International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

"Sign Language Recognition with Machine Learning"

Pranav A. Sonawane (IT Third Year), Yash B. Jaiswal (IT Third Year),

Prof.Hiralal Salunkhe(IT Dept.)

G.H.Raisoni College of Engineering and Management Jalgaon

Department of Computer Science & Engineering

Abstract

There is an undeniable communication problem between the Deaf community and the hearing majority. Innovations in automatic sign language recognition try to tear down this communication barrier. Our contribution considers a recognition system using the convolutional neural networks (CNNs) and GPU acceleration. Instead of constructing complex handcrafted features, CNNs are able to automate the process of feature construction. The predictive model is able to generalize on users and surroundings not occurring during training with a cross-validation accuracy of 91.7%. Conversing to a person with hearing disability is always a major challenge. Sign language has indelibly become the ultimate panacea and is a very powerful tool for individuals with hearing and speech disability to communicate their feelings and opinions to the world. It makes the integration process between them and others smooth and less complex.

Introduction-

Very few people understand sign language. Moreover, contrary to popular belief, it is not an international language. Obviously, this further complicates communication between the Deaf community and the hearing majority. The alternative of written communication is cumbersome, because the Deaf community is generally less skilled in writing a spoken language. Furthermore, this type of communication is impersonal and slow in face-to-face conversations. For example, when an accident occurs, it is often necessary to communicate quickly with the emergency physician where written communication is not always possible. The purpose of this work is to contribute to the field of automatic sign language recognition. We focus on the recognition of the signs or gestures. There are two main steps in building an automated recognition system for human actions in spatio-temporal data. The first step is to extract features from the frame sequences. This will result in a representation consisting of one or more feature vectors, also called

descriptors. This representation will aid the computer to distinguish between the possible classes of actions. The second step is the classification of the action. A classifier will use these representations to discriminate between the different actions (or signs). In our work, the feature extraction is automated by using convolutional neural networks (CNNs). An artificial neural network (ANN) is used for classification.

