ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology
(IJARTET) Vol. 5, Special Issue 14, April 2018

REAL TIME SIGN LANGUAGE RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS

VIGNESHWAR.L
Department of Electronics and Communication,
SRM Institute of Science and Technology
M.KARTHIKEYAN
Department of Electronics and Communication,
SRM Institute of Science and Technology
HANNAH PAULINE S
Department of Electronics and Communication,
Assistant Professor,
SRM Institute of Science and Technology

ABSTRACT

Incapability to speak is considered to be true infirmity. People with this infirmity use various modes to communicate with others, there are several approaches available for their communication and one such common mode of communication is sign language. Developing a system which can act as an interpreter translator between the sign language and the spoken language dynamically and can make the communication between people with hearing impairment and normal people can be both effective and useful . Our project aims at taking the basic step in bridging the communication gap between normal people and deaf and dumb people using sign language. We hereby the implementation of a S ign Language Recognition system based on Convolutional Neural Networks. The architecture which is being proposed using the Convolutional Neural Networks identification and tracking to interpret the sign language to a voice/text format. This makes the system more effective and hence communication of the hearing and speech impaired people easy without any delay.
Keywords- Convolutional neural network, Deep Learning, Gesture recognition, S ign language Recognition.

1. INTRODUCTION

Sign language is a form of physical communication by which a deaf and dumb person can convey his/her idea through hand/arm gestures, alignment of fingers or facial expressions etc. They can resort to written communication but it is cumbersome and even impractical in certain emergency situation. In order to overcome this hurdle, we propose an ASL recognition system that uses Convolutional Neural Networks (CNN) in real time to translate a video of a user's signs into text. This will enable a dynamic communication among hearing impaired people and bridge the gap .

A sign may be considered as a method of information transmission in such a way that it is reconstructed by the receiver. It can be broadly classified as Static Signs and Dynamic Signs. The static signs are the ones that include Poses and Configurations while the latter one includes movement of body parts. Dynamic gestures comprises of strokes, postures and phases.

2. LITERATURE SURVEY

In the recent years, there has been tremendous research on the hand sign recognition.

2.1. Vision Based

In this approach, the web camera is used to capture the images of hands or fingers [5]. This approach requires only a camera, therefore recognizing a natural collaboration between humans and computers without the use of any extra devices. These methods tend to accompaniment biological vision by describing artificial vision systems that are realized in software and/or hardware. To achieve real time performance, these systems need to be background invariant, lighting insensitive, person and camera independent . M oreover to meet the requirements, including accuracy and robustness, these systems must be optimized. Vision based analysis, is based on how people recognize information about their surroundings, it is the most difficult to implement in a reasonable way . Various different methods have been tested so far.

All Rights Reserved © 2018 IJARTET

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology
(IJARTET) Vol. 5, Special Issue 14, April 2018

First is to build a 3-D model of the human hand. The model is matched to images of the hand with the help of cameras, and parameters corresponding to palm orientation and joint angles are calculated. These parameters are then used to perform gesture classification.

Second method is to capture the image using a camera then extract some feature and those features are used as input in a classification algorithm for classification.

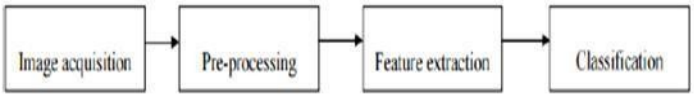

Figure2.1: Block Diagram of vision based recognition system

2.2. Sign Language Recognition Based on Histogram Of Oriented Gradient and NN [8]

2.4. Hand Gesture Recognition Using Flex Sensors

In this paper, an approach for hand gesture recognition of Indian sign language is proposed. Here the hand gesture plays a fundamental role in developing the hand gesture recognition system. Here they have proposed an approach to recognize the characters using Histograms of Oriented Gradients (HOG) features extraction method. To implement this approach, a simple web camera is used to capture the input hand gesture images. The main objective is to extract Histogram of Gradient Orientation (HOG) features from these images and to use these features for training in neural networks and thereby test it for the gesture recognition purpose.

