International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 01, 2020
ISSN: 1475-7192

A Survey of Approaches for Sign Language Recognition System

N. Kaushik, Vaidya Rahul and K. Senthil Kumar

Abstract--- *Sign language is a form of communication for the deaf and dumb community with the rest of the world. But most of the people do not know or understand the sign language of these people which makes it very difficult for them to communicate with the rest of the world. The sign language recognition can have different level of success when it is based on different image processing techniques. There are a large number of sign languages which differ according to regions like the Indian Sign Language, Taiwanese Sign Language etc. The sign language can be represented as text by making use of Convolutional Neural Networks, Support Vector machines and other methods.*

Keywords--- *Sign Language, Communication, Convolutional Neural Network, Computer Vision, Support Vector Machines.*

I. INTRODUCTION

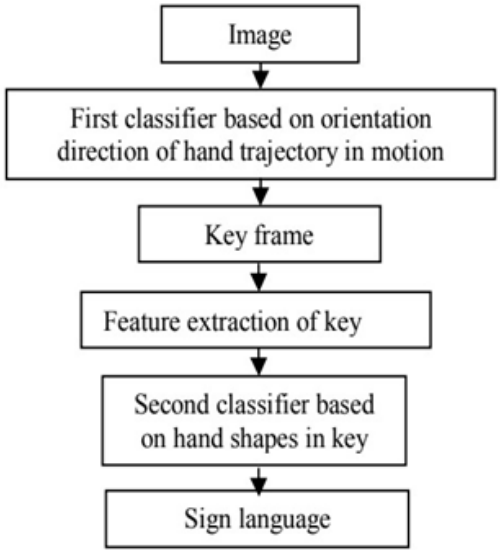


Figure 1: Flow diagram of proposed sign language recognition system

Currently, around 17% of the adults in the world are deaf and dumb. It becomes very difficult for us to identify a deaf person unlike a blind person who can be easily spotted. One way of communicating with the rest of the world is by using interpreters. But hiring an interpreter is a very costly option which the poor people cannot abord. A sign language recognition will ease their communication by translating their gestures to the rest of the world enabling them to communicate with the rest of the world.

N. Kaushik, SRM Institute of Science and Technology, Kattankulathur.
Vaidya Rahul, SRM Institute of Science and Technology, Kattankulathur.
K. Senthil Kumar, SRM Institute of Science and Technology, Kattankulathur.

DOI: 10.37200/IJPR/V24I1/PR200278
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

Most of the sign language recognition systems take the hand gestures as input and based on the features of the gesture, determine the meaning for the gesture. The output can be expressed either as text by displaying the data on the screen or they can be expressed as voice output where we can use a raspberry pi or a speaker device to communicate the meaning of the gesture.

II. CURRENT TECHNIQUES

There are various techniques which have been explored for performing sign language recognition starting from the basic image processing, Support Vector Machines, Convolutional Neural Networks and 3-D Convolutional Neural Networks. The s e techniques are being discussed below.

A) Sign Language Recognition using image processing

This system in [2] makes use of the basic image processing algorithm to recognise the gestures of the Taiwanese Sign Language. The process involves two layer classification: movement of hand and the gestures and selecting the right key frames as shown in figure 1.

The history of movement of the image along with fourier descriptors are used by [2] and [7] for movement detection recognition and for selecting the key frames respectively. Feature Extraction is done from the key frames using a generic cosine descriptor (GCD). The GCD is independent of scale and rotation of the hand shapes which makes the device more efficient to various factors.

