ISBN 978-9937-0-4872-9
September 27, 2018

NEPALI SIGN LANGUAGE TRANSLATION USING CONVOLUTIONAL NEURAL NETWORK

Drish Mali
Department of Computer and Electronics Engineering
Kantipur Engineering College, T.U.
Lalitpur, Nepal
drish01771@gmail.com

Rubash Mali
Department of Computer and Electronics Engineering
Kantipur Engineering College, T.U.
Lalitpur, Nepal
rubash9849@gmail.com

Sushila Sipai
Department of Computer and Electronics Engineering
Kantipur Engineering College, T.U.
Lalitpur, Nepal
sipaisushila@gmail.com

Sanjeeb Prasad Pandey (PhD)
Department of Computer and Electronics Engineering
Pulchowk Campus, Institute of Engineering, T.U.
Lalitpur, Nepal
sanjeeb@ioe.edu.np

***Abstract*—Sign language is a basic mode of communication between people who have difficulties in speech and hearing. If computers can detect and distinguish these signs, communication would be easier and dependency on a translator reduces. This paper provides the structure of the system which translates Nepali signs from Nepali Sign Language (NSL) into their respective meaningful words. It captures the static hand gestures and translates the pictures into their corresponding meanings using 2-D Convolutional Neural Network. Red glove is used for segmentation purpose. Data set is obtained by manually capturing images using a front camera of a laptop. The system got higher accuracy for the model that recognizes 5 signs than the 7 signs model. It also facilitates the users to search for the signs using their corresponding English words. Its core objective is to make easy communication between differently-abled people and who do not understand sign language without involvement of a translator.**

***Index Terms*—NSL, translate, 2-D Convolutional Neural Network, differently-abled, easy communication**

I. Introduction

Sign language is a basic mode of communication between people who have difficulty in speech and hearing. However, normally, people do not have much idea about sign language and find it hard to communicate with sign language users. There is still a big communication gap which can cause confusion and demands for a translator to fill this gap.

In Nepal, the prevalence of deafness is about 2.8% which accounts to about 7 lakhs population. About 10% of them are unaware of their problem because of mild impairment and 7% of them are suffering with disabling impairment. There are approximately around two hundred sign languages in use around the world today. And several efforts have been made to translate these languages into English language and other languages. There are many apps available for American Sign Language(ASL). However, there has not been significant progress in case of Nepali Sign Language(NSL).

So, this system is an initiation to translate NSL to English words and help to solve the problem of communication gap that people are facing now. This system is used to translate sign language into its meaning. The user clicks picture of the hand signs and then the app processes it and its meaning is presented on the screen as a text. This app also shows the image of the signs with their meaning and provides the feature of searching the required sign by its English meaning.

II. Related Works

Most of the recent works related to hand gesture interface technique has been categorized as: glove-based method and vision-based method. Glove-based method requires a user to wear a cumbersome device, and generally carry a load of cables that connect the device to a computer. Vision-based method consists of various techniques such as: Model-based and State-based.

Huang et al. [1] used 3-D neural network method to develop a Taiwanese Sign Language (TSL) recognition system to recognize 15 different gestures. David and Shah [2] proposed a model-based approach by using a finite state machine to model four qualitatively distinct phases of a generic gesture. Hand shapes were described by a list of vectors and then matched with the stored vector models. Starner et al. [3] described an extensible system which used one color camera to track hands in real time and interpreted American sign language (ASL). They used hidden Markov models (HMMs) to recognize a full sentence and demonstrate the feasibility of recognizing a series of complicated series of gesture.

Feng-Sheng Chen, Chih-Ming Fu, Chung-Lin Huang [4] developed a hand gesture recognition using a real-time tracking method and hidden Markov models where they introduced a system to recognize continuous gesture before stationary background. The system consisted of four modules: a real time hand tracking and extraction, feature extraction, HMMs training, and gesture recognition. They first applied a real-time hand tracking and extraction algorithm to trace the

moving hand and extract the hand region, then used the Fourier descriptor (FD) to characterize spatial features and the motion analysis to characterize the temporal features. They then combined the spatial and temporal features of the input image sequence as their feature vector. After having extracted the feature vectors, they applied HMMs to recognize the input gesture. The gesture to be recognized was separately scored against different HMMs. The model with the highest score indicated the corresponding gesture. In the experiments, they tested the system to recognize 20 different gestures.

