ISSN: 2586-7652 (Print) Vol. 04, No. 01, March 2021
ISSN: 2635-7607 (Online)
International Journal of Advanced Engineering
Source: http://ictaes.org
Manuscript received: December 26, 2020; Revised: February 2, 2021; Accepted: February 6, 2021

Sign and Speech Interchanger using Convolutional and Recurrent Neural Network

Rajbanshi.M[1], Singh.N[2], Paudel.MP[3] , Guragai.MK[4] , Kafley.S[5]

[1,2,3,4,5]*Department of Electronics and Computer Engineering, Purwanchal Campus, Institute of Engineering, Tribhuvan University, Dharan, Nepal*

mausam073bex@ioepc.edu.np[1], nitesh073bex@ioepc.edu.np[2], mayank073bex@ioepc.edu.np[3], mkguragai@gmail.com[4], sabinkafley@ioepc.edu.np[5]

Abstract

Very few people understand sign language. Obviously, this complicates communication between the deaf community and the hearing majority. The alternative of written communication is cumbersome, because the deaf community is generally less skilled in writing a spoken language. To elaborate more, the difficulty of speaking with deaf people has left them behind in the socioeconomic development of their own, society and country. The paper proposes an idea of creating a platform to understand the language the deaf speak i.e. the sign language and make them understand our language creating a friendly environment with the impaired people. Since the people with hearing disability talk with the help of gestures of hand, therefore vision based automatic sign language translator is required to extract relevant hand features from real life image sequences to allow correct and stable gesture classification. This paper is purposed to ease the communication between normal & deaf and blinds as well, fill up absence of translator and also to reach hearing impaired audience during speeches and campaigns.

Keywords: *Sign to speech; speech to sign; sign language; CNN; RNN*

1. Introduction

Based on World Health Organization statistics in 2018, over 6.1% of the world's population – 432 million people – has hearing disability. In particular, unemployment rate for deaf workers is about 75% and thus the social burden. One of the reason of that is the ordinary people are difficult to communicate with hearing impaired people. Sign language is a visual language used by deaf people as their mother tongue.

Corresponding Author : Kafley S, Guragai MK.
Author's affiliation: Dept. of Electronics and Computer Engineering, Purwanchal Campus, IoE, TU, Dharan, Nepal
Email: sabinkafley@ioepc.edu.np, mkguragai@gmail.com
Copyright © ICT-AES

Unlike acoustically conveyed sound patterns, sign language uses body language and manual communication to fluidly convey the thoughts of a person. It is achieved by simultaneously combining hand shapes, orientation and movement of the hands, arms or body, and facial expressions. It can be used by a person who has difficulties in speaking and by normal people to communicate with hearing disabled people. As far as a deaf person is concerned, having access to a sign language is very important for their social, emotional and linguistic growth [1]. It should be recognized as the first language of deaf people and their education can be proceeded bilingually in the national sign language as well as national written or spoken language. It is used by deaf people for communication by showing signs using different parts of body. Gestures can be static or dynamic [2]. Although hand postures and gestures are often considered identical, there are distinctions between them. A hand posture is defined as a static movement [3]. This paper blends with the idea to translate the sequence of hand postures to text or speech format what actually it means and the audio to its corresponding sign language. The framework provides a helping-hand for speech impaired to communicate with the rest of the world using sign language. This leads to the elimination of the middle person who generally acts as a medium of translation. This would contain a user-friendly environment for the user by providing text or further to speech output for a given sequence of hand postures input and sign output for speech or text input.

2. Related Work

A lot of attempts has been made in the past to generate a system to make the communication between the normal and speech impaired people. Our work is certainly not the first to develop application. It can be found from the literature survey that the solution for classifying gestures has evolved from 70's till the date. Mostly earlier solutions were based on Human Computer Interactions (HCI) [4]. Glove development was the first work start by the MIT in the year 1977 called Sayre glove [4] where all works were purely based in sensors and mostly come under Human Computer Interactions (HCI) system. Microsoft Kinect [4] used ASL which was later further advanced with CNN and a new model was developed in the year 2014 by Lionel Pigou [5] but experimented with Italian sign language instead of ASL. Jie Huang et al [6] have proposed Latent Space hierarchical attention network (LSHAN) which is an advancement of Long Short Term Memory (LSTM) has been used for video-based sign recognition for German sign language and CSL for classification. Most of the researchers classified gesture recognition system into mainly three steps: Extraction method, features estimation and extraction [7]. A Fuzzy C-Means (FCM) clustering method was used to classify hand gestures with a recognition accuracy of 85.83% [8].

