A

Synopsis Report

on

American Sign Language Detection Using Image Processing

submitted in partial fulfillment of the requirements
for completion of AI LAB
of

TY COMP

in

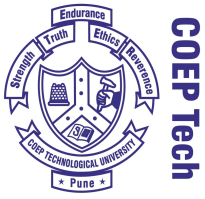
Computer Engineering

by

Vanshita Jain	**Shivani Joshi**	**Juhi Shekokar**
112003052	**112003058**	**112003059**

Under the guidance of

Suraj Sawant
Professor
Department of Computer Engineering



Department of Computer Engineering,
COEP Technological University (COEP Tech)
(A Unitary Public University of Govt. of Maharashtra)
Shivajinagar, Pune-411005,Maharashtra, INDIA

December, 2022

1 Introduction

Signs probably have been the first form of communication used by man since his evolution's. Babies make noises or use gestures to express what they want or how they feel. People who are not natives use signs to communicate with others as language is a barrier. We still use signs when we need to communicate with each other without being noticed or heard. Nobody can say with certainty as to where and when did the first sign evolve. Different groups of people have developed their own particular sign languages over the years. Just like other spoken languages, sign language has evolved in different parts of the world and today we have more than 50 sign languages such as Indian, American, Danish, Chinese, and French.

An array of gestures made using hands, fingers, arms, head and facial expressions besides symbols constitute a sign language. It helps the deaf and the dumb to communicate with the people around them and vice versa. It enables them to understand the world around them through visual descriptions and in turn, contribute to society. The American Sign Language (ASL) is based on ideas rather than words. It cannot be written and comprises of 26 hand symbols each representing one alphabet. These symbols are conveyed using fingers and each word is spelt. Words or names are expressed using a combination of finger spellings as well as gestures. Many deaf schools use a technique called total communication which implies the use of ASL, oral speech and lip reading.

This paper aims to search for the available technique that would provide the best sign language translator. Which means it has to be efficient and accurate while being accessible to everyone.

The structure of this paper is as follows: First we cover the introduction of our project and understand the motivation behind this topic. We then get into the literature review and research gaps. We specify the exact problem statement we are working on and state the objectives of this paper. The methodology, and the hardware and software requirements needed to carry out this methodology are specified. We end with conclusions.

2 Motivation

Sign language is a form of communication used by people with impaired hearing and speech. People use sign language gestures as a means of non-verbal communication to express their thoughts and emotions. But non-signers find it extremely difficult to understand, hence trained sign language interpreters are needed during medical and legal appointments, educational and training sessions. Over the past five years, there has been an increasing demand for interpreting services. Other means, such as video remote human interpreting using high-speed Internet connections, have been introduced. They will thus provide an easy to use sign language interpreting service, which can be used, but has major limitations.

Communication is mainly done verbally, however not all people can do so because of muteness or deafness. Deafness can be caused by genetics, complications at birth, infectious diseases, chronic ear infections, use of drugs, excessive noise exposure, and aging 1, while muteness can be caused by endotracheal intubation, tracheostomy, or damage to the vocal cords from disease or traumas 2. Muteness in a way can also be an effect of deafness. Currently, about 466 million people worldwide have hearing loss, 34 million among which are children, and by 2050, 900 million people are estimated to have hearing loss 1. With such a staggering number of people suffering from hearing loss, roughly the same number of people will have lost the ability to speak. Despite many workarounds and prevention, those who are unfortunate with these disabilities primarily communicate using sign language. The use of sign languages have dated back since the 5th century B.C. as stated by Socrates ”If we hadn’t a voice or a tongue, and wanted to express things to one another, wouldn’t we try to make signs by moving our hands, head, and the rest of our body, just as dumb people do at present?” 3. Since the 5th century B.C., there have been numerous versions of signed languages. However, the version focused on this project is American Sign Language (ASL). Using sign language, the deaf and the mute can somehow communicate. Only some people understand sign language. Should the need for the deaf or mute to speak publicly arise, people usually employ the help of a translator. This paper aims to search for the available technique that would provide the best sign language translator. Which means it has to be efficient and accurate while being accessible for everyone.

3 Literature Review

Author	Appr	DataSet	Acc	Advantages	Limitations	Remark
Qutaishat Munib	HMM	Massey	98.75 [1]	Robust against changes	Low Precision	Static signs
Huy B.D Nguyen	CNN	Massey (2524)	98.30 [2]	Good Accuracy	Only CNN	.
Huy B.D Nguyen	ANN	Massey	95 [2]	Fast and easy to comprehend	m and n produce errors	Simple and better results
Murat Taskiran	CNN	Massey	98.05 [3]	High accuracy over letters with same gestures	Complex architecture	converts RGB space YCbCr while many opt for grayscale.
Jason Isaacs	ANN		99.9 [4]	Genetic algos to optimize the image pre-processing	Restricted vocab -ulary to 24 ASL	Use of GA along with ANN increased accuracy.