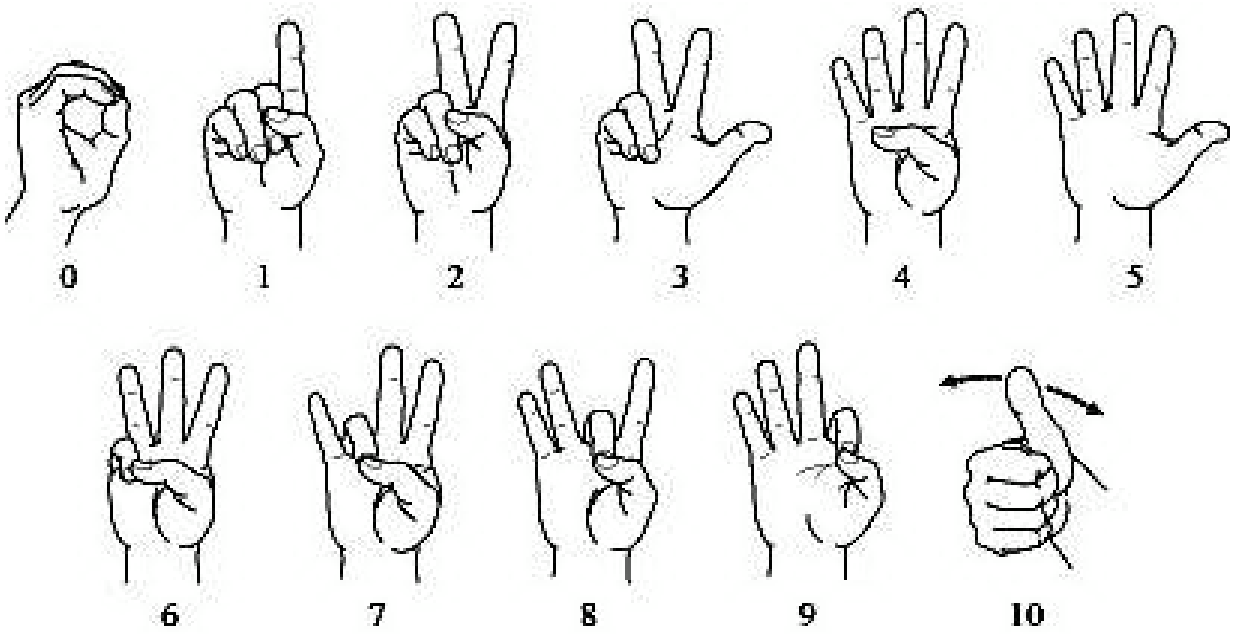

Fig.(1) Sign language

Review of Literature-

For the past decades, research on SLR has been explored. Many studies used sensor-based devices such as SignSpeak. This device used different sensors such as flex and contact sensors for finger and palm movements and accelerometers and gyros for the hand movement; then, by Principal Component Analysis, the gloves were trained to recognize different gestures, and each gesture was then classified into alphabets in real time. The device also used an Android phone to display the text and word received. ASL static word signs. gloves via Bluetooth. SignSpeak was found to have 92% accuracy. There are other means of capturing signs by using motion sensors, such as electromyography (EMG) sensors, RGB camera, Kinect sensors, and leap motion controller or their combinations. Although these sensors provide accurate parameters in measurement of data, they also have limitations; first is their cost, as they require large-size datasets with diverse sign motion they going toned a high-end computers with powerful specifications; next is aesthetics,

as the sensors are attached to the fingers and palms of a user, the user can encounter difficulties in setting up the device; ambient lighting conditions or backgrounds in real-world settings may also affect the recognition. Therefore, many researchers jumped from sensor-based to visual-based SLR. Several methods have been developed in visual-based SLR. Because sign language includes static and dynamic movements, image, and video processing was explored by many.

Methodology-

- **Preprocessing-**

Our first step in the preprocessing stage is cropping the highest hand and the upper body using the given joint information. We discovered that the highest hand is the most interesting. If both hands are used, they perform the same (mirrored) movement. If one hand is used, it is always the highest one. If the left hand is used, the videos are mirrored. This way, the model only needs to learn one side. Furthermore, the noise in the depth maps is reduced with thresholding, background removal using the user index, and median filtering.

(a) Original (b) Noise reduction (c) 4 Input channels

Fig.(2) Preprocessing.

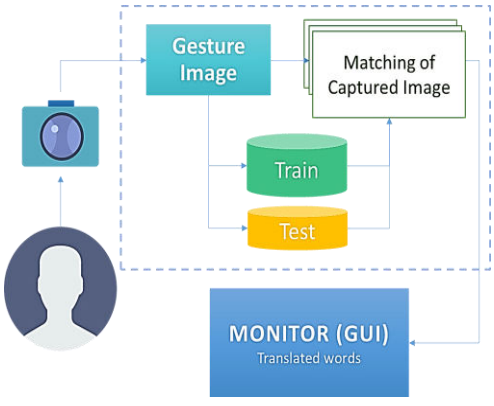

Fig.(3) Methodology.

- **Convolutional Neural Network (CNN)**

CNNs are feature extraction models in deep learning that recently have proven to be to be very successful at image recognition. As of now, the models are in use by various industry leaders like Google, Facebook and Amazon. And recently, researchers at Google applied CNNs on video data. CNNs are inspired by the visual cortex of the human brain. The artificial neurons in a CNN will connect to a local region of the visual field, called a receptive field. This is accomplished by performing discrete convolutions on the image with filter values as trainable weights. Multiple filters are applied for each channel, and together with the activation functions of the neurons, they form feature maps. This is followed by a pooling scheme, where only the interesting information of the feature maps are pooled together. These techniques are performed in multiple layers

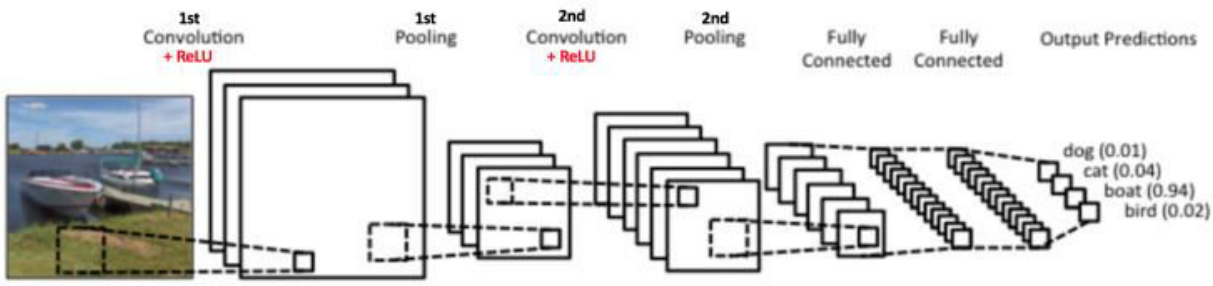

Fig.(4)-CNN.

- **Proposed Architecture-**

For the pooling method, we use max-pooling: only the maximum value in a local neighborhood of the feature map remains. To accommodate video data, the max-pooling is performed in three dimensions. However, using 2D convolutions resulted in better validation accuracy than 3D convolutions. The architecture of the model consists of two CNNs, one for extracting hand features and one for extracting upper body features. Each CNN is three layers deep. A classical ANN with one hidden layer provides classification after concatenating the outcomes of both CNNs. Also, local contrast normalization (LCN) as in is applied in the first two layers and all artificial neurons are rectified linear units.

