Hand Gesture Recognition for Emoji and Text Prediction

Usha Kiruthika, Mayank Mohan, Neil Abraham

Abstract– Emoticons' are ideograms and smileys utilized in electronic messages and website pages. Emoticons exist in different classifications, including outward appearances, regular items, places and kinds of climate, and creatures. They are much similar to emojis, however emoticons are real pictures rather than typo graphics. This undertaking perceives the emoticons utilizing hand motions. We are detecting hand gestures and preparing a Convolutional Neural Network (CNN) model on a training dataset. We will make a database of hand gestures and train them. The system utilized here is a CNN. We are utilizing the SIFT filter to identify the hand and CNN for preparing the model. SIFT filter give a lot of highlights of an image that are not influenced by numerous factors, for example, object scaling and rotation. The SIFT filtering procedure comprises of two areas. The first is a procedure to identify intrigue focuses in the hand. Intrigue focuses are the points in the image in a 2D space that surpasses some limit measure and is better than straight forward edge recognition. The second segment is a procedure to make a vector like descriptor and this is the most special and prevalent part of the SIFT filter.

***Index Terms*: Hand gesture recognition, Convolutional neural networks, deep learning, emoticons, text prediction**

I. INTRODUCTION

Emoticon utilization has turned into another type of social correspondence, which is significant in light of the fact that it can improve correspondence frameworks, for example, chat applications. They have turned into a significant piece of our lives to express our feelings. In this paper, we have utilized the SIFT Filter to distinguish our hand gestures and Convolutional Neural Network (CNN) to prepare the model. We are making a database of various hand gestures. After which emoticons are to be shown as we demonstrate our hand gestures on the screen. This work is prepared on a vast dataset of pictures; we can make a framework with programmed attention based learning. This methodology of learning the highlights and making the machine self-sufficient is not quite the same as most different methodologies accessible in the market today and whenever finished effectively will make ready for a lot more noteworthy works later on.

Hand gesture recognition is one obvious approach to construct easy to understand interfaces among machines and their clients. Hand gesture recognition takes into account the activity of complex machines and fast gadgets through only arrangement of hand stances, finger and hand movements, dispensing with the need for physical contact among man and

machine. Gesture recognition on pictures from single camera is a troublesome issue because of impediments, contrasts close by life systems, varieties of stance appearance, and so on. In the most recent decade, a few ways to deal with signal acknowledgment on shading pictures were proposed. As of late, Convolutional Neural Systems (CNNs) have turned into the cutting edge for object detection in computer vision. In spite of high capability of CNNs in article identification and picture division, just a few papers report promising outcomes an ongoing review close by motion acknowledgment reports just a single huge work. A few impediments to more extensive utilization of CNNs are high Computational requests, absence of adequately huge datasets, just as the absence of hand locators appropriate for CNN based classifiers.

In a paper, CNN has been utilized for grouping of six hand signals communicated by people to control robots utilizing smart gloves. In later work, a CNN has been actualized in Theano and executed on the Nao robot. In an ongoing work, a CNN has been prepared on one million of models. Be that as it may, just a subset of information with 3361 physically marked casings in 45 classes of communication through signing is freely accessible. Various explores available gesture recognition have connected element extraction strategies which manage singular casings in a picture arrangement of a hand motion. For each casing, the highlights are remove d from the static hand picture, and these different information are consolidated together to be made sense of what kind of motion was performed. Be that as it may, since this technique takes just the static data at a particular point, it may not be anything but difficult to perceive the definite importance of the dynamic hand motion all in all, which could be differed by the unique situation.

The main objective of this paper is to keep the emoji prediction as accurate to the defined emojis and display it effectively and at an efficient speed. This can be achieved by training a large number of data sets to attain as much perfection to the right emoji prediction on the screen without having to search for that emoji. The prediction will be based on the hand gestures and to capture the gestures to at most precision by training for various positions of the hand movement.

Revised Manuscript Received on September 22, 2019.

Usha Kiruthika, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India.
Mayank Mohan, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India.
Neil Abraham, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India.