There are two common deep neural network architectures:the Convolutional Neural Network (CNN) and the Recurrent Neural Network (RNN). CNN is a special form of the Feedforward Neural Network (FNN), also known as the Multi-layer perceptron (MLP), trained with back-propagation. CNN are used to recognize visual patterns directly from pixel images with variability. There has been a large amount of recent efforts devoted to the understanding of CNNs. Examples include scattering networks, tensor analysis, generative modeling, relevance propagation and Taylor decomposition etc. Divya Deora and Nikesh Bajaj [5] proposed Indian Sign Language Recognition system using red masking technique and PCA for predciting 25 alphabets and 9 numbers .

This paper presents the idea of recognizing the NSL hand gestures. The authors have used red glove for easier segmentation and 2D CNN for recognition.

III. Data Acquisition

Data is acquired through a 1 Mega pixel camera. The image data set consists of NSL gestures referenced from Nepali Sign Language Dictionary and Nepali Sign Language Alphabet provided by National Deaf Federation - Nepal (NDF-N). A total of 7 static signs were selected. The data set consists of 400 red masked images(300 training and 100 test set) for each sign. These images were captured by 3 people under different lighting conditions. Each image in the training set were re-sized into 64*64 pixel image and was further augmented using the operation shear, re-scale, zoom, horizontal flip. Thus, creating 1500 images(300*5) for training set of each sign.

IV. Methodology

The methods used for the accomplishment of the recognition of signs of NSL can be divided into 3 steps: image segmentation , developing a CNN model and prediction.

A. Image Segmentation

Initially, an image was captured using the camera of a laptop. Image was modeled using the HSV (hue, saturation, value)color model to obtain red mask of the image as shown in figure 1. For all shades of red, Hue ranges from 160-180 while Saturation and Value range from 0 to 255.

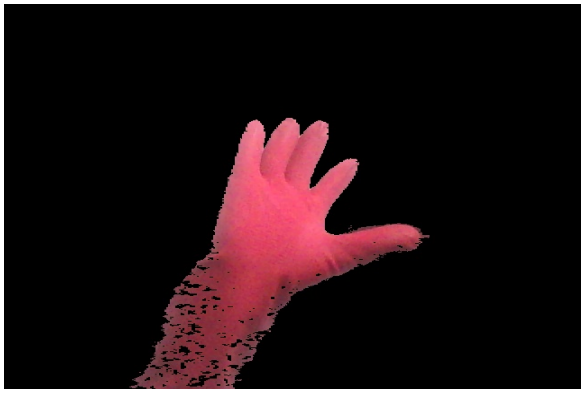

Fig. 1. Image Segmentation

B. The Convolutional Neural Network Model

Convolutional Neural Network (CNN, or ConvNet) is a class of deep, feed-forward articial neural networks, most commonly applied to analyze visual imagery. CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing. Convolutional Neural Network Model was proposed by LeCun et al. [6], and has proved to be a significant milestone in the area of detection and classification of images.

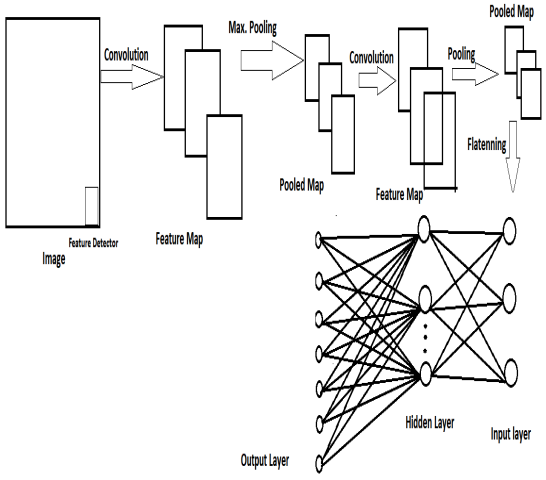

Fig. 2. Convolutional Neural Network Architecture

1) Proposed architecture: The architecture of model consists of two convolutional layer, pooling layer and one Artificial Neural Network (ANN). Convolutional layer, pooling layer are used to extract features and ANN is used to classify the features into classes. The architecture is depicted in figure 2.

