

# **Sign Language Detection using Deep Learning**

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE  
AWARD OF DEGREE

OF

**MASTER OF TECHNOLOGY**

IN

**COMPUTER SCIENCE AND ENGINEERING**

Submitted By:

**MANISHA SINGH**

**2K18/CSE/25**

Under the supervision of

**Dr. MANOJ KUMAR**

(Associate Professor  
& Head (Computer Centre))



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

JUNE 2020

# DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

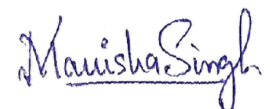
Bawana Road, Delhi-110042

## DECLARATION

I, Manisha Singh, Roll No. 2K18/CSE/25 student of M. Tech (Computer Science & Engineering), hereby declare that the Project Dissertation titled “**Sign Language Detection using Deep Learning**” which is submitted by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi. Report of the Major II which is being submitted to Delhi Technological University, Delhi, in partial fulfillment for the requirement of the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associate ship, Fellowship or other similar title or recognition.

Place: DTU, Delhi

Date: 24-10-2020



Manisha Singh

(2K18/CSE/25)

# **DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

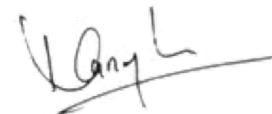
Bawana Road, Delhi-110042

## **CERTIFICATE**

I hereby certify that the Project Dissertation titled “**Sign Language Detection using Deep Learning**” which is submitted by Manisha Singh, Roll No. 2K18/CSE/25, Department of computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment for the requirement of the award of degree of Master of Technology (Computer Science and Engineering) is a record of a project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

**Place : Delhi**

**Date: 26/10/2020**



**(Dr. Manoj Kumar)**

**SUPERVISOR**

**Associate Professor  
& Head (Computer Centre)  
Department of Computer Engineering  
Delhi Technological University**

## **ACKNOWLEDGEMENT**

I am most thankful to my family for constantly encouraging me and giving me unconditional support while pursuing this research.

I am extremely grateful to **Dr. Manoj Kumar** Associate Professor, Department of Computer Science Engineering, Delhi Technological University, Delhi for providing invaluable guidance and being a constant source of inspiration throughout my research. I will always be indebted to his for the extensive support and encouragement he provided.

I also convey my heartfelt gratitude to all the research scholars of the web Research Group at Delhi Technological University, for their valuable suggestions and helpful discussions throughout the course of this research work.



Manisha Singh

Roll No. 2K18/CSE/25

## ABSTRACT

Sign language is a language which allows dumb people to communicate. People with this disability use different modes to talk with others, there are number of methods available for their communication one such common method of communication is sign language. But the challenge is to understand the sign language who is not aware of it. Then it is hard for mute people to communicate with them.

Developing sign language application for deaf people is very important, as they'll be able to communicate easily with even those who don't understand sign language. Sign Language Recognizer took the basic step in bridging the communication gap between normal people, deaf and dumb people using sign language.

The focus of this work is to create the best vision-based system to identify sign language gestures from the video sequences. 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. In this report, 46 different gestures have been considered.

Video sequences contain both the temporal as well as the spatial features. Using of two different models to train both the temporal as well as the spatial features. To train the model on the spatial features of the video sequences Inception model [14] is used which is a deep CNN (convolutional neural net). CNN was trained on the frames obtained from the video sequences of train data. RNN (recurrent neural network) is used to train the model on the temporal features. Trained CNN model was used to make predictions for individual frames to obtain a sequence of predictions or pool layer outputs for each video. Now this sequence of prediction or pool layer outputs was given to RNN to train on the temporal features. The data set [7] used consists of Argentinian Sign Language (LSA) Gestures, with around 2300 videos belonging to 46 gestures categories. Using the predictions by CNN as input for RNN 93.3% accuracy was obtained and by using pool layer output as input for RNN an accuracy of 95.217% was obtained.

# LIST OF CONTENTS

<b>DECLARATION</b>	<b>ii</b>
<b>CERTIFICATE</b>	<b>iii</b>
<b>ACKNOWLEDGEMENT</b>	<b>iv</b>
<b>ABSTRACT</b>	<b>v</b>
<b>LIST OF CONTENTS</b>	<b>vi</b>
<b>LIST OF FIGURES</b>	<b>viii</b>
<b>LIST OF ABBREVIATIONS AND NOMENCLATURE</b>	<b>ix</b>
<b>CHAPTER 1</b>	<b>1</b>
<b>1.1 Background</b>	<b>2</b>
1.1.1 Introduction to Machine Learning and its Technique	2
1.1.2 ANN(Artificial Neural Network) using an MLP (Multilayer Perceptron)	2
1.1.3 Feature Selections	5
1.1.4 Image segmentation applying Machine Learning	5
<b>1.2 Image Segmentation</b>	<b>5</b>
<b>1.3 Image segmentation using CNN</b>	<b>6</b>
<b>1.4 Methods of Image Segmentation</b>	<b>6</b>
1.4.1 Region Based Image Segmentation (RBIS)	6
1.4.2 Edge-detection Based Image Segmentation	7
1.4.3 Cluster Based Image Segmentation	9
1.4.4 CNN Based Image Segmentation	10
<b>1.5 Applications of Image Segmentations</b>	<b>11</b>
<b>1.6 Deep Learning</b>	<b>11</b>
<b>1.7 CNN (Convolutional Neural Network)</b>	<b>12</b>
1.7.1 Convolution Layer	12
1.7.2 Pooling Layer	13
1.7.3 Activation Layer	14
1.7.4 Dropout Layer	14
1.7.5 Fully Connected Layer	14
1.7.6 Data Augmentation and Pre-processing	15
1.7.7 CNN Architectural Design	15
1.7.8 OUTPUT OF CNN	16
<b>1.8 Recurrent neural network (RNN)</b>	<b>17</b>
1.8.1 Architectures of RNN	18

<b>CHAPTER 2</b>	<b>20</b>
2.1 Approach Used in above described Papers	24
2.1.1 ProbSom used in Argentinian Sign Language (LSA) [4]	24
2.1.2 Sign Language Recognition in Continuous Video Sequence [3]	24
2.1.3 Sentence Formation using Gesture Recognition [1]	25
2.1.4 Recognition of Gesture in Real Time [3]	26
<b>CHAPTER 3</b>	<b>27</b>
3.1 Convolution Neural Network (CNN)	27
3.1.1 Procedure followed in CNN	27
3.1.2 Establishment	29
3.2 Recurrent Neural Network (RNN)	30
3.2.1 Loops present in RNN	30
3.2.2 Memory carry forward by previous inputs	31
3.2.3 Vanishing Gradient Problem	31
3.2.4 Long Short-Term Memory Units (LSTMs)	33
3.2.5 RNN Model	33
<b>CHAPTER 4</b>	<b>34</b>
4.1 Data Set	34
4.2 Approach First	35
4.2.1 Methodology	35
4.2.2 Limitations	38
4.3 Approach Second	38
<b>CHAPTER 5</b>	<b>39</b>
5.1 Comparison	39
5.1.1 Advantage	39
5.1.2 Disadvantage	39
<b>CHAPTER 6</b>	<b>40</b>
REFERENCES	41

## LIST OF FIGURES

<b>S. NO.</b>	<b>TOPICS</b>	<b>PAGE NO.</b>
1.1	Architecture of ANN	4
1.2	Sobel-Filter (Horizontal)	8
1.3	Sobel Filter (Vertical)	8
1.4	Laplace Filter	8
1.5	Blurring Filter	8
1.6	Sharpening Filter	9
1.7	Architecture of CNN	16
1.8	Workflow diagram of CNN	17
1.9	Unfolded Basic RNN	18
1.10	The Eleman Network	18
2.1	Block diagram of Hand Gesture Recognition System for LSA	24
2.2	Block diagram of Vision-based recognition system	24
2.3	Block diagram of System Overview	25
2.4	Gesture of Sentence	26
2.5	Methodology for Real Time ISL Classification	26
3.1	Convolution Neural Network	27
3.2	Inception V3 Model Architecture	29
3.3	Recurrent Neural Network Chunk	30
3.4	An Unrolled Neural Network	30
3.5	Memory of Previous inputs carried forward	31
3.6	Vanishing Sigmoid	32
3.7	RNN model	33
4.1	Block Diagram	35
4.2	Extract and Filter Frame	36
4.4	Structure of Models	37-38



## **LIST OF ABBREVIATIONS AND NOMENCLATURE**

1. CNN- Convolutional Neural Network
2. ANN- Artificial Neural Network
3. R –CNN- Region based Convolutional Neural Network
4. RNN: Recurrent Neural Network
5. LSA: Argentinian Sign Language
6. LSTM: Long Short-Term Memory
7. BPTT: Back Propagation Through Time
8. ISL: Indian Sign Language
9. OH: Orientation Histogram
10. PCA: Principal Component Analysis
11. H3DF: Histogram of 3D Facets
12. GPU- Graphics Processing Unit
13. ReLu- Rectified Linear Unit
14. MRI- Magnetic Resonance images
15. MOCNN- Multiple Output Convolutional Neural Network
16. IDE- Integrated Development Environment

# CHAPTER 1

## INTRODUCTION

---

The present dissertation aims to explore hand gesture recognition using deep learning approaches applied to sign language media. Sign language has its own linguistic structure, grammar and characteristics, and independent of rules that govern spoken languages. Sign languages in different countries are vastly different from one another. Gesture can be performed by movement in any body part like face, Hand or head. Here for gesture recognition, image processing and computer vision are used. Understanding of human actions is done by gesture recognition and also acts as an interpreter between computer and human. Without the use of mechanical devices, human has the potential to work alongside naturally with computers. Gestures are made for an ease for deaf and dumb community to perform sign language. Mute people use sign language for their communication when broadcasting audio is impossible and writing or typing all the time at random locations are very difficult, but there is the vision possibility. Sign language is the only way for exchanging information between people. Normally sign language is used by everyone when anyone does not want to speak; some make facial expression some does other. To understand these, we need different dictionary to decrypt it, but this is the only way of conveying their message to people for mute people. Sign language also serves the same meaning as spoken language does. This is used by mute people all over the world but in their regional form like ISL, ASL. Hand gesture recognition is used to detect sign language which can be done by either one hand or two hands. This classifies sign language into two parts, isolated sign language and continuous sign language. Isolated sign language consists of Single gesture having single word while continuous ISL or Continuous Sign language is a sequence of gesture that generate a meaningful sentence. In this report, Isolated ASL gesture recognition technique with providing sequence of images in a form of video and provide the best accuracy using Deep learning approaches. Identifying regions in the sequence of images which is in motion and labeling them to match with the set of data and predicting it accordingly.

## **1.1 Background**

In present world, images are main source of information distribution. We can present in the utmost all domains of work. But images can be perused precise only by humans as they require to be elucidated in a natural perspective way. At present, as we see the technology has pushed forward to a proportion such that it can match the potential of a human brain. Now conversion of images into information is possible by computer. Now images are in readable form and objects present in it can be identified automatically. Based on developed outcome analysis and automatic decision making is feasible now these days. Several Machine Learning technologies have been proposed for such jobs, but CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network) proves to be the best technique to do image segmentation and movement detection as well as prediction of the sign language.