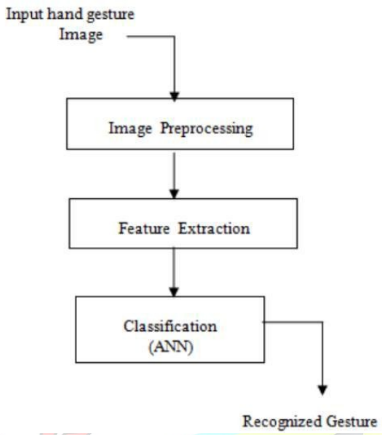

Figure2.2: Block Diagram of Hand Gesture Recognition System

In this system an electromechanical robot is designed and controlled using hand gesture in real time. Hand gesture recognition is done with the help of flex sensor which is based on the principle of resistance change. The input to the system is given by the sensors which are incorporated in the hand gloves. This system consists of two sections namely transmitter and receiver. The transmitter section will be present in the hand gloves from which the data is recognized and is being processed through PIC16F7487. The data is sent serially to the receiver section. At the receiver section the RF technology is used to transmit the data at a frequency of 2.4 GHz. The data is received in the ARM 7 (LPC2148) processor. Here from the received data, the character is predicted and matched with the closest character from which the character is identified and displayed on LCD. The various case studies is prepared for the designed system and tested in real time.

All Rights Reserved © 2018 IJARTET

2.3. Automatic Indian Sign Language Recognition for Continuous Video Sequence

The proposed system comprises of four major modules: Data Acquisition, Pre-processing, Feature Extraction and Classification. Pre-processing Stage involves Skin Filtering and histogram matching after which Eigenvector based Feature Extraction and Eigen value weighted Euclidean distance based Classification Technique was used. 24 different alphabets were considered in this paper where 96 % recognition rate was obtained.

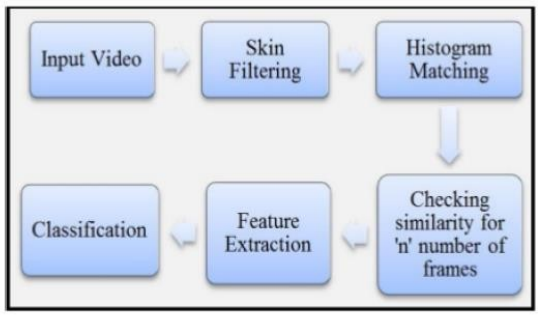

Figure2.3:System Overview of Recognition System in Video Sequences

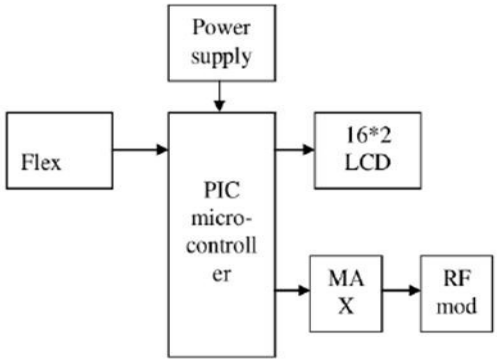

Figure2.4.1: Transmitter Section of Hand Gesture Recognition System

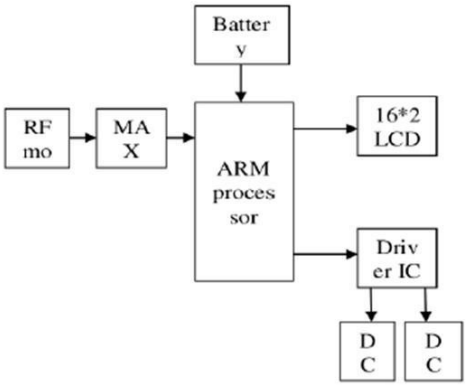

Figure2.4.2: Receiver Section of Hand Gesture Recognition System

3. APPROACH AND METHODS 3.1 Data Set Used

The data set used for this approach consists of 2524 American Sign Language (ASL) Gestures, with 70 gestures belonging to each of the 36 categories: 26 categories for English Alphabets (A-Z) and categories for Numerals (0-9). 1975 (nearly 55 images per category) images were used for training.

