# AMERICAN SIGN LANGUAGE RECOGNITION SYSTEM

SERC
INDIAN INSTITUTE OF SCIENCE

# INTRODUCTION

- Inability to speak is considered to be true disability.
- People with this disability use different modes to communicate with others, there are a number of methods available for their communication one such common method of communication is sign language.

- The main focus of this work is to create a vision based system to identify Finger spelled letters of ASL.
- The reason for choosing a system based on vision relates to the fact that it provides a simpler and more intuitive way of communication between a human and a computer.

The two approaches used for the classification of sign language :

1. In the first approach, features were extracted from the images using SIFT(scale invariant vector transform) which were then plotted into histograms and used for training hierarchical SVM.The accuracy of the model obtained using this approach was 44.259%.

2. In the second approach, Convolutional neural network was used. The accuracy of the model obtained using Convolutional Neural Network was 95.50.

# DATASET

Dataset was taken from two different sources which provided different types of images.

Dataset 1: In this dataset the images are non segmented images taken with webcam

Size: 78000 images

Number of classes: 26 (A-Z alphabets)

Resolution of each image: 200 * 200

Images per class: 3000

Dataset 2: In this dataset the images are segmented images

Size: 2671 images

Number of classes: 36 (A-Z alphabets & whole numbers)

Resolution of each image: 600 * 670

Images per class: 60

# METHODOLOGY & IMPLEMENTATION

- **SIFT** - Scale invariant feature transform (SIFT) is an algorithm in computer vision to detect and describe local features in images.
- **K Means Clustering** - The k-means algorithm takes as input the number of clusters to generate, k, and a set of observation vectors to cluster. It returns a set of centroids, one for each of the k clusters.
- **SVM** - In this algorithm, each data item is plot as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.
- **LeNet** - The LeNet-5 architecture consists of two sets of convolutional and average pooling layers, followed by a flattening convolutional layer, then two fully-connected layers and finally a softmax classifier.
- **AlexNet** - The AlexNet architecture contains 5 convolutional layers and 3 fully connected layers. Relu is applied after very convolutional and fully connected layer. Dropout is applied before the first and the second fully connected year.

# SEGMENTATION

In computer vision segmentation is the digital image into multiple segments.

The goal of segmentation is to simply add or change the representation of the image that is more meaningful and easy to analyze.

Segmentation was done using three stages:

- Applying constraints on HSV
- Applying constraints on YCbCr colormaps.
- Output fed to Watershed algorithm.

| Model | Dataset | Train-Validation-Test Split | Train Accuracy | Validation Accuracy | Test Accuracy | K-Fold Cross Validation |
|---|---|---|---|---|---|---|
| LeNet | 2517 images(36 images) | 80-15-5 | 100% | 95.48% | 96.50% | 90.3% |
| Modified AlexNet | 2517 images(36 classes) | 64-16-20 | 100% | 99.75% | 100% | 97.91% |
| | 78000 images (26 classes) | 80-15-5 | 96% | 94% | 51% | 73.75% |
| | 2517+154(real and segmented) | 64-16-20 | 100% | 95% | 95% | 91.37% |
| | 78000 images (26 classes) | 72-18-10 | 91% | 89% | 93.31% | 79.4% (+/- 2.60%)(K=5) |

# CONCLUSION

- Implemented and trained an American Sign Language recognition system based on a CNN classifier
- Satisfactory results on character recognition using 78000 images dataset
- With the help of a dataset with images taken in different environmental conditions, the models would be able to generalize with considerably higher efficacy and would produce a robust model for all letters

# THANK YOU