### **1.1.1 Introduction to Machine Learning and its Technique**

Machine learning is the quintessential accomplishment of this digital era. As we dissect the technique how a machine learns to categorize the information provided or the raw materials required for learning the features of the desired task, features or attributes build the base of what we provide to the learning algorithm. Machine learning plays a crucial part in the place of image processing and objects identification. There might be numerous methods present to do such task. Discussion of various methods, which can be used to achieve better accuracy to perform such task in this report, is present.

In this world of digitization, images play a very crucial part in various stages of life including scientific computing and visual persuasion tasks. Images technically can be classified into different conversion methods like binary images, RGB images, grey scale images, hue saturation value or hue saturated lightness images etc. Each dataset can be represented via an enormous number of features. But all the features aren't necessarily considerable for analysis or classification. Thus, feature selection and its extraction are striking research areas.

### **1.1.2 ANN(Artificial Neural Network) using an MLP (Multilayer Perceptron)**

An artificial neural network feed forward's part is a multilayer perception. Artificial neural network is influenced by methods and anatomy of biological neural network. Traditionally, by

biological neural network that implies the complexity and working of human intellectual capacity. ANN attempts to emulate the operations of human mind using its fundamental concepts. Human brain is consisting of nerve cells which are known as neurons. In the same way, an artificial neural network is made up of enumerable artificial neurons which are known as nodes that act nearly the similar way a human brain's living neuron does. Human brain's neurons are resided of three parts.

Dendrites – Accepts the output of the previous layer.

Axon – Axon connect every Neuron with each other.

Synapses – Transfer of information to the next layer neurons.

The data is obtained by dendrites and delivered to the nucleus, then the decision is made by nucleolus about the creation of output signals, generation should be done or not. Provided that nucleus sent the output, which delivered to synapses with the help of axon which will be forwarded to the next layer neuron. The dendrites provide the information, which is further forwarded onto the cell nucleus, nucleus take the decision of transferring stimuli to the other layer. If it feels it is informative then stimuli are transferred to the next level otherwise it will be rejected provided threshold value. Similarly, an ANN made up of neurons called nodes, every node interconnected with each other by links and each link is corresponded with some weight. A node receives data from other nodes and dependent on the activation function used in a node and based on the results action is taken or ignored. The aggregate of input values approaches to a threshold value on a condition than it will passed onto the next level input. Otherwise, the input will be rejected or ignored and no action against is taken.

### **1.1.2.1 Advantages of ANN**

- Non-Linear Complex relationships can be created when ANNs start learning themselves. All the real- life relationship between input and output are non-linear relationships that are very useful to be built Non-Linear input-output graph.
- ANN Matches the intellectual capabilities of human, after learning from the big data set, as the dataset increases the performance of artificial intelligence increase. The key point, which is hard to humans to detect, it's easy for ANN to build the relation. That's why, Prediction on never seen input data's output become easy.

- No restriction is applied on input dataset by ANN. A study shows that ANN is capable of producing better results on heterogeneous data which contain a high very high volatility and non-constant variance.

### 1.1.2.2 Applications of ANN

- Recognition of Characters and Image processing – An image contains a non-linear instructed pattern; this is the reason ANN efficiently recognize the unseen objects and characters in images base on past learning. Based on calculus and predicting ANN, finds the cluster which is most compatible.
- Forecasting – Forecasting in any region is always unpredictable so based on experience ANN finds the unpredictable behavior and forecast it before it actually occurs with good accuracy. E.g. Weather, Stock market forecasting, score prediction, etc.

### 1.1.2.3 Disadvantages of ANN

- Long detaining times – Prediction of never seen data requires a lot of time due to the presence of huge dataset, training time increases to train a model. It mostly happens when CPU is used in place of GPU.
- Huge dataset required – Complex model requires a lot of data to form neural network model. A huge number of connections along with weights are present, so adjustment of weight is required according to the input provided.
- Architecture must be tuned fast – Solely obtaining good data set doesn't give an assured superior result. In a network optimally adjusted weights are required so that future prediction will be accurate.

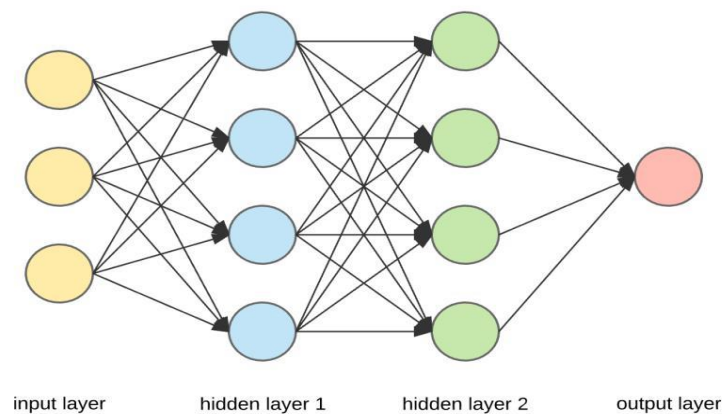


Figure 1.1 Architecture of ANN [1]

### **1.1.3 Feature Selections**

Feature selection is a dilemma of picking the small set of features which makes it feasible to locate the problem in a more compact, efficient and computationally effective manner. Feature selection includes generating new features, eliminating redundancy and needless features, amalgamating various features to a minimal count, also partitioning a feature to many other features.

### **1.1.4 Image segmentation applying Machine Learning**

Image segmentation is a major task in computer vision and machine learning techniques. Images are classified into predefined classes and image's meaningful parts get labeled to those classes using semantic segmentation. For e.g. robot controlling system, soil quality monitoring system, fruit and vegetable monitoring system, self-driving systems, background detection, production line quality management and body cells diagnosis system etc. In image segmentation, every pixel in the image is tagged to different classes. This process which indulges with pixel labeling part is known as dense prediction. Assume in an image, different objects present like cats, dogs, signals, tress and cars etc. In this case image segmentation classifies all cats in one class, all dogs to another single class and all cars to another single class. The most important thing in image segmentation is considering two different objects of the same type as a single class. For example, one tall tree and one thick short tree will be considered as a one class tree. Instance segmentation is used to differentiate between objects of same type.

## **1.2 Image Segmentation**

Imagine, a situation where you are waiting to crossroad. What you see around yourself before stepping forward and taking decision? Your eyes see numerous objects around like traffic lights footpath, vehicles, pedestrians and zebra-crossing. Now while crossing the road your eyes instantly identify every object and process their situations to make decision that moment to step forward or not.

Can computer or system perform this task? Well initially the answer to this problem was “No” a long ago before breakthrough inventions of Object Identification and Computer vision. At present, provided the help of image segmentation and object identification techniques it is

possible for computers to visualize the real-world objects in reel world objects and based on their location, orientation and positions necessary actions can be taken. In diagnosis and identification of cancer cells, image segmentation possesses an important role. It is crucial to identify the shape of blood cells and unusual growth in blood cell which can be diagnosed for the existence of cancer cell. Recognition of cancer at early stage is important so that it can be cured within time.

### **1.3 Image segmentation using CNN**

For Un-sequential or stable set of images, CNN (Convolutional Neural Network) is the most powerful, trending and useful Neural Network. Presently, everything works on Artificial Intelligence, Artificial Vision and automation. Everyone wants to ease their work by making self-driven cars, self-soil checking agriculture field system, self-temperature controlling system, based on facial expression song playing system or self-travelling robots. CNN made it possible. In image processing CNN is the best method in the identification of objects like human face etc.

### **1.4 Methods of Image Segmentation**

- i. Region Based Image Segmentation (RBIS)
- ii. Edge-detection Based Image Segmentation (EDIS)
- iii. Clustering Based Image Segmentation (CBIS)
- iv. CNN based Image Segmentation. (CNNIS)

#### **1.4.1 Region Based Image Segmentation (RBIS)**

Region based image segmentation divides the object in an image into different regions based on some threshold value(s). In this technique, Image pixel intensities are used. This technique separates the homogeneous intensity areas out from the higher intensity regions. To identify the difference threshold value is chosen. Those pixels having less intensity then the threshold comes under one region and higher intensity pixels come under different region. In the similar way, more than one threshold can be selected, and an Image can be divided into several parts depending on its pixel intensities.

This is the simplest way of segmenting an image that is to be based on its pixel values. This process makes the use of image pixel intensity difference as there is a huge difference between background pixels and object pixels. In this case we can allocate threshold value, the difference

in pixel values which comes below threshold value can be segmented accordingly. One object is present in single background then only one threshold value will be set. For many objects or overlapping objects, multiple threshold values will be needed. This Process is known as threshold segmentation.

#### **1.4.1.1 Advantages of Region Based Image Segmentation**

- Uncomplicated calculations.
- Object and Background have high intensity difference, this method gives better results.
- Speed of Operation performing is high.

#### **1.4.1.2 Limitations of Region Based Image Segmentation**

- Overlapped Image and background edge pixels create problems.
- Object and background have same pixel intensity or approx. same difference. Then it's hard to identify object.

### **1.4.2 Edge-detection Based Image Segmentation**

An Edge is the key to differentiate between object and its background as well as object from other objects. It is the fact that an Edge separate one object from another. This technique makes use of this fact to identify the object while scanning an image. Convolutions and filters help to detect the edge of new objects. Every filter is assigned with specific edge to detect, in this technique we have filter matrix which works on every image to detect the edge. Edge detection-based image segmentation technique is the most popular and easy to use technique. Mostly task in image processing is image identifications, which can be achieved by edge detection in an image. In an image, discontinuous local features are used to detect edges and boundary of the object. But this technique works well only for less not of objects in the image like one or two provided there is no overlapping between the objects. Rest of the cases it is hard to identify the edges of the object, so this technique fails at this point. These filter matrixes detect edges also known as kernels. These can be of multiple types. Some types are follows:

- Sobel Filter (horizontal) – Sobel filter is used for detecting edges. Gradient of image intensity is calculated at every pixel within the image. Direction of the largest increase from light to dark is found and the rate of change in that direction.



-1	-2	-1
0	0	0
1	2	1

Horizontal

Figure 1.2 Sobel Filter (horizontal)

- Sobel Filter (vertical) – Detection of vertical edge is done by the filter in the image.

-1	0	1
-2	0	2
-1	0	1

Vertical

Figure 1.3 Sobel Filter (vertical)

- Laplace Filter (both horizontal and vertical) – Detection Horizontal edges are done by the filter as well as vertical edges in the image.

1	1	1
1	-8	1
1	1	1

Figure 1.4 Laplace Filter (Both horizontal and vertical)

- Blurring filter – Blurring out an image is done by filter.

1/9	1/9	1/9
1/9	1/9	1/9
1/9	1/9	1/9

Figure 1.5 Blurring Filter

- Sharpening filter – Sharpening an image is done by filter.

0	-1	0
-1	5	-1
0	-1	0

Figure 1.6 Sharpening Filter

#### 1.4.2.1 Advantages of Edge-detection Based Image Segmentation

- Calculations are fast.
- Images having better contrast between objects give better result.

#### 1.4.2.2 Limitations of Edge-detection Based Image Segmentation

- Not suitable, if there are many objects in the image.
- If the object has less contrast between them, result won't come accurate.
- Overlapped objects boundaries in an image don't work properly

#### 1.4.2.3 Edge-detection Based Segmentation Algorithm

- **Step 1:** Take a weight matrix.
- **Step 2:** Put it on top of the image.
- **Step 3:** Process element-wise multiplication and get the output.
- **Step 4:** Move the weight matrix as per the stride chosen
- **Step 5:** Convolve until all the pixels of the input are used.