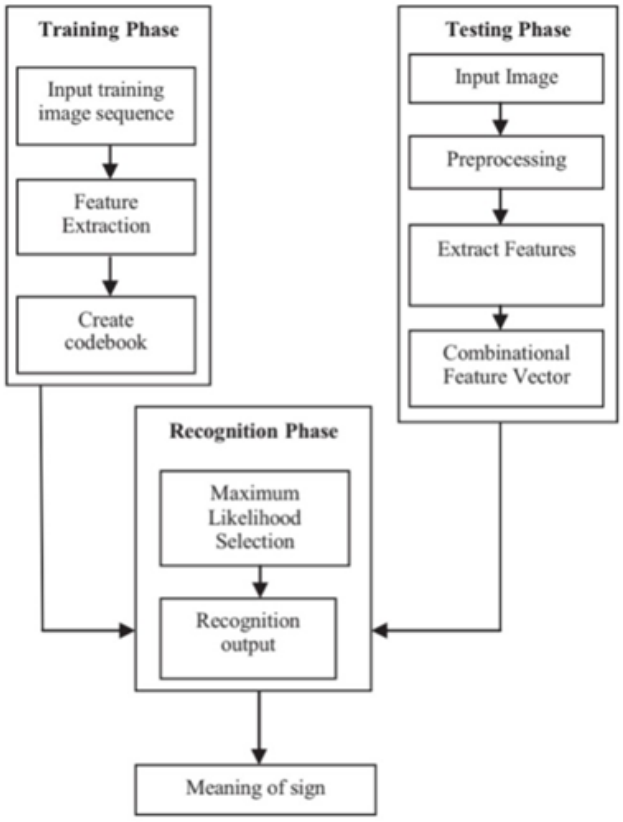


Fig. 2: Block diagram for the hand gesture recognition system

DOI: 10.37200/IJPR/V24I1/PR200278
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

It makes use of two different types of images: Motion Energy Image which records where the motion occurs inside the image and a Motion History Image which stores the pixel values within the image at different time periods. The images are also classified into four types according to their motions: up and down, down and up, left and right and right and left. According to [7], based on the motion descriptors, feature selection is performed using a fourier transform with 25 different harmonics which we use to find the distance between the fourier descriptors given by

$$Dis(S_a, S_b) = \sqrt{\sum_{i=2}^{25} |s_a(i) - s_b(i)|^2} \quad \text{——— 1}$$

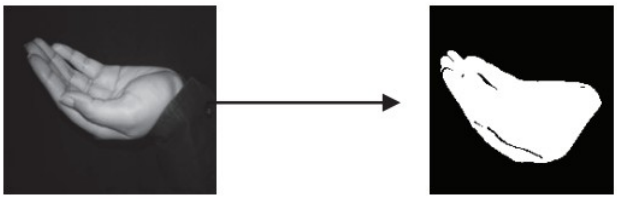
Where Sa and Sb are the fourier descriptors which represent the edges of the hand and the palm.

The system in [2] trains images of 15 different hand gestures of 10 different people of the Taiwanese Sign Language. The model gives a training accuracy of 96% and a testing accuracy of 91%. Further pre-processing of the data and the use of machine learning methods on the data are some of the upgradations that could be possibly done to the system.

B) Sign Language Recognition using multi- class Support Vector Machines

The system in [4] and [5] makes use of a multiple class Support Vector Machine (SVM) to recognise the meaning of the gesture. [4] Contains three stages: train, test and recognise phases (as shown by figure 2). Combinational parameters of Hu invariant movement and structural shape descriptor are created to form a new feature vector to recognise sign.

[5] Suggests two different approaches that have been used by the system: glove based approach where user will be wearing a device that has been connected to the computer and a vision based approach which requires a camera and deals with recognising the gesture. There are two steps in these systems. The first step is to perform sign capturing where we obtain the gesture and the second step is to perform sign analysis to recognise the gesture.



Segmentation OF gray scale image

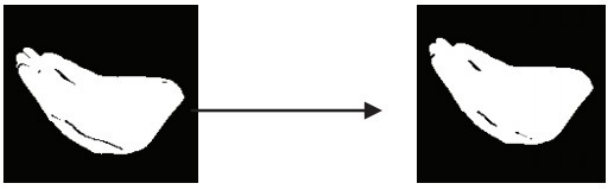


Figure 3: Filtering of segmentation image

DOI: 10.37200/IJPR/V24I1/PR200278
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

There are three phases in the implementation of the system. During the first phase, each class is trained using a multi class SVM. Second phase is the testing phase where the trained model is tested with a sample dataset fro m each class. Third phase is the recognising phase where the various classes of gestures are tested.

The data processing contains two steps: segmentation is done first followed by filtering as shown in figure 3. The grey scale image is converted into binary image where the pixels containing the hand are labelled as 1 and the pixels not containing the hand are labelled as 0.

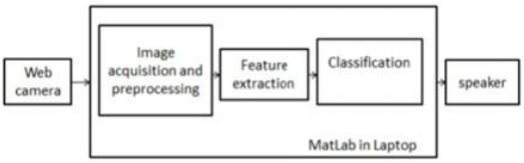


Fig. 4: Frame work of the hand gesture recognition system

A median specific filter is used to eliminate the extreme values from segmented image. The Support Vector machine will then convert the hand gesture into multiple classes. The feature extraction is then done using the Hu invariant movement. A binary classifier can be converted into a multi class classifier i.e. if there are three features X,Y,Z, then the feature vector contains XxY, YxZ, ZxX values with respect to X, Y, Z.

The main advantage of these systems is that features can be extracted from the disjoint shapes and that it can handle the invariant movements. The system provides an overall accuracy of 93% for [5]. Scope for future development include signer independent, larger vocabulary systems in both isolated and continuous recognition systems.

