Indian Sign Language Numeral Recognition Using Region of Interest Convolutional Neural Network

Conference Paper · April 2018

DOI: 10.1109/ICICCT.2018.8473141

2 authors:

Sajanraj T D
Jyothi Engineering College
5 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Beena M V
Vidya Academy of Science & Technology
4 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018)
IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

Indian Sign Language Numeral Recognition Using Region of Interest Convolutional Neural Network

Sajanraj T D
Dept. of Computer Science & Engineering
Vidya Academy of Science and Technology
Thrissur - 680501, Kerala, India
sajanrajtd.wordpress.com
sajanraj.t.d@gmail.com

Beena M V
Assistant Professor
Dept. of Computer Science & Engineering
Vidya Academy of Science and Technology
Thrissur - 680501, Kerala, India
beena.m.v@vidyaacademy.ac.in

***Abstract*: - Communication provide interaction among the people to exchange the feelings and ideas. The deaf community suffer a lot to interact with the community. Sign language is the way through which the people communicate with each other. In order to provide interaction with normal people there is a system which can convert the sign languages to the understandable form. The purpose of this work is to provide a real-time system which can convert Indian Sign Language (ISL) to the text. Most of the work based on handcrafted feature. In this we are introducing a deep learning approach which can classify the sign using the convolutional neural network. In the first phase we make a classifier model using the numeral signs using the Keras implementation of convolutional neural network using python. In phase two another real-time system which used skin segmentation to find the Region of Interest in the frame which shows the bounding box. The segmented region is feed to the classifier model to predict the sign. The system has attained an accuracy of 99.56% for the same subject and 97.26% in the low light condition. The classifier found to be improving with different background and the angle of the image captured .Our method focus on the RGB camera system.**

Keywords: - Deep learning, Convolutional neural network, Region of interest, Real-time system.

I. INTRODUCTION

In the world different languages are used among the people to provide communication, while talking about physically impaired people both deaf and dump community also use different sign languages. The different languages are American Sign Language, Chinese sign language, Indian sign language etc. In each case symbols are vary with the involvement of motion, single handed and double handed representations. The static symbols are used to represent letters and some cases dynamic symbols are used for the words like “hello”,”Hai”,etc. A real time system among these community will allows the communication barrier among them. Once it has been converted to using the Computer Vison approach then it can be converted to any language. There are many researches undergone by in this are to build an efficient and accurate system. The previous works done by the researchers using the hand crafted feature but having limitation and used special conditions.

The most works are based on the pattern recognition, feature extraction based on HOG, SIFT, LBP, etc. But the system using a single feature is not sufficient in most of the cases and the Hybrid approach are introduced to solve this problem. But for a real time system we need faster methods to solve our problems. Now a days ours computers are improved with the speed of processing using parallel implementation. In most of the time our system utilize a single core for solving a problem. Using the GPU system the problems can be solved by parallel computing and the number of cores is higher than the CPU system. Using the Deep Learning approach we can model a self-learning system for our needs. Convolutional neural network is one of the trending deep learning system which is capable of solving any kind of computer vision problem. In our method we used a Region of interest -Convolutional neural network for the real time implementation.

II. LITERATURE SURVEY

Computer Vison has become one of the trending technology used in most of the AI based system such as Robots, Cars, Markets, etc. The system has more impact on image classification problems and object detection. The sign language system can be implanted using this method. There are many other methods used in the prior systems.

In literature [1] used a framework for ISL Recognition system. A glove based color Segmentation is used and the

Recognition is done using PCA (Principal Component analysis). The real-time data frames in every 20th frame is taken as input to be recognized. The sign with both overlapping and motion crated problems in this method. Fingertip algorithm along with PCA is used for recognition purpose. Recent research works have focused on static signs of ISL [2] from images or video sequences that have been recorded using data glove, colored glove under controlled conditions like single background and special hardware devices. The light and position has more importance in the system. The signer must be aware of the system to work on this conditions.

