Pose Invariant Hand Gesture Recognition using Two Stream Transfer Learning Architecture

Anjali R. Patil, S. Subbaraman

Abstract*: *The hand gesture detection problem is one of the most prominent problems in machine learning and computer vision applications. Many machine learning techniques have been employed to solve the hand gesture recognition. These techniques find applications in sign language recognition, virtual reality, human machine interaction, autonomous vehicles, driver assistive systems etc. In this paper, the goal is to design a system to correctly identify hand gestures from a dataset of hundreds of hand gesture images. In order to incorporate this, decision fusion based system using the transfer learning architectures is proposed to achieve the said task. Two pretrained models namely 'MobileNet' and 'Inception V3' are used for this purpose. To find the region of interest (ROI) in the image, YOLO (You Only Look Once) architecture is used which also decides the type of model. Edge map images and the spatial images are trained using two separate versions of the MobileNet based transfer learning architecture and then the final probabilities are combined to decide upon the hand sign of the image. The simulation results using classification accuracy indicate the superiority of the approach of this paper against the already researched approaches using different quantitative techniques such as classification accuracy.

Keywords: Convolutional Neural Networks; edge map images; hand gestures; ROI, Transfer learning; YOLO;

I. INTRODUCTION

Human motions establish a space of movement communicated by the body, face, as well as hands to pass on some important data. Among an assortment of signals, hand gesture is the most expressive and most widely utilized gesture. Motions have been utilized as an elective type of info methodology to interact with computers in a simple manner. The as of now utilized machine association instruments viz. Mouse, joystick, console, electronic pen and so forth are not adequate with the improvement and acknowledgment of virtual condition. Hand signal has the characteristic capacity to speak to thoughts and activities in all respects effectively. Consequently utilizing these diverse hand shapes provides progressively common interface to the computer system. This kind of normal collaboration is the center of vivid virtual situations. Despite the fact that the hand motions change significantly among various regions and diverse social viewpoints with setting or data to be passed on, they are normally utilized in correspondence. The noteworthy utilization of signals in our day by day life as a method of connection propels the utilization of gestural interface in wide scope of use through computer vision.

Many machine learning techniques have been utilized and revealed in the literature to solve the issues involved in real time hand gesture recognition. These include template matching algorithm [1], Hidden Markov Model [2], Neural Networks [3], Finite State Machines [4], Fuzzy [5], Genetic algorithms [6] and Support Vector Machine [7]. Most machine learning methods extract the complex and crafted features from the raw input images/videos and train the system, while convolutional neural network, which is the type of deep learning model works on the images itself bypassing the feature extraction stage. Convolutional neural network are popularly utilized for pattern recognition. This paper aims to design a system to correctly identify hand gestures using deep learning techniques. In this paper we propose the transfer learning architecture for training the system. The decision fusion based system using the transfer learning architectures to achieve the said task has been presented. For this purpose MobileNet and Inception V3 Models are used. The region of interest in the image is found out using the YOLO (You Only Look Once) [8] architecture to decide upon the type of model. An approach of learning of the edge map images and the spatial images separately using the two versions of the MobileNet based transfer learning architecture and then combining the final probabilities to decide upon the hand sign of the image is presented. The results obtained through the simulation demonstrates the highest accuracy when compared with distinctive other state of the art techniques using different quantitative analysis indicate that the proposed framework is better.

The remainder of the paper is sorted out as pursues: Section II provides the review of work done in hand motion recognition utilizing deep learning techniques. Section III highlights on the theoretical background of the proposed architecture followed by the experimental set up and discussion in Section IV. The outcomes are talked about in Section V. End and future work bearings are referenced in Section VI.

II. LITERATURE REVIEW

Deep learning, a generally ongoing way to deal with AI, including neural systems with more than one hidden layer, has been utilized with much achievement in face recognition, speech recognition, action recognition and natural language processing tasks and so forth [9]. Deep learning offers new potential outcomes to oblige with machine and to plan progressively normal and increasingly instinctive associations with processing machines.

Revised Manuscript Received on October 20, 2019.

***** Correspondence Author

Mrs. Anjali R. Patil, Assist. Prof., Electronics Engineering, DKTE's Textile and Engineering Institute, Ichalkaranji, Maharashtra, India E-mail: anjalirpatil@gmail.com

Dr. S. Subbaraman Professor, Electronics Engineering Department, Walchand College of Engineering, Sangli, Maharashtra, India E-mail:s.subbraaman@gmail.com

It also provides new possibilities to interact with machine and to design more natural and more intuitive interactions with computing machines. In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Models are prepared by utilizing an enormous arrangement of labeled data and neural network information and neural system designs that contain numerous layers. Among these deep networks, convolutional neural networks (CNN or ConvNet) are popular [10]. A CNN convolves features with input information, and utilizes 2D convolutional layers, making this design appropriate to handling 2D information, for example, images. The CNN works by extracting features directly from images thereby eliminating the need for manual feature extraction generally required for classification of images. The relevant features are not pretrained; they are found out while the system prepares on a collection of pictures. This automated feature extraction makes deep learning models exceptionally accurate for computer object classification mostly used in computer vision applications.