The recurrent neural network used here is similar to the bidirectional RNN used by Hannun et al. [9], but with multiple changes to enhance its scalability. Neural network acoustic models and other connectionist approaches were first introduced to speech pipelines in the early 1990s [10, 11,12]. Fingerspelling has also been introduced into certain sign languages by educators [13].

3. Methodology

3.1 Dataset Collection

Dataset for image classification consists of 52,650 images belonging to 27 classes out of which 47,250 images are for training and 5,400 images for validation. Each class has 1750 training images and 200 validation images. Dimensions of each image is 64 * 64 pixel size. Dataset has 26 classes for English alphabets and 1 for whitespace with arbitrary hand posture.

To train largest language models, an extensive dataset consisting of 5000 hours of read speech from 9600 speakers is collected and evaluated using Hub5'00 (LDC2002S23) dataset. Table 1 shows the comparison of the available labeled datasets [14].

Table 1. A summary of the datasets used to train Deep Speech. The Wall Street Journal, Switchboard and Fisher corpora are all published by the Linguistic Data Consortium

Datasets	Type	Hours	Speakers
WSJ	read	80	280
Switchboard	conversation	300	4000.
Fisher	conversation	2000	23000
Baidu	Read	5000	9600

3.2 Convolution Neural Network (CNN) & Recurrent Neural Network (RNN)

A convolution neural network is the convectional ANN that is very effective and efficient in performing image recognition and classification tasks. CNN resembles the human way of visual recognition. It receives pixels as image inputs, transforms them using mathematical computations and activation function then builds a classifier for the input classes. CNN has an encode-specific feature in its architecture that suits it to perform image processing and computer vision task [15]. CNN is popularly used to train machines to recognize and classify images.

CNN architecture consists of three essential layers namely: convolutional layer, the pooling layer and the fully connected layer. Our CNN architecture has three convolutional layers each followed by pooling layers which is then flattened and is further extended with fully connected layers to output layers.

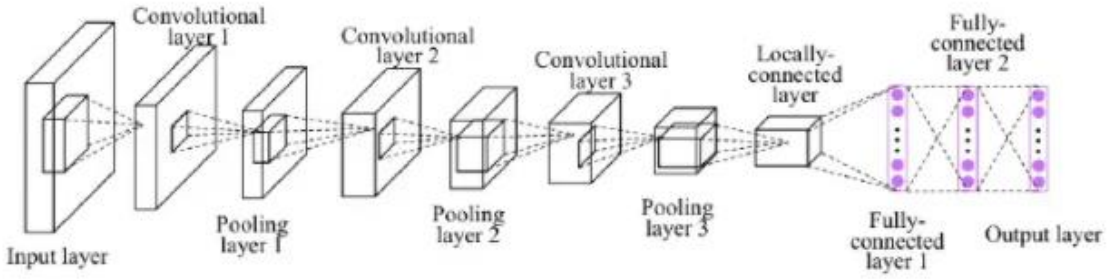


Figure 1. CNN Architecture

Recurrent Neural Networks are a form of neural networks which have their own output fed to their own input. This practice of using output as input creates loop in the Networks and helps in preserving some memory. Previous input or output has the very good weightage in predicting the output. Dropout is done to reduce the loss in the neural network [16, 17,18]. The recurrent neural network was trained to ingest speech spectrograms and generate English text transcriptions [14]. The pre-trained RNN model is composed of 5 layers of hidden units out of which fourth layer is bidirectional recurrent layer and remaining layers are non-recurrent [14, 16, 19].

