# Hand Gesture Recognition and its conversion to text

Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala

| | | | |
|---|---|---|---|
| Devansh Kaushik | Meghna Sinha | Utkarsh Chauhan | Bhoomica Gupta |
| Department of CSE | Department of CSE | Department of CSE | Department of CSE |
| Thapar University | Thapar University | Thapar University | Thapar University |
| Patiala ,India | Patiala ,India | Patiala ,India | Patiala ,India |
| dkaushik_be20@thapar.edu | msinha_be20@thapar.edu | uchauhan60_be20@thapar.edu | bgupta_be20@thapar.edu |

***Abstract*-- Hand Gesture Detection Systems have advanced significantly in recent years due to its effective machine interaction. Humanity seeks to replace multi-touch technology, which necessitates touching motion on screens, in order to merge human gestures into contemporary technology. This study provides a summary of various approaches for achieving hand gesture identification using three key modules: camera and segmentation, detection, and feature extraction. Depending on the benefits, a variety of techniques can be utilised to achieve the desired effects.In this study, we develop a hand gesture detection system that effectively tracks static hand gestures. Our system converts the hand gesture it detects into text. This study also includes a summary of prior research, findings of hand gesture approaches, and a comparison of gesture recognition.**

## I. INTRODUCTION

A natural way for people to communicate is through gestures.Gestures that accompany speech can reveal a speaker's intentions, interests, feelings, and ideas. In noisy settings, from a distance, and for those who have hearing loss, gestures are even more crucial. In these situations, gestures take the place of speech as the main method of communication, becoming both more prevalent and structured .

The relationship between a human and a computer (machine) is known as human-computer interaction (HCI) and is symbolised by the development of the computer itself. The key goal of a hand gesture recognition system is to make meaningful information exchange between a human and a computer feel natural. Functionality and usability are the two key factors that should be taken into account while creating an HCI system.It is possible to create a system that performs effectively and effectively.

However, it is first necessary to distinguish between hand postures and hand gestures in order to develop a successful hand gesture detection system. Hand gesture is described as a dynamic movement made up of successions of hand postures over a brief period of time, whereas hand posture is a static hand configuration depicted by a single image without any involvement of motions. A hand gesture is anything you do with your hands, such as waving goodbye, while a hand posture is something you do with your hands.Automatic gesture detection is consequently an important domain of computer vision research, with applications in Human Computer interactions (HCI) . Unsurprisingly, there is now a substantial body of literature on gesture recognition;

While system usability is defined by the level and extent of the system in which the system can be effectively utilised to achieve certain specific user goals, system functionality refers to a collection of activities or services that are supplied to its users. By achieving a balance between usability and functionality.

## II. RELATED WORK

In earlier research on hand gesture identification, non-geometric variables such as multivariate Gaussian distribution were used to identify hand movements.

The input hand image is segmented using two separate methodologies: clustering-based thresholding techniques and skin color-based segmentation utilising the HSV colour model. In order to extract hand features, some operations are used to capture the shape of the hand. The modified Direction Analysis Algorithm is adopted to find a relationship between statistical parameters (variance and covariance) from the data, and is used to compute object (hand) slope and trend by determining the direction of the hand gesture, as shown in the fig. below.

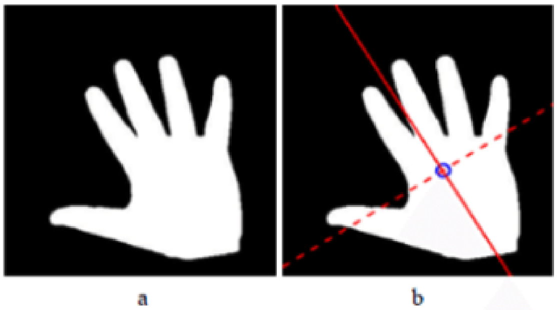


Fig 1. Computing Hand Direction

The image has been divided into circular sections using the resulting Gaussian function, or, to put it another way, those regions are produced in the shape of a terrace to counteract the effect of rotation. The shape is split into 11 terraces, each with a width of 0.1 . Nine terraces are created as a result of the 0.1 width division, and they are as follows: (1-0.9, 0.9-0.8, 0.8-0.7, 0.7-0.6, 0.6, 0.5, 0.5-0.4, 0.4-0.3, 0.3-0.2, 0.2-0.1), as well as one terrace for the terrace with a value less than 0.1 and the final terrace for the external area that extended out from the outer terrace .Figure explains this split and provides an example.

