Hand Gesture Recognition System Using Computer Vision And Neural Networks

Akhil Dixit, Arush Agarwal

Abstract— A gesture can be recognized as a form of nonverbal communication in which certain bodily actions represent and communicate particular messages. These actions may be represented by facial features or by hand movement. The hand movements are conveniently deciphered to particular messages that are universal in nature. This research work presents a model system that allows the differently abled people to communicate with their counterparts effectively. It focuses on the problem of predicting the appropriate response to a particular hand sign gesture effectively. The system can be used as a support application to facilitate efficient communication in medical and other emergencies. In this regard, the human hand is segmented from the complicated background. Then, the area of hand gesture which has been detected in real time is recognized by convolutional neural network so as to realize the recognition of 26 alphabet digits. Experiments show 98.1% accuracy.

Index Terms—Computer vision, Deep learning, Gesture Recognition, Feature Extraction, Keras, Neural Network

1 Introduction

Hand gesture recognition refers to the task of finding and interpreting hand signs and movement through mathematical algorithms. It is an active area of research as the interaction between computers and human is becoming more prominent. The task of interaction and communication between people over the globe is done through internet, hence it becomes necessary for the differently abled to communicate effectively and efficiently. In the last decade, several methods of potential applications in the advanced hand gesture interfaces have been suggested but these differ from one another in their models. Some of these models are Neural Network [1], Hidden Markov Models (HMMs) [2], [3], [4], [5], Dynamic Time Warping (DTW) [6], [7] and CRFs [8], [9], [10] . This paper focuses on using the neural network to build and train a suitable model which recognizes the gestures and present its verbal counterpart.

The recognition and prediction system is composed of three parts as hand gesture segmentation, neural network modeling and gesture recognition.

The segmentation part deals with identifying desired area of interest in the frame, recognizes and differentiates hand from the background. It mainly segments on the basis of skin color, corner detection and motion sensitivity. The proposed algorithm converts the entire image in a two dimensional matrix which is further changed to HSV format to compare each pixel with the skin tone. The compared pixels are then rendered on a threshold frame.

These masked images are stored in a set of 1400 images per gesture. Total 14 hand gestures from 1-14 in convoluted background are acquired for this research work. With 1400 images collected to build the model, we have a total of 19600 images. 15,700 images are taken for training model and 3900 taken for testing to have a ratio of 80% to 20% dataset. The experiment utilizes Keras library to build the Convolution Neural Network with the TensorFlow as the backend to train the model. The features are extracted from the masked image data converted to a 1 dimensional matrix with the first column being the class of each gesture.

- *Akhil Dixit, Undergraduate Student B.Tech, Electronics and Communication, Delhi Technological University, Bawana,Delhi, E-mail: akhil.dixit.ec1998@gmail.com*
- *Arush Agarwal, Undergraduate Student B.E, Electronics and Communication, Netaji Subhas University of Technology, Dwarka,Delhi, E-mail: agarwal.arush98@gmail.com*

The recognition and prediction of the gestures is performed by utilizing the neural network model formed. Each frame of a video sample is taken to get the predicted gesture. The classifier with the maximum feature match is taken to be the optimum gesture result.

The structure of the procedure is represented in fig

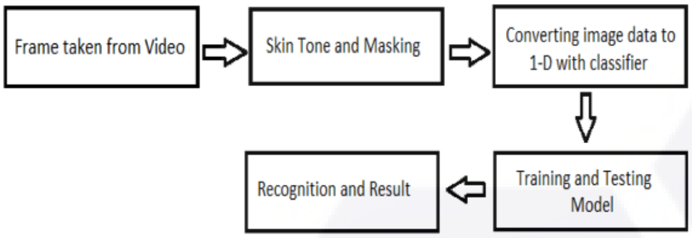

Frame taken from Video
Skin Tone and Masking
Converting image data to 1-D with classifier
Training and Testing Model
Recognition and Result

Fig. 1. Structure of proposed hand gesture system

In this research work focus is made to recognize hand gestures without markers and special gloves.

2 AVAILABLE GESTURE SYSTEM ANALYSIS

Over the years several gesture recognition systems have been proposed and albeit they recognize the gestures efficiently they are not able to overcome the gap between traditional and sign language communication. This research work takes into care the practicality of the problem and solution in such a way as to implement a practical approach towards overcoming the barrier.