### 1.4.3 Cluster Based Image Segmentation

This technique divides the data points of an image into clusters having indistinguishable pixel values. Numbers of Clusters in an image are selected based on the objects available in the image. Throughout the Cluster processing at homogenous kind of data points were classified into single Cluster. In the end of the process one image get classified into multiple regions. Despite the fact that the process is time taking but it delivers the accurate results on small datasets. Cluster based image segmentation is done on the principle of splitting the pixels of an image into similar clusters. This process chooses a random cluster from the center of the image, and then keeps on expanding the size the cluster based on the similar characteristics. The image is divided into regions which are equal to the number of clusters chosen. Homogenous kind of data points

classified into single cluster. Then at last image gets classified into several regions. This process is time consuming, but it provides accurate results on small datasets.

#### **1.4.3.1 Advantages of Cluster Based Image Segmentation**

- Small datasets are suitable, and it creates excellent clusters.
- On small dataset gives accurate results

#### **1.4.3.2 Limitations of Cluster Based Image Segmentation**

- Very expensive because of too large computation time.
- K-means is not suitable for clustering non-convex clusters which is a distance-based algorithm.
- It is hard for selecting K's optimal value.

#### **1.4.4 CNN Based Image Segmentation**

This technique is a booming topic of image segmentation in the field of research. Images are converted into three dimensions i.e. height, width and number of channels. Height and width tell us the image resolution and the number of channels represents the intensity value of red, blue and green (RGB) colors. To reduce the processing time, the image is fed to a neural network and then reduced in dimensions which neglect the problem of underfitting. An image of size  $224 \times 224 \times 3$  (height \* width \* number of channels) is taken and converted into one dimension which will create an input vector of 150528. The input vector is too large to be fed as an input in the neural network.

This image segmentation technology is very important and widely used because it gives better results. It is basically used for semantically segment an image into regions which is used where we need self-driven systems like cars and many other places. To do this we need a collection of images and then manually label all the images respective to the objects. Labels like ball, bat, traffic signals, footpaths, cars, bikes, cycles, trucks, etc. are used using different masks and further this labeled data is used to train the CNN. After this if any image is fed into the neural network then it will be easy for the neural network to identify the object and can take the decisions. On a survey, it is found that deep learning is a very important part of image segmentation algorithm and technology. CNN is continuously contributing in making computer vision and automatic surveillance systems well trained and smart.

#### **1.4.4.1 Advantages of CNNBS (CNN Based Segmentation)**

- It performs traditional image segmentation methods.
- Any size of input can be executed.
- Pixel wise predictions can be done on trained model.

#### **1.4.4.2 Limitations of CNNBS (CNN Based Segmentation)**

- Large dataset is required.
- Step takes a long time for deep network.
- Parameter tuning is the main factor for performance of the network.

### **1.5 Applications of Image Segmentations**

- Autonomous Driving
- Motion Detection
- Object Recognition
- Video Surveillance
- Parking sensors
- Robotic path detection
- Facial segmentation- for identifying facial features, emotions or estimation or gender or age etc.
- Automatic braking
- Geo sensing – In Satellite images, it identifies target resource.
- In light microscopic images, its segmented cell
- Identification of Tumors and cancer cell detection in medical fields.
- Categorizing clothing items.
- Identification of neurons in electron microscopic recordings.

### **1.6 Deep Learning**

All the Machine Learning Techniques uses training data and based on the trained data, it processes new input and take decisions. With the help of deep learning with neural networks, system can make the decisions better without the need of labeled data. This method is useful for self-driven systems which require human intellectual capabilities. It uses differentiation to find the slope and predict properly between two images. Neural network uses algorithms side by side in which output of one algorithm will be the input of another algorithm. This creates a system

which can interact with human beings and act like one of us with its super high computation power as well as high capability of data storage. This model is perfect artificial intelligence systems which can take decisions just like humans as well give accurate results.

## **1.7 CNN (Convolutional Neural Network)**

CNN classifies images. Three-dimensional input images are feed into CNN system i.e. height, width and Number of channels. Height and width tell us about the image resolution and the numbers of channels which represent (RGB) system tell us about the intensity of red, green and blue colors respectively. Images that are fed into neural network are reduced in dimensions which reduce the processing time and decrease the problem of image shape like under fitting or over fitting. Although if we take an image of  $244*244*3$  which will be converted into 1 dimension and make an input set of 150528. This number is too large to feed it as an input to the neural network.

Layers of CNN are:

- Convolution layer
- Activation Layer (e.g. ReLu)
- Pooling layer
- Batch norm layer
- Drop out layer
- Fully connected layer

### **1.7.1 Convolution Layer**

A filter or kernel is executed on image in fixed gap intervals called strides. Size of stride selection is crucial to achieve desired results. Calculation is done during of the filter on the image, in filters dot product is done with part of image. Map matrix featured by convolving using the sum of all values of product matrix is reprinted to corresponding position. Thus, features of image decreased and mapped to an image. Different kind of filters are available; every filter is used to extract features from image. For ex., one feature is extracted using one filter based on its shape and edges and another filter used to extract feature based intensive color.

### **1.7.1.1 Adjusting CNN's Performance based Parameters**

- Filters – Filters are also called as kernels. It comes in many types. Every filter expands the depth of the output created after convolution. That's why; even we are using 3 filters then the 3 output. Convolution output depends upon 3 parameters i.e. Depth, Stride and Padding. We tuned these parameters to get desired results.
- Stride – The Number of Pixels by which we must move our filter over the image so we can anchor on a new set of pixels while doing convolution. Range of stride's value ranges from 1 to 3 based on the amount of loss which can be achieved during convolution. The value of stride is directly proportional to loss in image.
- Padding – Padding is Process of adding zeros in the image's border to make it symmetrically. This will help to map the output features matches to the size required. After convolution it preserves the edges of the images.

### **1.7.1.2 CNN Dimension Reduction**

In the process of convolution, the dimensions of the input image decrease to the filter image which runs over the image provided. This process increases the depth of the images as it helps to identify the regions in the image. E.g. An image which is of  $9 \times 9$  RGB with 3, kernels of  $3 \times 3$  at stride 1 (per channel), this will create a feature containing RGB which is of  $7 \times 7 \times 3$ . Here 3<sup>rd</sup> dimension depicts the depth of the convolved image.

### **1.7.2 Pooling Layer**

The Pooling Layer works on a small kernel on the image where stride is fixed. In this method highest intensity pixels are picked up and discard other pixels. Dimensional matrix of feature image is created by reducing the resultant matrix. Unnecessary sparse cells of image will be reduced with this method which is of no use in classification. Dimensionality of the network or image can be reduced with the help of Max Pooling, but this process can lead to information loss. The main concept of this method is that nearby or adjacent pixels can be approximate by the high intensity pixels which contain the Maximum information carrier.

### **1.7.3 Activation Layer**

The Activation layer uses the ReLu function as an Activation function. It is a function which put all the negative values to zeros and keeps

All the negative values to zeros and keeps the positive value as it is. This is done after the convolution layer and pooling layer. This step helps in many fields of Computer vision or Artificial intelligence as it helps to take better decisions as it eliminated the negative part which is not required in the image. This step is so powerful that it can lead to near human intelligence results or sometimes better than that. Well it does sounds interesting, but the designing of CNN is a grueling task. There is no fixed formula for designing of CNN. Many researchers have come up with different algorithms, suggested many techniques. But they don't hold for every from we face where CNN is applied. In CNN the task is based on data applied or provided and on algorithm. If the dataset provided for training contain any default, then the outcome provided by it will be faulty also no matter how accurate the algorithm is. The task in Deep leaning technique is data driven. CNN utilizes the occupied space hierarchical features of data withdraw features and assist in classifying them into unique classes. This need has led to collection of augmented datasets and pre-processed to increase the data, the more data increases the possibility of better training and shunning over fitting. Robust models can be built when we provide more generalized data to avoid noise in the training phase.

### **1.7.4 Dropout Layer**

After fully connected network which contain neurons in the layer, dropout layer is applied. Regularization layer is dropout layer. This layer basically randomly drops or decreases the weight of some neurons in alternative iterations so that the other neurons will get more weightage in that particular iteration which will give better results to enhance the feature of the image. Generally, people take dropouts of around 20-50% with the help this dropout layer. It increases the performance of the system by 1 to 2 percent.

### **1.7.5 Fully Connected Layer**

Pyramid structure is created by top to bottom model structure. Top down structure contains different layers in which the parameters of these layers continuously merge till they finally reach

the number of desired output classes. To increase the learning ability of CNN, the number of hidden units in the layers are increased which helps to understand the problem more precisely. There is no fixed formula yet defined for CNN to solve a particular problem, although it works on hit and trial method mostly. The subset of network which is fully connected usually piled up. It is found out in researches that the networks which are multiple of 64 work well. For some layer like two- or three-layers network are excellent if there are ample patterns being forwarded to the network after compress the outputs of the Convolutional layers.

### **1.7.6 Data Augmentation and Pre-processing**

CNN absorbs in special hierarchical features; the presence of pattern limits the capacity of algorithm that exists in the data. Generalization is one of the features needed in a good algorithm. Trained data is prone to over fit and it gives high accuracy results on the training dataset while it gives poor performance on the untrained or untested data. To improve it, data augmentation and pre-processing is used. To make more robust to real world unknown or unseen data, many techniques used as channel drop, rotation, translation, flipping, cropping and Gaussian noise help to create more synthetic data and classify it with better accuracy. Optimizing the parameters of CNN in a more accurate manner can be done by the help of more data which turns out to be beneficiary for the system. In this Study, we study the case of video analysis. There are many techniques which provide image free of noises like Gaussian Noises to provide better clearer image. Noises which can be present are such as dark image and video recorded in dark lightning, super resolution and Blurred image reconstruction.

### **1.7.7 CNN Architectural Design**

The Architecture of CNN design structure is based on data and optimizing technique helps most of the CNNs with low dependency on the structure. In this section, we are going to look at the techniques of a reverse manner which is alike to the flow of a back propagation to signify the learning cycle of a CNN. Architecture of CNN includes the flow of all the steps taken in creating convolution neural network. Which includes CNN input that collect data and store it in a data base then Convolution Layer, followed by Activation Layer that include ReLu layer, Dropout Layer, followed by Batch Norm Layer, which create Fully connected Layer and then conclude the Output Layer. Following is the flow chat of the Architecture of CNN.