4 Research Gaps

Sign language is usually a combination of hand gestures, body language and facial expressions. Most models in the past eliminate the significance of facial expressions and concentrate the detection system on the hand gestures only. The data that is being fed to the classifier mostly consist of one body part (hand).In some scenarios communication via one body part is insufficient to communicate a message. Efficient algorithms that can process images that depict complex sign language need to be developed. This development will affect the way in which data is processed and fed to the classifier.This will demand a better processing technique which will broaden the spectrum of segmentation and feature extraction methods.

Method	Advantage	Disadvantage
Cross-Correlation Coeff	Low computation	Too simple for classification. Does not learn.
Support Vector Machine	Memory efficient. Effective for classification.	Low performance when noise is in
Artificial Neural Network	Able to learn.	Slow convergence speed. Robust fault-tolerant network.
Hidden Markov Model	Efficient learning algorithm.	High computation. Need many training data.
CNN	Highly accurate for image classification Can work well even without segmentation or pre-processing.	High computation. Need strong hardware.

5 Problem Statement and Objectives

Through this project, we are developing a model which can recognize Finger spelling-based hand gestures in order to form a complete word by combining each gesture.

The major initial objectives of this research are:

1. Develop a set of feature models for each ASL image
2. Design algorithms for ASL feature vector recognition.

6 Methodology

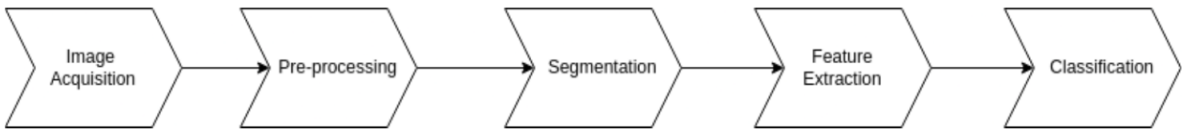


Figure 1: Steps used in processing images

6.1 Image Acquisition

The word sign language is similar to the language phrase, many of both are spread around different world territories . Similar to language, sign language evolves over a long period of period sign language grammar and vocabulary, so it is considered a legitimate language.Because no perception of hearing is needed to understand sign language and no voice is needed to produce sign language, it is the common language among the deaf. Sign languages are usually constructed by using simultaneous compilation associated with hand shapes, orientations, and moves of the hands, palms, or body, along with facial expressions to fluidly express a speaker's thoughts. (Refer Figure 2). The signs must be captured to provide input for the sign language recognition system. To capture images of hand gestures through a camera which handles single-hand signs, double-hand signs, and finger-spelling.[5]

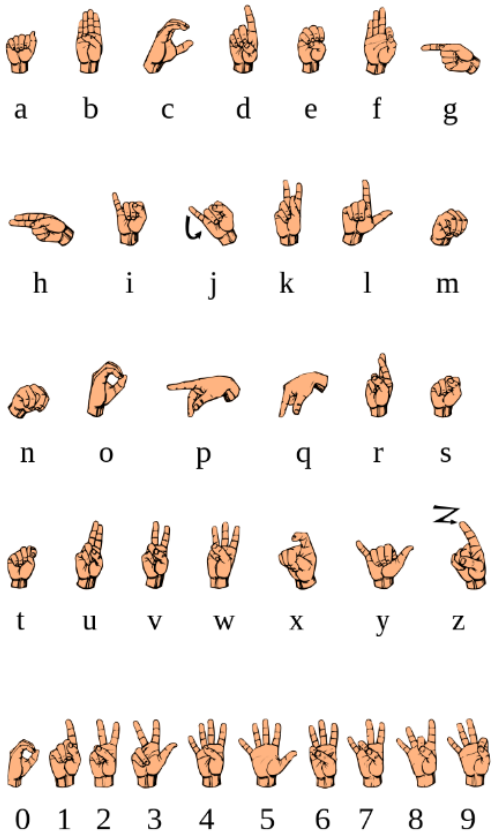


Figure 2: American Sign Language

6.2 Pre-Processing

Pre-processing is the stage where the data is changed after being obtained. It's also to help the classifier to perform better by reducing some bad data that may cause an inaccuracy. Some of the

well-known techniques are Gaussian Filter and Median Filter that will reduce unwanted noise in a 2-dimensional image to improve edge detection on the segmentation stage.