2.1 RESEARCH REVIEW

AEl-Sawah, N. Georganas, et.al [11] ,explained designing a prototype for three dimensional tracking and dynamic gesture recognition. Objective is to be able to continuously track the hand in a background and to be able to recognize dynamic gestures in real time.

P. Mekala, Ying Gao, et.al [12], proposed a research article for an architecture that uses neural networks identification and tracking to translate sign language to a voice/text format.

J.Singha, K. Das [13], proposed a system where weighted Euclidean distance is used as a classifier technique for recognition of various sign Languages. The system comprises of four parts are Skin Filtering, Hand Cropping, Feature Extraction and Classification.

Qing Chen and N.D. Georganas [14], proposed a new approach to solve the problem of real-time vision-based hand gesture recognition with the combination of statistical and syntactic analysis.

G. R. S. Murthy and R. S. Jadon [15] designed method for recognizing specific hand gestures and use them to convey information and data

Jing-Hao Sun,Ting-Ting Ji,Shu-Bin Zhang,Jia-Kui Yang [16]proposed a method using CamShift algorithm for hand gesture tracking and deep learning for classification.

M.Elmezain,Ayoub Al-Hamadi,B. Michaelis [17] proposed a new approach of gesture recognition using Conditional Random Fields algorithm.

2.2 PROPOSED SOLUTION

To establish an efficient communication between deaf and hearing persons a system must be established that could facilitate and decipher gestures. A system that could identify hand movements and gestures and then be able to interpret them based on previous model analysis is required. Such a model should be inexpensive, and universally available.. Creating such a system in Python language and deploying the software as a back-end of the modern application world would have dual effects. The differently abled people would be able to utilize this tool as an online web page that would allow them to communicate with people that don’t comprehend sign languages. Second being that now this software would have a huge amount of data through which better training dataset can be created. This allows to have entire sentences being formed in gestures instead of only relying on the American Sign Language and their individual characters. The software would be available online as a web page and so reduces the download time and takes into fact the necessity of the situation so that output be displayed in real time and dynamically.

3 IMPLEMENTATION AND RESEARCH

The proposed solutions work in three stages. First stage being able to recognize and segment the hand gesture to form image dataset. Second stage being, converting the image dataset to matrix format to apply mathematical algorithms. Third stage focuses on training, testing the model to predict gesture.

3.1 HAND SEGMENTATION

Segmenting and differentiating the hand from its dynamic background is necessary to extract optimum features from the gesture. Traditionally several methodologies exist to extract the data. Edge detection and motion detection are able to establish gesture recognition and tracking but are not optimum to segment the gesture as certain parts would be static. The proposed method of segmentation deals with masking the input frame in range with the skin tone. The input frame is in RGB model which is converted to HSV color model,

$R' = \frac{R}{255}$

$G' = \frac{G}{255}$

$B' = \frac{B}{255}$

$Cmax = \max(R', G', B')$

$Cmin = \min(R', G', B')$

$\Delta = Cmax - Cmin$

Hue Calculation:

$$If\ Cmax = R',\ H = 60 \times \frac{(G'-B')}{\triangle} mod6$$
$$If\ Cmax = G',\ H = 60 \times \frac{(B'-R')}{\triangle} + 2$$
$$If\ Cmax = B',\ H = 60 \times \frac{(R'-G')}{\triangle} + 4$$

Saturation Calculation:

$$If\ Cmax = 0\ ,\ S = 0$$
$$If\ Cmax \neq 0\ ,S = \frac{\triangle}{Cmax}$$

Value Calculation:

V=Cmax

The RGB values are divided by 255 to change their range from 0-255 to 0-1.

The HSV for skin tone is taken and is masked to segment the region of interest.. Gaussian Blur and Morphology are performed on the segment to reduce noise and imperfections.

In one dimension, the Gaussian filter mathematically represented as,

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{\sigma^2}}$$

Where σ is standard deviation of the distribution. The distribution is assumed to have a mean of 0.

The Gaussian filter works by using the 2D distribution as a point-spread function. The image space Gaussian filter is an NxN-tap convolution filter that weights the pixels inside of its footprint based on the Gaussian function. The features are,

Convolution filter – An algorithm that combines the color value of a group of pixels.