The input image is resized into 64*64 size to reduce calculation complexity and feed to the input of convolution layer. In convolution layers, 32 feature map is sampled of the input to generate feature map using convolution operation which is further sampled by pooling matrix using max pooling technique. The convolution and max pooling operation are

further repeated once. The output is then attened into 1D vector and supplied to the ANN as an input. Then, the ANN uses the input and the weight which are generated from training and the 128 neurons in the fully connected (hidden) layer for the image classication into one of the class.

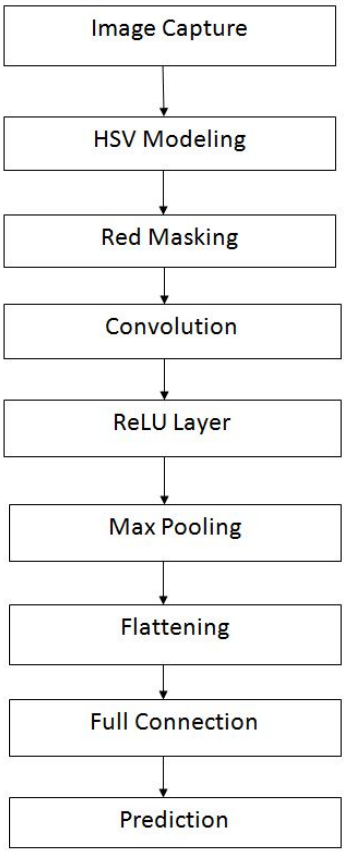

Fig. 3. The CNN model flow

2) CNN model flow: Initially the image is captured using the front camera of a laptop and the red mask of the image is produced. Then the image is re-sized to 64*64 size and supplied to the input of CNN. In the convolution layer convolution operation is performed between the input image and 32 feature detector of size 3*3 to produce 32 feature map. The output of the convolution network is further sampled using 2*2 pooling matrix with the max pooling technique to reduce the feature map size by 75% but still preserve the important features and provides the spatial variance. Rectifier Linear Unit (RELU) activation function is used in the convolutional layer and pooling layer to remove negative pixel values in the feature maps .In the context of artificial neural networks, the rectifier is an activation function defined as the positive part of its argument:

$$f(x) = x^+ = \max(0, x) \tag{1}$$

where x is the input to a neuron. This is also known as a ramp function. The pooled feature map were converted into 1-D vector by flattening and then supplied to the input of the neural network.128 hidden nodes were used in fully connected layer(hidden layer) which used RELU function as an activation function.7 nodes were used in the output layer for prediction of an image and used softmax function as an activation function and categorical cross entropy as loss function due to the multiclass output nature of the model. The softmax function takes all output neurons weight as input and converts them into probability. The output neuron with highest probability is the predicted class. The softmax function for any output neuron with weight x_j is expressed mathematically as:

$$\sigma(\mathbf{x}_j) = \frac{e^{x_j}}{\sum_i e^{x_i}} \tag{2}$$

where i is total number of output neurons.

3) Prediction: The class with the highest probability is predicted as the class of the input image by the model.

The snapshots of the result obtained are shown in figure 4 5 and 6.

Fig. 4. Correct prediction for Camera symbol

V. Result and Discussion

In this project, a Dell laptop with the specifications of 6 GB memory(RAM), 2.4 GHz i7 processor and another HP laptop of 8 GB memory and 2.4 GHz i5 processor were used to train and test the model. The number of epochs used for training the model was 100. Mini Batch Gradient method with batch size 3 was used for adjusting the weight of CNN model. The CNN model was compiled using Adam Algorithm for gradient descent process and categorical cross entropy as loss function. It took 22 minutes for the Dell laptop to train the model and 20 minutes for the HP laptop to train the model. The accuracy of this model was 95.4% on the data set of 500 images (100 of each class. The confusion matrix of the the 5 signs model is shown in figure 7.

Again, the CNN model was trained for predicting for 7 signs classes with 1500 images of each class after augmentation.