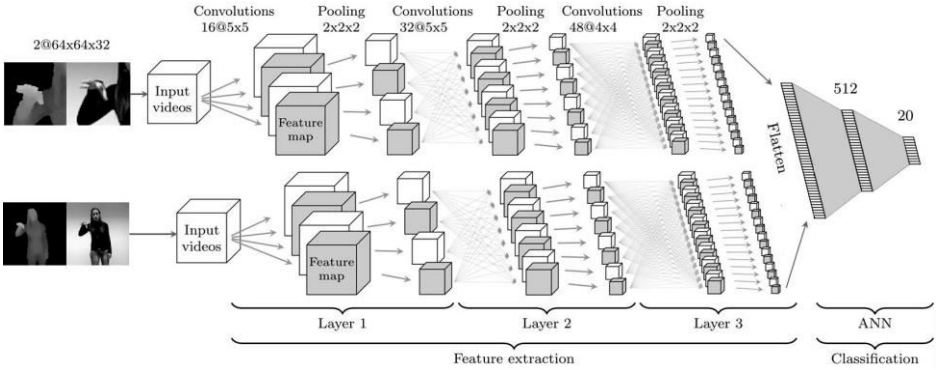

Fig.(5)-Architecture.

- **Generalization and Training-**

During training, dropout and data augmentation are used as main approaches to reduce overfitting. The data augmentation is performed in real time on the CPU during the training phase whiles the model trains on the GPU as in [12]. This consists of zooming up to 10%, rotations up to (-)3◦, spatial translations up to (-)5 pixels in the x and y direction, and temporal translations up to (-)4 frames. We use Nesterov's accelerated gradient descent (NAG) [16] with a fixed momentum-coefficient of 0.9 and mini-batches of size 20. The learning rate is initialized at 0.003 with a 5% decrease after each epoch. The weights of the CNNs are randomly initialized with a normal distribution with $\mu = 0$ and $\sigma = 0.04$, and $\sigma = 0.02$ for the weights of the ANN. The biases of the CNNs are initialized at 0.2 and the biases of the ANN at 0.1. Experiments are conducted on one machine with a hexa-core processor (Intel Core i7-3930K), 32GB SDRAM and a NVIDIA GeForce GTX 680 GPU with 4096MB of memory.

- **Temporal Segmentation-**

The CLAP14 challenge consists of spotting gestures in video samples. Each video sample is an unedited recording of a user signing 10 to 20 gestures, including noise movements that are not part of the 20 Italian gestures. The goal of the temporal segmentation method is to predict the begin and end frames of every gesture in the video samples. We use the sliding windows technique, where each possible interval of 32 frames is evaluated with the trained model (as previously described). Consecutive intervals with identical classes and sufficiently high classification probability (thresholding) are considered as a gesture

segment. The validation set of CLAP14 is used to optimize the thresholding parameters. Furthermore, an extra class is added to the classifier to help identify video intervals without gesture.

SYSTEM ARCHITECTURE –

A CNN model is used to extract features from the frames and to predict hand gestures. It is a multilayered feedforward neural network mostly used in image recognition. The architecture of CNN consists of some convolution layers, each comprising of a pooling layer, activation function, and batch normalization which is optional. It also has a set of fully connected layers. As one of the images moves across the network, it gets reduced in size. This happens as a result of max pooling. The last layer gives us the prediction of the class probabilities. A. Classification In our proposed system, we apply a 2D CNN model with a tensor flow library. The convolution layers scan the images with a filter of size 3 by 3. The dot product between the frame pixel and the weights of the filter are calculated. This particular step extracts important features from the input image to pass on further. The pooling layers are then applied after each convolution layer. One pooling layer decrements the activation map of the previous layer. It merges all the features that were learned in the previous layers' activation maps. This helps to reduce overfitting of the training data and generalizes the features represented by the network. In our case, the input layer of the convolutional neural network has 32 feature maps of size 3 by 3, and the activation function is a Rectified Linear Unit. The max pool layer has a size of 2×2. The dropout is set to 50 percent and the layer is flattened. The last layer of the network is a fully connected output layer with ten units, and the activation function is Softmax. Then we compile the model by using category cross-entropy as the loss function and Adam as the optimizer.