NxN-tap filter – A filter that uses a square shaped footprint of pixels with the square's side length being N pixels.

N-tap filter – The filter uses an N-pixel footprint.

Filter kernel – A collection of relative coordinates and weights that are used to combine the pixel footprint of the filter.

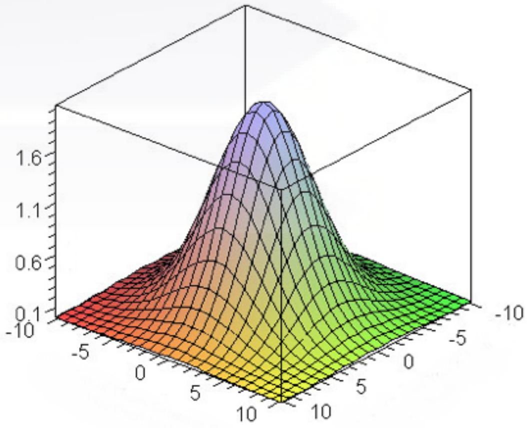


Fig. 2. Representation of 2 dimensional Gaussian Representation

The segmented masked contour is then compared with minimum contour area as to remove any noise or background matching the observed HSV range. The feature region is then taken and stored.

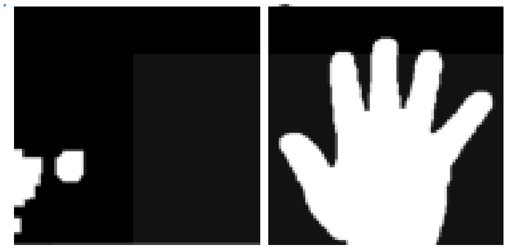


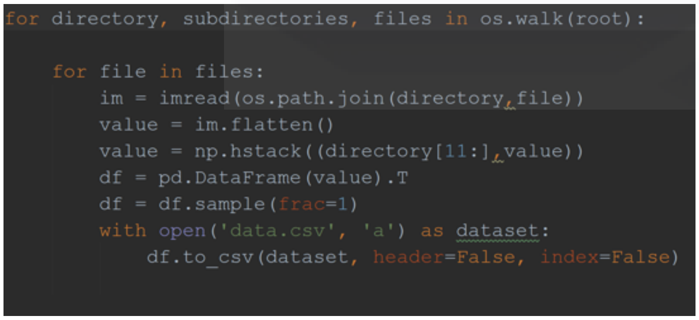
Fig. 3. The noise contour is ignored and region of interest is explored.

These images are then stored in a set of 1400 images per gesture.

3.2 DATA EXTRACTION

The images stored represent the threshold segmented region of interest. The data has to be converted to a form that is efficient to create a high accuracy model. The following steps illustrate how the proposed method works,

1. Read the data from each directory and sub-directory where images are stored.
2. Flatten the data into 1 Dimensional matrix.
3. Stack the matrix in heap with the first column being the classifier.
4. Frame the data according to axis and randomize it.



```
for directory, subdirectories, files in os.walk(root):

    for file in files:
        im = imread(os.path.join(directory,file))
        value = im.flatten()
        value = np.hstack((directory[11:],value))
        df = pd.DataFrame(value).T
        df = df.sample(frac=1)
        with open('data.csv', 'a') as dataset:
            df.to_csv(dataset, header=False, index=False)
```

Fig. 4. Implementation of proposed algorithm

3.3 CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Networks are analogous to traditional Artificial Neural Networks in that they are comprised of neurons that self-optimise through learning. Each neuron will still receive an input and perform a operation (such as a scalar product followed by a nonlinear function) - the basis of countless ANNs. From the input raw image, vectors to the final output of the class score, the entire network will still express a single perceptive score function (the weight).

They have proven to be very successful frameworks for image recognition. In the past few years, variants of CNN models achieve increasingly better performance on the renowned ImageNet dataset for object classification, starting from AlexNet from [18], OverFeat [19], GoogLeNet [20], and a recent model by [21] with classification accuracy surpassing human-level performance.

In this research work CNN architecture is used to form a neural network and Keras library is used to realize it. The model takes into factor different layers and activation parameters. It uses sequential model which is a linear stack of neural layers and each CNN layer is then specified.