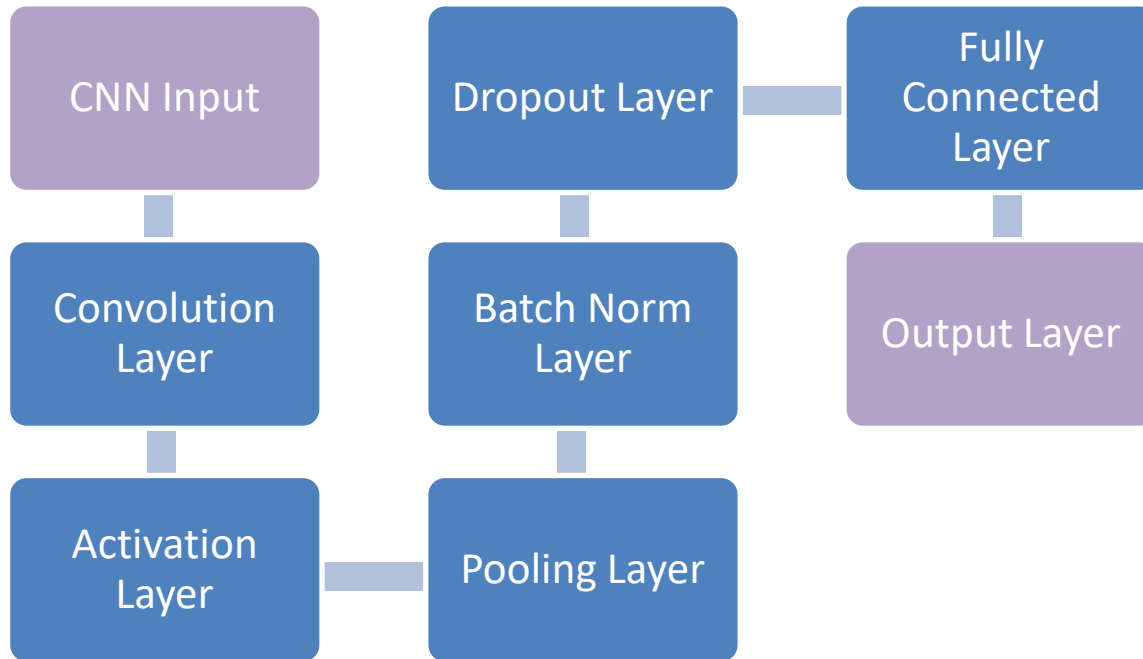


Figure 1.7 Architecture of CNN

### 1.7.8 OUTPUT OF CNN

The probability of different outputs classified the output of CNN in tasks. Examples which are present are the final output of a CNN which include 4 units (denoted as classes), it is normalized into the probability of most occurrences of such classes and then maximum probability is marked as the output class. The tagging of class may actually differ for task specific cases such as face recognition. In this case the output is basically encoded from the input rather than the representation of output for particular class. Reduction of loss/cost function minimizes the error which uses Back propagation. With the help of different filters, the image matrix is converted into features Matrices that will give different output based on the different filters which will converge, all the feature matrix then added to produce convolved feature map.

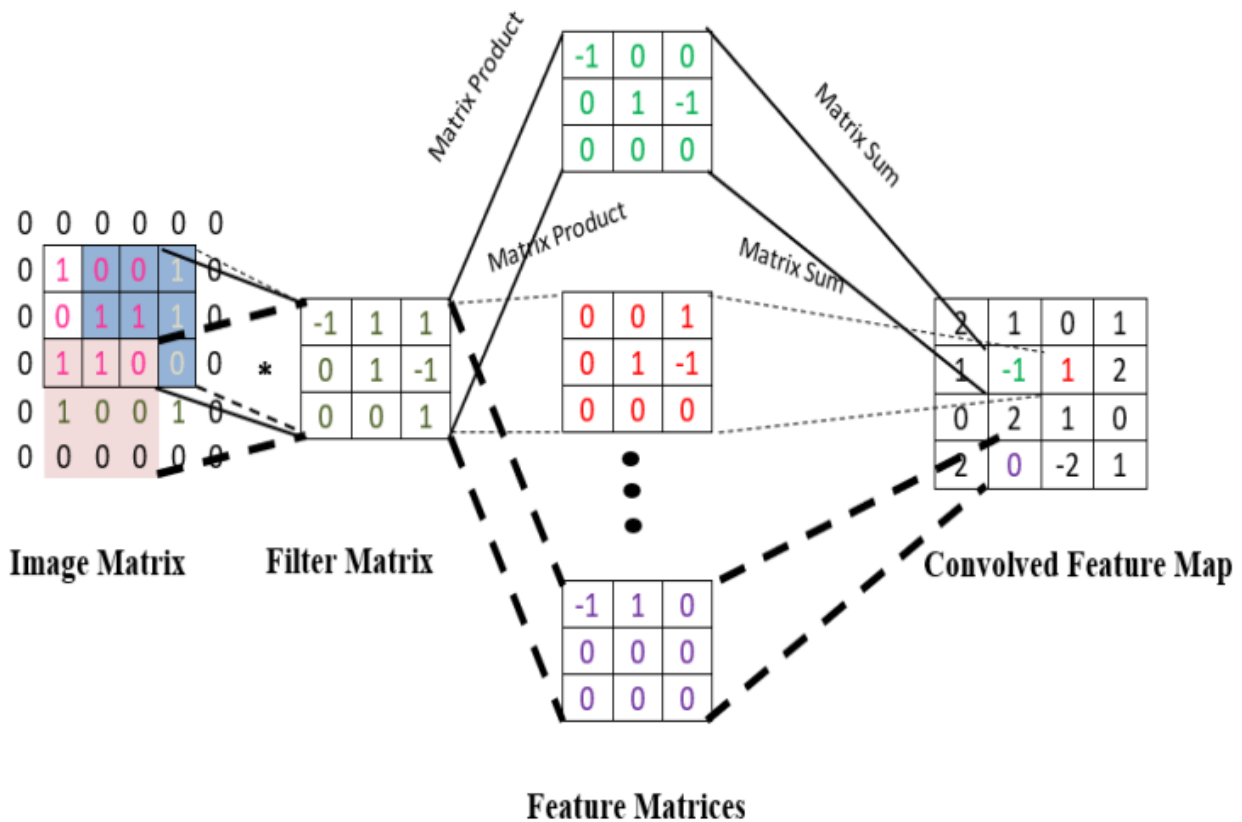


Figure 1.8 Workflow diagram of CNN [1]

## 1.8 Recurrent neural network (RNN)

RNN is another class of ANN. It allows expressing the temporal dynamic behavior. It is derived from feed forward neural networks, RNNs uses their hidden clusters or intermediate mode to process irregular length sequences of input data stream. That's why it is applicable to tasks such as un-segmented, connected sign language or handwriting recognition or speech recognition.

Recurrent neural network is used indiscriminately is classified into two major classes where one is finite impulse and the other is infinite impulse. Temporal dynamic behavior can be shown by both class of network. A Directed acyclic graph is used by finite impulse recurrent network with a strictly feed-forward neural network, on the other a directed cyclic graph is used by an infinite impulse recurrent network that can be unrolled. Additional storage is present in both Infinite and finite impulse recurrent networks and neural network has a direct control on it. Graphs and another network replace the storage. Provided that it works with time delays or contain feedback loops. These controlled states are labeled as gated state or gated memory, and

these are the part of LSTM network (Long Short-Term Memory) and gated recurrent network. This is also called as Feedback Neural Network.

### 1.8.1 Architectures of RNN

RNNs come in many variants.

- Fully recurrent:** Neuron-like nodes are present in the basic architecture of RNNs, which is organized into successive layers. In this architecture each and every node is connected with other successive node by one way directed connection. Each value has real time varying real valued activation function in it.

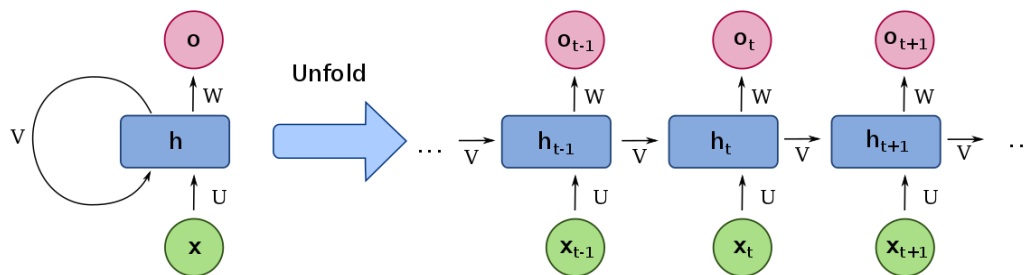


Figure 1.9 Unfolded Basic Recurrent neural networks [1]

- Elman networks d Jordan networks:** This network contains 3-layer which is arranged horizontally as x, y, and z in the illustration which include set of context units. The hidden layer is attached with the context units which are fixed with a weight. At every step of layers, the input if fed forward and a rule is applied for learning. Multilayer perceptron performs the task such as sequence-prediction.

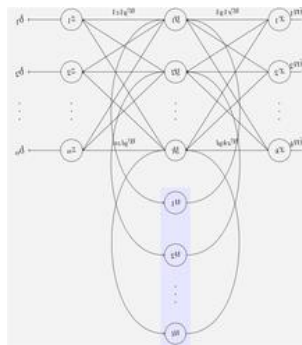


Figure 1.10 the Elman network [7]

- **Hopfield:** All the connections are symmetric in this type of RNN. It is not general RNN because it requires stationary inputs; it doesn't process sequence of patterns. When the network is trained using Hebbian learning then this RNN provide robust content-addressable memory.
- **Independently RNN (IndRNN):** In the fully connected RNN problems were addressed by gradient vanishing and exploding problems. Context information is being sent to the neurons layer where it receives its own data therefore it does not depend of each other's history. To keep long or short- term memory it is required to regulate the gradient back-propagation to keep away from gradient vanishing and exploding. The cross-neuron information is regulated to next layers. This system can become robust using different non-saturated non liner functions such as ReLU.
- **Second order RNNs:** it uses higher order weights in place of standard weights where state in standard weight is 'j' while in second order RNNs it can be the product of k 'j \* k'. Training, stability and representation are allowed by direct mapping to a Finite state machine. LSTM is an example but doesn't have any formal mapping or proof of stability.
- **Long Short-Term Memory:** It is a deep learning system which keeps away from the vanishing gradient problem. "Forget gates" augmented the normal LSTM. Vanishing and exploding creates backpropagation errors that is prevented by LSTM. LSTM is made to acquire very large amount of task that needs memories of the happened events which occur millions of distinct time steps before. Different type of applications is used which uses LSMT RNNs and trained on them by Connectionist Temporal Classification (CTC). LSTM works differently from Hidden Markov Models (HMM) that recognize Context-sensitive language.
- **Neural Network Pushdown Automata:** It works more like NTMS, in which analogue stacks replaces the tapes which are differentiable. Differentiable neural computers (DNCs) allows the usage of fuzzy amounts. Its complexity is similar to recognizers of context free grammars (CFGs).

## CHAPTER 2

### LITERATURE REVIEW

---

**Kumud Tripathi** and her team have used image processing and computer vision. In his research, both hands are used for gesture recognition. Where gesture recognition difficult problem is elucidated by gradient based key frame extraction method [1]. Key frame extracted from this method is very helpful for dividing continuous sign gestures into sequence of signs and useless frames. The split gestures, each sign treated as separate gesture. Then these gestures are used to create features after processing on which Orientation Histogram (OH) with Principal Component Analysis (PCA) is implemented for shrinking the dimension of features acquired after OH. Dataset is created by the canon EOS camera in Robotics and Artificial Intelligence laboratory [1]. Research is examined using different types of compartmentalization like Euclidean distance, city block distance, Manhattan distance, Correlation etc. The experiment is performed using 10 type of sentences and each sentence contain more than 2 gesture. But the range did not exceed 5. Where every continuous gesture is created using static and dynamic gesture. While creating the dataset, every sentence is recorded ten times, six times for training and four times for testing. In training set, every sentence is created by 20 frames where these frames are extracted from the set of n numbers of frames (n is the variable differ from sentence to sentence). In testing set, each sentence is created by 10 frames. This research purpose based on comparative analysis of distance classifier. From their analysis it is observed that better results are provided by Euclidean and Correlation than other classifiers. In the proposed framework for real time gesture recognition in which continuous gestures are used, gives adequate performance including features obtained using OH with PCA where both the hands been used for creating any gesture. In this process hand segmentation is applied before applying OH for extracting features for training dataset. In this research the accuracy is measured by the maximum number of frame matches. Euclidean distance and correlation give satisfactory result when results created experimentally.

