IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⋇ Impact Factor 7.39 ⋇ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11609

# Gesture Recognition to Text and Voice

**Varsha S[1], Vidya D[2], Mrs. Asma Begum[3]**

[1,2]UG Student, Department Of Information Science And Engineering, Atria Institute Of Technology,

Bangalore, Karnataka, India.

[3]Professor in the Atria Institute of Technology's Department of Information Science and Engineering,

Bangalore, Karnataka, India.

**Abstract:** Sign language is a main mode of communication for vocally disabled. This language uses a variety of symbols, including finger signs, expressions, and a combination of the two, to express information. This system offers a novel method for translating sign action analysis, recognition, and generation of a written description in English language using mobile applications. Training and testing are two crucial procedures that are used. Each domain in the training set has 5 video samples, and each video sample has a class of words associated with it that will be stored in a database. Pre processing on the test sample is done using the median filter, the clever operator for edge detection, and the HOG for feature extraction.The text description will finally be produced in English. The performance efficiency over the traditional model is validated by the minimal average computation time, acceptable recognition rate, and performance.

## INTRODUCTION

The goal of activity recognition is to identify one or more agents' actions and objectives from a collection of observations on those actions and the surrounding circumstances. This research area has drawn the interest of numerous computer science communities since the 1980s.

The text description will finally be produced in English. The performance efficiency over the traditional model is validated by the minimal average computation time, acceptable recognition rate, and performance.

In image processing, the input is initially an image, which is then processed in accordance with the specifications. Image processing requires a variety of inputs, including video, images, and the collection of frames from videos. The output is usually in the form of an image or a group of parameters connected to images.

The use of image processing for improve the image quality and gather useful information in the image this process is called as feature extraction. This image processing techniques can be used for detecting the hand gesture or analysing actions or many other purposes in various fields. These image processing techniques are employed in the system created for communication purposes in a community where people with hearing and vocal impairments live. The primary distinguishing feature that distinguishes a group of hard of hearing people is thought to be the body language. Making sure that hard of hearing persons have communication of chance and full commitment in the public sphere is a crucial component of communication by gestures acknowledgment plots in general society human advancement. Fundamentally, communication by gestures is talked to by an underwriter continuously varying various hand forms.

Using our hands, arms, fingers, and eyes to make signs, we may physically communicate with hearing-impaired and illiterate persons. Seeing human action from picture game plans is a champion among the most troublesome issues in computer vision with various imperative applications, for instance, tuning video perception, content-based video recuperation, human-robot collaboration, and splendid home. The task is challenging not only because of the differences between classes, camera advancements, establishment confusion, and fragmented impediment, but also because of some commonalities between classes, including sprinting instead of walking or running. Prior attempts at affirming human movement in videos typically used general representations.

The dynamic signals verification applications need for the acquisition of a high data rate of hand positions typically provided by development-following gloves that are configured to do unambiguously record finger joint developments through flex sensors in an immovably fitted glove. A distinctive and distinguishing correspondence philosophy for human–computer collaboration is provided by the hand flag. To enable the computer to ostensibly see sustained hand gestures, effective human-computer interactions (HCIs) must be developed. However, because to the erratic nature of hand signs, which are rich in variety as a result of the high Degrees of Flexibility (DOF) required by the human hand, vision-based hand following and movement affirmation is a challenging issue. The hand movement HCIs must satisfy the requirements with regard to progressing execution, affirmation precision, and vigour against alterations and jumbled foundation in order to successfully complete their role.

The gesture-based communication correspondence comprehension entails a semantic analysis of hands following, hands shapes, hands presentations, sign verbalization, as well as basic etymological information discussed with head movements and external manifestations. Motion-based communication is seen from various angles and reveals a variety of language structures, such as grammatical intricacies, references in virtual checking spaces, and facial and hand expressions. When compared to talk-based correspondence affirmation, the major drawback of signal-based correspondence affirmation is having to simultaneously observe a guarantor's many correspondence features, such as hands and body advancement, external appearances and body act. For an affirmation structure that is superior than the norm, these qualities must be considered simultaneously. The researchers' attempt to establish a model for spatial information that contains the components made in tail able in the movie is the second major problem that motion based correspondence affirmation structure engineers are up against. Many professionals are researching this as a checking space. Through marking discussion, a crucial test stood in the middle of the correspondence. This study tackles the issue of action recognition, or how to identify the kind of action taking place in a film. Here, the issue of video representation—specifically, how to encode videos in a reliable manner—is taken into account. Which representation type is appropriate for a wide range of action classes, tasks, and video types? This study demonstrates a system for recognising sign language hand gestures and translating them into their English equivalents. Therefore, the suggested approach offers numerous opportunities for a person with a vocal disability who can express their thoughts and talents on paper.