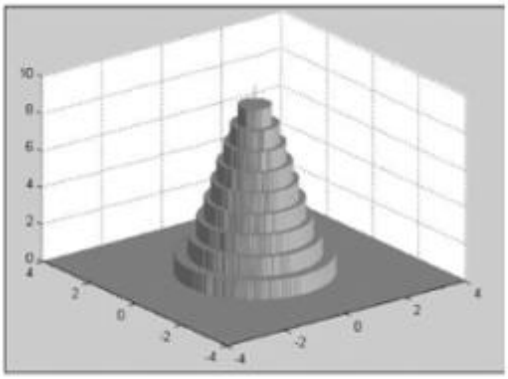


Fig 2. Terraces Division

Two types of features—local features and global features—are retrieved once the hand shape has been captured to create the feature vector. Using geometric central moments that produce two distinct moments, local features As indicated by equation (1)

$$\mu_{pp} = \sum_{x}\sum_{y}(x-\mu_x)^p(y-\mu_y)^p f(x,y) \qquad (1)$$

$$\mu_{pp}^{(k)} = \sum_{y}\sum_{x}\left(x^{(k)}-\mu_x^{(k)}\right)^p\left(y^{(k)}-\mu_y^{(k)}\right)^p f\left(x^{(k)},y^{(k)}\right) \qquad (2)$$

$$\forall\, k \in \{1, 2, 3, \ldots, 88\} \;\&\; \forall\, p \in \{0, 1\}$$

The input image is represented by 88*2 features, as detailed in equation, where x and y is the mean value for the input feature area, x and y are the coordinated, and (2). The first and second moments are two features that are computed for the entire hand features area, while the global features are two features. These feature areas are calculated by multiplying the intensity of the feature area by the position of the feature area on the map. Any input image is represented in this scenario with

178 characteristics. The system used 20 different gestures, 10 samples for each gesture, 5 samples for training, and 5 samples for testing. The system had a 100% identification rate, although this percentage dropped when there were more than 14 motions. In.

The heterogeneous networks of Wang et al.are the approach that comes the closest to ours. To identify gestures in videos, they employ 3D ConvLSTMs and CNNs, two different types of networks to distinguish between motions in dynamic images created by rank pooling. They use these networks at the body and hands, two different spatial scales. The networks are performed on RGB and depth data, and the combined scores from the 12 modalities. In order to prevent overfitting to the background, Wang et al. use F-RCNN to identify bounding boxes around the hands in every frame. They then remove any scene elements that are outside of the bounding box encircled by the hands. The bounding boxes nearly fill the entire image for actions involving two hands and/or large movements, undermining the purpose of the hand channel. Wang et alhand .'s level networks are intended to remove background noise rather than to divert attention to the hands. Our focus nets, on the other hand, are always focused on just the hands when we detect the right and left hands and choose attention windows surrounding them. A similar concept of training a global and focus net was put forth by Karpathy et al. They rely on camera bias to draw attention to the centre of the picture, though. On the ChaLearn and NVIDIA data sets, when the subjects are not in the centre of the frame, this will not function.

## Global Channels

Global Channels: Based on 50 different countries,network layers with deep residuals On the ImageNet task, ResNet-50 performs well. Despite the fact that ResNet has more advanced iterations (ResNet-101,better performing ImageNet designs like Inception-V4 and Squeeze, as well as ResNet-152 and ResNet-1001).

For practical reasons, Excitation Network and ResNet-50 were chosen: we must train numerous channels, and eachIn our lab, ResNet-50 can fit on a single GPU. Four modalities are taught to global channels: RGB,depth, RGB-derived optical flow fields, and optical flow fields.

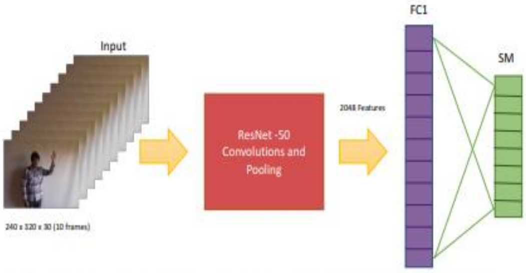

Input
240 x 320 x 30 (10 frames)
ResNet -50 Convolutions and Pooling
2048 Features
FC1
SM

Figure 3. Network Architecture of Global Channels. The input to the network is a stack of 10 images resulting in a 240 × 320 × 30 volume. The input volume is passed through ResNet-50 convolution and pooling layers resulting in 2048 features. A fully connected layer on top produces a vector of softmax scores.

## III. METHODS AND MODELS

### Experimental Design

Participants were instructed to categorise movies as one of 249 gestures for the 2017 ChaLearn IsoGD challenge. A collection of 35,878 training movies with labels were made available to participants. and another batch of 5,784 videos for validation that have labels. With the help of the training films and validation movies, participants were urged to create the best system they could. A previously hidden collection of 6,271 categorised test videos was made available to participants after the challenge. They were asked to assess their system using the test movies as-is.

We mimicked this experimental concept as precisely as we could given that our system was created after the challenge deadline. We kept the test videos to ourselves and did not use them to test our system while it was being developed.

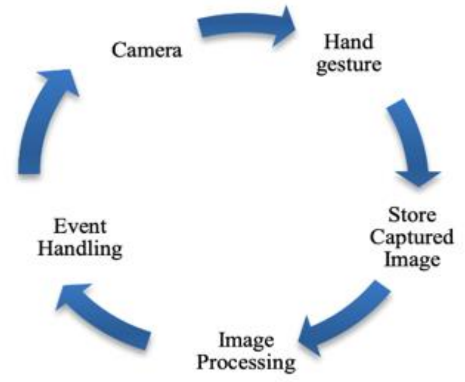


Fig 4. Steps/Workflow of the model proposed.

