One should keep an eye out to understand the feasibility, methodology, and accuracies discovered in some papers.**

Keywords — Hilbert Curve, Support Vector Machines, Random Forests, Artificial Neural Network, Feed-forward Backpropagation, Hough Transform, Convolutional Neural Networks, Stacked deionised decoders, Multilayer Perceptron neural network, Adaline neural network

I. INTRODUCTION

Using statistics of the World Health organization, there are 466 Million hearing disabled people and a million people that are speech impaired. This rounds up to over 5% of the World's Population that cannot be communicated with by using conventional speech based approaches.

Sign languages have been the most widespread method of communicating with members of the deaf-mute community throughout history, even being mentioned by Socrates in Plato's Cratylus.

Multiple books and scholarly articles were written from the 16th to 18th century in European countries regarding instructions on how to communicate with and teach deaf-mute people. These books formed the basis for multiple sign languages ,like the British Sign Language (BSL) and the French Sign Language (FSL) , American Sign Language (ASL) (based on the FSL),New Zealand Sign language and the Sign Language used in Spain and Mexico. However, until the 19th century most of these sign languages were mainly based on fingerspelling systems that transferred spoken language to a sign language and vice versa. Sign languages have evolved since then to develop more complex relations with the languages spoken in the land. Hence developing multiple dialects and variations from country to country.

Some notes about Sign Language that readers must be aware of are :

a.) Sign languages have an equally vibrant vocabulary as spoken languages and exhibit all fundamental structures that exist in all spoken languages.

b.) Just like in spoken languages words do not have any onomatopoeic relation with the referent they are describing, sign languages do not have any visual relevance with what they convey.

c.) Just like spoken languages use grammar to turn words into meaningful sentences, Sign languages have semantics that organize elementary meaningless units into meaningful units/phrases.

d.) Unlike spoken languages, sign languages convey meaning by simultaneous meanings by the main articulators i.e.the head and the hands.

Given the significant percentage of people that rely upon Sign Languages to be their primary mode of communication, it is imperative for the wider public outside the deaf-mute communities to be aware of sign languages to some extent at least. However, normal individuals have very less incentive to learn even basic Sign language. For eg. In India there are only 250 certified sign language interpreters that translate for a community of up to 7 million people. The current situation creates an overwhelmingly exclusionary society for the deaf-mute community . Given the increasing impetus of communication skills in the workforce, the deaf mute community are presented with a very high barrier of entry to participate in society as a fully functioning member. These people are dependent on their people close to them that have taken the effort to be able to understand and converse with them to be able to interact with society.

It has been increasingly evident that a technological solution is needed to bridge the communication gap that exists between the members of the speech impaired community and society as this community is most vulnerable to being left behind in the technical revolution of the past few decades.

Section II highlights the different approaches that can be adopted for capturing videos and/or images for interpretation. Section III and IV talk about the literature survey and the proposed methodology respectively. The paper is concluded in section V.

II. APPROACHES

Camera-based image or video capturing has been one of the widest implemented and effective methods used in sign language interpretation systems. Using this technique, researchers have been successful in interpreting sign language by capturing gestures of only one hand, both hands and static or dynamic images. The signs could therefore be either isolated signs or continuous signs. In case of videos, they are first captured and broken down into frames of images that can then be passed onto the system for further analysis and interpretation. Hence, overall a stream of images is passed to the system, after which different techniques as per application are utilized to obtain results.

Using Kinect is another approach that has started receiving recognition from the research community. Microsoft Kinect is a motion camera device that captures users' movements in real-time. Kinect has been primarily used for gaming purposes in the recent past. It [1] provides a significant advantage over the camera-based approach as it is not restricted to 2D image/video capturing, but can also capture depth information such as color depth, etc. effectively. However, the maintenance and overall costs pose a higher overhead than the camera method and is hence not so commonly adopted for commercial purposes.

The armband is a technique that depends on Electromyography (EMG) Signals. These signals are generated in our muscles whenever there is any movement. The data is collected [2] from the signers arm through a band in the form of signals and then processed to interpret sign language. One of the greatest advantages this method assures over camera- and Kinect-based methods is zero dependency on light. However, to detect signals effectively, a lot of wires need to be connected to the band and the portability is also an issue and proves to be a setback over the former two approaches.