There are different methods for preprocessing Otsu's thresholding [3][4] skin color and motion based segmentation and background subtraction [5].The feature extraction part uses wavelet decomposition, Fourier descriptors, scale invariant feature transform. There are many classifiers used to classify sign which are K Nearest Neighbor (KNN) Hidden Markov Models (HMM), Multiclass Support Vector Machines (SVM)[6], Fuzzy systems, Artificial Neural Networks (ANN) etc.

In another paper[7] implemented an edge detection method for hand gesture recognition .The frame features are extracted using edge detection and sorting features in the database. Apply template matching using the created database to predict the gesture. Here the template matching is based on the minimal distance. The system is capable of identifying the dynamic gestures as well as static symbols. A fuzzy [8] based method is introduced, the system extract the spatial features of signs using a fuzzy membership function. A suitable symbolic similarity measure is calculated and matched with the Nearest Neighbor classifier.

Reheja [9] et.al introduced a gesture recognition system on Indian sign language using the Microsoft Kinect sensor device. They experimented on both RGB and Depth images from Kinect .The research is found to be increasing the accuracy of the system while using RGB-D images. The model extract HU-Moments which are angle, location and shape invariant moments and feed those features to the SVM classifier. Pranali Loke[10] et.al developed an android app based system for Indian sign language .The android system collect images and sent to the server .The server system forward these images to Matlab application where the feature extraction takes place using Sobel operator and the system has trained using Neural network .The system apply pattern recognition and classification on the images and create an output as text.

Beena M.V.et.al [11] developed a system to recognize American Sign Language (ASL) from the depth images of Kinect sensor. The system has trained using 1000 images of each numeral signs. The algorithm extract features from the block processed images and trained using the Artificial Neural Network (ANN) and obtained an accuracy of 99.46% for the depth images. The system has been trained on GPU for the faster execution. As an extension to the work uses Convolutional Neural Network [2017-2] (CNN) with softmax classification for 33 static symbols of Kinect depth images. The implementation shows that while the number of classes increases the handcrafted feature are become insufficient for classification purpose. The CNN structure is capable of learning from the given training set and it will outperform the accuracy related to other traditional methods.

III. PROPOSED SYSTEM

Indian sign language is a complex system which include the involvement of two hands. The efficient way is to perform convolutional neural network on the image to increase the efficiency of classification and for real life application. The basic steps involving the proposed system is shown below.

Steps:

1. Input the image (video frame).
2. Find the hand object.
3. Extract the feature.
4. Classification and prediction.

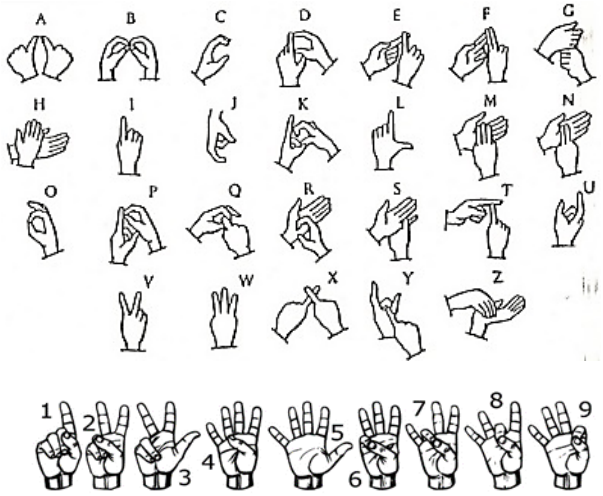

Figure 1: Indian Sign Language (Alphabets and Numerals).

Most of the object detection problems train the model using the image data set along with a bounding box mapping. The marking of bounding box for each image is costly. In addition with that we proposed a region of interest predictor using the skin segmentation. From the segmented bounded region we crop the image and feed to the classifier for prediction.