Data Flow Diagram-

Data Flow diagram (DFD) is a traditional demonstration of the view of information flowing within a system. A clean and clear DFD can clearly show the right amount of system requirement. It can be manual, automatic, or a combination of both. Indicates how data enters and leaves the system, what changes the data, and where the data is stored. The purpose of the DFD is to indicate the size and parameters of the entire system. It can be used as a communication tool between a program analyst and any person who plays a role in an order that serves as the starting point for program rebuilding. DFD is also called data flow graph or bubble chart

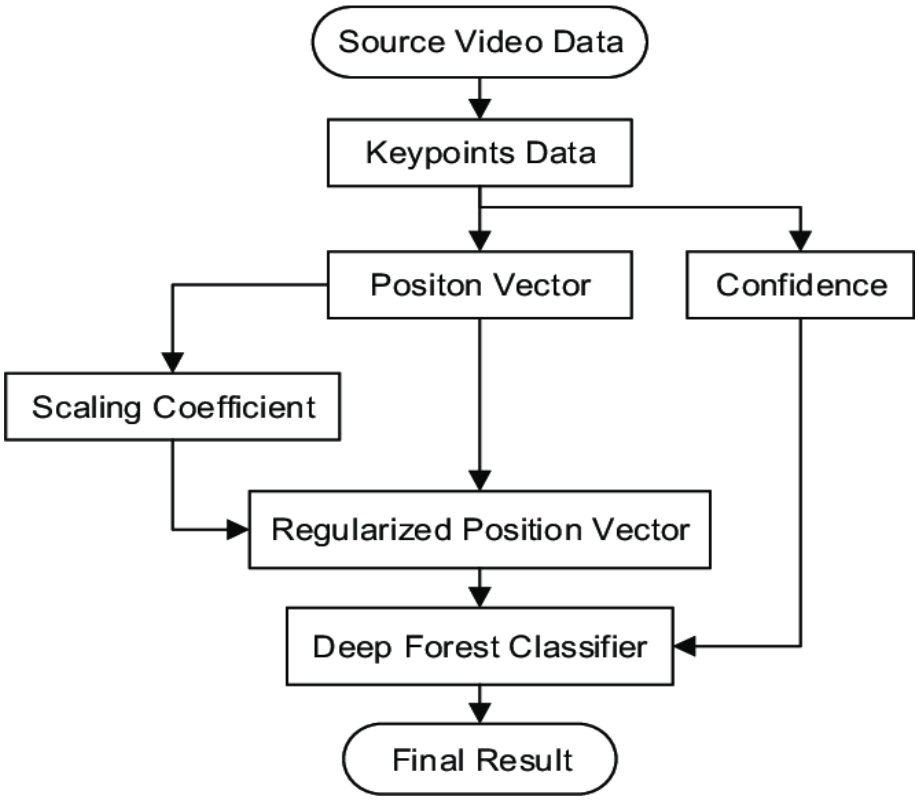

Fig 6-Flow chart

CONCLUSION-

Many breakthroughs have been made in the field of artificial intelligence, machine learning and computer vision. They have immensely contributed in how we perceive things around us and improve the way in which we apply their techniques in our everyday lives. Many researches have been conducted on sign gesture recognition using different techniques like ANN, LSTM and 3D CNN. However, most of them require extra computing power . On the other hand, our research paper requires low computting power and gives a remarkable accuracy of above 90%. In our research, we proposed to normalize and rescale our images to 64 pixels in order to extract features (binary pixels) and make the system more robust. We use CNN to classify the 10 alphabetical American sign gestures and successfully achieve an accuracy of 98% which is better than other related work stated in this paper.

PROBLEMS-

Sign languages are very broad and differ from country to country in terms of gestures, body language and face expressions. The grammars and structure of a sentence also varies a lot. In our study, learning and capturing the gestures was quite a challenge for us since the movement of hands had to be precise and on point . Some gestures are difficult to reproduce. And it was hard to keep our hands in exact same position when creating our dataset.

FUTURE WORK-

We look forward to use more alphabets in our datasets and improve the model so that it recognizes more alphabetical features while at the same time get a high accuracy. We would also like to enhance the system by adding speech recognition so that blind people can benefit as well

REFERENCES-

1. Sign Language Recognition Using Convolutional Neural Networks Lionel Pigou(B), Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen
2. American Sign Language Recognition Using Leap Motion Controller with Machine Learning Approach Teak-Wei Chong and Boon-Giin Lee *
3. Sign Language Recognition System using Convolutional Neural Network and Computer Vision Mohammad Elham Walizad Mehreen Hurroo Computer Science & Engineering Department Delhi Technological University (DTU) Delhi,
4. Static Sign Language Recognition Using Deep Learning Lean Karlo S. Tolentino, Ronnie O. Serfa Juan, August C. Thio-ac, Maria Abigail B. Pamahoy, Joni Rose R. Forteza, and Xavier Jet O. Garcia