## LITERATURE SURVEY

### Paper I: Implementation of gesture based voice and language translator for dumb people

Dumb persons communicate through gestures which are not understood by the majority of people. A gesture is a motion of a body part, most often the hand or the head, to convey a thought or meaning. This paper suggests a system that translates the English voice output from user gestures made in the form of English alphabets into any other Microsoft supported languages. The system is made up of a speaker, a three-button keypad, and MPU6050, which detects gesture movement. The algorithm for trajectory recognition is used to recognise alphabets in its implementation. Using voice RSS and Microsoft Translator, the Raspberry Pi produces voice output for the text in a variety of languages. During testing, the system identified the A-Z alphabets and produced vocal output in a variety of languages using the motions.

### Paper II: A Gesture-to-Emotional Speech Conversion by Combining Gesture Recognition and Facial Expression Recognition

In order to address the communication issues between speech impairments and healthy persons, this article suggests a facial expression integrated sign language to emotional speech conversion approach. First, a deep neural network (DNN) model is used to derive the properties of sign language and the features of face expression. Secondly, a support vector machine (CNN) are trained to classify the sign language and facial expression for recognizing the text of sign language and emotional tags of facial expression. At the same time, a hidden Markov model-based Mandarin-Tibetan bilingual emotional speech synthesizer is trained by speaker adaptive training with a Mandarin emotional speech corpus. Finally, the identified text of sign language and emotional tags is converted into Mandarin or Tibetan emotional speech. According to the objective tests, static sign language is recognised 90.7 percent of the time. The recognition rate of facial expression achieves 94.6% on the extended CohnKanade database (CK+) and 80.3% on the JAFFE database respectively. According to subjective analysis, emotional speech that has been artificially sythesized can get a mean emotional opinion score of 4.0. The PAD values for both facial expression and synthesised emotional speech are assessed using the pleasure-arousal-dominance (PAD) tree dimensional emotion model. The findings indicate that the PAD values of facial expression and synthetic emotional speech are comparable. This means that visual expressions of emotion can be expressed through emotional speech synthesised in speech.

### Paper III: Quantifying the effects of varying light-visibility and noise-sound levels in practical multimodal speech and visual gesture (mSVG) interaction with aerobots