The system translates gestures made in sign language (ASL) into English. The signs that have translated include numbers, alphabets and few phrases. The algorithm first performs data acquisition, and then the pre-processing of the images is performed to remove the noise in the background. After that the relevant features are being extracted for training and learning process. The trained models are tested so that the gestures can be recognized as text.

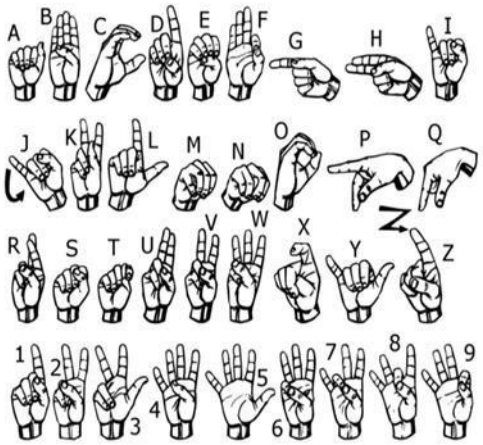

Figure 3.1: Sign Language Gestures used for dataset.

3.2 Convolutional Neural Networks

Convolutional Neural Networks (**ConvNets** or **CNNs**) are a type of Neural Networks which are made up of neurons with learnable weights and preferences. They have been established in image recognition and classification. It consists of 25 different layers whose neurons are formed three-dimensionally (width, height, depth). The special feature about CNN is that it requires less memory than other Neural Networks and also it uses reduced number of parameters which leads to a lowered training time.
There are four layers present in CNN: Convolution, Subsampling, Activation and Full Connectedness.

3.2.1 Convolution Layer

Convolution layer involves the process of convolution operation between an input and a filter to give an output .The result from the output will be passed on to the next layer. The convolution is a special operation that extracts different features from the input. The first it extracts low-level features like edges and corners. Then higher-level layers extract higher-level features. For the process of 3D convolution in CNNs. The input is of size N x N x D and is convolved with the H kernels, each of them sized to k x k x D separately .

3.2.2 S ubsampling

Subsampling helps in reducing the dimensionality of each feature but also preserves the important information. Sub sampling is also known as Spatial Pooling. The different types of spatial pooling are M ax, Average, and Sum etc. The M ax Pooling defines a spatial neighbourhood and takes the largest element from the rectified feature map within that window. The Average Pooling takes the average of all the elements in the window and Sum takes the sum of all the elements in the window. Practically, M ax Pooling has been widely used than the other pooling types because of its performance.

3.2.3 Activation

Activation Layer is a layer which decides the ultimate value of a neuron. It finds out the value with the use of an activation function. The activation function performs an element -wise procedure over the input measurements and therefore the function makes sure that the measurements of the input and the output are same.

3.2.4 Fully Connected

The term “Fully Connected” indicates that every neuron in the previous layer is connected to every neuron on the next layer. The idea of the Fully Connected layer is to use the features from the output of the convolutional layers for classifying the input image into several classes based on the training database. The Fully Connected layer is used as a classifier in Convolutional Neural Networks.

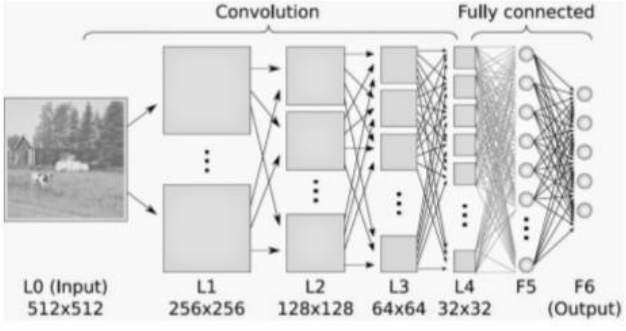

Figure 3.2: Convolutional Neural Networks.