C) Recognising sign language gestures using Artificial Neural Networks

The system given by [1] and [10] performs sign language recognition by training a supervised Artificial Neural Network (ANN) using a matlab simulation where the hand gestures are fed to the network and the meaning of the gesture is obtained as a result of the neural network.

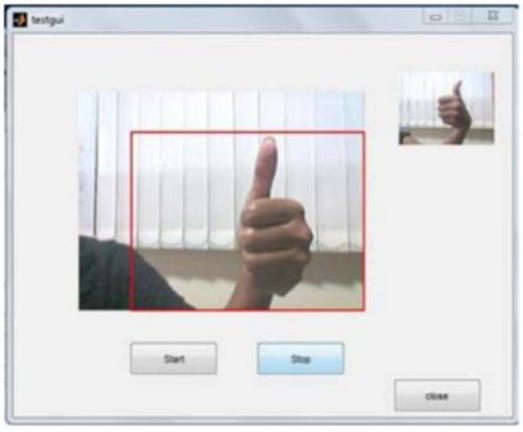


Fig. 5: Real time gesturing and matching it with image in the database

DOI: 10.37200/IJPR/V24I1/PR200278
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

The system first takes in the gestures through the web camera as shown in figure 5. Once the presence of hand is detected on the web camera, it is being sent into the artificial neural network which will recognise the gesture. The artificial neural network is trained using a database containing the hand gestures in different lighting conditions.

Once the gesture is obtained through the camera, the system has to identify the skin region as well. So, segmentation of skin is performed. The captured image is in the RGB domain and it is not efficient for skin segmentation, the image is converted into YbCr domain. But, the problem with the YbCr domain is that it always results in spots indicating noise duet ovaryinglight intensities.

Fig. 6: Preprocessing for recognising hand gestures via CNN

The next step is to perform the feature extraction. The system makes use of a feature called as edge feature. This feature is obtained through a canny edge detector. A canny edge recognition algorithm is a multistage algorithm which detects different types of edges in an image. The canny edge algorithm is used since it can adapt to the different environments.

The Artificial Neural Network is used for recognising the gestures that are obtained via the web camera in real time. Symbol recognition is the process of mapping the segmented gestures or symbols with a library of pre-defined gestures. Here, a feed forward ANN is being used which takes the image as a

6x7 frame. Thus, there are 42 inputs to the Artificial Neural Network layer. The network will compare the input with the required output and returns a value ranging from 0 to 1.

The meaning of the hand gesture is conveyed through an audio output through a raspberry pie or a speaker. [1] trains the neural network using a database containing 25 images for 5 different gestures. The system has been found to be robust and efficient under different lighting conditions as a result of which it gives a 90%. However, the system can further be expanded to include more hand gestures and different types of sign languages.

D) Sign Language Recognition using Convolutional Neural Networks

This system suggested by [3] makes use of the Convolutional Neural Network (CNN) and the Graphical Processing Unit (GPU) acceleration and is implemented on the Microsoft Kinect platform. The Convolutional Neural Networks are able to automate the feature extraction process. It is implemented on the It alien Sign Languaged at a set containing 20 gestures.

The system has to pre-process the data before feeding it to the neural network as shown in figure 6. The pre-processing begins with cutting the high arm and the top part of body using the given information. The pre-processing gives four video samples of 64x64x32 resolution i.e. 32 frames of 64x64 size.

DOI: 10.37200/IJPR/V24I1/PR200278
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

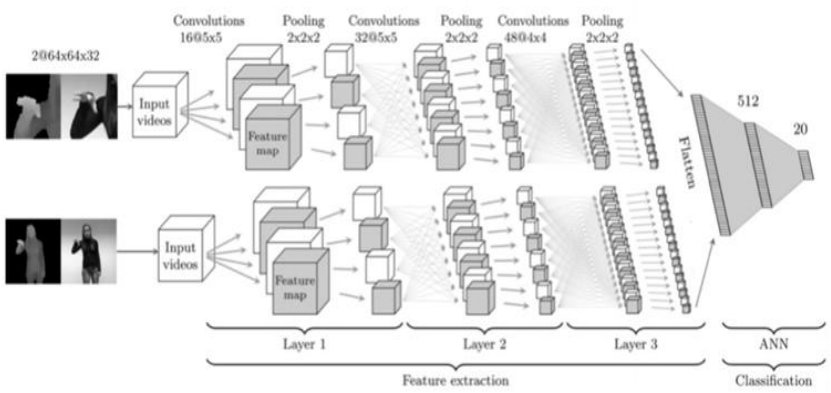


Fig .7: Architecture of CNN for sign language recognition system

The system architecture is given in figure 7. The CNNs are neural networks which have proved to be successful especially in training images. The artificial neurons of CNN are connected to a local region of the visual field which is also called as receptive field. This is done by performing discrete convolutionals on the given image with the values given by the filter as the training weights. Multiple filters are applied to each channel, and along with the activation functions of all the neurons, they form feature maps.