## Hand Detection

We utilise Liu et alhand .'s detection results for ChaLearn. In order to detect hands, they employ a two stream Faster RCNN.

Convolutions are applied to RGB and depth videos individually. Depth maps and RGB feature maps are stacked one on top of the other. On the stacked feature maps, a Region proposal network and an object classifier are employed. Liu et alhand .'s detection results do not distinguish between right and left hands. To distinguish between the left and right hands, skeletons taken from RGB frames using Cao et al .'s multi-person posture estimate code are employed.

Estimates of the right and left wrist skeletons are extrapolated and interpolated as necessary to fill in the gaps.

## Image processing

Different image processing methods, including as colour conversion, noise removal, and thresholding, are applied to the output from the camera module before contour extraction. If there are flaws in the image, convexity flaws are discovered in accordance with the gesture is recognised. If there are no flaws, the gesture is detected by classifying the image using the Haar cascade.

The detection module performs the following actions in relation to dynamic gestures; The dynamic gesture swipe is identified if Microsoft PowerPoint has been opened with a slideshow enabled and the webcam has picked up palm movement for five consecutive frames.

## Deep Learning based recognition

Artificial intelligence provides a good and trustworthy method employed in a variety of contemporary

due to the adoption of the learning role principle. Multilayer learning was utilised in deep learning.

data and produces a reliable prediction. The biggest difficulties with this method are necessary.

algorithm to learn from dataset that may impact time processing.

Seven common hand gestures that are captured by smartphone cameras and produce 26498 picture frames. The deep convolutional neural network that was modified and extracted features (ADCNN) used for classifying hands. 100% of the training data and testing results are evaluated in this experiment.

99% of the data were processed in 15,598 seconds [90]. While Webcams were employed in other planned systems in order to follow hand. then removed using morphology and the skin colour (Y-Cb-Cr colour space) method.

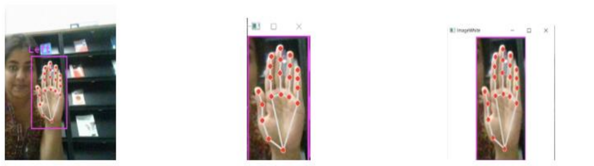
the history. Additionally, ROI is tracked using kernel correlation filters (KCF). The final picture

is a deep convolutional neural network (CNN). where to compare using the CNN model

two modified Alex Net and VGG Net performances. The frequency of training data and

Test results from [91] show 99.90% and 95.61%, respectively. a fresh approach utilising deep convolutional neural network in which the scaled image is sent into the network without being segmented or otherwise processed.

Direct classification of hand gestures requires no detection phases. The technology provides a real-time evaluation and with a complex background 97.1% with a simple background result.



Hand Detection

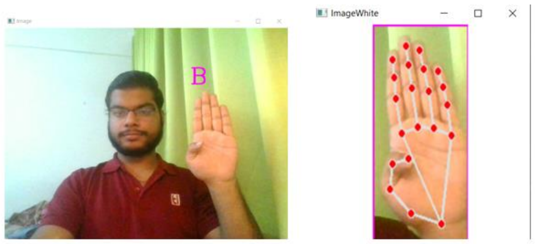
cropped image

Centered against a backgrounc

Fig 5. Hands detection and cropping

## IV. RESULTS

To validate the method proposed in this paper, We have taken a sample from our webcam and it gives the wonderful result and able to detect our hand gesture which provide the information through which it convert into text. All the result attached below,



Alphabet Recognised from Hand Gesture

Cropped Image

Fig 6. Alphabet Recognition

we can notice that the method proposed in this paper has a high accuracy. The overall accuracy has reached 98.41%, which shows the effectiveness of the method. Furthermore, we also report the number of frames per second that can be processed. It can reach the speed of 10fps, which can achieve real-time gesture recognition.

## V. CONCLUSION AND FUTURE WORK

Our project aims for a robust recognition system that will not utilize any markers, hence making it more user friendly and low cost. In this system we aim to provide all the 26 alphabets as per ASL.

However our project is limited to only single hand gestures. In the future we would like to extend our project to both hands in order to formulate letters and also aim at improving accuracy of our current model.

We expect FOANet will continue to advance. The

In a current architecture, temporal fusion is not addressed.

with sophistication. Most networks for gesture recognition overlap.

RNN-based information over time. Despite having a tendency to overfit on brief training sessions,

However, based on empirical evidence, RNNs appear to improve performance, thus we want to incorporate them into FOANet.

## REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248–255. IEEE, 2009.

[2] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, and H. J. Escalante. Chalearn gesture challenge: Design and first results. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, pages 1–6. IEEE, 2012.

[3] D. Pathak, R. Girshick, P. Dollar, T. Darrell, and B. Hariha- ´ ran. Learning features by watching objects move. In Computer Vision and Pattern Recognition (CVPR), 2017.

[4] Pradyumna Narayana, J. Ross Beveridge, Bruce A. Draper
Narayana_Gesture_Recognition_Focus_CVPR_2018_paper.