The experiment uses the following CNN architecture.

TABLE 1

CNN Architecture, width taken to be 50

Layer	**Layer Type**	**Size**	Output Shape
1	Convolution + ReLU	32 5x5 filters	
1	Max Pooling	2x2, stride 2	(32,23,23)
2	Convolution + Sigmoid	64 5x5 filters	
2	Max Pooling	5x5, stride 5	(64,10,10)
3	Dense + ReLU	1024 units	
3	Dense + Softmax	15	15

The model uses convolution 2 dimensional with activation 'ReLU'. This layer has 32 filters denoting the output shape of this layer having 5x5 kernel size. This layer creates a convolution kernel that is convolved with the layer input over a single spatial dimension to produce a tensor of outputs.

Max pooling is a sample-based discretization process. The objective is to down-sample an input representation (image, hidden-layer output matrix), reducing its dimensionality and allowing for assumptions to be made about features contained in the sub-regions binned.This is done to in part to help over-fitting by providing an abstracted form of the representation. As well, it reduces the computational cost by reducing the number of parameters to learn and provides basic translation invariance to the internal representation. We use a stride of 2 to down sample it.

Dense refers to the number of neurons in hidden layer with the input parameters.

The algorithm to create proposed model involves following steps:

1. *Read dataset from the data extracted CSV*

2. *Shuffle data to have random occurrence of classifier*.

3. *Create a training and testing set in 80% to 20% of available data*.

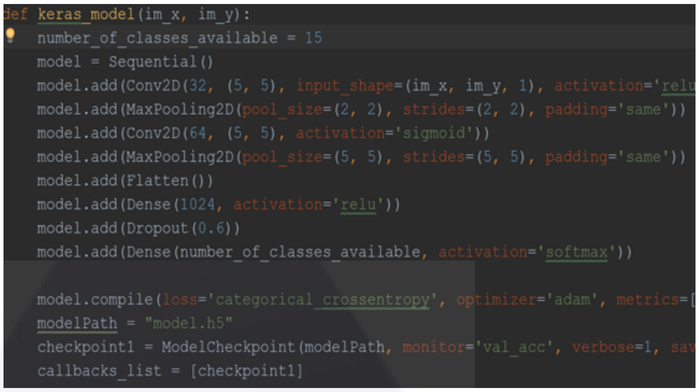
4. *Create a CNN model by following table 1 architecture to develop the layers*.

5. *Train and test the model and take epoch to be 10*

6. *Save the model*

The 1 dimensional data is read from the CSV file. The experiment takes images of 50 X 50 size. A matrix of single row and 2500 columns is created. As the CSV file stored dataset in stack so the number of row in data set equals 1400 X 1400 14 that is 19600. The classifier is in the first column and is extracted to Y_Label accordingly. The data set is shuffled to have a random training and testing data. This data is fed to the CNN model created and trained

The following snippet shows the model,



```
def keras_model(im_x, im_y):
    number_of_classes_available = 15
    model = Sequential()
    model.add(Conv2D(32, (5, 5), input_shape=(im_x, im_y, 1), activation='relu
    model.add(MaxPooling2D(pool_size=(2, 2), strides=(2, 2), padding='same'))
    model.add(Conv2D(64, (5, 5), activation='sigmoid'))
    model.add(MaxPooling2D(pool_size=(5, 5), strides=(5, 5), padding='same'))
    model.add(Flatten())
    model.add(Dense(1024, activation='relu'))
    model.add(Dropout(0.6))
    model.add(Dense(number_of_classes_available, activation='softmax'))

    model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=[
    modelPath = "model.h5"
    checkpoint1 = ModelCheckpoint(modelPath, monitor='val_acc', verbose=1, sav
    callbacks_list = [checkpoint1]
```

Fig. 5. Implementing CNN model

3.4 INTERPRETING GESTURE

The CNN model is prepared to predict and interpret the gestures.

The date for the work in the paper includes 14 categories of hand gestures ranging from 1 to 14 in the indoor environment, with 4000 images for each gesture. The video source provides the input frames. Each frame is segmented and the acquired masked image is processed with the neural network model to get maximum predicted classifier.