**Anup Nandy**, have done Image processing and gesture recognition using both type of frames, static and dynamic hand gestures. In his work most of the gestures are formed by using both

hands and the creation of dataset and its consumption is done which obtained several videos, for a large number of signs. The main problem in image processing occur is illumination and orientation. For such problem Direction Histogram feature is used for classification. In this research, two main approach is used for gesture recognition are K-nearest neighbor metrics and Euclidean distance. Here features extraction is done by extracting frames from video and the converting every frame into grey-scale image. Then resize all the images into 160 X 120 pixels after that normalization to be done. On that normalized image 3 tap derivative filters are applied, U direction =  $\{0 -1 1\}$ , V direction =  $\{0 1 -1\}$  and Gradient of image X is calculated at every point of (U, V). this process creates Quantized angle. These angles are stored after normalization of values is done. For collecting feature vector, two different classifiers are used which are Euclidean and K-nearest neighbor. Where K-nearest neighbor gives very good classification result [2]. Approximately 100% recognition is obtained. 36 bins give more accurate result during classification. Which time consumed by 36 bins using direction histogram is large. Few ISL considered during testing were complex and repeated frames were used which gave quite a poor result. To handle such situation where two gestures or words are similar, different features are used to improve the recognition rate. Recognition of gestures are entertained by choosing the closed pattern among all the training set available [1].

**Sepp Hochreiter**, done the research on storing information by recurrent backpropagation over extended time intervals, which takes quite a long time, mostly happens because of the insufficient, decaying error backflow [3]. This research reviews Hochreiter's (1991) analysis. Then this problem I solved by gradient based method which is known as long short-term memory (LSTM). In LSTM it truncates the gradient where it won't affect. LSTM can learn to decrease the time lags in surplus of 1000 discrete-time steps by enforcing constant error flow through constant error carousels [3]. Constant error flow can be managed by the multiplicative gate units which can open and close the access for the same. The space and time complexity of LSTM per time step and weight is  $O(1)$  which is also known as local in space and time. This experiment, which is done on artificial data includes distributed, local, noisy valued and real-valued representations. Research includes the comparison between different methods with LSTM which includes recurrent cascade correlation, recurrent learning, Elman nets, backpropagation through time and neural sequence chunking that provided the result that LSTM is the more successful method in all the runs. It can learn faster with large data provided. It can solve more

complex task that can never been solved before by different recurrent algorithms. This paper discusses the limitations and advantages of this process under different condition and how algorithm works. LSTM does not provide finite number of state while the static model like Hidden Markov models works on definite state automata, it deals with unlimited state numbers. That's why it is best suitable for real time gesture recognition. The problem occurs when each memory cell needs two additional units for input and output gate. But it does not increase more than a factor of 9; each hidden conventional layer is replaced by 3 units in LSTM which increases the weight by a factor of  $3^2$  in fully connected case [3]. The experiment contains 6 different types of problems and their results with the comparison of different method on that problem.

**Franco Ronchetti**, and his team created an Argentinian Sign Language Dataset. The signs differ for different languages which the technique used for sign language recognition is almost same. Training the different language requires entire new dataset for that language. The paper presents 64 signs. And the dataset is called as LSA64. Contains 3200 videos of 64 different signs recorded by 10 subjects [4]. For better performance colored gloves are used to identify properly, this way it provides ease in identifying hand and segmenting. There is a pre-processed version also available from which handshape of sign, position and statistics of movement is computed. The database is two sets which include the hand gesture of both hands separately. One dataset contains one hand sign gestures and the other contain two hand sign gestures. Every set of videos they have used light background, preferably white wall and dark dress background including black dress and hair and to distinguish between both the hands. Different colored hands are used.

**David E. Rumelhart**, in his research work, a new learning procedure is proposed which is back propagation. In this process the weight of the edges in neural network regulate their weights to decrease the difference between the actual output value to the expected value. The hidden layers which are not contributing anything in the output and hold important role will be eaten up by the collision of these units. Evolving new feature's ability defines back-propagation. The main motive of this method was self-maintaining neural network.

**Lionel Pigou**, done a research on human computer interaction using CNN. In the research the gesture recognition system used GPU acceleration, Convolutional neural networks (CNNs) and MK (Microsoft Kinect). In which 20 gestures were recognized accurately. In different

surrounding the model obtain 91.7% accuracy with challenging lightning condition and obtain 0.789 score in Jaccard index. Which conclude that CNN is very powerful which recognizing motion gesture recognition. In this model they have used “ChaLearn looking at people 2014” dataset. This dataset contains 20 different gestures contributed by 27 different people from 37 different background. 6600 gestures are used in 1<sup>st</sup> stage. Then separated into training set includes 4600 and test set contain 2000 videos. The test set environment is different from the train set by user and background. The model includes the comparison between different methods like data augmentation, dropout, ReLU and LCN. The validation accuracy is coming out 91.7% with 8.30% error rate due to change in weather condition and background as well as light situation. The accuracy on the validate set turned out to be 95.68% and error rate comes to be nearly 4% sue to the noise present in the video.

**Chenyang Zhang**, in this research 3D sensors are used to measure the depth of real time gesture recognition maps. In this method the depth of information is retried by grey images and the strong 3D sharpness is ignored. In the paper they have described inventive characteristics, i.e., vividly encode the 3D shape data from depth maps using Histogram of 3D Facets (H3DF) [7]. With the help of strong coding and pooling facets from depth map, the H3DF can effectively represent the shape and motion of hand gesture. This method used two types of classification in making hand gesture recognition, one is defined as numbers and the other is defined as alphabet. The dataset used here is the ASL Finger Spelling dataset [7] and NTU hand Digit dataset [7]. Both featured by same specified cam. In digit hand gesture, 1000 samples are used having depth maps of 10 subjects with 10 digits (from 0 to 9) and every digit will have 10 videos. While in Alphabet it contains 60,000 gestures obtained from 5 subjects. Which is having 24 alphabets excluding j and z. Traditional 2D HOG descriptor gives lesser accuracy with 7% as compare to this method.

The technology used in gesture recognition in this dissertation is vision-based method.

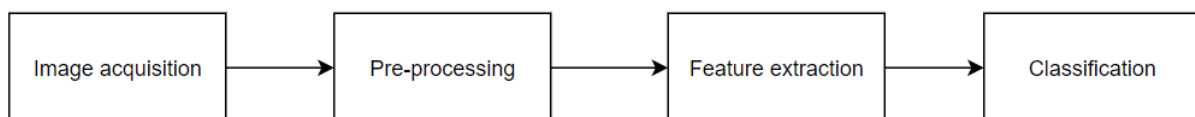


Fig 2.1: Block Diagram of vision based recognition system



Many different methods are used till now in with different combinations of algorithms and image capturing method is used. There are two main methods which are used widely.

1. Image of hand or moving object will be converted into 3D model. To obtain that image one or more cams are used. And its dimensions are approximated using different joint angles and its orientation.
2. Image is captured using a cam then converting those images using different filters to create features that will work as gestures inputted in different algorithm.

## 2.1 Approach Used in above described Papers

### 2.1.1 ProbSom used in Argentinian Sign Language (LSA) [4]

Dataset which is used is Argentinian Sign Language (LSA). Research is based on two approaches which uses one's methods output as the input of the second method that include CNN and RNN. ProbSom is a terminology used for the techniques of image processing, image shape classification and its extraction in a self-organized supervised learning. ProbSom based neural network gives 90% above accuracy if it follows the below term.

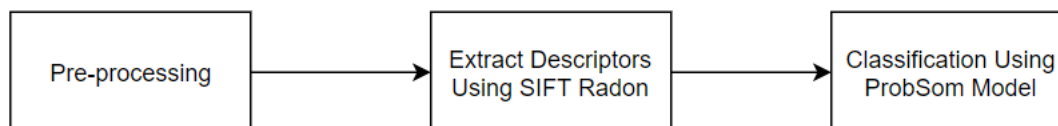


Fig 2.2: Block Diagram of Hand Gesture Recognition system for LSA

### 2.1.2 Sign Language Recognition in Continuous Video Sequence [3]

In continuous Video Sequence, while exacting features from separated frames four major steps are taken place to perform the actions. LSA contain 24 different alphabets and using this method 96% accuracy is acquired. the major modules are:

1. Data Acquisition
2. Preprocessing
3. Feature Extraction
4. Classification

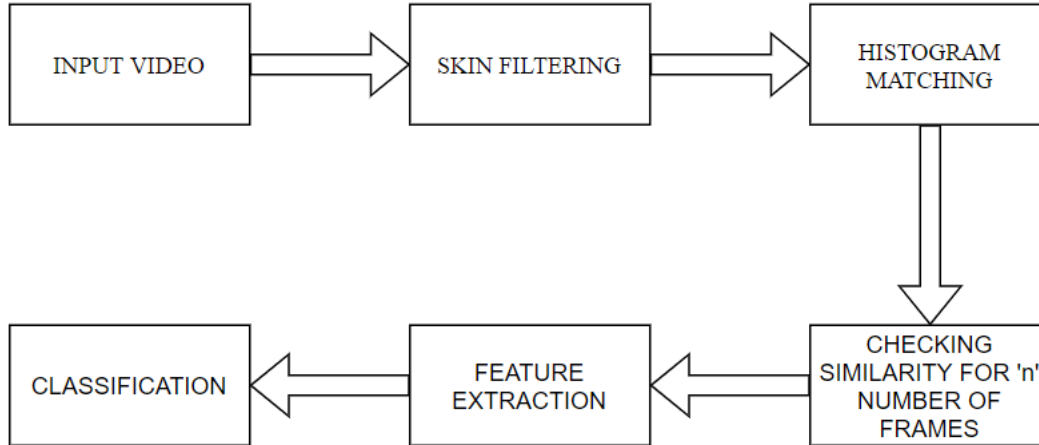


Fig 2.2: Block Diagram of System Overview [2]

### 2.1.3 Sentence Formation using Gesture Recognition [1]

Gesture recognition in real time is very difficult task. These problems are solved by researchers with the help of gradient based key frame extraction. This method removes the useless frames and gives the set of frames which stores the features and those set of sequenced features converted into sign. Signs obtained from that process to make features and further processed to make sentence. These features are obtained by OH with PCA which optimize the dimensions.

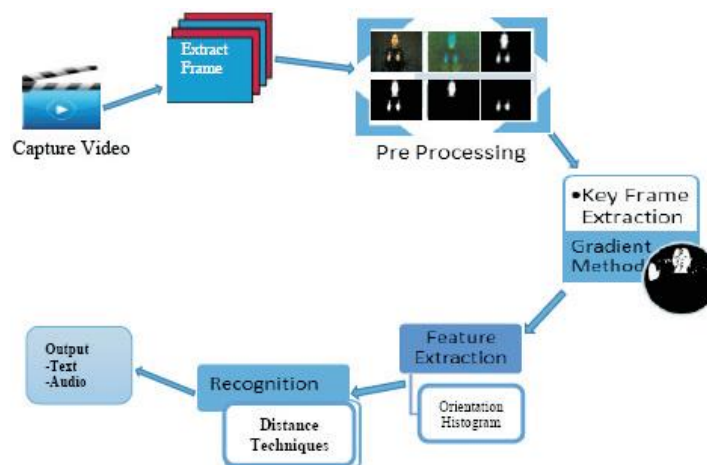


Fig 2.3: General Diagram [1]



Fig 2.4: Gesture of Sentence [1]

#### 2.1.4 Recognition of Gesture in Real Time [3]

Recognition of Gesture in Real Time which uses both the hands requires statistical techniques. Two different approaches are used for recognition of gestures, which are Euclidean and K-nearest neighbor metrics. In which K- nearest neighbor gives higher accuracy.