4. IMPLEMENTATION

4.1 Image Augmentation and Resize

The images in the dataset were of a varying size and shape. Therefore the first step was to read and resize each of the images to the similar size of 256x256 pixels. Only images in the dataset, which are of the same size, can be fed into a neural network for training. To combat this challenge the dataset was augmented to produce several images from each image, thus increasing the size of the dataset and also tackling the problem of over fitting.

Maximum ranges or degrees for shear, zoom, horizontal and vertical shifting were specified. Then random values within the maximum range each of the mentioned parameters were applied on the image, thus generating new images but conserving the class or the category of the image.

Image augmentation not only helped to provide a larger dataset or prevent the classifier from over fitting, but also helped in developing a more robust classifier which could classify much more efficiently even if some changes are brought in owing to the fact that different camera modules might be used and also, there might be a shift in the position of the device.

4.2 Image Pre-processing

The mean value of RGB for all pixels was subtracted from each pixel value which serves to center the data. M ean subtraction is done in such a way that training the model involves processes like multiplying weights and adding biases to the primary inputs to produce activations which are then circulated back with the gradients to train the model. Here each feature must have a similar range, to stop the gradients from getting out of control. Also CNN's involve sharing of parameters and if the inputs are not scaled to have similar ranged values sharing will not happen easily because one part of the image will have large value of weights while the other will end with smaller values.

4.3 VGG 16

VGG 16 as depicted is a deep convolutional neural network model. The input to the ConvNets is a RGB image of size 224 × 224. The image is then passed through a stack of convolutional layers, where filters with a very small receptive field of 3 × 3 are used. The convolution stride is fixed to 1 pixel. The spatial padding of convolution layer input is 1 pixel for 3 × 3 convolutional layers thus preserving the spatial resolution. Spatial pooling has been carried out using five max-pooling layers, which follow some of the convolutional layers but not all the convolutional layers. M ax-pooling is implemented with a window size of 2 ×2 pixel and stride - size of 2. The stack of convolutional layers is followed by three fully - Connected layers: the first two layers have 4096 channels each and the third layer performs 1000-way ILSVRC classification. The final layer is the softmax layer. All hidden layers are equipped with the rectification (ReLU (Krizhevsky et al., 2012) nonlinearity .

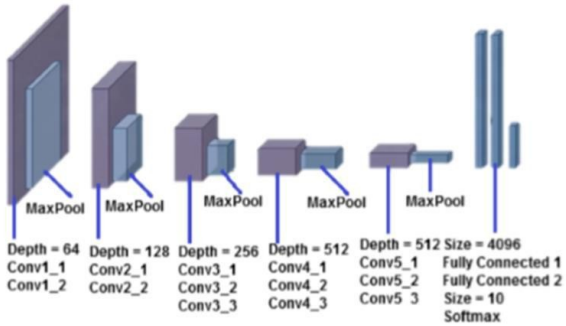

Figure 4.1: VGG 16 Architecture.

4.5 Training and Testing

Pre-trained weights which were obtained by training the VGG16 model on the ImageNet database were used to initial the weights of the model. However, the original model contained 1000 channels signifying the 1000 categories which it aimed to classify . But here only 36 classes are being targeted. Therefore, the final layer was popped and replaced with a fully -connected softmax layer with 36 channels to perform the 36 way

classifications. The remaining model was retained as it is. The training was carried out using stochastic gradient descent with momentum. Batch size of 128 and momentum of 0.9 was used. The learning rate was initialized with 0.001 and the decay rate was set to10^{-6}.It was observed that in spite of a having a very deep model, the model required very few epochs on the dataset to converge. This is due to the fact that pre-trained were used for weight initialization, which reduced the learning time by a great extent. The pre-trained were obtained on the ImageNet database, but still the model converged with very few epochs despite the fact that the project shares no category with ImageNet.

5. RESULTS AND DISCUSSIONS

The system was tested on 10 different images for each sign. The accuracy was checked as per correctness of every gesture made i.e. Alphabets and Words. The maximum accuracy is 85% for all the alphabets. This implies that the system works efficiently for most of the alphabetic character recognitions.