The research ran several experiments and the most suitable coefficients and features are utilized to use the experiment in a suitable environment.

The average accuracy of hand gesture recognition is 98.1%. The recognition rate for alphabet 'K' is not high. The highest accuracy is registered for the 'I' alphabet with 99.75% accuracy.

The result show that the model system is able to predict the hand gesture and then convert it into sentence form. The previous result is stored and displayed in textbox.

Fig 6. Shows the training and testing accuracy of the model in 20 epochs. Fig 7. Shows the loss percentage while training and testing the model.

The accuracy of the model increases in the initial epochs. As the dataset available is formed under controlled indoor conditions, segmenting one of the skin tone, the accuracy is high. With subsequent iterations some loss is observed due to changes in the light and testing conditions.

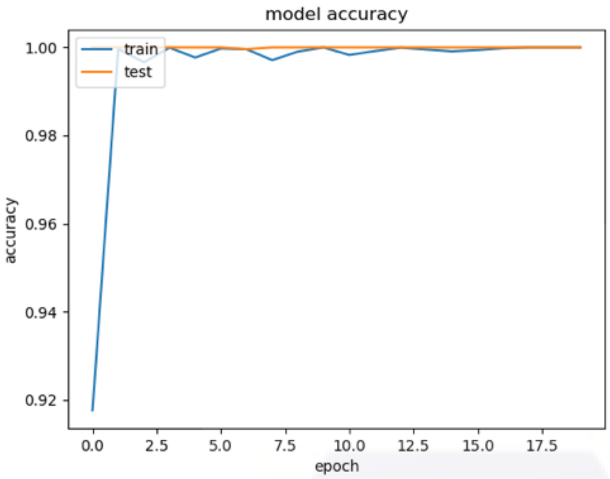


Fig. 6. The accuracy of the model increasing epoch, the model accuracy is high as experiment is performed in controlled environment.

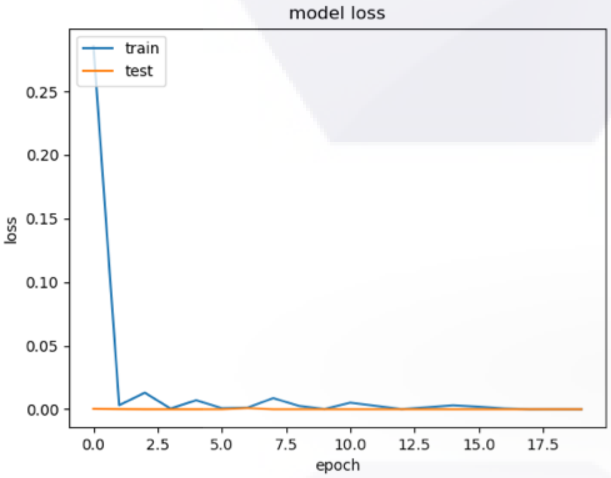


Fig. 7. Model loss decreases with each epoch and the model is carefully fit.

The loss curve follows a similar trend of accuracy curve in each epoch.

The result of categories of hand gestures are,

The model works efficiently and is able to interpret the hand signs and gestures. These results are then passed through a text to speech recognition which gives audio feedback to the results found.

TABLE 2

The predicted results compared to the actual value

S	1	2	3	4	5	6	7	8	9	10	11	12	13	14	A %
1	392	0	0	0	1	0	0	0	0	0	5	0	2	0	98
2	0	398	0	0	0	0	0	0	0	0	2	0	0	0	99.5
3	0	0	385	0	0	0	0	0	0	0	0	0	15	0	96.3
4	0	0	0	398	0	0	2	0	0	0	0	0	0	0	99.5
5	9	0	0	0	386	0	2	0	0	0	0	0	3	0	96.5
6	0	12	0	0	0	385	0	0	0	0	3	0	0	0	96.25
7	0	0	0	4	0	1	392	2	0	0	0	1	0	0	98
8	0	0	1	3	0	0	5	391	0	0	0	0	0	0	97.75
9	0	0	0	0	0	0	0	0	399	1	0	0	0	0	99.75
10	0	0	0	1	0	0	0	0	2	397	0	0	0	0	99.25
11	5	0	0	0	1	0	0	2	0	0	392	0	0	0	98
12	0	0	0	0	0	0	0	3	0	0	0	397	0	0	99.25
13	4	0	0	0	0	1	0	0	0	0	0	0	393	6	98.25
14	5	1	2	0	0	0	1	0	0	0	0	0	3	388	97