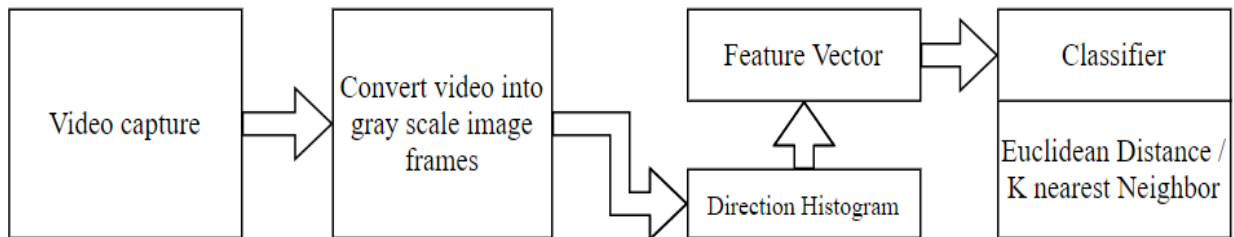


Fig 2.5: Methodology for Real Time ISL Classification [3]

# CHAPTER 3

## IMPLEMENTATION

### 3.1 Convolution Neural Network (CNN)

Neural network has a structure as brain. It contains same units for learning known as neurons. These neurons convert input into output by learning the automated recognition. Feed- forward architecture is followed by CNN. It has monotonous blocks of neurons that is applied for audio as well as real-time image processing. The neurons convert into image and then treat them like 2D convolution kernels, applied on each image. And for audio, it is converted into one dimensional convolutional kernel that works on time difference. Training includes these repeated units are shared among all nodes.

#### 3.1.1 Procedure followed in CNN

These are the main steps followed in CNN in ANN.

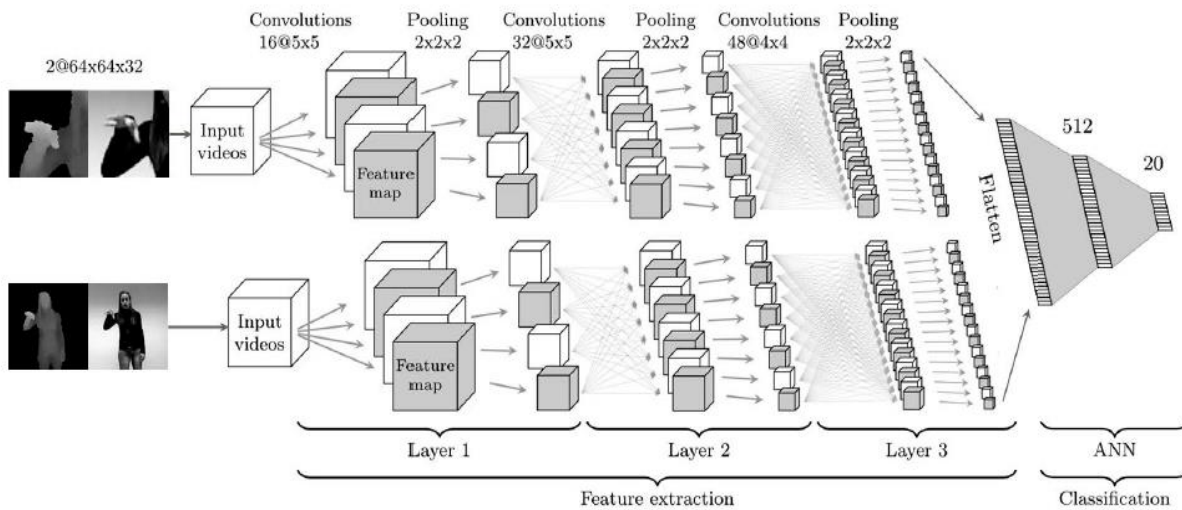


Fig 3.1: Convolutional Neural Network [10]

### **3.1.1.1 Convolution**

Convolution layer is the first layer that receives the input and works as filter. In this layer the blocks work according to their previous experience they learned. If the image matches with the previously tackled image, then it is labeled with the previous image signals and then the result is pass on to the next stage. It has different layers and blocks; blocks are known as convolutions. These filters show features, convolution layer is translational invariant. Which helps to maintain the orientation of the object detected easily.

### **3.1.1.2 Subsampling**

Subsampling layer removes the unwanted action to the noise and differentiation. This process works as smoothening and know subsampling. Subsampling takes the average of all the frames present as the output of convolution layer and then take the maximum of it. In this process we reduce the size or color of image.

### **3.1.1.3 Pooling**

This layer reduces the spatial size of rendering the dimensions and computation in the network. This layer works on every feature independently. In this process the maximum of all the region is taken out from all the features and collaborated.

### **3.1.1.4 Activation**

This layer takes cares of the control flow from one layer to another. The signals which will be pass on to the next layer will be decided by the activation layer. Past reference helps to activate the next layer of neurons to activate them. These signals identify more effectively. There are wide variety of activation function present in CNN. For example, ReLU (Rectified Linear Unit).

### **3.1.1.5 Fully Connected**

This layer is the result come from the other layers, that's why it is known as fully connected. Here all the possible paths are connected. Last layer which provides the expected output.

CNN algorithms are homologous to the work of playing chess. With every step it will be easier for a system to predict which path will lead to the victory and with every play you will learn about the player and its move. So, it will be easier to play the next game. Same as in playing cards. Every move it will be easier to learn and for the cards which are not helpful to predict the results or make you win you can discard them or directly add in the cart. This algorithm is called as gradient descent. Training process and learning new feature requires feedback.

Feedback is done by set of validation where prediction is done by CNN and compared with the test values or set of true labeled values. The predictions which does not match with the expected output treated as error or useless and those feed back to the CNN known as backward-pass. And this algorithm is called as Error's Backpropagation. CNN are very difficult to develop from base. So different tools are available like Tensor flow, Torch, Caffe and Mat Conv Net.

**3.1.2 Establishment**

Tensor flow libraries are used or Inception of version 3. For establishing this CNN's classification model, Inception [11] is used to differentiate the huge number of images. Training of last layer is to be done and this step takes the most chunk of time. Inception version 3 model is trained for Image and gave least error rate of nearly 3.46%. trained model is available, using that model training of final layer is being done.

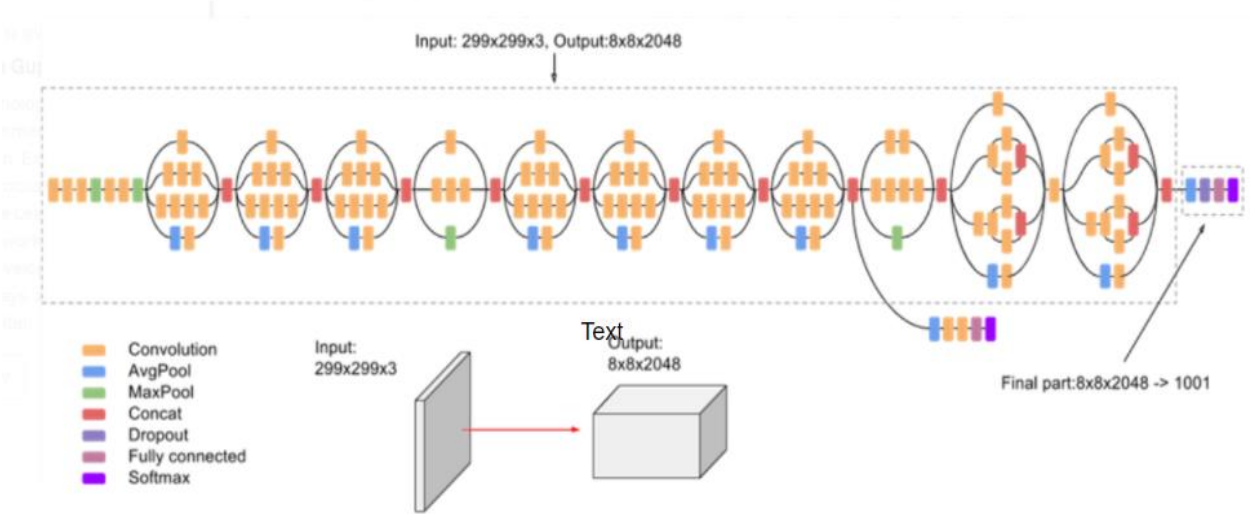


Fig 3.2: Inception V3 Model Architecture

### 3.2 Recurrent Neural Network (RNN)

Our thinking process doesn't start from the basics of every problem. We always take actions on every problem of ours based on previous problem faced by us. Human mind stores the action performed as well as the output of the problem and its solution. Traditional Neural network can't do the saving procedure, but the RNN works on this principal. The process of solving a problem is also a big information about what problem to be solved how and the expected problems during its way. RNN contain feedback loop which insert the output of the layers as the input of the same layer. This process is very effective when the data provided is sequential. Here it is very effective because each neuron stores the information about the previous layers or input. Because every time updating the current working state with the previous one commit request it, is useless.

#### 3.2.1 Loops present in RNN

Loops are present in RNN that allow data to be carried out between neurons while accessing it. The data when pass from one layer to another, it passes through the layer itself using a loop.

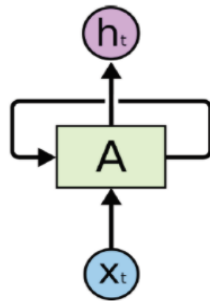


Fig 3.3: Recurrent Neural Network's chunk

Here, in diagram A is the RNN's chunk and  $X_t$  is the input and  $h_t$  is the output and here is input which loop down the chunk. So, the previous  $t-1$  step does affect the decision.

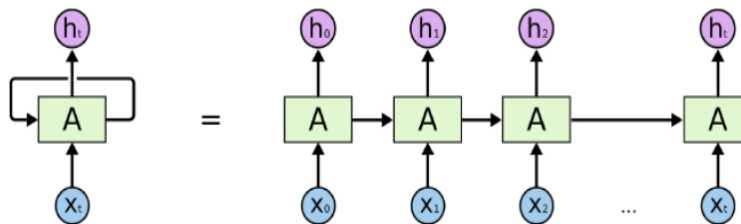


Fig 3.4: An Unrolled Recurrent Neural Network

### 3.2.2 Memory carry forward by previous inputs

$$h_t = \phi(Wx_t + Uh_{t-1}) \quad 1$$

$h_t$  is the hidden layer of RNN at time  $t$ . in this function the input  $x_t$  is provided and it modified the weight  $W$ . which will be added the previous time step  $t-1$  and multiplied by its own transition state. The matrix of weight tell the significance of present and past input. Back Propagation Through Time (BPTT) helps to manage the error created during the transition.