II. RELATED WORK

In their work [1], the paper presents a convolutional neural system based calculation for acknowledgment of hand poses on pictures procured by a solitary shading camera. The hand is extricated ahead of time based on skin shading appropriation. A neural system based regressor is connected to find the wrist. At last, a convolutional neural system prepared on 6000 physically marked pictures speaking to ten classes is executed to perceive the hand act in a sub-window decided based on the wrist. The model accomplishes high classification exactness, including situations with an alternate camera utilized in testing. It is demonstrated that the convolutional network accomplishes better outcomes on pictures pre-filtered by a Gabor filter.

In another work [2], visual Elucidation of motions are used in achieving regular Human-Computer Interactions (HCI). In their paper, a technique for perceiving hand motions are proposed. A structured framework which can recognize explicit hand signals is presented and they are used to pass on data. Whenever, a client can display his/her hand completing a particular signal before a web camera connected to a PC. The hand motion of a client is caught and put away it on circle. At that point, we read those recordings caught one by one, changed over them to double pictures and made 3D Euclidian Space of double qualities.

So as to perceive the dynamic signal, hand has to be followed the hand and the movement direction needs to be obtained [3]. The progression previously following hand is to decide the position of hand. Be that as it may, hand arrangement has more than 26 degrees of possibilities, and its appearance changes radically. The proposed strategy depends on such an intuitive mode: individuals demonstrate outstretched hand to the camera; at that point find the position of hand by recognizing of the situation of outstretched hand; and after that start to follow hand while hand can be in any shape to connect with the PC. So this strategy improves the accuracy and precision of the dynamic signal acknowledgment framework.

In another related paper [4], an example acknowledgment model for dynamic hand motion acknowledgment is proposed. Their proposed model consolidates a convolutional neural system (CNN) with a weighted fuzzy min-max (WFMM) neural system; each module performs highlight extraction and highlight examination, separately. The information portrayal proposed in this look into is a spatiotemporal layout which depends on the movement data of the objective article. To process the information, the paper builds up a changed CNN model by broadening the responsive field to a three-dimensional structure. To increment the productivity of the example classifier, they utilize a component investigation system using the WFMM calculation. The trial results demonstrate that the proposed technique can limit the impact brought about by the spatial and transient variety of the element focuses.

Other important works in this field includes recognition in real-time [5-9]. Each of these works use several techniques such as image processing techniques, Principal component analysis and image segmentation. Another work uses self organizing map for hand gesture recognition [10]. In our work we use CNN along with SIFT (Scale-invariant feature transform) for emoji prediction. This has not been used so far in any of the literature that we reviewed. It has been shown to achieve good results in prediction and accuracy.

III. PROPOSED METHOD

Motion innovation pursues a couple of essential states to influence the machine to perform in the most enhanced way. These are:

- Pause: In this express, the machine is trusting that the client will play out a motion and give a contribution to it.
- Gather: After the signal is being played out, the machine assembles the data passed on by it.
- Control: In this express, the framework has assembled enough information from the client or has been given an info. This state resembles a handling state.
- Execute: In this express, the framework plays out the undertaking that has been asked by the client to do as such through the motion.

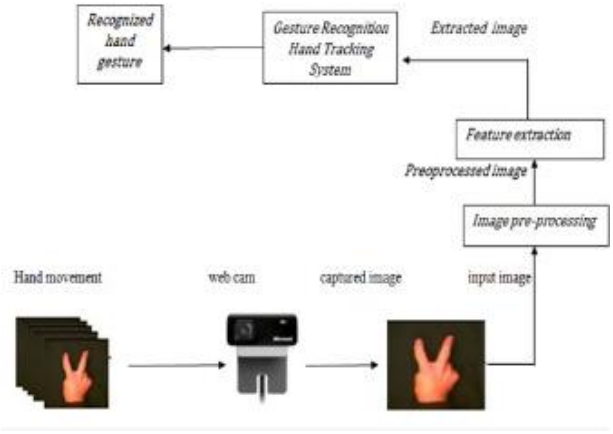

Fig. 1: Overview of the system: The image depicts the flow of the model where it accepts the gesture and processes it by feature extraction

The primary steps that are being followed for emoji prediction are:

- Compression of the image
- Encoding the image
- Finding key points in hand using SIFT filter.
- Comparing with data set using CNN.
- Prediction of Emoji based on detected gesture.