6.3 Segmentation

For SLR, the most important part is the hand gesture which is acknowledged by the hand movements. Therefore, image segmentation is applied to remove unwanted data like the background and other objects as the input, which might interfere with the classifier calculation. Image segmentation works by limiting the region of the data, so the classifier will only look at the Region of Interest (ROI). One of the methods in segmentation is Gray scaling. It turns RGB/Colored Image to a gray-scale image. This will make the classifier avoid taking color as consideration for the calculation. And the important part is that gray scaling also helps identifying the background and foreground. The most common method in segmentation is Thresholding. Usually applied after the image is gray-scale, thresholding turns the image into binary form. Since gray scaling turns an image black and white, thresholding is applied to identify the background and foreground where black represents background and white represents foreground.

Another method is Skin Segmentation. This method is popular because of the simplicity. It takes a predetermined color range or color histogram of human skin so the image will only display the selected color. This way, the classifier won't receive any unwanted information such as the background or objects.The downside is that sometimes objects with similar color to the skin (e.g. other body parts or face) will also get detected. Skin segmentation is sensitive to illumination.Furthermore, a Morphological filter or operation is usually applied to the binary image. Morphological filters help reduce an error either from the foreground or the background, effectively increasing the region of interest to fit the object. Moreover, there's a Canny edge detection method which is used to extract edges from the image. This method is good to map out the hand's edges and determine the Region of Interest. Some also use Background Subtraction 18,23. Background Subtraction requires video input with the static objects like the backgrounds removed. Since the signer's hand will move around, this method is used for an easy segmentation process. However, if there are many moving objects like cars, television, etc., it will also take it as the foreground.

6.4 Feature Extraction

Feature Extraction is the part where relevant information from the data is taken and enhanced. It cuts out the redundant information from the region of interest and starts taking the features to be used for the classifier calculation. For SLR, that features might vary depends on what the researcher thinks can be taken for recognizing gestures. The most common feature extraction is by using the convolution layers of Convolutional Neural Network (CNN).The convolution layers can extract important features from the image. This feature is not just taken, but also enhanced by the MaxPooling layer and increases the performance of the calculation. Which is why CNN is the most commonly used because of its robustness.Another method, Principal Component Analysis (PCA), is used to reduce the features/data dimension, transforming the variable to be uncorrelated.

To extract the external boundary of objects in images, Fourier descriptors are used. The sequence of coordinates forms boundaries to identify objects in an image. The Horn-Schunck optical flow algorithm extracts tracking points for both arms in every frame. SURF is a patented descriptor for finding local features in a video. In the Hough transform , the elements are arranged into pairs (q, h) since we utilize polar directions to identify lines. It is used to find two-hand communication features for the recognition of SLR which is broadly utilized in the segmentation stage. Haar classifiers have been used for object recognition and used for initial real-time face detectors. Local binary patterns (LBPs), find the surface and shape in gray-scale images. LBP is, by all accounts, good with different facial expressions and rotation of an image. Therefore, it is reasonable for extraction in gesture-based recognition. Other feature extraction methods are tracked particle filters, 121 points used as basic descriptors, Zernike moments for keyframe extraction, and the distance algorithm, which are used to extract features for classification in the final step.

6.5 Classification

A convolutional neural network (CNN) is perceived as the most important deep learning neural network model to perform with regards to recognising and classifying images. It uses multi-layer superposition to extract low-level features into relevant features, resulting in a hierarchical structure similar to simulations of human brain activity. The procedure of lengthy manual feature extraction can be prevented because the recent features are passed from the past layer. CNN

combines both feature learning and classification. A convolutional neural network is made up of many layers in general. In the convolution layer, a convolution operation is used to extract features from an input layer or a prior layer. The pooling layer can constantly shrink the data's space size, reducing the number of features and computations. In the CNN, the fully connected layer serves as a "classifier." CNN automatically learns the values from these layers.

In the context of image classification, our CNN may learn to identify edges, identify shapes, and help boundaries identify higher-level features such as face structures, respectively with the first, second, and highest layers of the network by applying convolution filters, nonlinear activation functions, pooling, and back propagation. A CNN architecture can be built by multiple layers of convolution and pooling in an alternating fashion. 2D CNNs are applied to image datasets to classify them and extract spatial features. Anyway, for SLR in videos, both spatial and temporal information are required to be captured. 3D CNN carries out convolution in videos to extract both spatial and temporal information. Our version learns and extracts each spatial and temporal capability through the performance of 3D convolutions.