A. Region of Interest Detection

In the first part we input the video frame, apply CLACHE (Contrast Limited Adaptive Histogram Equalization) on the image to equalize the lightness in the image frame using LAB color system. In the next step apply blurring on the original image using Gaussian blurring. In order to obtain the skin we apply thresholding operation using the HSV color space. Some situation where the light variation is high, we can adjust the threshold values on the run. The last step is to find the largest contours in the segmented images and draw a

rectangle box around the part which shows the output classified result as text. To get the sign as a text the bounding box is feed to the Model created using the Convolutional Neural Network .The models consists of different layers with learned weights during the training of the network. The model generation and recognition part is shown below.

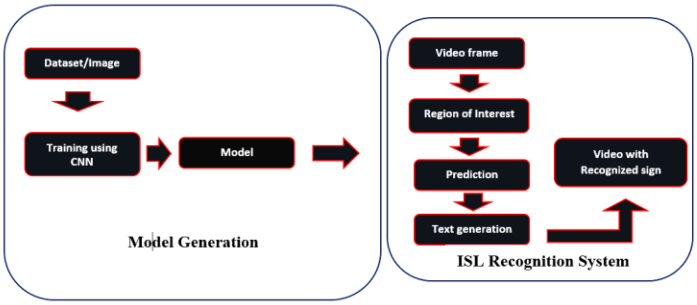

Figure 2. General diagram for the System

B. Convolutional Neural Network

Most of the image processing problems start with the extraction of relevant features which are capable of classifying the images to our desired classes. The main problem of choosing a handcrafted feature is that whenever new classes are added the system has to choose other methods. As a solution to this problem in 1990's LeNet architecture is introduced which is running on the convolutional architecture.

In our architecture different layers are included and the following layers are given below.

1. Convolutional layers
2. Pooling layers.
3. RELU(Rectified Linear Unit)
4. Fully connected layer
5. Softmax layer.

The 1st layer is the input layer which accept the region of interest in the given video frame. The input layer size is 128x128x3. Apply convolution on the given input frame using 3x3x3 kernel with 32 filters. On the next layer apply pooling of size 2 x 2 on the convoluted image and reduced this size into 64 x 64 size. After passing through the 3 convolutional layers and 3 pooling layers the size is reduced to 16 x 16 .An activation function called RELU is used in between the convolutional and pooling layers.

In the convolutional layer our entire image is considered as a multidimensional array and apply convolution operation using convolution matrix or kernel. A convolution operation is an addition of its neighboring elements along with its weights. In our model we use a filter size of 3 x 3 on each convolution layer.

The pooling layer is the layer which reduce the dimension of the image. Depending upon the pool size, in each image a single pixel is selected from the selected mask. Here we have used a pooling size of 2 x 2, which will reduce the actual size of the image in to half with a stride 2. The pooling has been implemented by max-pooling layer .This operation take the maximum value in the 2 x 2 kernel. In order to enables the positive values we have used an activation function called RELU .It is called as a positive function since it only enables the positive values and return zero for negative values and given as

$$F(x) = x+ = max\,(0, x)$$

After all convolution and pooling operation the system will flatten the entire processed images to a liner array which becomes the node of the next layer. Each layer is connected to next layer with corresponding weights which is known as fully connected layer or dense layer. The output os the dense layer is called the scores and these scores are given to the classification layer. Here we have added softmax layer as classification layer. The classification layer is an exponential function which normalize our scores to corresponding probability.