Most deep learning applications use the transfer learning approach, a process that involves fine-tuning of pretrained model. These work with an existing network, such as AlexNet, VGGNet, GoogLeNet, MobileNet etc. and feed in new data containing previously unknown classes. After making some tweaks to the network, a new specific task, such as classifying hand gestures instead of 1000 different objects is obtained. This also has the advantage of requiring comparatively less data (e.g. processing thousands of images, rather than millions), so as to result into faster computation Transfer learning requires an interface to the internals of the pre-existing network, so that it can be modified and enhanced for the new task [11].

This section highlights on some of the deep learning techniques utilized in the area of hand gesture classification by the earlier researchers.

F. Florez et al. [12] proposed a structure capable of characterizing hand posture and its movement dependent on self organizing neural network topology. It determined posture, while its adjustment elements all through time decided motion. They have approved their framework with 12 signals, some of which were fundamentally the same as, and have acquired high success rates.

H.J. Kim et al. [13] proposed dynamic hand motion recognition by consolidating a CNN with a weighted fuzzy min-max (WFMM) neural system; every module include extraction and analysis of relevant features individually dependent on 'Motion History Images' (MHI). They revealed normal exactness of 87.5% for all out six unique motions. Nagi J. et al. [14] reported max pooling CNN (MPCNN). Colored gloves were used for detecting hand gestures. The hand shape was recovered by shading division and afterward smoothened by morphological operations. The accuracy reported of the MPCNN for 6 motion classes was 96%.

P. Molchanov et al. [15] proposed a methodology utilizing 3D Convolutional Neural Networks (3D-CNN) for drivers' hand signal identification, working with VIVA hand motion dataset. The dataset was caught utilizing a Kinect gadget under certifiable driving settings incorporating variation in the background, illumination and with occlusions. Their outcomes uncovered that the proposed 3D-CNN accomplished 77.5% right order.

V. John et al. [16] implemented the hand motion identification framework for sensible vehicle. Deep learning techniques were utilized to extract the representative frames from the video sequence. Deconvolution neural network (Deconvnetn) and long-term recurrent convolution network used in the HGR system. Their outcomes uncovered an order precision of 91%.

In another CNN based approach, X. Yingxin et al. [17] reported the method which automatically extracts the spatial and semantic feature of hand gesture. Their examinations were led with both the Cambridge Hand Gesture Dataset (9 classes) and also the self developed dataset (5 classes). Their results discovered that the CNN approach achieved higher recognition accuracy than the SIFT+SVM predefined feature approach. Static hand gestures whose options varied in scale, rotation, translation, illumination, noise, and background were recognized with 98.2% accuracy.

Rocco et al.[18] planned six CNN styles to understand three categories of hand motions: "open", "shut" and "obscure". Six structures were executed with varying hyper parameters and depth. All the architecture was observed and the best was selected. The author reported 73.7% accuracy for the best designed neural n/w.

Y. Kim et al. [19] utilized deep learning procedure with Doppler radio detection and ranging for hand-based signal identification. Direction and separation among hand and radar devices were obtained. Their investigations accomplished an arrangement precision of 85.6% for 10-class hand signal characterization, and 93.1% for 7-class hand motion classification.

Oyeb Ade et al. [20] applied deep learning-based systems, for example, CNN and stacked denoising auto encoder (SDAE) to the assignment of perceiving twenty four American Sign Language (ASL) hand motions got from an open database in particular Prima. They reported the accuracy 91.33 % (CNN) and 92.83 % (SDAE).

Strezoski G. et al. [21] displayed ConvNets comprising of various layers of neurons that are spread hierarchically. Classification was obtained with different connection patterns and with varying weights. The experiment was performed on Sebastian Marcel dataset comprising of six gestures. They proclaimed most noteworthy accuracy for GoogleNet of 90.41%.

S. Hussain et al. [22] implemented transfer learning architecture which was implemented with the pretrained CNN on large dataset. A strategy was proposed which controls the computer through six static and eight unique hand signals. The three principle steps were: hand shape identification, tracing of the hand and converting the gesture into appropriate commands. Analyses show 93.09% precision rate.

The adapted version of CNN (ADCNN) was proposed by Ali A. Alani et al.[23]. ADCNN `model which was engaged by the nearness of system instatement (ReLU and Softmax) and L2 Regularization to take out the issue of information over fitting. The authors claimed high accuracy of 99.73% consolidated varieties in features, for example, rotation, translation, scale, and noise.