A. Camera and Gesture Input

We use a camera to track real time gestures created by the user. The user can create any hand gestures which will be recorded by the camera and will be sent further for processing in MATLAB. The recorded pictures are taken in and compared with the dataset for the prediction of that gesture.

B. CNN Algorithm

Using the Convolution Neural Network, we can train the dataset of hand gestures and convolution is the first layer to extract features from an input image. Convolution preserves the relationship between pixels by learning image features using small grids of input data. Convolution is a mathematical operation that takes two inputs. The 2 inputs are: image matrix and a filter.

C. SIFT Algorithm

The SIFT process consists of two sections. The first section is to detect interest points in an image. Interest points are where the signal in 2D space has variation that exceeds some threshold criterion and is better than simple edge detection. The second section is a process to create a vector like descriptor. To create scale invariance, the interest points are scanned at a wide range of scales and the scale where the interest point features.

D. Emoji and Text Prediction

After using the CNN and SIFT algorithms, we can successfully and accurately predict the gesture and text based on the requirement specified. The extent of accuracy is very much high as compared to other algorithms used to predict the outcome.

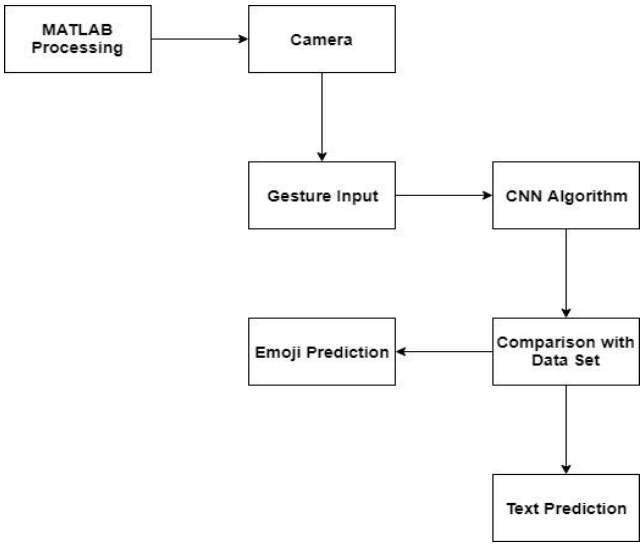

Fig. 2: System Architecture: The above image depicts the different modules of the model that it undergoes to detect gestures for predictions.

IV. TECHNIQUES USED

Convolutional neural system (ConvNets or CNNs) is one of the primary classifications to do pictures acknowledgment, pictures orders. Items location, acknowledgment faces and so on, are a portion of the regions where CNNs are broadly utilized. CNN image classification takes an info picture, processes it and categorizes it under specific classifications. Machines see an info picture as cluster of pixels and it relies upon the picture goals. In light of the picture goals, it will see h x w x d (h = Height, w = Width, d = Depth).

CNNs utilize moderately little pre-handling contrasted with other picture arrangement calculations. This implies the system learns the channels that in conventional calculations were hand-built. This freedom from earlier information and human exertion in highlight configuration is a noteworthy favorable position. A convolutional neural system comprises of an information and a yield layer, just as different concealed layers. The shrouded layers of a CNN commonly comprise of convolutional layers, RELU layer for example initiation work, pooling layers, completely associated layers and standardization layers. Portrayal of the procedure as a convolution in neural systems is by show. Numerically it is a cross-connection as opposed to a convolution. This just has centrality for the lists in the framework, and in this manner which loads are put at which index.

Filter makes scale, rotational and introduction invariant picture descriptors, however depictions of SIFT don't come in all scales. They are either paltry or need full detail, potentially why the inquiries was inquired. Here is a shot at a less paltry yet minimal depiction in any case.