$$\text{softmax}\, Xij = \frac{exp^{Xij}}{\sum_k exp^{Xik}}$$

In each epochs the softmax function finds the accuracy of

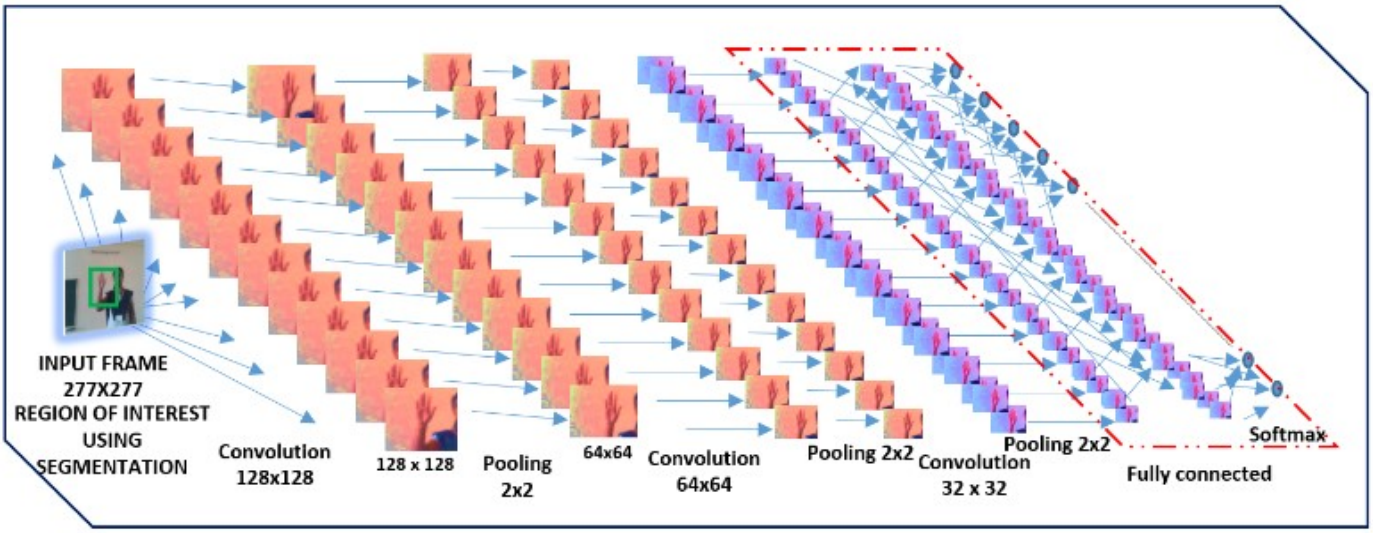

Figure 3. Layered Architecture of R-CNN

classification and a loss function. Here we use categorical cross entropy as loss function.

$$\text{Cross Entropy Loss} = -\log\left(\frac{exp^{Xij}}{\sum_k exp^{Xik}}\right)$$

By finding the output of the network and true value, the total error is calculated. Based on this error the weight has been update along with its learning rate and this process continues on each epochs called as back propagation .In the final stage the learned model has been saved on the disk, using this model the system predict new image inputs. In our system the input structure is an RGB image.

IV. EXPERIMENTAL RESULTS

The system has been trained using 300 images of each Indian Sign language numerals captured using RGB camera. The images are trained on the GPU system NVIDIA GeForce 920MX having 2GB of graphics memory, i5 processor of speed 2.7 GHz and 8GB of RAM. The system takes 28 Min to train a model using the 300 images of Numeral signs. The system trained using a batch size of 16 and initial learning rate 0f 0.001. The system attained 99.56% accuracy in 22 epochs .The system has experimented with different learning rate changing from 0.01 and activations has been updates during the training state.

Python implementation with the Deep Learning tool are used in the system. The method use Keras API with Tensor Flow as backend. The model has checked with static symbols and showing good results while testing with 100 images of test dataset. The experimental results has been shown in the figure 4.The figure consists of 3 frames ,one with the recognized sign with bounding box,2nd one is the processed image in HSV range and last one shows the segmented part using the thresholding. There is a track bar panel has been created and it can be used for adjusting the thresholding range in the real time system. Most of the part the system has to be adjusted the threshold as the lightness conditions changed. The system will not work in the situation where the background and skin color has same color field. The system is created so that it will use only 3 frames in a seconds for predicting the sign even though there is 30 frames are created in each second. The skipping of intermediate frames are used to increase the speed of the system since real-time system needs the prediction on the run.