The system makes use of max-pooling i.e. it maintains only the maximum value in the local neighbourhood of the feature map. For accommodating video data, the max-pooling is performed using three dimensions. The architecture of the model contains two CNNs, for extracting the features of the hand and for extracting features of the upper body. Each CNN has three layers. The Local Contrast Normalisation (LCN) is also applied for the first two layers and all artificial neurons are rectified linear units (reLU).

While training, the data augmentation and drop-out are the main approaches used to reduce overfitting. The data augmentation is performed on the CPU while the model trains on the GPU simultaneously. Random weights are assigned for the CNNs using normal distribution. Temporal segmentation is also done on the images. The main aim of the temporal segmentation method is predicting the start and end frames of all the gestures in the given video samples. The system uses a sliding window technique where each window consisting of 32 frames is evaluated using the trained model.

The system is robust and is also found to have a very high accuracy of 92%. Future scope for further improvement of the system includes adding more data and also including diff erent sign languages apart from only the Italian sign language system.

E) Sign Language Recognition using 3D Convolutional Neural Networks

The system suggested by [9] is similar to the one suggested by [3] except that it makes use of a 3D convolutional neural networks while the previous system makes use of the 2D Convolutional Neural Networks. In 2D CNNs, the 2D feature maps maps are used but they calculate only the spacial dimensions. In order to effectively incorporate the information for motion which can also be used in the video analysis, we can make use of a system that performs the 3D convolution on convolutional layers of the Convolutional Neural Network so that the discriminative features are also being captured.

DOI: 10.37200/IJPR/V24I1/PR200278
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

Based on the 3D convolution, different CNN architectures can be designed as shown in figure 8. The system make use of Microsoft Kinect which accepts the input gestures while also providing depth and colour video stream. It is also capable of keeping track of the users body movement simultaneously. The colour information contains three channels for RGB. So it obtains five different input data thereby, adding to the depth and the body skeleton.

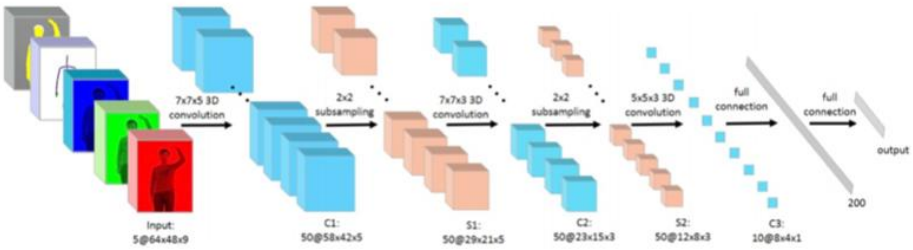


Fig. 8: 3-D CNN for sign language recognition

Table 1: Comparison of all the methods listed in the paper

Algorithm	Advantages	Disadvantages	Accuracy
ANN[1]	1. Matlab is used which makes it easier to execute Artificial Neural Networks due to the extensive packages that it contains. 2. The model was robust under various lighting conditions for the samples it was trained for. 3. Uses a Canny edge detector algorithm which is adaptable to various environments.	1. Uses a very narrow dataset for training so the model hasn't been tested for a wide variety of data which is often found in a real life scenario 2. Supervised learning is used which restricts the accuracy of the ANN	90%
Multi-class SVM[4]	1. Features can be extracted from disjoint shapes 2. It can handle the invariant movements.	1. Support Vector Machines are not constant in prediction.	92.5%
Image Processing[2]	1. It is invariant to scale, rotation and translation. 2. Proposed method gives a 100% accuracy for the given dataset.	1. Enlarging the library of stored models 2. Extracting the moving hands from the complex backgrounds.	91%
CNN[3]	1. High level of accuracy obtained 2. It uses an extensive dataset containing 6000 different images for 20 different signs in the Italian sign language.	1. Only 20 gestures are being considered for this paper. 2. The test result is more than the validation result.	91.7%
3-D CNN[9]	1. It can extract multiple types of input from adjacent frames 2. Better performance than the rest of the discussed models	1. It is complex	87.9%

DOI: 10.37200/IJPR/V24I1/PR200278
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

For each visual source, the system considers nine different frames of 64×48 size centred on the frame currently used as the input to 3D CNN. This gives five different feature maps which are denoted by depth, body skeleton, colour-R, colour-G, colour-B. Each feature map contains nine frames in the form of a stack from the respective channel as a cube. Making use of more number of feature maps as input generally leads to better performance compared to using only grey-scale intensity input.