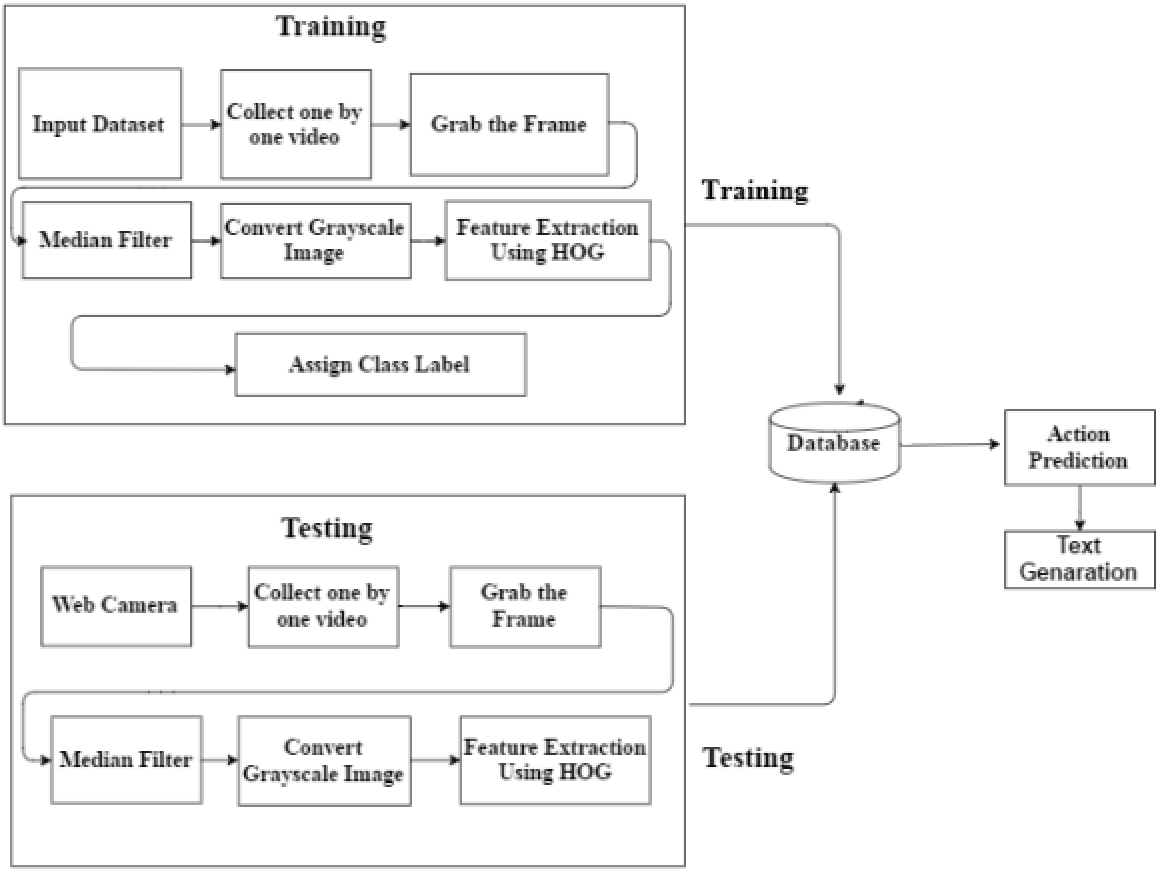
In order to support the design and development of a multimodal speech and visual gesture (mSVG) control interface for the management of a UAV, this paper covers the research work done to quantify the effective range of lighting levels and ambient noise levels. Under controlled lab conditions, noise levels ranging from 55 dB to 85 dB are studied in order to pinpoint the locations where verbal commands for a UAV fail, think about why, and perhaps offer a remedy. To define a range of effective sight levels, lighting levels are also changed within the control lab environment. The work's limitations and some follow-up research were also discussed.

DOI: 10.17148/IJARCCE.2022.11609

**Paper IV: A Survey of Hand Gesture Recognition methods in Sign Language Recognition.**

HMM-based SLR approaches have been shown to achieving good recognition accuracy especially in small to medium sized datasets. Efforts on improving the performance of HMM-based approaches have also been proposed by modifying the standard HMM method. Designing invariant sign features can be tedious works highly dependent on the type of input data being used. Moreover, feature extraction also contributes to the computational load and the classification performance often depends on the quality of the extracted sign features. Combinations of CNN and HMM have also been proposed to improve the performance of the SLR system.

## FLOW CHART



## IMPLEMENTATION

The method is made to visually identify all static English Sign Language (ESL) signs and all alphabetic signs made with bare hands. It is not necessary for the user or signers to utilise any equipment or gloves in order to engage with the system. But because different signers have distinct hand shapes, body sizes, operating styles, and other characteristics, this makes it harder to recognise them. In order to increase the system's stability and applicability in the future, it recognises the need for signer independent sign language recognition.

A approach based on recognition rate is suggested by the system after comparing the three feature extraction techniques utilised for ESL recognition. It is predicated on the gesture being represented as a translation, rotation, and scale invariant feature vector. The ESL recognition system has been successfully developed as a result of the successful integration of the feature extraction method with superior image processing and neural network capabilities. The feature extraction phase and the classification phase are the two phases of the system. The system can handle photographs with a uniform background because they were created using portable document format. In order to extract the desired features from the digitised sign, an image processing technique called feature extraction was used. During this phase, each coloured image is resized and then converted from RGB to grey scale one. The next step is an edge detection method. The aim of edge detection is to identify the locations in an image where the intensity abruptly changes. Sharp changes in image qualities typically correspond to significant discoveries and modifications to the physical world. The next important step is the application of proper feature extraction phase and the next is the classification phase, a three-layer, feed-forward back propagation neural network is constructed.