Here selective search algorithm has tried on the system, but it is found to be complex and more number of bounding box are created apart from the object and not useful in this case. Another way of doing this work is to use object detection model to predict the bounding box. But the training state has to include the images with bounding box labeled. The way of labelling bounding box in each image is a tedious task and needs much time and workers to do this. In this situation our system has the advantages over the above problems.

In the dataset part the images are captured from the students by teaching the signs. The dataset has collected from different background and orientations. The only restriction to the development of the system is the lack of high performance GPU .In normal system the time for performing the learning is very high and the batch size has to be chosen less than 32. Since it does not fit in the memory. The system can also be trained on CPU systems, but the time will be higher than the GPU may be 10 times slower.

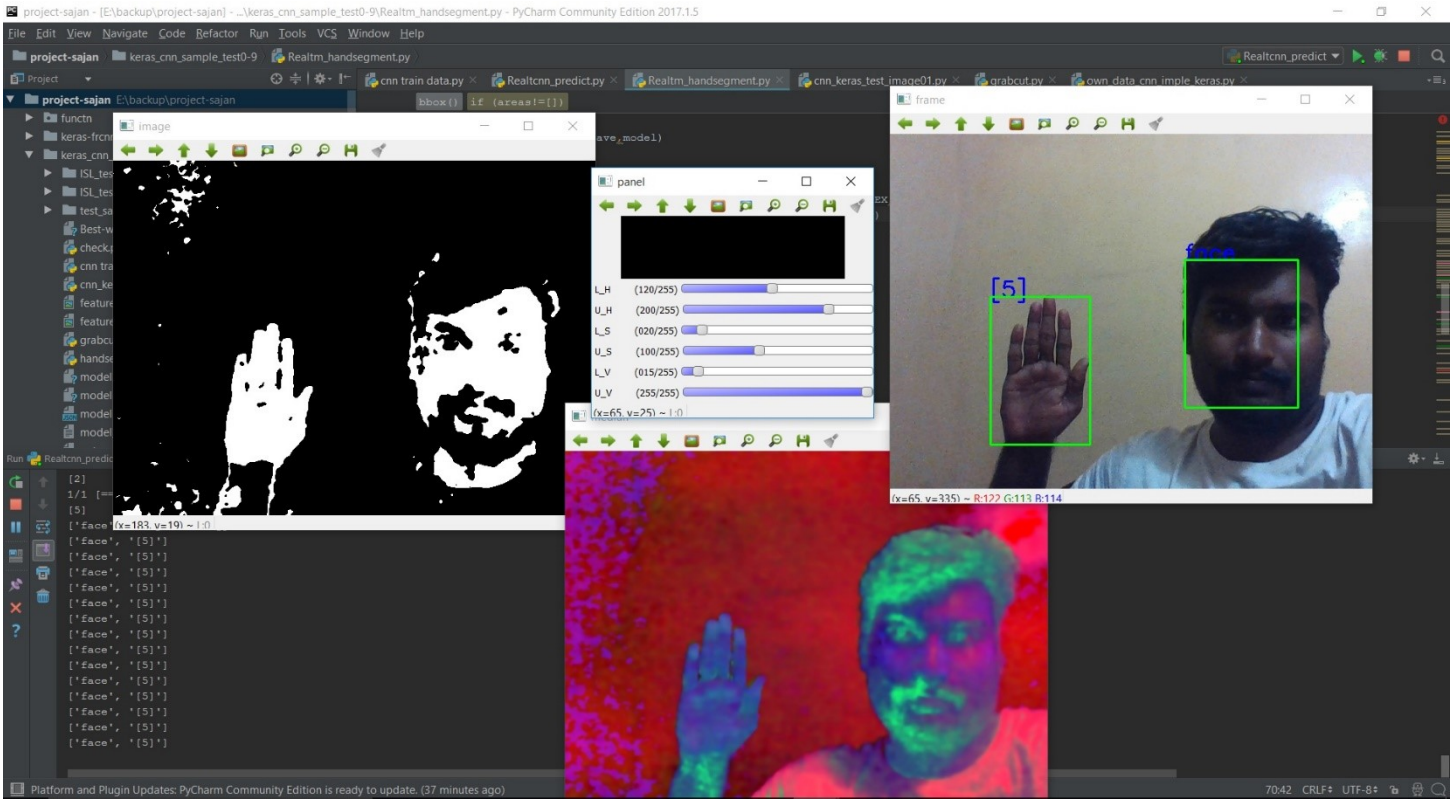

Figure 4. The Real-time System for Indian Sign Language

V. CONCLUSION

As an initial part of Indian Sign Language recognition the real time system has developed for numeral signs from 0-9. The system has been trained using the 3000 statics symbols of RGB images captured using the normal camera. The system has used 100 images for each symbols for testing. The model has been created by the successful implementation of Deep Learning system using Region based Convolutional Neural Network. The system has attained an accuracy of 99.56% for the same subject while testing and the accuracy reduced to 97.26% in the low light condition.

In future include more symbols from alphabets of static symbols of Indian sign language which include double hand notation. The low light problems has to be solved by increasing the dataset.

VI. REFERENCE

[1] Divya Deora, Nikesh Bajaj, *Indian Sign Language Recognition, 2012 1st International Conference on Emerging Technology Trends in Electronics, Communication and Networking, IEEE 2012-978-1-4673-1627-9/12.*