The SIFT procedure (Fig. 3) comprises of two areas. The first is a procedure to recognize intrigue focuses in a picture. Intrigue focuses are the place the flag in 2D space has variety that surpasses some limit foundation and is better than straightforward edge identification. The second segment is a procedure to make a vector like descriptor and this is the most remarkable and predominant part of SIFT. To make scale invariance, the intrigue focuses are examined at a wide scope of scales and the scale where the intrigue point highlights, i.e. the 2D space savvy flag fluctuations, meet certain dependability criteria, is chosen as the one that gets coded by the descriptor. Along these lines when a similar procedure is kept running on a competitor intrigue point for acknowledgment, it will get a similar portrayal paying little respect to real scale, as the scale with the most steady nearness of the premium point highlights will be coded. Along these lines we have scale invariance. Next, a similar intrigue focuses are assessed for a course. This is the 2D bearing in which the flag fluctuation highlights have the most elevated changeability.

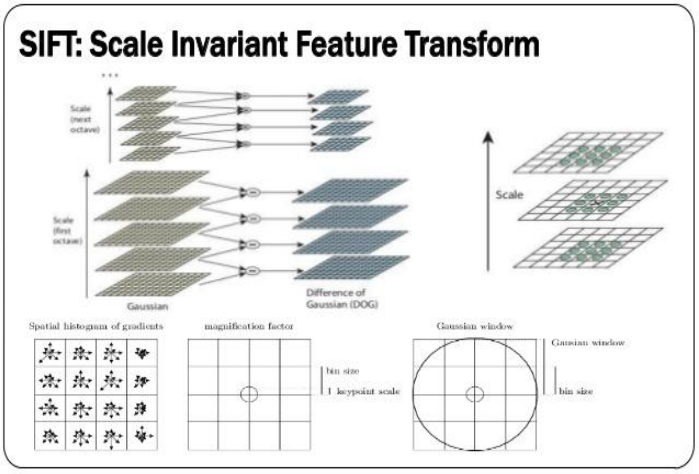

Fig. 3: A Typical SIFT Filter: The image shows the various feature extraction layers a SIFT Filter uses.

In this paper, to implement the system, different technologies have been used for different. We used MATLAB, TensorFlow and Large scale hand gesture Database.

V. RESULTS AND DISCUSSION

A. Feature Extraction

Not all highlights will use in the classification learning process. Just highlights that have a standard deviation and a unique normal chose for preparing. The GMM will be used to see which highlights will be chosen as preparing information. The explanation behind utilizing GMM is its capacity to order fragmented information. There are ventures to utilize the Gaussian appropriation in checking highlights. The means are:

1. Scanning for the normal estimation of a component,
2. finding the standard deviation estimation of a component,
3. finding the ordinary dispersion esteem, trailed by plotting information on a graph utilizing the likelihood thickness function (PDF), and aggregate conveyance work (CDF).

B. Classification Learner

The following stage is making expectations, we have to calculate the likeness between two instances of accessible information. This is necessary so we can find the most comparative information tests in the preparing dataset. The forecast will utilize cosine, Euclidean and cubic separation metric. It will be using k-NN with Euclidean distance metric. The number of neighbours for this metric is 1. This defined as the square root of the sum of squares difference between two numbers arrangement.

C. Feature Comparison

It takes place in two stages:

Pause: In this express, the machine is trusting that the client will play out a motion and give a contribution to it. Gather: After the signal is being played out, the machine assembles the data passed on by it. Control: In this express, the framework has assembled enough information from the client or has been given an information. This state resembles a handling state.

Execute: In this express, the framework plays out the undertaking that has been asked by the client to do as such through the motion.

