International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com

ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395

Volume 5, Issue 8 - August 2017 - Pages 78-83

American Sign Language Recognition using Convolution Neural Network

Markandeshwar Jerabandi[1]
[1]Dept. of CSE, R. T. E. Society's Rural Engineering College, Hulkoti, Gadag Karnataka, India-582205
mark.jerabandi@gmail.com

***Abstract*—Sign language is a form of communication language to connect a deaf-mute person to the world. It involves the uses of hand gestures and body movement in order to express an idea. Nevertheless, general publics are mostly not educated to comprehend the sign language. For this reason, there is a need to have a translator to facilitate the communication. This paper would like to present a Convolutional Neural Network (CNN) model for predicting American Sign Language. There are 34,800 images were captured in American sign language dataset in which each gesture consists of 1200 images to train and validate the proposed model. There were total of 29 gestures which consists of 26 Alphabets and space, delete and nothing for which 99% recognition accuracy was attained in experiment, which shows robust performance in recognition 26 static American Sign Language Alphabets and space, delete and nothing. The successful development of this model can be served as the basis to develop a more complicated sign.**

Keywords—American Sign Language; Convolution Neural Network; Deaf-Mute Person; Translator; Hand Gestures;

I. INTRODUCTION

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics. The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field. A collection of such fields overlap to cover the entire visual area. A ConvNet is able to successfully capture the Spatial and Temporal dependencies in an image through the application of relevant filters.

The architecture performs a better fitting to the image dataset due to the reduction in the number of parameters involved and reusability of weights. In other words, the network can be trained to understand the sophistication of the image better.

American Sign Language (ASL) is a complete, natural language that has the same linguistic properties as spoken languages, with grammar that differs from English. ASL is expressed by movements of the hands and face. It is the primary language of many deaf and hard of hearing, and is used by many hearing people as well. The exact beginnings of ASL are not clear, but some suggest that it arose more than 200 years ago from the intermixing of local sign languages and French Sign Language (LSF, or Langue des Signes Française).

According to the World Health Organization (WHO), the number of people having hearing or listening disability increased from 278 million in 2005 to 466 million in early 2018. It is assumed that this number will be increased to 400 million by 2050. This deaf community uses a set of signs to express their language (called sign language). In other words, a sign language (SL) is a nonverbal communication language, which utilizes visual sign patterns made with the hands or any parts of the body, used primarily by the people who have the disability of hearing and/or listening.

Among the works developed to address this problem, the majority of them have been base on basically two approaches: contact-based systems, such as sensor gloves; or vision-based systems, using only cameras. The latter is way cheaper and the boom of deep learning makes it more appealing.

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 5, Issue 8 - August 2017 - Pages 78-83

The rest of the paper is organized in the following sections. The related work is discussed in section 2. The proposed methodology for American Sign Language Recognition is presented in section 3. Section 4 includes results and its discussion. The concluding remarks and future scope is presented in section 5.

II. Related Work

The important step in project development process is literature survey where the concentration is on existing system‟s methodology to determine the reliability factor and accessibility. By using some of the components and technology of existing system, we propose a system that can be useful for the future purpose with ease of accessibility. Before developing the proposed system, it is necessary to take all literature survey for consideration to perform further future work. The following are some of the references for the existing system.

A Convolutional Neural Network (CNN) model for predicting American Sign Language is presented in [1]. There are 4800images were captured to train and validate the proposed model. 95% recognition accuracy was attained in experiment.