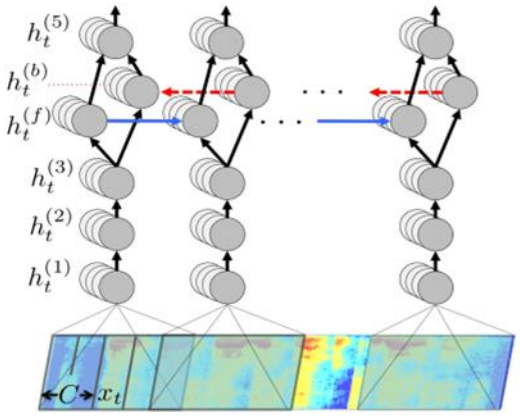


Figure 2. Structure of RNN model used.

3.3 Image Preprocessing

Images for each alphabet were generated using webcam during dataset preparation. Each of the images were preprocessed to some extent during creation resulting to speed up training. Preprocessing includes resizing of image to desired dimension, converting to HSV color format, masking of image so as to eliminate background & to focus on subject and complex calculation during training. Above preprocessing tasks are also carried out in the image to be classified before feeding to CNN model for prediction.

3.4 System Solution

The design results in producing speech from the sequence of hand postures of ASL and is also true for the reverse. In producing speech from sequence of hand postures, each video frame is preprocessed and fed as input to deep learning CNN model [5] to classify which English alphabet it corresponds iteratively. These alphabets are appended sequentially on user command to form desired text which is further generated to speech using SAPI5 backend engine.

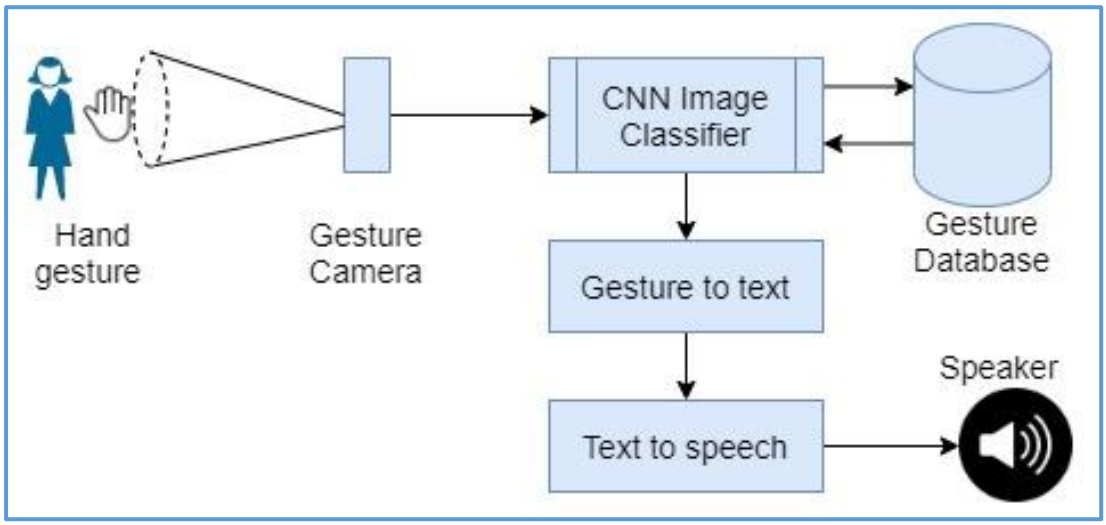


Figure 3. Sign to speech conversion.