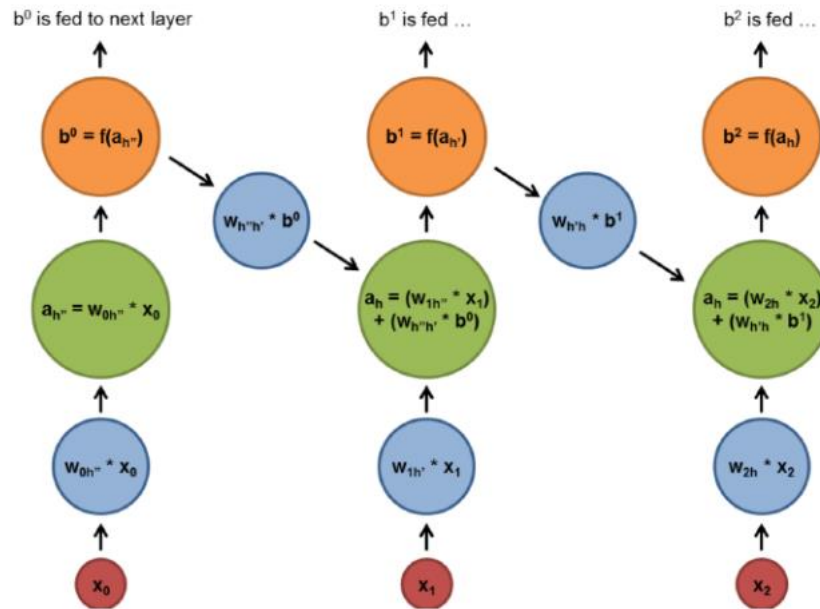


Fig 3.5: Memory of Previous inputs carried forward

### 3.2.3 Vanishing Gradient Problem

Theoretically, RNN can solve all the problem that depends on Long- term dependencies. But RNN doesn't able to solve such problems as provided in [6]. The Gradient changes with the change in error present in every layer. Deep neural network are related to each other by multiplication so the gradient is capable of vanishing.



### 3.2.3.1 Vanishing Gradient

The output gradient present in the layer vanishes as compared to the dimensions of the next layer. Effect on the output is not visible even if there is a large change present in the previous layer. So, the parameters are not available to the next layer which creates the problem.

Activation functions (tanh and sigmoid) creates such problem because they provide very small value as input. For example, sigmoid shows the input in the range of  $[0,1]$ . Because of this a very large range input mapped to very small range of sigmoid function. In this case a huge change in input will reflect negligible change in the output function.

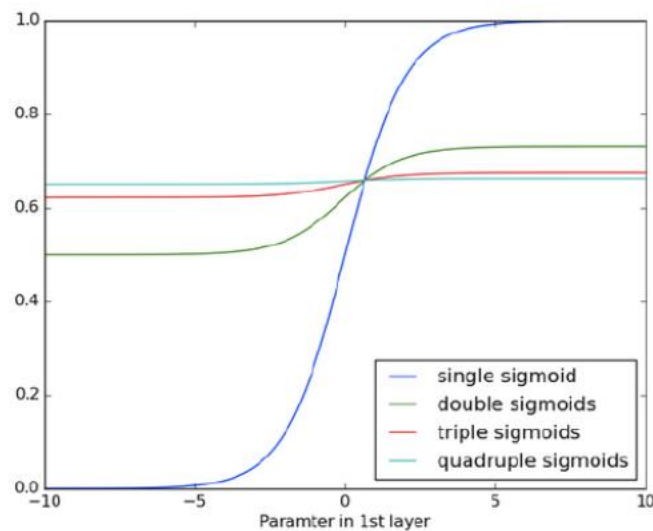


Fig 3.6: Vanishing sigmoid (Vanishing Gradient)

This problem increases when the huge input converges into a small output. That small output used input and merge into even smaller output. And this process continues. It is clearly visible in 3.1 that in the end the data flattened. Just like this the gradient vanishes.

### 3.2.3.2 Exploding Gradient

Those gradients which becomes different from others on high end. That are assumed very strong. These gradients can be solved easily because it truncates or ended. It will be difficult to work with so small data. The small the input the better result it will provide.

### 3.2.4 Long Short-Term Memory Units (LSTMs)

The research paper written by Sepp Hochreiter and Juergen Schmidhuber [6] proposed a solution of vanishing gradient problem. This method helps to find the solution of problem created by the backpropagation through time and layers. This allows it over more than 1000 steps. Long-term dependency problem is removed by the LSTMs.

### 3.2.5 RNN Model

RNN model's creation is based on Long Short-Term Memory model. The input is provided to the 1<sup>st</sup> layer of the model and the capacity of weight is defined by the input provided. In one single layer 256 units are present of LSTM. Then the activation layer is followed. Every layer neuron relates to the previous layer neurons. Then the output is transferred to soft\_max layer which is fully connected. Then it reaches to the last layer which contain minimum loss function on regression layer is applied.

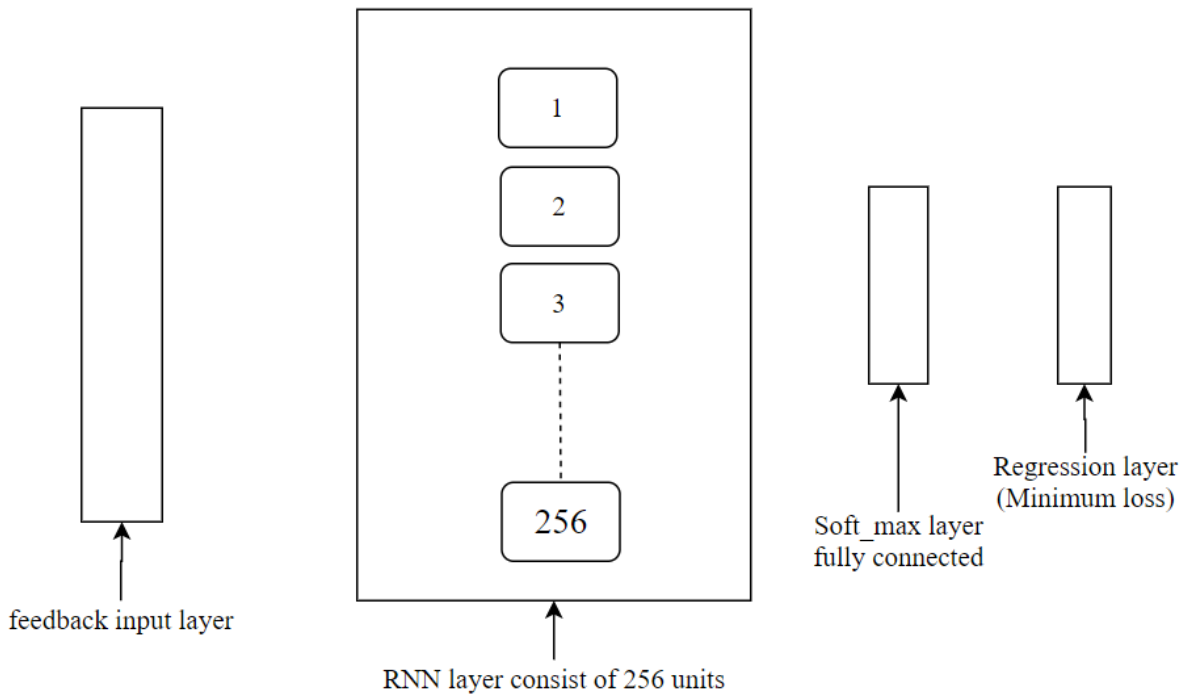


Fig 3.7: RNN Model

# CHAPTER 4

## EXPERIMENTAL DESIGN

---

Two types of approaches are used to train the model. The difference is with the input provided.

### 4.1 Data Set

LSA (Argentinian Sign Language) is the database is used to compare the efficiency of the approaches. This database contains 2300 videos. Here used the unprocessed video in which both the hands have been used.

<b>Id</b>	<b>Name</b>	<b>Id</b>	<b>Name</b>	<b>Id</b>	<b>Name</b>	<b>Id</b>	<b>Name</b>
1	Rice	13	Bathe	25	Deaf	37	Thanks
2	To-Land	14	Country	26	Enemy	38	Son
3	Patience	15	Drawer	27	None	39	Cat
4	Yellow	16	Red	28	Dance	40	Music
5	Food	17	Spaghetti	29	Man	41	Help
6	Give	18	Call	30	Green	42	Trap
7	Photo	19	Barbeque	31	Yogurt	43	Born
8	Away	20	Run	32	Coin	44	Accept
9	Chewing-gum	21	Uruguay	33	Name	45	Perfume
10	Copy	22	Bitter	34	Where	46	Opaque
11	Sweet Milk	23	Milk	35	Catch	47	Colors
12	Skimmer	24	Map	36	Breakfast	48	Water

Out of these gestures which contain 48 categories in which 25% used for testing and rest used for training.

## 4.2 Approach First

CNN model obtains the dimension feature of every frame and RNN obtains the temporal feature. Sequence of frames makes a video and every frame represent a gesture that will be predicted in sequence. The output of CNN is provided as input to RNN.

### 4.2.1 Methodology

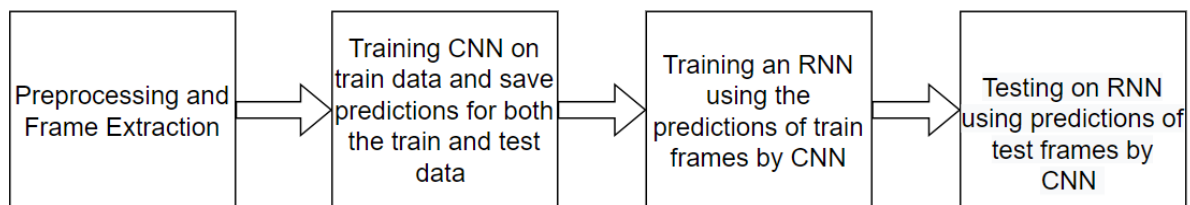


Fig 4.1: Block diagram of the Process

- From video, multiple frames are extracted in sequence and from those frames, sequence of gesture is obtained.
- Filters are applied on those gesture to remove noise. Apart from hand, other objects which are present in frame are removed.
- The obtained filter frame is provided to the CNN as input. Inception model is used in this.
- The input data pass through different layer of the above model and the train and test frame predictions are saved.
- The output of CNN model, the predicted data will become trained temporal feature when provided to the RNN. Which is also known as LSTM model.

#### 4.2.1.1 Processing and Frame Extraction

Every video is divided into set of sequential frames. Then while processing all background unwanted objects are removed. Different gloves are used to distinguish between hands.