A glove can also be used which primarily relies on the path-breaking innovation [3] done in 1993 called a Cyber Glove. For getting data, the signers wear this glove which comes with a number of sensors for each finger attached to it. A motion tracker [4] is also employed along with the glove to track the orientation and position of hands which is then connected to a computer via serial ports. It provides an easy way of detecting sign language, however, a lot of equipment needs to be appropriately set and configured for use. This is not feasible in real-world situations such as on roads, ships, shopping centres, etc. Moreover, it is also unable to capture facial features and symbols which can be easily done in camera-based systems.

The leap motion technique [5] makes use of a cost-effective sensor, called Leap Motion Controller, based system. Here the information about hand and finger movements is captured by this sensor via APIs designed for the same. This is done by performing the movements a few feet above the horizontally positioned sensor. This data is then sent to a computer via USB. This approach ensures a cheaper solution to the glove- and kinect-based approaches however still faces the same challenges as the two mentioned.

Brain-Computer Interfacing is an advanced approach to identifying sign language. Electroencephalogram [6] brain activities are obtained for the recognition of sign-language. This approach goes one-step further by completely eliminating the need of any physical movements for detection of sign language. Here, brain waves are made use of, which are then directly transmitted to a computer with the help of Bluetooth. Other techniques like functional magnetic resonance imaging i.e. fRMI [7] and Electrocorticography [8] are also used in a similar fashion. They face a major problem of implementation complexity and still rely on using devices connected to the head to detect signals.

III. LITERATURE SURVEY

[9]M. Qutaisht et al developed a system for automatically translating the static gestures in the American Sign language(ASL). They used Hough transform and neural networks which essentially deals with images of bare hands and thus allows the user to interact with the system in a natural way. The image was converted into a feature vector which is compared with the feature vector of the training set. The striking feature of this method is that it is not affected by rotation, scaling or translation, thus making the system more flexible. The system was implemented and tested against 300 samples of hand gestures with 15 images for each sign and an accuracy of 92.33% was achieved.

[10] Hardik Rewari et al proposed a sign language interpreter that automatically converts the sign language video input into audio output. They essentially worked on the Indian Sign Language (ISL) to aid the deaf and dumb Indian people.They used various compoenents like flex sensors, MPU6050, HC-05, SD-Card module and tested around 90 words out of the 3000 words present in the ISL. Remaining words can be interpreted during the same algorithmic techniques.

[11]Rajaganapathy. S et al converted the human sign language to audio with the help of gesture understanding and motion capture using a motion capture device called Microsoft Kinect. The device captures 20 human joins and gestures. The input to the device would be live actions of human gestures. Once the skeleton is identified, it keeps a track of the gestures and matches with the user defined gestures.Their range was 40 cm to 4m and the frame could identify gestures of maximum 2 people at a time.

[12]Sarbjeet Kaur et al provided a solution to interpret the Indian Sign Language(ISL) which involves recognition of gestures of alphabets.An image of hand gestures is captured, processed and converted to an Eigenvector. These Eigenvectors are then compared with those of the training set of signs. MATLAB coding is used for feature extraction in the form of Eigenvectors.A dataset of 650 samples of hand signs was implemented and tested with 25 images for each sign.

[13] Amira Ragab et al explore in this paper a new method for the representation of hand-based images. It is based on the Hilbert space-filling curve. The hands are segmented, which is then followed by the application of the Hilbert space-filling curve to extract a feature vector. After which, classifiers like Support Vector Machines (SVM) or Random Forests (RF) classify the gestures. The accuracy is 99% when it comes to images with Uniform Backgrounds but

heavily falls to 69% when introduced to noise, non-uniform backgrounds.

[14] W. Tangsuksant, S. Adhan and C. Pintavirooj et al discuss the following procedure: one or more Standard Definition (SD) cameras capture the subject and extract 3D Maker Coordinates using the well-known DLT algorithm. This paper then proposes a new feature to compute all feasible triangle area patches constructed from 3D coordinate triplets. For training the model, an Artificial Neural Network is used with feed-forward backpropagation training. The training process uses around 2,100 images and the average accuracy of the algorithm turns out to be 95%.