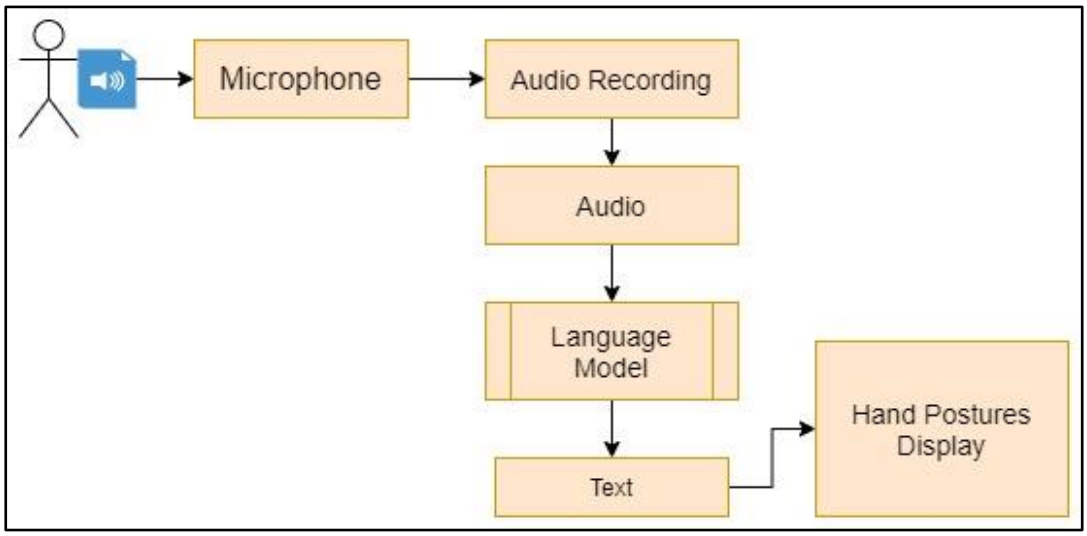


Figure 4. Speech to sign conversion.

On other hand, in inducing finger-spelled hand postures from speech, recorded speech is fed to language model trained using RNN algorithm which predicts the text form of audio. Finger- spelling algorithm aids in switching text so induced from prediction into sign language representation. This sub-system uses pre-trained language model trained using homogenous RNN algorithm [20] and Nesterov's Accelerated gradient method for training [21]. The complete structure of RNN model is illustrated in figure 2. The structure is considerably simple as limited to single recurrent layer and no LSTM are included in the circuit [20].

3.5 Frontend Software Design

The Graphical User Interface is designed using Qt-designer later converted in a python file. Layout is provided with two options: sign to speech and speech to sign conversion. In sign-speech conversion process, image capturing event is carried out with user control. Image so captured generates backend method call. These operations are executed in incremental iterations. In speech-sign conversion process recording can be initiated clicking respective button. On completion, backend method call comes in action for further processing.

3.6 Backend Software Design

Image captured is fed to sequential Convolution Neural Network model which predicts the corresponding label and gets stored as first character. The following images in upcoming iterations are predicted returning respective labels which gets stored as second, third character and so on. Once desired word/ phrase/ sentence is formed, speech is generated with sapi5 backend engine facilitated with python API.

On the other hand, recorded speech is fed to language model trained with Recurrent Neural Network (RNN) algorithm which predicts the corresponding text of speech which then further calls a method based on finger-spelling algorithm that organizes the displaying of hand postures.

3.7 System Design

The entire system is presented in the flowchart below. The alphabet gets predicted by the CNN model. User confirms the alphabet to be used in building the desired text repeatedly until desired text is formulated. Then, external command is used to generate speech of text formulated.