[2] Anuja V. Nair, Bindu V.*, A Review on Indian Sign Language Recognition, International Journal of Computer Applications (0975 – 8887) July 2013, Volume 73– No.22.*

[3] Jorge Badenas, Josee Miguel Sanchiz, Filiberto Pla, *Motion-based Segmentation and Region Tracking in Image Sequences*, Pattern recognition 2001, 34, pp. 661-670.

[4] Ping-Sung Liao, Tse-Sheng Chen, Pau-Choo Chung, 2001, A Fast Algorithm for Multilevel Thresholding, Journal of Information Science and Engineering 17, pp. 713-727

[5] Dr. Alan M McIvor*, Background subtraction techniques, Image and Vision Computing*, Newz Zealand 2000 (IVCNZ00).

[6] Aseema Sultana, T. Rajapushpa, *Vision Based Gesture Recognition for Alphabetical Hand gestures Using the SVM Classifier*, International Journal of Computer Science and Engineering Technology, Volume 3, No. 7, 2012.

[7] Purva A. Nanivadekar, Dr. Vaishali Kulkarni, *Indian Sign Language Recognition: Database Creation, Hand Tracking and Segmentation,* International Conference on Circuits, Systems, Communication and Information Technology Applications, IEEE 2014,978-1-4799-2494-3/14.

[8] Nagendraswamy H S, Chethana Kumara B M, Lekha Chinmayi R, *Indian Sign Language Recognition: An Approach Based on Fuzzy-Symbolic Data, 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, 978-1-5090-2029-4/16.*

[9] J. L. Raheja , A. Mishra, A. Chaudhary, *Indian Sign Language Recognition Using SVM, Pattern Recognition and Image Analysis,* 2016, Vol. 26, No. 2, pp. 434–441.

[10] Pranali Loke, Juilee Paranjpe, Sayli Bhabal, Ketan Kanere, *Indian Sign Language Converter System Using An Android App.* ,International Conference on Electronics, Communication and Aerospace Technology, 2017 IEEE ,978-1-5090-5686-6/17.

[11] M.V. Beena and M.N. Agnisarman Namboodiri, *ASL Numerals Recognition from Depth Maps Using Artificial Neural Networks,* Middle-East Journal of Scientific Research 25 (7): 1407-1413, 2017,ISSN 1990-9233.

[12] Beena M.V., Dr. M.N. Agnisarman Namboodiri, *Automatic Sign Language Finger Spelling Using Convolution Neural Network: Analysis, International Journal of Pure and Applied Mathematics, Volume 117 No. 20 2017, 9-15.*