Hand Gesture Recognition

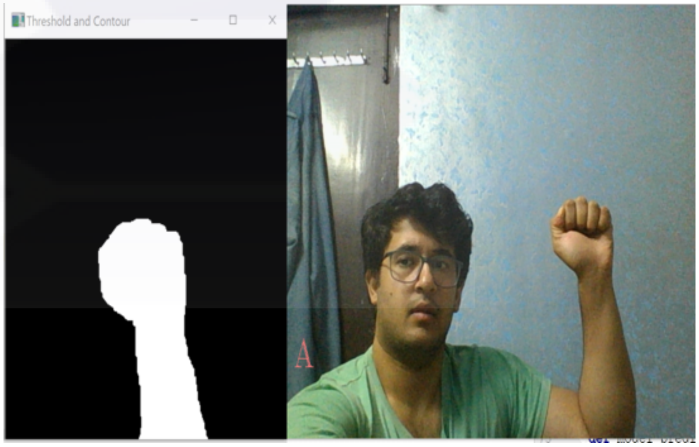


Fig. 8. Shows input video feed animating 'A' hand sign from ASL and in next frame the output is printed.

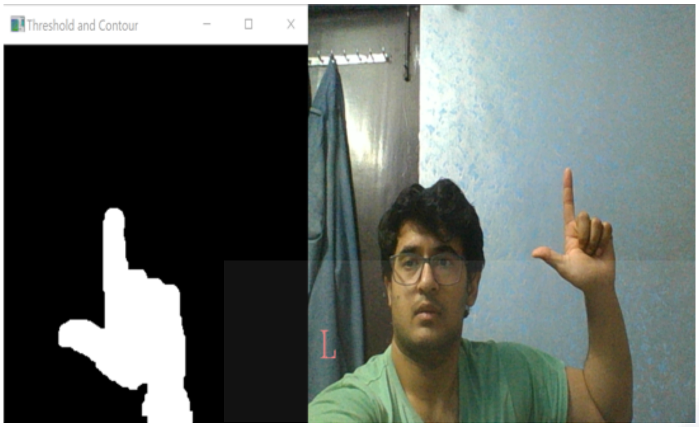


Fig. 9. Shows input video feed gesturing 'L' hand sign from ASL and in next frame output is printed.

4 APPLICATIONS, ADVANTAGES AND LIMITATIONS

4.1 APPLICATIONS

1. *The system can be used at public places like Hotels, Banks, Railway Stations.*

2. *The system can be used at emergency services like Police, Fire Department and Hospitals to facilitate efficient communication with the differently abled people.*

3. *The system deployed on web services can be used in video sources such as Skype, Messaging platforms as a translation device.*

4.2 ADVANTAGES

1. Efficient way of communication for differently abled people.

2. Increase in communication and reduction in time delays.

3. Quick response time.

4. An automate system, inexpensive and has low power usage.

4.3 LIMITATIONS

1. The light intensity exposure affects the accuracy as the model uses skin color tone as a masking parameter.

2. Due to currently low amount of dataset, proper hand gesture images should be taken.

3. As it uses video source the resolution should be adequate for the model to recognize gestures.

5 FUTURE SCOPE AND DEVELOPMENT

The proposed system is able to recognize the gestures with high probability and interpret them by their ASL code. The implementation of experiment is performed on Python language, and is hosted locally. The scope of the experiment is not limited to local computing device. A back-end for web service that implements the system provides a viable and efficient way to impart emergency services and non-sign language comprehending people to point a mobile camera device and interpret the hand gestures. This would decrease the communication time period which is critical in emergency. The server would interpret and build a dataset which would further improve the model, the model would be self-enhancing in nature and with the large dataset predict, entire sentences to follow. The development of such system utilizes Artificial Intelligence at its core to facilitate communication in a virtual world where typing is superfluous and 80% of communication performed in a virtual platform of video conferencing rather than typing long sentences

6 CONCLUSION

The paper carries out research on a set of overall flow for the hand gesture recognition. It utilizes deep learning and neural network to realize the model. Segmentation for the hand gestures is performed by masking the frame with threshold. The proposed system performs hand gesture spotting and recognition tasks simultaneously. Furthermore it is suitable for real time applications and solves the issue of time delay between the segmentation and recognition tasks. The segmented gesture is classified by convolutional neural network. This work can be further extended to sign to speech conversion.