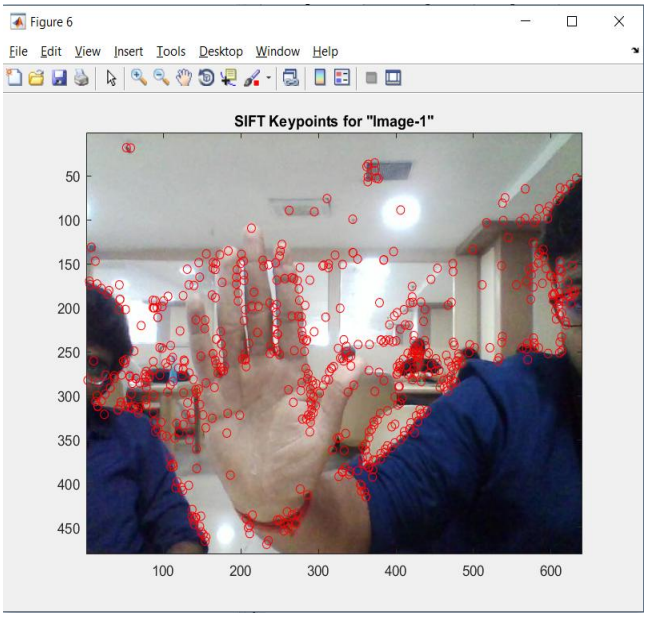

Fig. 4: SIFT Keypoints in the image of a hand

As can be seen from the figure (figure 4) the SIFT keypoints are detected from any image. This helps to identify the edges of the hand and to recognize the hand without the background images and other noise. The extracted image is then sent to CNN for recognition and getting the corresponding emoji.

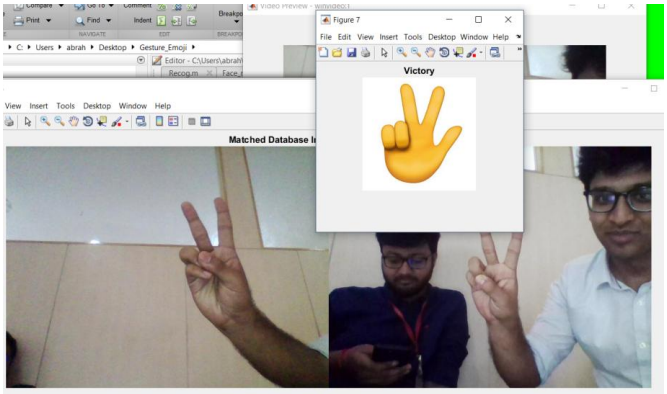

Fig. 5: Prediction of victory emoji from image of hand

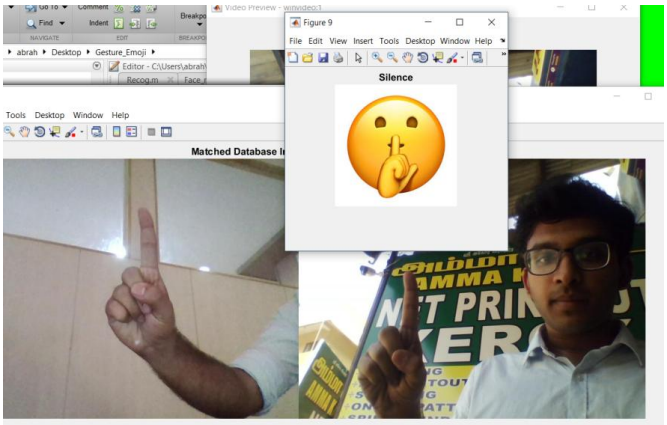

Fig. 6: Prediction of silence emoji from image of hand

As shown in figures (Fig. 5 and Fig. 6) our model correctly identifies the image of a hand and the gesture shown by the hand. Then it gives the correct emoticon that corresponds to the hand.

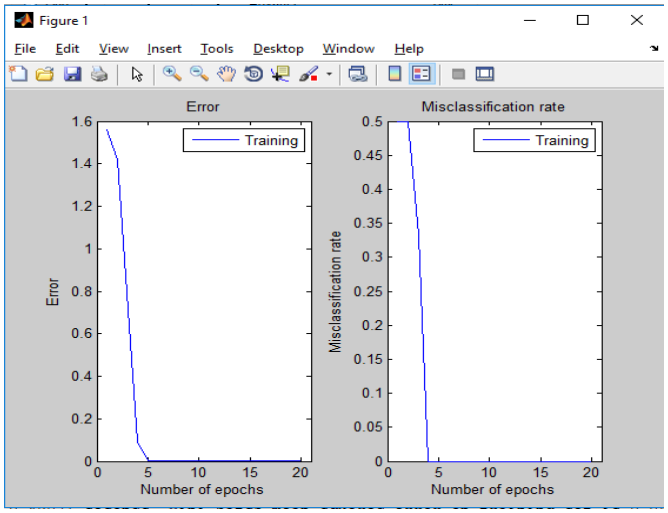

Fig. 7: Training Graph: The graph depicts the training discrepancies or errors with increasing number of epochs.