[2] proposed to use a CNN (Convolutional Neural Network) model named Inception to extract spatial features from the video stream for Sign Language Recognition (SLR). Then, by using a LSTM (Long Short-Term Memory), a RNN (Recurrent Neural Network) model, we can extract temporal features from the video sequences via two methods: Using the outputs from the Softmax and the Pool layer of the CNN respectively. The data set was further divided into training and test data sets, with 1800 as part of the training and the 600 as the test data set. The CNN model extracted temporal features from the frames which was used further to predict gestures based on sequence of frames. Metin Bilgin and Korhan Mutludogan [3] implemented sign language character recognition operation by using LeNet and capsule networks (CapsNet). In this study, three different experiments were made to recognize sign language characters. In [4], they proposed a novel method of multi view augmentation and inference fusion for ASL alphabet recognition from depth images using a Convolutional Neural Network (CNN). Multi view augmentation first retrieves the 3D information embedded in a depth image, and then generates more data for different perspective views. The result has shown that it outperforms the traditional image augmentation methods because it can simulate realistic perspective variations that the traditional methods cannot. [5] proposed a new user independent recognition system for American sign language alphabet using depth images captured from the low-cost Microsoft Kinect depth sensor. A real time vision-based static hand gesture recognition system for sign language, capable of identifying the position of hand and translating the gesture to the hearing in the form of text was introduced in [6]. [10] presented five distinct features such as fingertip finder, eccentricity, elongatedness, pixel segmentation and rotation are used for feature extraction. Hand Gesture Recognition (HGR) is a subfield of HCI. Today, many researchers are working on different HGR applications like game controlling, robot control, smart home system, medical services etc. The purpose of this paper is to represent a real time HGR system based on American Sign Language (ASL) recognition with greater accuracy.

From the study of above literature, it is proposed to design and develop an American sign language recognition system which would convert standard sign language to text followed by to speech, which will help deaf to understand many form of sign language and by using CNN.

III. Proposed Methodology

The primary objective of the proposed work is to establish the level of usability of the gesture recognition system. The following objectives were pursued in order to achieve the above primary objective:

- The main goal of our work is to classify 26 ASL alphabets and space, delete, nothing from static 2D images.
- To apply appropriate image pre-processing techniques in order to remove the noise and obtain the ROI.
- To design the model and architecture for CNN to train the pre-processed images and achieve the maximum possible accuracy.
- To develop an algorithm to predict the gesture in real time.
- To determine the satisfaction of the deaf participants on the use of the system.

DATASET:

Preparing enough dataset is a very prominent part in machine learning. This work would only focus on the static American Sign Language gestures which are letters A to Z as showed in Fig. 3.1. There were

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 5, Issue 8 - August 2017 - Pages 78-83

29 gestures images were collected using a smartphone camera to form the dataset. Images were captured for each gesture from two different users under the variation of background and lighting condition. To reduce the training time, all the images were resized. Some examples of the captured images were displayed in Fig. 3.2.

Fig. 3.1 Dataset of A to Z, space, delete and nothing

Fig. 3.2: Some example of captured and resized dataset. From left to right (top) A, B, C and (bottom) D, E, F

There were 80% of the dataset used as training set and 20% of the dataset were used to validate the accuracy of the sign language recognition system. The preprocessed images are shown in Fig. 3.3.

Fig. 3.3: American Sign Language Pre-processed Images

Convolution Neural Network:

The proposed architecture of CNN model is presented in Fig. 3.4.

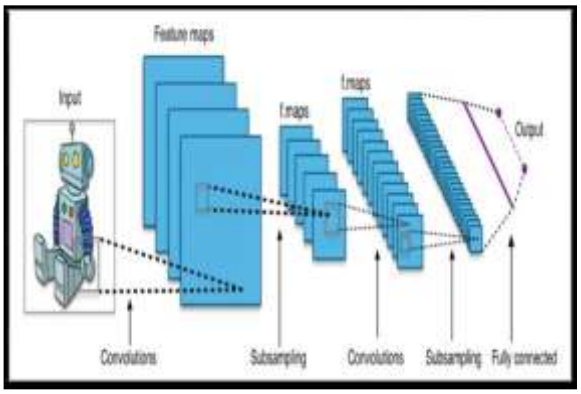

Fig. 3.4: Proposed Architecture of the Model

There are four major components needed to form a CNN model excluding the input and output layer. They are known as convolution layers, pooling/subsampling layers, flattening, and fully connected layers. The convolutional layers serve as feature extractors to learn the feature representations of their input images. It is otherwise known as filters to detect image features such as lines, edges, colors and other visual elements. There are convolutional neurons to perform convolutional operations by scanning over every pixels of the input image and make way for the result to the next layer. A

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 5, Issue 8 - August 2017 - Pages 78-83

convolutional layer can have many filters to detect different features. All the filters‟ weights will be updated in backpropagation. A complete CNN can have more than one convolutional layer to further process the extracted features to form more complex features.