Fig. 5. Correct prediction for House symbol

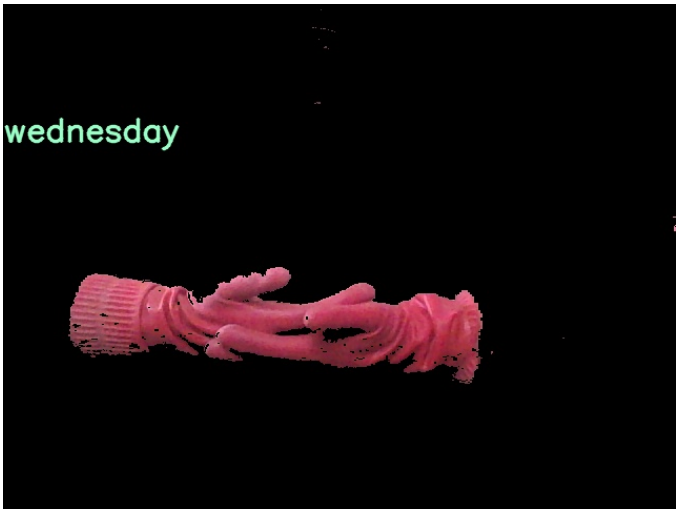

Fig. 6. Correct prediction for wednesday symbol

Predicted /Actual	Affirmative	House	Skin	Wa	Wednesday
Affirmative	98	-	1	-	1
House	-	96	3	-	1
Skin	-	1	98	-	1
Wa	3	2	-	95	-
Wednesday	2	-	4	4	90

Fig. 7. A confusion matrix for 5 signs model

Predicted /Actual	Affirmative	Binoculars	Camera	House	Skin	Wa	Wednesday
Affirmative	93	-	3	-	-	1	3
Binoculars	-	94	3	1	1	-	1
Camera	-	-	100	-	-	-	-
House	-	1	2	93	2	-	2
Skin	-	4	-	1	93	-	2
Wa	2	-	1	1	-	96	-
Wednesday	-	2	-	-	2	3	93

Fig. 8. A confusion matrix for 7 signs model

The accuracy of this model was 94.57% on the data set of 700 images (100 of each class).The confusion matrix of the 7 signs model is shown in figure 8.

VI. Conclusion

The proposed system has proven to demonstrate considerable result for the recognition of signs of Nepali Sign language.The accuracy of the model is high due use of red gloves for detecting hand of the signer. Even with the addition of two more gestures, it was observed that the accuracy of the model decreased by 0.83%. More new signs can be added to the model by adding the gestures in the data set and increasing the number of prediction classes. However, the number of training data should also be increased accordingly for the acceptable accuracy of the model.

Further enhancement to this work can be done by using Neural Network to detect the hand directly rather than using red gloves. Video processing can be done for motion based gesture recognition.

VII. Acknowledgment

The authors acknowledge National Deaf Federation - Nepal (NDF-N) for their kind co-operation and support.

References

[1] Chung-Lin Huang and Wen-Yi Huang, 'Sign language recognition using model-based tracking and a 3D Hopfield neural network', Machine Vision and Applications, vol. 10,pages 292-307,1998.

[2] J. Davis and M. Shah, 'Visual gesture recognition', IEE Proceedings on Vision, Image and Signal Processing, 141(2): pages 101106, 1994.

[3] T. Starner and A. Pentland, 'Real-time american sign language recognition from video using hidden markov models', In SCV95, pages 265270, 1995.

[4] Feng-Sheng Chen, Chih-Ming Fu, Chung-Lin Huang, Hand gesture recognition using a real-time tracking method and hidden Markov models , Institute of Electrical Engineering, National TsingHua University, Hsin Chu 300, Taiwan, March 2003.
[5] D.Deora and N.Bajaj, 'Indian sign language recognition', 1st International Conference on Emerging Technology Trends in Electronics, Communication and Networking (ET2ECN), pages 1-5, 2012.
[6] LeCun, Yann, Lon Bottou, Yoshua Bengio, and Patrick Haffner, 'Gradient-based learning applied to document recognition', Proceedings of the IEEE 86, vol. 11, pages 22782324, 1998.