## RESULT

The ability of the recognition system to categorise signs for both the training set and the testing set of data is tested to determine how well it performs. The impact of the neural network's input volume is taken into account.

### 1. User Input

The system for recognising sign language will take input from the user in the form of hand movements. During the training phase of the application, the user creates a data base of his /her hand sign gesture images. The training phase is complete when the system has captured enough gestures for which it knows the class and the system is then ready to recognize gestures. The gestures are captured by the webcam with a constant distance between the hand of the user and the camera. The illumination should also ideally be constant. a block colour board is kept behind to make the background constant and make the hand area identification process easier.

### 2. Data set

The data set used for training and testing the recognition system consists of grey scale images for all the KSL signs used in the experiments. Additionally, 8 separate participants will provide 8 samples for each sign. For each sign 5 out of 8 samples will be used for training purpose while remaining 5 signs were used for testing. Web cameras will be used to take the samples at various distances and in various orientations. In this manner, a data set with cases of various sizes and orientations will be obtained, allowing for the examination of the feature extraction scheme's capabilities.

## CONCLUSION

To produce grammatically sound words, the designed system is processed for real-time sign action. When features are collected from distinct samples, a class label is applied to the features that were extracted, and text is then produced. In comparison to a typical processing system, the purposed method yields improved retrieval accuracy. This approach achieved the goal of more accuracy and less processing overhead by producing lower descriptive features with fewer processing frames. A alternative technique can be used in future work with the system to minimise processing time and maximise recognition rate. Future plans include the creation of a system that will produce summaries and is signer independent.

People who are deaf or dumb rely on interpreters to communicate. However, they are unable to rely on interpreters on a daily basis, primarily because of the high fees and the challenge of finding and arranging capable interpreters. The quality of life for those with disabilities will considerably improve because to this method. The idea of automatically decoding sign language is intriguing; the technology is already available, and the prospective uses are intriguing and valuable. The system has been shown to be resistant to changes in gesture. We obtain the incorrect findings using the histogram technique. Hence histogram technique is applicable to only small set of KSL alphabets or gestures which are completely different from each other. It does not work well for the large or all 592 number of set of KSL signs. Segmentation is advised for a larger set of sign motions. How well one can differentiate using this method is the major issue. The photographs have the biggest role in this, but the algorithm also plays a role. It could be improved by utilising other image processing methods, such as edge detection, as demonstrated in the presenting work. To find the edges with various thresholds, we used well-known edge detectors like Canny, Sobel, and Prewitt operators. With a 0.25 threshold value, Canny performs well. Using edge detection along with segmentation method good recognition rate is achieved. Also the system is made background independent. As we have implemented sign to text interpreter in future reverse system can also possible to implement that is text to sign interpreter.

## ACKNOWLEDGMENT

Thanks to the project guide **Ms Asma Begum**, thank you for guiding us through such great support and research. She gave the paper the insights and expertise needed to make it look good. Her advice, professional insight, and encouragement proved to be invaluable as a guide.

DOI: 10.17148/IJARCCE.2022.11609

## REFERENCES

[1] L. Anusha, Department of ECE, Vignan's Lara Institute of Technology and Science, Guntur, A.P, India. Y. Usha Devi, Department of ECE, Vignan's Lara Institute of Technology and Science, Guntur, A.P, India. Implementation of gesture based voice and language translator for dumb people (2017)

[2] Nan Song, Hongwu Yang, Peiwen Wu, College of physics and electronic engineering, Northwest Normal University, Lanzhou,730070, China. A Gesture-to-Emotional Speech Conversion by Combining Gesture Recognition and Facial Expression Recognition (2018)

[3] Ayodeji O. Abioye, Stephen D. Prior, Glyn T. Thomas, Peter Saddington, Sarvapali D. Ramchurn, Sarvapali D. Ramchurn, Faculty of Engineering & the Environment, University of Southampton, UK. Quantifying the effects of varying light-visibility and noise-sound levels in practical multimodal speech and visual gesture (mSVG) interaction with aerobots (2018)

[4] Suharjito, Meita Chandra Ariesta, Fanny Wiryana and Gede Putra Kusuma, A Survey of Hand. Gesture Recognition methods in Sign Language Recognition. (2019)