7 REFERENCES

[1] X. Deyou, A Network Approach for Hand Gesture Recognition in Virtual Reality Driving Training System of SPG, International Conference on Pattern Recognition, pp. 519-522, 2006.

[2] M. Elmezain and A. Al-Hamadi and B. Michaelis, Real-Time Capable System for Hand Gesture Recognition Using Hidden Markov Models in Stereo Color Image Sequences, Journal of WSCG, Vol. 16, No. 1, pp. 65-72, 2008.

[3] M. Elmezain and A. Al-Hamadi and J. Appenrodt and B. Michaelis, A Hidden Markov Model-based continuous gesture recognition system for hand motion trajectory, International Conference on Pattern Recognition, pp. 1-4, 2008.

[4] D. Kim and J. Song and D. Kim, Simultaneous Gesture Segmentation and Recognition Based on Forward Spotting Accumlative HMMs, Journal of Pattern Recognition Society, Vol. 40, pp. 3012-3026, 2007.

[5] M. Elmezain and A. Al-Hamadi and B. Michaelis, A Novel System for Automatic Hand Gesture Spotting and Recognition in Stereo Color Image Sequences, Journal of WSCG, Vol. 17, No. 1, pp. 89-96, 2009.

[6] K. Takahashi and S. Sexi and R. Oka, Spotting Recognition of Human Gestures From Motion Images, In Technical Report IE92-134, pp. 9-16, 1992.

[7] J. Alon and V. Athitsos and Y. Quan and S. Sclaroff, A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 9, pp. 1685-1699, 2009. [8] H. Yang and S. Scharoff and S. Lee, Sign Language Spotting with a Threshold Model Based on Conditional Random Fields, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 31, pp. 1264-1277, 2009.

[9] J. Lafferty and A. McCallum and F. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling sequence Data, International Conference on Machine Learning, pp. 282-289, 2001.

[10] M. Elmezain and A. Al-Hamadi and B. Michaelis, Discriminative models-based hand gesture recognition, International Conference on Machine Vision, pp. 123-127, 2009

[11] A. El-Sawah, et.al "A prototype for 3-D hand tracking and gesture estimation," IEEE Transaction Instrumentation Measurement. volume 57,no. 8, pp. 1627–1636, August 2008.

[12] S. Pandita, and S. P. Narote "Hand Gesture Recognition using SIFT" International Journal of Engineering Research and Technology (IJERT) Volume 2, no.1, pp 1760-1764, January 2013.

[13] J. Singha and K. Das, "Indian Sign Language Recognition Using Eigen Value Weighted Euclidean Distance Based Classification Technique",(IJACSA) International Journal of Advanced Computer Science and Applications, Volume 4, no. 2 , pp. 188-195,July 2013.

[14] Qing Chen, N. D. Georganas, "Hand Gesture Recognition Using Haar Like Features And a Stochastic Context-Free Grammar", IEEE transactions on instrumentation and measurement, volume 57, no. 8, pp.113-117, August 2008.

[15] G. R. S. Murthy and R. S. Jadon, "Hand Gesture Recognition using Neural Networks", IEEE Transactions, Volume 6, pp.427-431, May 2010.

[16] Jing-Hao Sun, Ting-Ting Ji, Shu-Bin Zhang,Jia-Kui Yang, "Research on Hand Gesture Recognition based on Deep Learning", IEEE transaction vol. 43 no. 6A pp. 103-108 2016

[17] M.Elmezain,Ayoub Al-Hamadi,B. Michaelis, "A Robust Method for Hand Gesture Segmentation and Recognition Using Forward Spotting Scheme in Conditional Random Fields", IEEE Pattern Recognition, vol,43, pp. 3850-3854 August 2010

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[19] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R.

Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” CoRR, 2013.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” CoRR, vol. abs/1409.4842, 2014.

[21] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” arXiv, 2015

[22] The hand signs utilised are of American Sign Language.