The next layer in the convolutional network is generally known as max pooling layer. The main function of max pooling is to further reduce the size of images. For example, max pooling takes the largest value from one patch of an image, after that places the largest number into a new matrix, so that it discards the rest of the information that contains in activation map.

The next layer in a convolutional neural network is the flattening layer. The function of this layer is to convert all the pooled images into a continuous vector through flattening. For

example, it will convert all the two dimensional arrays into a single long continuous linear vector to serve as the inputs of the next fully connected layer. Lastly, the fully connected layer is the neural network to perform classification based on the extracted inputs from the convolution layers.

Experiments were all conducted on a laptop with Intel Core i5-7300HQ, 8GB SDRAM and a NVIDIA GeForce MX 150. The whole models were implemented using the Python programming. Dropout layer and Image Data Generator for data augmentation were utilized to avoid overfitting. A dropout ratio Of 0.5 was set. It means that one in 4 inputs will be randomly excluded from each update cycle. Dropout is the technique randomly selected neurons and setting a fraction rate to ignore during the training. It means that one of the neurons is temporally removed on the forward pass. Any updated weight are not applied to that neuron on the backward pass.

Therefore, the effect is that the network becomes less sensitive to the specific weight of the neurons and less likely to overfit the training data. The data augmentation was performed in real time on the CPU during the training phase while the model was being trained. The ImageDataGenerator for data augmentation consists of random zoom into images, randomly rotate the images in the range up to 0° - 180°, randomly shift the images horizontally according to the fraction of total width and randomly shift the images vertically according to the fraction of total height.

IV. RESULTS AND DISCUSSION

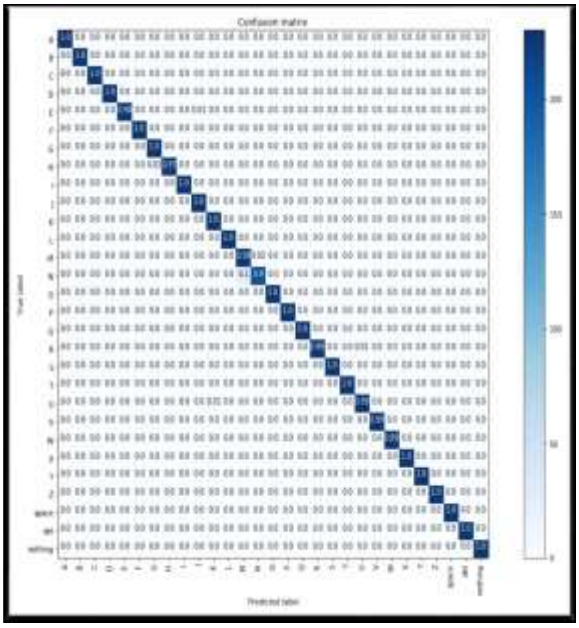

Fig. 4.1: Confusion Matrix

According to the Fig. 4.1, we can observe that the efficiency of N and H is least which is 90% and 97% compared to efficiency of other gestures. Whereas for E and M got the efficiency of 98% and for R, U, V and W got the 99% accuracy. This is due to the sign similarity and background noise.

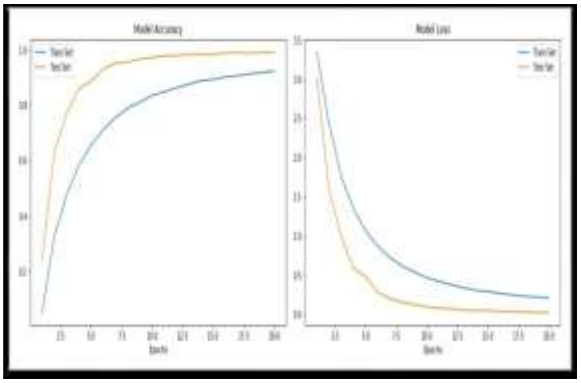

Fig. 4.2: Model Performance

According to the Fig. 4.2, the overall accuracy of the test set is 99% which is the best model obtained and we observed model loss is 2.7%.