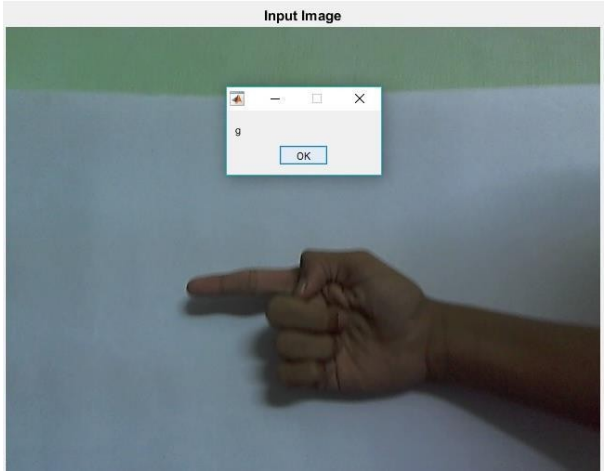

Figure 5: The recognized gesture.

6. CONCLUS IONS AND ENHANCEMENTS

This system will help to achieve high performance in recognizing the sign language, which is the main communication bridge between the deaf and dumb people and the normal people. It is hard for most of the people who is not familiar with the sign language to communicate without an interpreter. In this system, we have created an idea of translating the static image of sign language to the spoken language of hearing. The static image includes alphabet and some words, used in both training and testing of data. Feature representation will be learned by a technique known as convolutional neural networks. The new representation is expected to capture various image features and complex non-linear feature interactions.

REFERENCES

[1] S. C. W. Ong and S. Ranganath, ―Automatic sign language analysis: A survey and the future beyond lexical meaning,‖ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp . 873–891, Jun. 2005.

[2] L. Ding and A. M . M artinez, ―M odelling and recognition of the linguistic components in American sign language, *Image Vis. Comput.*, vol. 27, no. 12, pp . 1826–1844, Nov. 2009.

[3] D. Kelly, R. Delannoy, J. M c Donald, and C. M arkham, ―A framework for continuous multimodal sign language recognition,‖ in *Proc. Int. Conf. Multimodal Interfaces*, Cambridge, M A, 2009, pp . 351–358.

[4] G. Fang, W. Gao, and D. Zhao, ―Large vocabulary sign language recognition based on fuzzy decision trees,‖ *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 34, no. 3, pp . 305–314, M ay 2004.

[5] Simonyan, Karen,and Andrew Zisserman."Very deep convolutionalnetworksforlarge-scaleimage recognition." ArXivpreprint arXiv:1409.1556 (2014).

[6] www.cs231n.stanford.edu

[7] Automatic Indian Sign Language Recognition for Continuous Video Sequence Joyeeta Singha 1 , Karen Das 2 1 National Institute of Technology, Silchar, Silchar-788010Assam,INDIA. joyeeta_singha90@yahoo.com 2 School of Technology, Assam Don Bosco University Airport Road, Azara, Guwahati - 781017, Assam, INDIA karen.das[at]dbuniversity .ac.in .

[8] Neha V . Tavan, Prof. A.V . Deorankar : Indian Sign Language Recognition based on Histograms of Oriented Gradient. http://ijcsit.com/docs/Volume%205/vol5issue03/ijcsit20 140503220.pdf

[8] Hand Gesture Recognition Using Flex Sensors#1D.K.Barbole, Dr. D. V . Jadhav 1 barbole.dhanshree@gmail.com 2 dvjadhav@gmail.com #12 BSCOER, Narhe, Pune .

[10] Rajaganapathy, S., B. Aravind, B. Keerthana, and M . Sivagami. "Conversation of Sign Language to Speech with Human Gestures", Procedia Computer Science, 2015.

[11] Poppe, R.: A survey on vision-based human action recognition. Image and vision computing 28(6), 976–990 (2010).

[12] Simonyan, Karen,andAndrewZisserman."Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).