The system architecture contains eight different layers including an input layer as shown in figure 8. The input layer is followed by four layers that contains Convolutional layer (C1) under which is the sub-sampling (S1) besides the next convolutional layer (C2) which in turn is followed by sub-sampling (S2). Under these two layers is a third and final convolution layer (C3) which does not have any sub sampling following. Under these layers there are two fully-connected layers which contains the output layer. It is also essential to ensure proper size of kernels in different layers in the architecture.

Besides multiple stages of convolutions and sub-samplings, the system is also capable of extracting the spatial-temporal features from the input. It contains 45 consecutive input frames from 5 channels which are converted into a 320-D ((8×4×1)×10) feature vectors that capture the information on any movement in the input frames besides the convolutional layers. The last two fully connected layers will act as a multi layer perceptron classifier on the 320-Dimensional input.

The system presented by [9] trains the CNN with a dataset containing 25 vocabularies which are most commonly used. Each word is played by 9 players and every singer will play thrice for each word. Hence, each word has 27 samples and there are a total of 25x27 samples. The system gives an overall accuracy of 87.9% for the grey channel. There is a scope for further data to be used and more gestures to be included in the system. The accuracy can also be improved further by using more layers of the CNN.

III. Comparison Of Current Techniques

A comparison of the various techniques for sign language recognition system is described in table 1. It compares the advantages, disadvantages and accuracy of the different methods that have been explained in this paper. The CNN and 3-D CNN are found to be the best methods among the models discussed in the paper.

IV. Conclusion

This paper has presented a survey of all the techniques that exist for performing the sign language recognition system. We have seen a wide variety of techniques in detail starting with simple image processing, SVMs upto the ANN and CNNs. There are a lot of scopes for future improvement of the system. Most of the systems are limited to very few gestures and does sign language recognition only for any one sign language. We can come up with future systems where we can have a system which is capable of making predictions for more number of gestures and across different sign languages.

References

[1] Priyanka C Pankajakshan, Thilagavathi B, “Sign Language Recognition System”, IEEE International Conference on Innovations in Information, *Embedded and Communication Systems (ICIIECS)*, 2015.

DOI: 10.37200/IJPR/V24I1/PR200278
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

[2] Maryam Pahlevanzadeh, Mansour Vafadoost, Majid Shahnazi, "S i g n Language Recognition System", *9th International Symposium on Signal Processing and it's application,* 2007

[3] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, Benjamin Schrauwen, "Sign Language Recognition using Convolutional Neural Networks", *Ghent University, ELIS, Belgium,* 2015.

[4] Dixit, K., & Jalal, A. S. (2013, February). Automatic Indian sign language recognition system. In *2013 3rd IEEE International Advance Computing Conference (IACC)* (pp. 883-887).

[5] Anup Kumar, Karun Thankachan and Mevin M. Dominic, "Sign Language Recognition", *3rd International Conference on Recent Advances in Information Technology (RAIT),* 2016.

[6] Kanchan Dabre, Surekha Dholay, " Machine Learning Model for Sign Language Interpretation using Webcam Images", *International Conference on Circuits, Systems, communication and IT Applications,* 2014.

[7] Hassan, S. T., Abolarinwa, J. A., Alenoghena, C. O., Bala, S. A., David, M., & Farzaminia, A. (2017, October). Intelligent sign language recognition using enhanced fourier descriptor: a case of Hausa sign language. In *2017 IEEE 2nd International Conference on Automatic Control and Intelligent Systems (I2CACIS)* (pp. 104-109). IEEE.

[8] Anderson, R., Wiryana, F., Ariesta, M. C., & Kusuma, G. P. (2017). Sign language recognition application systems for deaf-mute people: A review based on input-process-output. *Procedia computer science*, *116*, 441-448.

[9] 9. Jie Huang , Wengang Zhou , Houqiang Li , Weiping Li, "Sign Language Recognition System using 3D Convolutional Neural Networks", *IEEE International Conference on Multimedia and Expo (ICME),* 2015.

[10] 10. Admasu, Y. F., & Raimond, K. (2010). Ethiopian sign language recognition using Artificial Neural Network. In *2010 10th International Conference on Intelligent Systems Design and Applications* (pp. 995-1000). IEEE.

DOI: 10.37200/IJPR/V24I1/PR200278
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020