Classification finds a function to determine the category to which the input data belongs. It can be two-category or multi-category. Factors including the classification construction method, the properties of the data to be classified, and the number of training samples all influence classification accuracy

7 Hardware and Software Requirements

7.1 Hardware Requirements

In order to acquire images, the most common method is using a standard video camera that will give a 2-dimensional image Joshi et al. 6, Jin et al. 7, and some others. This method is more common than the others because nowadays most people have a smartphone with built-in camera. However, the image quality depends on the camera so it's better to be pre-processed first. For example, a low-quality camera has too much noise on the captured image that might interfere with the feature extraction. Also, the resolution of the image depends on the camera so usually an image pre-processing needs to be applied so it can be consistent to the classifier. Some researches use Kinect 8,9,10. Kinect has sensors and a camera to capture colored images including the

depths of the objects. These provide more detailed data that can improve the classification. The depths feature can help the segmentation by identifying the background and foreground objects. There's also a unique method by Abdelnasser et al. that utilizes wi-fi signal strength to detect in-air hand gestures around the device, called WiGest 11. The changes in the returned signal becomes the input. This method is useful since the user does not have to carry an object and it's multi-directional. However, it is not easy to implement. Data can also be obtained from datasets already created on the internet, which are ready to use and usually come in a large amount with good quality. For example, Jalal et al. use the Kaggle dataset which contains about 27000 sign language images 12. With existing good quality datasets, the research can focus on other parts of the SLR. If the data collected is not enough, Data Augmentation can be done to acquire more data 8,13. Data Augmentation augments an existing data to create a new different data. This is used to save time and to improve the classifier accuracy by preventing over-fitting to some sign language while providing more invariant data to the classifier. In SLR that uses images as input, hand images are captured from different angles and distance, which cause invariance in scale and rotation. A common Data Augmentation example is the rotation, where the image angle is slightly rotated. In this project we have generated the dataset using the in-built camera in Asus Vivobook 14 M413 - Microsoft 2021.105.10.0. We will also be predicting and testing our model on this laptop to maintain highest level of accuracy.

7.2 Software Requirements

Most researchers create their own datasets for the training of their data. Because of the non-availability of sign language datasets in particular regions, researchers record the data from the signer to create a dataset. Researchers prepare their own sign language datasets because they usually do not have enough datasets to use for research. Numerous researchers have used predefined datasets from the American Sign Language like Image Dataset (ASLID), ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) – 2010, ChaLearn Looking at People 2014 (CLAP14), RWTH-Phoenix-Weather Multisigner 2014T], SIGNUM and The ArSL database , the Massey University dataset, and the AND VIVA challenge dataset.

We decided to generate a dataset of 3500 images, 350 images per letter/number to feed into the training model of CNN to bring out satisfactory results. A pre-trained model is a model that has been trained on a large benchmark dataset to solve a problem similar to the one that we want

to solve. Due to the computational cost of training such models, it is common practise to import and use models from published literature. The procedure of lengthy manual feature extraction can be prevented because the recent features are passed from the past layer. CNN combines both feature learning and classification. A convolutional neural network is made up of many layers in general. In the convolution layer, a convolution operation is used to extract features from an input layer or a prior layer. The pooling layer can constantly shrink the data's space size, reducing the number of features and com putations. In the CNN, the fully connected layer serves as a "classifier."

8 Conclusions

In this study, we have reviewed multiple research articles and papers related to Sign Language Recognition and Detection (SLR/SLD). The process can be broadly classified into 5 steps: Data acquisition, pre-processing, segmentation, feature extraction, and classification. A Standard Video Camera is most commonly used for data acquisition to get image captures of the user's hand from different angles, lightings, backgrounds, and sizes. Convolutional Neural Network (CNN) is the most widely used classifier, which is popular because of its accuracy, which can reach 90% or more for SLR tasks. In our project, we have achieved an accuracy of 94%. We observe that those researches which have the most accuracy have static images of their input instead of dynamic data acquisition. A few alphabets in ASL require active hand movement, and thus need a real-time, live translator. Overall, this study research has been insightful since it shows different approaches for sign language translation.

References

[1] Q. Munib, M. Habeeb, B. Takruri, and H. A. Al-Malik, "American sign language (asl) recognition based on hough transform and neural networks," *Expert systems with Applications*, vol. 32, no. 1, pp. 24–37, 2007.

[2] H. B. Nguyen and H. N. Do, "Deep learning for american sign language fingerspelling recognition system," in *2019 26th International Conference on Telecommunications (ICT)*. IEEE, 2019, pp. 314–318.

[3] M. Taskiran, M. Killioglu, and N. Kahraman, "A real-time system for recognition of american sign language by using deep learning," in *2018 41st international conference on telecommunications and signal processing (TSP)*. IEEE, 2018, pp. 1–5.

[4] J. Isaacs and S. Foo, "Hand pose estimation for american sign language recognition," in *Thirty-Sixth Southeastern Symposium on System Theory, 2004. Proceedings of the*, 2004, pp. 132–136.

[5] S. Sharma and K. Kumar, "Asl-3dcnn: American sign language recognition technique using 3-d convolutional neural networks," *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 26 319–26 331, 2021.