Fig 4.2: Extracted Frame

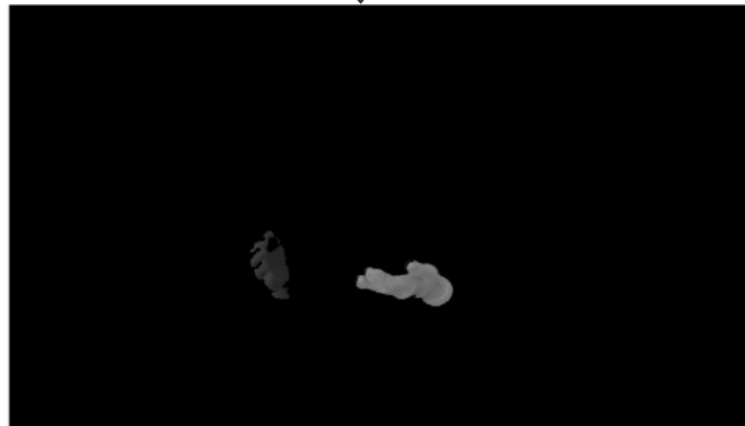
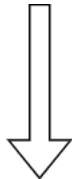


Fig 4.3: Filtered Frame (Background removal)

### 4.2.1.2 CNN Training and Prediction

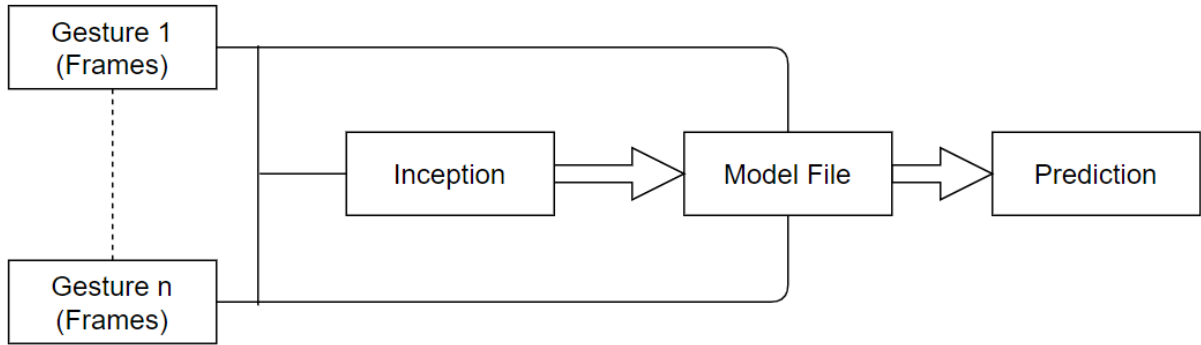


Fig 4.4: Structure of model 1

In below diagram, it is shown that initial layer having all the data, frame is selected then prediction of CNN is done. After that Predicted output transferred to the next level.

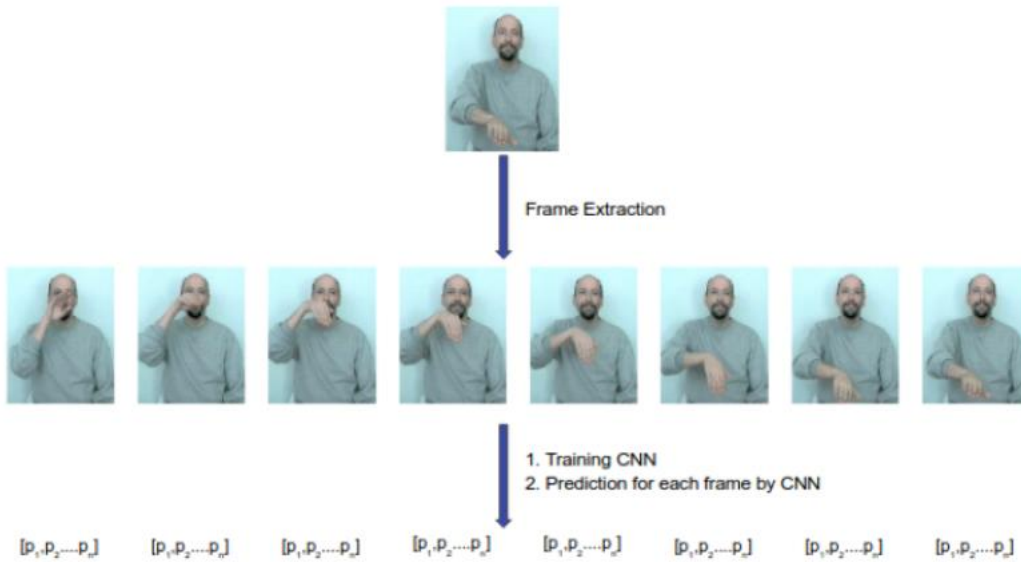


Fig 4.5: Block diagram of Frames

### 4.2.1.3 RNN Training

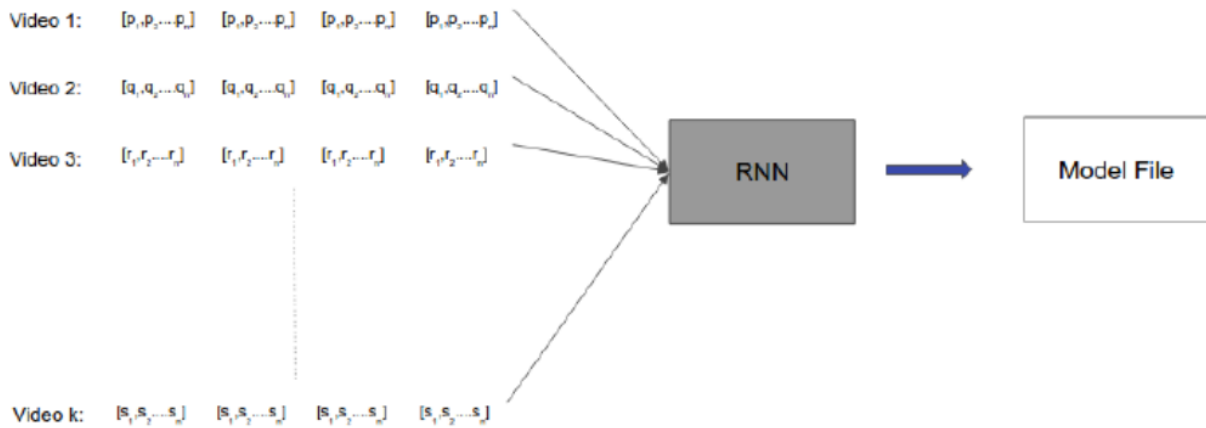


Fig 4.6: RNN Model

### 4.2.2 Limitations

The features depend upon the number of classes present in the model. If the number of classes is less and limited, then the model depends on those things only. When CNN creates predictions based on limited data then RNN model also be dependent on the limited data only. So, less features are being noticed.

### 4.3 Approach Second

Approach second contain same models only the data transferred storage and passing creates a difference of prediction and increases the efficiency. In this model the output of data before it becomes prediction, it is saved in different pool. That pool layer represents CNN layer with the dimensional vector of 2048 used for images. That doesn't treat as prediction. After this the predictions pass to the RNN model.

# CHAPTER 5

## RESULT AND ANALYSIS

---

### 5.1 Comparison

Various methods are used for Prediction and Sentence creation. By comparing we can have vivid overview of the techniques used and what procedure used by them with their efficiency on different kind of situation. Here for prediction we have used two approaches in which the techniques are same, just the process through which the inputs are provided to them are different.

#### 5.1.1 Advantage

- CNN is providing high accuracy which is nearly 93.333% when the image is still and very less movements are required while the it provides the real time independent prediction.
- RNN provides much higher accuracy rate than CNN. Its accuracy rate is nearly 95.217%. it works on real time system in much better way. The predicted outputs provided as the input to the RNN model.

#### 5.1.2 Disadvantage

- CNN's output depends on the limited classes present in it. And due the gradient the input decreases and with very a smaller number of inputs, prediction becomes more difficult.
- RNN works on the input of the CNN predicted gesture. Its works same as CNN for steady image or where the background is not changing.



## CHAPTER 6

### CONCLUSION AND FUTURE WORK

---

Communication is very important part of our life. But those who can't do, it is very important to provide the same facility by human computer interaction. Traditional method of human computer interaction contains less advance over Vision-based. Hand gesture recognition in real time is a very difficult task to achieve. Very less work has been done in this field. This dissertation is based on Isolated hand gestures provided to a vision-based system.

The database used here is LSA (Argentinian Sign Language) which contain 46 gestures classes and both the hand distinguished with the help of 2 different gloves color. In this report two different models are used in which one method follows CNN method to create the prediction of the gesture while other uses that same output of the CNN model and feed it into the RNN model. Which provide 95.217% accuracy. According to this, to achieve more accuracy, use of RNN with CNN is better than CNN alone. Thee input provided to the RNN is the only difference between the two models.

Future work contains better accuracy of achieve in the field of sign language detection so that more gestures will be recognized and sentence formation with the product will be easy. This project can be extended to the next level. Next level contains sign language into sentence formation and speaking those sentences through device.

## REFERENCES

---

- [1] Tripathi, Kumud and Neha Baranwa GC Nandi. “*Continuous Indian Sign Language Gesture Recognition and Sentence Formation.*” *Procedia Computer Science* 54 (2015).
- [2] Nandy, Anup, Jay Shankar Prasad, Soumik Mondal, Pavan Chakraborty, and Gpra Chand Nandi. “*Recognition of isolated Indian Sign Language gesture in real time.*” *Information Processing and Management* (2010).
- [3] Hochreiter, Sepp, and Jurgen Schmidhuber. “*Long short-term memory.*” *Neural computation* 9, no 8 (1997): 1735-1780
- [4] Ronchetti, Franco, Facundo Quiroga, Cesar Armando Estrebou, Laura Cristina Lanzarini, and Alejandro Rosete. “*LSA64: An Argentinian Sign Language Dataset.*” in XXII Congreso Argentino de Ciencias de la Computacion (CACIC 2016). 2016
- [5] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. “*Learning representations by back-propagation errors.*” *Cognitive modeling* 5, no. 3 (1988):1.
- [6] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen, “*Sign Language Recognition Using Convolutional Neural Networks.*” *Information Processing and Management* (2015).
- [7] Zhang, Chenyang, Xiaodong Yang, and YingLi Tian. “*Histogram of 3D facets: A characteristic descriptor for hand gesture recognition.*” In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pp. 18. IEEE, 2013.
- [8] Sirshendu Hore, Sankhadeep Chatterjee, V. Santhi, Nilanjan Dey, Amira S. Ashour, Valentina Emilia Balas and Fuqian Shi. “*Indian Sign Language Recognition Using Optimized Neural Networks.*” *2016 10th IEEE International Conference and Workshops on*, pp. 23. IEEE, 2016.
- [9] M.K. Bhuyan, Mithun Kumar kar, Debanga Raj Neo. “*Hand Pose Identification from Monocular Image for signLanguage Recognition*” *2011 IEEE International Conference on Signal and Image Processing Applications.*

- [10] Abadi, Maartin, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado et al. "*Tensorflow: Large-scale machine learning on heterogeneous distributed system.*" arXiv preprint arXiv:1603.004467 (2016).
- [11] Cooper, Helen, EngJon Ong, Nicolas Pugeault, and Richard Bowden. "*Sign language recognition using subunits.*" *Journal of Machine Learning Research* 13, no. Jul (2012): 22052231.
- [12] Kingma, Diederik, and Jimmy Ba. "*Adam: A method for stochastic optimization.*" arXiv preprint arXiv:1412.6980 (2014).
- [13] Ronchetti, Franco, Facundo Quiroga, César Armando Estrebou, and Laura Cristina Lanzarini. "*Handshape recognition for Argentinian sign language using ProbSom.*" *Journal of Computer Science & Technology* 16 (2016).
- [14] Singha, Joyeeta, and Karen Das. "*Automatic Indian Sign Language Recognition for Continuous Video Sequence.*" *ADBU Journal of Engineering Technology* 2, no. 1 (2015).
- [15] Cooper, Helen, Brian Holt, and Richard Bowden. "*Sign language recognition.*" In *Visual Analysis of Humans*, pp. 539562. Springer London, 2011.
- [16] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake, "*Real-time pose recognition in parts from single depth images*", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, page 7, (2011)
- [17] Z. Ren, J. Yuan, C. Li and W. Liu, "*Minimum near-convex decomposition for robust shape representation,*" IEEE International Conference on Computer Vision (ICCV), pp. 303-310, (2011)