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 5, Issue 8 - August 2017 - Pages 78-83

Fig. 4.3: Recognized Gesture as Alphabet 'C'

Fig. 4.4: Recognized Gesture as Alphabet 'A'

From the Fig. 4.3 to Fig. 4.7, we can observe that the system has successfully recognized the sign as C, A, T, Space and Delete in the American Sign Language to form a new word CAT Delete. Then finally it translates the text into speech using gTTs(Google Text TO Speech) library.

Fig. 4.5: Recognized Gesture as Alphabet 'T'

Fig. 4.6: Recognized Gesture as Alphabet 'Space'

Fig. 4.7: Recognized Gesture as Alphabet 'Delete'

V. CONCLUSION AND FUTURE SCOPE

We described a CNN architecture to recognize 29 letters in American Sign Language. The experimental results show the proposed method is effective in predicting static alphabetical gestures, which in return can be served as a beginning step to bridge the communication gap between the deaf-mute person and the community. As compared to most previous works which were using Microsoft Kinect, this work was merely using the dataset captured by widely available smartphone camera which gives more flexibility and convenience.

This work can be continued with real-time video-based sign language recognition to give more usability This can also accomodate more sophisticated sign language which involves hand movements. To minimize the environmental noise interference, video processing and recognition involves region of interest segmentation and hand

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 5, Issue 8 - August 2017 - Pages 78-83

tracking will be the next research. Besides, image occlusion is not been studied in this work due to limitation of self created database. This can be a challenging issue when part of the performing signs is not in view.

References

[1] Deafness and hearing loss, URL: https://www.who.int/news-room/factsheets/detail/deafness-and-hearing-loss, Accessed on 15 Nov 2018.

[2] Abhishek, K.S., Qubeley, L.C.F. and Ho, D. "Glove-based hand gesture recognition sign language translator using capacitive touch sensor." IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC) (pp. 334-337), August 2016.

[3] Wu, J., Sun, L. and Jafari, R. "A wearable system for recognizing American sign language in real-time using IMU and surface EMG sensors." IEEE journal of biomedical and health informatics, 20(5), pp.1281-1290, 2016.

[4] Yun, L.K., Swee, T.T., Anuar, R., Yahya, Z., Yahya, A. and Kadir,

[5] M.R.A. "Sign Language Recognition System Using SEMG and Hidden Markov Models" (Doctoral dissertation, Universiti Teknologi Malaysia), 2012.

[6] Cheok, M.J., Omar, Z. and Jaward, M.H. "A review of hand gesture and sign language recognition techniques." International Journal of Machine Learning and Cybernetics, 10(1), pp.131-153, 2019.

[7] Molchanov, P., Gupta, S., Kim, K. and Kautz, J. "Hand gesture recogni- tion with 3D convolutional neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 1-7), 2015.

[8] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J. and Chen, T. "Recent advances in convolutional neural networks." Pattern Recognition, 77, pp.354-377, 2018.

[9] Pramada, S., Saylee, D., Pranita, N., Samiksha, N. and Vaidya, M.S., "Intelligent sign language recognition using image processing," IOSR Journal of Engineering (IOSRJEN), 3(2), pp.45-51, 2013. AMERICAN SIGN LANGUAGE RECOGNITION 2021 Department of CSE, REC HULKOTI Page 30

[10] Bheda, V. and Radpour, D. "Using deep convolutional networks for gesture recognition in American sign language." arXiv preprint arXiv:1710.06836., 2017

[11] Pigou, L., Dieleman, S., Kindermans, P.J. and Schrauwen, B. "Sign language recognition using convolutional neural networks." European Conference on Computer Vision (pp. 572-578). Springer, Cham., September 2014.

[12] Garcia, B. and Viesca, S.A. "Real-time American sign language recognition with convolutional neural networks." Convolutional Neural Networks for Visual Recognition, 2016.