The Fig. 7 shown below , is obtained after the training of our model. The graph represent the error rate and misclassification rate during training of our model. As the no. of epochs get increases the training model gets more accuracy as both the error rate and misclassification rate is reduced. The Error rate becomes constant after a given period due to the implementation of SIFT Filter.

VI. CONCLUSION

We have demonstrated the feasibility of using the SIFT (Scale Invariant Feature Transform) algorithm for feature extraction along with CNN (Convolutional Neural Network) is one of the most accurate ways for precise gesture recognition. The demonstration has focused on three elements, first is about the common observable properties that are associated with neural activity play a direct and obvious role in our capacity to recognize gestures under our method. It is inferred that there are some drawbacks of each algorithm used but steerable filter performed the best. Second, that it maintains that the central challenge with our method is the accuracy and consistency of the results. Finally, it is shown that SIFT (Scale Invariant Feature Transform) along with CNN produced the most satisfactory results in our method. In future work, we would like to work on more gestures to obtain best possible results.

There are several applications which can use our method:

Smart TVs: Hand gesture makes lives easier by giving us the ability to move past through channels, menus and other functions of televisions with the help of simple hand gestures such as swipes, etc. Gaming Applications: Gaming without the hassle of incorporating unnecessary wires and joysticks. Other than this it can be used in E-commerce, gesture prediction in mobiles, etc.

REFERENCES

1) D. N. Fernández and B. Kwolek, "Hand Posture Recognition Using Convolutional Neural Network," in *Iberoamerican Congress on Pattern Recognition*, 2017, pp. 441-449.
2) G. Murthy and R. Jadon, "Hand gesture recognition using neural networks," in *2010 IEEE 2nd International Advance Computing Conference (IACC)*, 2010, pp. 134-138.
3) X. Jiang, *et al.*, "A dynamic gesture recognition method based on computer vision," in *2013 6th International Congress on Image and Signal Processing (CISP)*, 2013, pp. 646-650.
4) H.-J. Kim, *et al.*, "Dynamic hand gesture recognition using a CNN model with 3D receptive fields," in *2008 international conference on neural networks and signal processing*, 2008, pp. 14-19.
5) Z.-h. Chen, *et al.*, "Real-time hand gesture recognition using finger segmentation," *The scientific world journal*, vol. 2014, 2014.
6) Y. Fang, *et al.*, "A real-time hand gesture recognition method," in *2007 IEEE International Conference on Multimedia and Expo*, 2007, pp. 995-998.
7) R. Agrawal and N. Gupta, "Real time hand gesture recognition for human computer interaction," in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, 2016, pp. 470-475.
8) E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE transactions on intelligent transportation systems*, vol. 15, pp. 2368-2377, 2014.
9) J. A. Zondag, et al., "Practical study on real-time hand detection," in 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009, pp. 1-8.
10) N. N. Bhat, *et al.*, "Hand gesture recognition using self organizing map for Human Computer Interaction," in *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2013, pp. 734-738.

AUTHORS PROFILE

Usha Kiruthika is currently working as an Assistant Professor in the Department of Computer Science and Engineering in SRM Institute of Science and Technology, Kattankulathur Campus. She has completed her Ph.D at Anna University and she has many publications to her credit. Her areas of interest include artificial intelligence and deep learning.

Mayank Mohan is a B.Tech student in the Department of Computer Science and Engineering at SRM Institute of Science and Technology in Kattankulathur.

Neil Abraham is a B.Tech student in the Department of Computer Science and Engineering at SRM Institute of Science and Technology in Kattankulathur.