[15] Md. Mohiminul Islam, Sarah Siddiqua and Jawta Afnan propose a real time hand gesture recognition system based on the American Sign Language. It achieves higher accuracy by using a novel approach in the feature extraction step which includes combining K Curvature and Convex Hull Algorithms allowing for better detection of fingertips in sign language gestures. This allows their Artificial Neural network to recognize 37 signs of the ASL with a 94.32% accuracy.

[16] In contrast to earlier research which focuses on identifying hand gestures from the American Sign Language from a set of fairly distinguishable gestures which makes the classification easier and seem more robust, Oyebade K. Oyedotun and Adnan Khashman work on distinguishing 24 Signs that are modelled into sets of relatively similar gestures. By deploying Convolutional Neural Networks and Stacked deionised decoders, they achieve an accuracy of 92.8% on test data that the model has not been trained for. They further opine that the problems with using CNNs for an increasing depth could be overcome by using rectified linear activations in the hidden layers and hence control the effect of neuron saturation and vanishing gradients.

A Multilayer Perceptron neural network was also used by Karayilan and Kiliç [17] on the Marcel Static Hand Posture Database which is an American Sign Language dataset. They used the camera-based approach and were successful in extracting histogram and raw features from their data. They use these raw and histogram features on two different classifiers. 70% and 85% accuracy was obtained on these classifiers respectively.

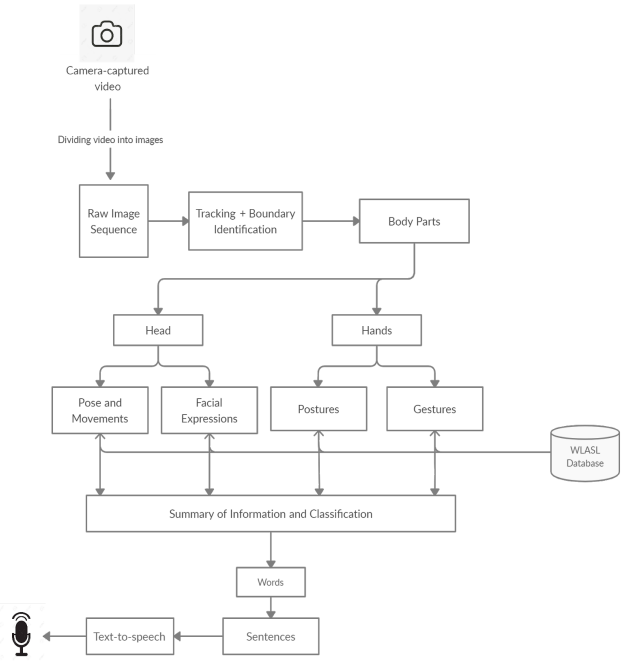
Saha et. al [18] proposed a novel approach to sign language recognition by using a modification over the traditional Adaline neural network. They performed recognition on English alphabets. Adaline networks are not capable of working on non-linear classification on their own and hence they were extended by using multiple Adaline neurons. The classification was done by using a voting technique, finally achieving an accuracy of about 94% with an F-Score of almost 90.

IV. PROPOSED METHODOLOGY

Our project aims to capture sign language performed by signers on a real-time basis and interpret the language to produce textual and audio output for the illiterate. For this, a camera-based approach will be made use of, owing to the ease of portability and movement that the camera-based method offers over other techniques.

The video of the signer will be first captured by a camera-enabled device. This video will then be processed by our application. The video would be divided into a number of frames which will convert the video into a raw image sequence. This image sequence will then be processed to initially identify the boundaries. This will be useful to separate the different body parts being captured by the camera into two major subparts - head and hands.

The head subpart will be further categorized into pose and movements as well as facial expressions. Postures and gestures will be extracted from the movement of the hands. All of the data will then be matched against the WLASL Dataset which would then be used for classification purposes. The classification will result in generation of words.



Words generated from sign language will not adhere to grammatical rules of English. Hence semantically correct sentences will be generated by the sentence generation module. For this Google's T5 model [19] will be put into use. Finally, this output will be sent through an audio generator to generate speech from the same. This provides support for illiterate people, who are not entitled to understanding written text.

V. CONCLUSION

In this survey paper, we go over a vast variety of camera-based implementations for one hand, two hands, static and dynamic images. We also go over other novel approaches like using Microsoft Kinect, Electromyography (EMG) Signals from armbands, Gloves, Motion Trackers, the cost-effective Leap Motion Controller, and much more. As most common people do not know sign language, we believe that our research could pave the way for making society more inclusive towards the historically isolated and disenfranchised speech impaired people. Applications of this research at scale, provide a simple, seamless, and highly available means to communicate with other members of society.