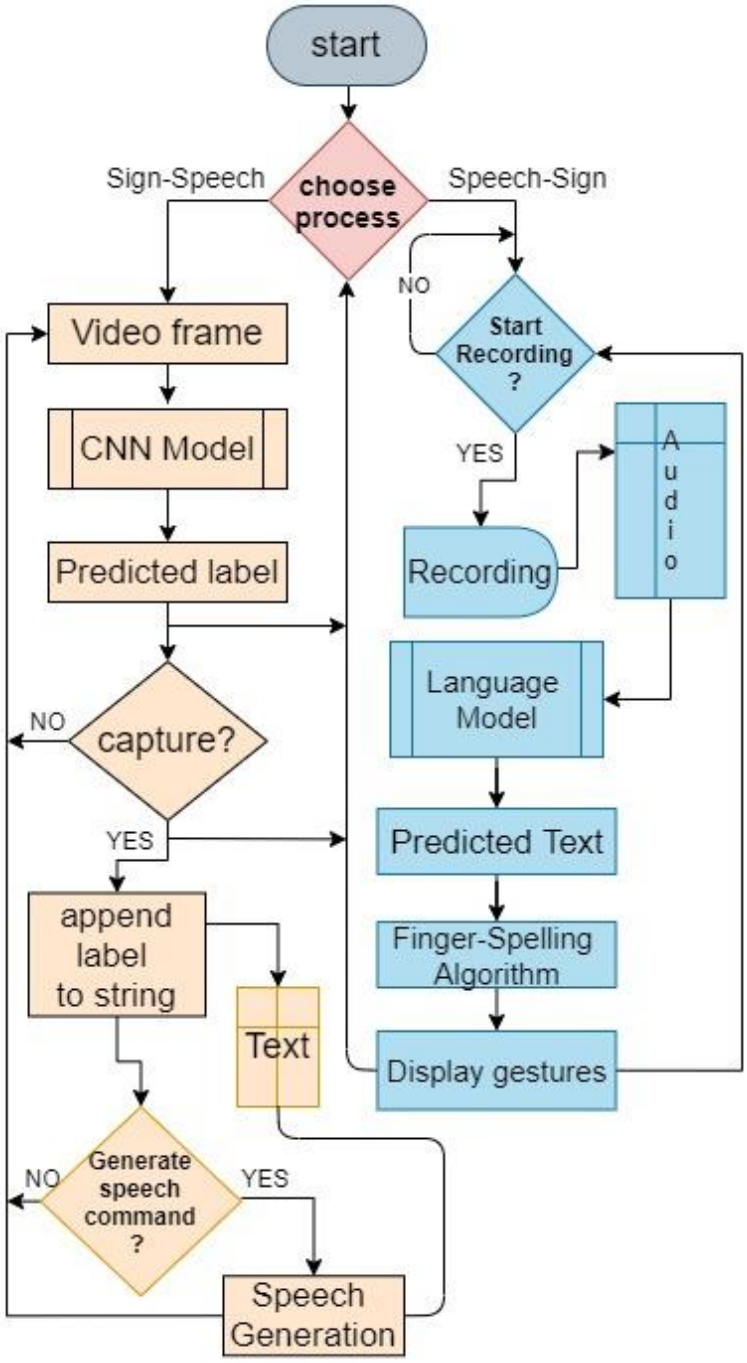


Figure 5. Flowchart of the system.

3.8 The Fingerspelling Algorithm

Fingerspelling is the process of spelling out words by using signs that correspond to the letters of the word. In many ways finger spelling serves as a bridge between the sign language and the oral language that surrounds it. It may be used to represent words from an oral language which have no sign equivalent, or for emphasis, clarification, or when teaching or learning a sign language. The typical "these things are spelled" list includes such items as people name, places, titles, brands, etc. [13]. The fingerspelling algorithm proceeds as [13]

1) Start
2) Initialization
    a) Initialize data dictionary containing each letter of ASL image.

   b) Initialize input text file that contains string.
3) Read the input written text file.
4) Preprocess the text file.
   a) Split the text into sentences, on occurrence of full stop ".".
   b) Split the sentences into words.
5) For each letter in the word do
   a) Convert to the uppercase.
   b) Search in dictionary to find ASL image corresponding to the letter.
6) Displaying:
   a) Concatenate all images to be saved on picture file such that each row is word in the text.
   b) Display images.
7) End

4. Results

The major goal of this paper was to build a system that translates sign to speech and vice versa but not both simultaneously. Two deep learning models: CNN & RNN were used.

The image classifier was trained using CNN algorithms on 47,250 image data and 5,400 validation image data to produce a baseline model with training accuracy of about 66% and validation accuracy of about 59%. On applying image augmentation, accuracy of training and validation raised to about 72% and 81% respectively. Adding dropout gave training & validation accuracy of 87.67% & 96.65% respectively. Training took about 8+ hours in laptop just with 4GB RAM for 20 epochs. The summary of the CNN classifier is shown in Figure 6.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 62, 62, 32)        896
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 31, 31, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 29, 29, 32)        9248
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 14, 14, 32)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 12, 12, 64)        18496
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 6, 6, 64)          0
_________________________________________________________________
flatten_1 (Flatten)          (None, 2304)              0
_________________________________________________________________
dense_1 (Dense)              (None, 256)               590080
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 27)                6939
=================================================================
Total params: 625,659
Trainable params: 625,659
Non-trainable params: 0
```

Figure 6. CNN classifier summary.

RNN algorithm for generating Language model was trained on the 300 hours Switchboard conversational telephone speech dataset and Fisher (FSH), a 2000 hours corpus collected in a similar manner as Switchboard total of 2300 hours of data and tested using Hub5'00 dataset by Deep Speech. Using optimizations techniques as data parallelism, model parallelism and striding the system was able train the RNN algorithm in few hours in power computer facilitated with GPU [14]. RNN was trained on all the datasets listed in table 2 for more than 7000 hours with newly synthesized noise in each pass to handle the noisy data [14].

For sign recognition, adequate amount of day light should be available. Any background with the color similar to skin may affect the classification process. High quality audio input is preferred for speech recognition. There is significant delay in text prediction from speech. Recording the audio with American ascent also helps to increase the accuracy in prediction.

Table 2. Overview of both models

Models	Algorithms	Optimizing Techniques	Validation Accuracy	Validation Loss
Image Classifier	CNN	Augmentation, Dropout, Keras Adam	96.65%	10.88%
Language model	RNN	Striding, Data and Model Parallelism	75%	-

The CNN model accuracy loss graphs are depicted in Figure 7 and 8.

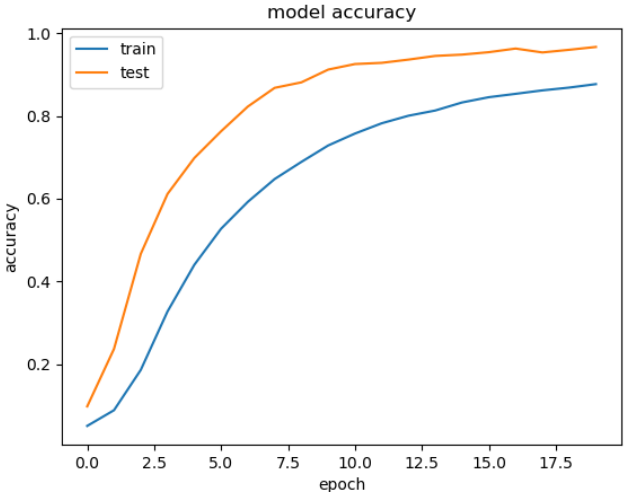


Figure 7. CNN model accuracy graph.

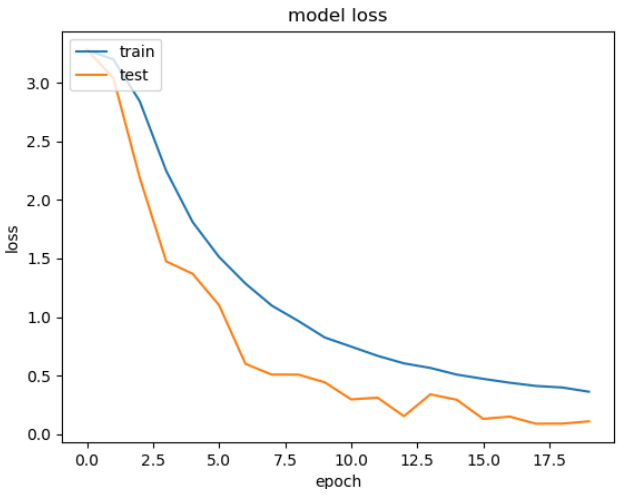


Figure 8. CNN model loss graph.

5. Conclusion

The proposed paper can be useful in bringing comfort to understand the language of speech impaired people and takes lesser time to understand the hand postures in real world application. Prior to this system many systems were developed as individual sign to speech and speech to sign translator but not both as integrated. Our paper integrates both of the functionality. Google has developed speech recognition to convert into text but our paper further extends in generation of sequence of signs (hand postures) corresponding to each of the alphabets. Entire system consists of two sub-systems that don't have full duplex operation so future tasks can be making it full duplex system, compensating lighting condition, deploying in web application and further increasing prediction accuracy. Since language model we used was pre-trained, we can train the model of own.