REFERENCES

[1] Sun C, Zhang T, Bao BK, Xu C (2013a) Latent support vector machine for sign language recognition with Kinect. In: 20th IEEE international conference on image processing (ICIP), pp 4190–4194

[2] Savur C, Sahin F (2016) American Sign Language recognition system by using surface EMG signal. In: IEEE international conference on systems, man, and cybernetics (SMC), pp 002872–002877

[3] S. S. Fels and G. E. Hinton, "Glove-talk: a neural network interface between a data-glove and a speech synthesizer," IEEE Transactions on Neural Networks, vol. 4, no. 1, pp. 2–8, 1993.

[4] Oz C, Leu MC (2011) American Sign Language word recognition with a sensory glove using artificial neural networks. Eng Appl Artif Intell 24(7):1204–1213

[5] Chuan CH, Regina E, Guardino C (2014) American Sign Language recognition using leap motion sensor. In: 13th IEEE international conference on machine learning and applications (ICMLA), pp 541–544

[6] AlQattan D, Sepulveda F (2017) Towards sign language recognition using EEG-based motor imagery brain computer interface. In: 5th IEEE international winter conference on brain–computer interface (BCI), pp 5–8

[7] N. A. Mehta, T. Starner, M. M. Jackson, K. O. Babalola, and G. A. James, "Recognizing Sign Language from Brain Imaging, " in 2010 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 3842-3845.

[8] M. G. Bleichner and N. F. Ramsey, "Give Me a Sign: Studies on the Decodability of Hand Gestures Using Activity of the Sensorimotor Cortex as a Potential Control Signal for Implanted Brain Computer Interfaces, " in Brain-Computer Interface Research, C. Guger, T. Vaughan, and B. Allison, Eds. Springer International Publishing, 2014, pp. 7-17.

[9] Munib, Qutaishat & Habeeb, Moussa & Takruri, Bayan & Al-Malik, Hiba. (2007). American sign language (ASL) recognition based on Hough transform and neural networks. Expert Systems with Applications. 32. 24-37. 10.1016/j.eswa.2005.11.018.

[10] Rewari, Hardik & Dixit, Vishal & Batra, Dhroov & Nagaraja, Hema. (2018). Automated Sign Language Interpreter. 1-5. 10.1109/IC3.2018.8530658.

[11] Rajaganapathy, S. & Aravind, B. & Keerthana, B. & Sivagami, M.. (2015). Conversation of Sign Language to Speech with Human Gestures. Procedia Computer Science. 50. 10.1016/j.procs.2015.04.004.

[12] Kaur, Sarabjeet and V. Banga. "Vision Based Static Hand Pose, Hand Movement Recognition System For Sign Language Using EigenVector Theory in MATLAB." viXra(2014).

[13] Ragab, Amira & Ahmed, Maher & Chau, Siu-Cheung. (2013). Sign Language Recognition Using Hilbert Curve Features. 7950. 143-151. 10.1007/978-3-642-39094-4_17.

[14] Tangsuksant, Watcharin & Adhan, Suchin & Pintavirooj, Chuchart. (2014). American Sign Language recognition by using 3D geometric invariant feature and ANN classification. 1-5. 10.1109/BMEiCON.2014.7017372.

[15] M. M. Islam, S. Siddiqua and J. Afnan, "Real time Hand Gesture Recognition using different algorithms based on American Sign Language," 2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Dhaka, 2017, pp. 1-6, doi: 10.1109/ICIVPR.2017.7890854.

[16] 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.

[17] Karayılan T, Kılıç Ö (2017) Sign language recognition. In: IEEE international conference on computer science and engineering (UBMK), pp 1122–1126

[18] .Saha S, Lahiri R, Konar A, Nagar AK (2016) A novel approach to American Sign Language recognition using MAdaline neural network. In: IEEE symposium series on computational intelligence (SSCI), pp 1–6

[19] Exploring Transfer Learning with T5: the Text-to-Text Transfer Transformer by Adam Roberts and Colin Raffel on 24/2/2020, Google AI Blog, https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html