References

[1] Destreza, Froilan. (2012). SIGN LANGUAGE TO VOICE TRANSLATOR. IJACCT.

[2] S. Mitra, and T. Acharya. (2007). "Gesture Recognition: A Survey" IEEE Transactions on systems, Man and Cybernetics, Part C: Applications and reviews, vol. 37 (3), pp. 311-324, doi: 10.1109/TSMCC.2007.893280

[3] Joseph J. LaViola Jr., (1999). "A Survey of Hand Posture and Gesture Recognition Techniques and Technology", Master Thesis, Science and Technology Center for Computer Graphics and Scientific Visualization, USA.

[4] Goswami, Tilottama & Javaji, Shashidhar. (2020). CNN Model for American Sign Language Recognition. 10.1007/978-981-15-7961-5_6.

[5] Pigou L, Dieleman S, Kindermans P-J, Schrauwen B Sign language recognition using convolutional neural network In: European Conference on Computer Vision (ECCV 2014) Workshops, Springer, pp 572-578.

[6] Huang J, Zhou W, Zhang Q, Li H, Li W (2018) Video-based sign language recognition without temporal segmentation. In: 32nd AAAI Conference on Artificial Intelligence (AAAI-2018) New Orleans, Lousiana, USA, 2-7 February 2018. ArXiv:1801.10111.

[7] Khan, Rafiqul Zaman & Ibraheem, Noor. (2012). Hand Gesture Recognition: A Literature Review. International Journal of Artificial Intelligence & Applications (IJAIA). 3. 161174. 10.5121/ijaia.2012.3412.

[8] Xingyan Li. (2003). "Gesture Recognition Based on Fuzzy C-Means Clustering Algorithm", Department of Computer Science. The University of Tennessee Knoxville.

[9] A. Y. Hannun, A. L. Maas, D. Jurafsky, and A. Y. Ng. First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs. abs/1408.2873, 2014.

[10] H. Bourlard and N. Morgan. Connectionist Speech Recognition: A Hybrid Approach. Kluwer Academic Publishers, Norwell, MA, 1993.

[11] D. Ellis and N. Morgan. Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition. In ICASSP, volume 2, pages 1013–1016. IEEE, 1999.

[12] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco. Connectionist probability estimators in HMM speech recognition. IEEE Transactions on Speech and Audio Processing, 2(1):161–174, 1994.

[13] Karim, Alia. (2015). A Proposed Algorithm for Translating English Written Text to Fingerspell Language. Eng. & Tech.Journal. 33Part(B).

[14] Hannun, A.Y., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. (2014). Deep Speech: Scaling up end-to-end speech recognition. *ArXiv, abs/1412.5567*.

[15] O'Shea, K., & Nash, R. (2015). An introduction to convolution neural network.arXiv preprint arXiv:1511.08458.

[16] Liu, Pengfei, XipengQiu, and Xuanjing Huang. "Recurrent neural network for text classification with multi-task learning."arXiv:1605.05101 2016.

[17] Lai, Siwei, et al. "RecurrentConvolutional Neural Networks for Text Classification." AAAI. Vol. 333. 2015.

[18] Mittal, Nikita, and AkashSaxena. "Layer Recurrent Neural Network Based Power System Load Forecasting." Indonesian Journal of Electrical Engineering and Computer Science 16.3 (2015): 423- 430.

[19] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11):2673–2681, 1997.

[20] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of momentum and initialization in deep learning. In 30th International Conference on Machine Learning, 2013.

[21] Varshini, Chava & Hruday, Gurram & Chandu, Guguloth & Sharif, Shaik. (2020). Sign Language Recognition. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS050781.