According to research work described in above section, it has been found that CNN are very effective when they are used to solve problems related to computer vision. The above reports by earlier researchers gave the vital foundation to leading our experiments and triggered our work. This paper implements the detection of hand gestures under uncontrolled environment and with different poses and different illumination conditions. This paper implements the detection of the region of interest (ROI) using YOLO architecture. ROI given by the YOLO is being given to the different models. The contribution proposes to use the inception V3 based architecture for images with larger ROI and a novel decision-based algorithm for images with smaller/no ROI. The pretrained networks MobileNet 0.5 , MobileNet 1.0 for the smaller ROI images with edge map of the image Probabilities obtained by both models are used to calculate the joint probability which is used as decision parameter to classify the images of hand gestures.

III. THEORETICAL BACKGROUND

This area gives the essential hypothesis of Convolutional Neural Network (CNN). CNN are utilized in an assortment of areas, including speech, image, and pattern recognition, normal language processing, and video processing. These areas of deep learning are winding up increasingly significant. Previously in almost all the pattern recognition techniques, feature extractors are hand designed. While in CNN, the weights of the convolutional layer utilized for feature extraction. While fully connected layers of CNN are used as classifier. The improved system structures of CNN lead to memory necessities and computational complexity prerequisites and, simultaneously, give better execution for applications where the information has the local connection. An exemplary CNN architecture is exhibited in Figure 1, CNN design comprises of 4 layers: in particular convolutional layer, pooling layer, Rectification Linear unit layer and fully connected layer.

Fig.1: Architecture of CNN

Layers of CNN-

By stacking multiple different layers in a CNN, complex designs are worked for classification issues. Various layers of CNN are examined in the accompanying subsections:

A. Convolution layers

The purpose of this layer is to extract significant feature information from the input image. There can be many layers. Low level features such as edges, lines, corners etc are extracted in the very first layers. Later stages of the convolutional layers are used to extract higher-level features. Following equations represents the typical Convolution process:

$$f[X,Y] * g[X,Y] = \sum_{n1=0}^{N} \sum_{n2=0}^{N} f(x,y).g[(x-n1),(y-n2)] \quad (1)$$

In (1), $f(x,y)$ is input image and $g(x,y)$ is kernel function. Kernel can be any filter which is used for edge detector, color detector, and curves detector etc.

B. Pooling/Subsampling Layers

CNN layer provide more number of features. If they are directly given to the classifier require very large computational overhead, particularly for large-size high-resolution images. The high dimensional features extracted from high resolution images require massive computational resources and may turn to heavy over fitting issue. However, since the image include "static" attribute, the feature obtained in a local region of the image is highly correlated with the features in another local area. Thus, it is potential to carry out comprehensive statistical operations on attributes of the various areas in a neighborhood the picture, which is alluded to as "pooling".

The main task of pooling layer is lowering the resolution of the features extracted in the convolutional layer. It makes the features sturdy against noise and distortion. Pooling is obtained using two ways: max pooling and average pooling. In each case, the input is split into non-overlapping 2-D spaces. For instance, let's say the input is of size 4x4 then for 2x2 subsampling, a 4x4 image is split into four non-overlapping matrices of size 2x2. Within the case of max pooling, the maximum value of the four values in the 2x2 matrix is the output. In the case of average pooling, the average of the four values is the output. Figure 2 demonstrates the case of max pooling and average pooling.

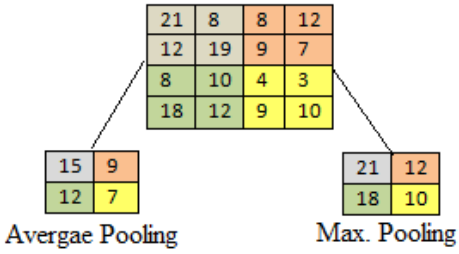


Fig. 2: Representation of Max Pooling and Average Pooling

C. ReLU Layers

The next layer is Rectification Linear Unit layer. The ReLU has yield output zero if the input information is smaller than zero and raw output otherwise. If the input is larger than zero the output is adequate to input.

$$ReLU = \begin{cases} 0 & if \ x < 0 \\ x & if \ x \geq 0 \end{cases} \quad (2)$$

It expands the nonlinear properties of decision function. It also enhances the characteristics of the overall network without touching the receptive fields of the convolutional layer. In contrast to the opposite non-linear function employed in CNN, the advantage of a ReLU is that the networks train repeatedly quicker.

D. Fully Connected Layers

Last layers of CNN are typically fully connected layers. These layers numerically aggregate a weighting of the past layer of features, showing the precise blend of "ingredients" to decide a regular target outputs.

In fully connected layer, all the elements of all the features of the previous layer get utilized in the estimation of every component of each output feature.

IV. EXPERIMENTAL SET-UP AND WORK

From the literature it is observed that CNNs are hardly trained from scratch (with random initialization), since it is generally uncommon to have a dataset of adequate size. Rather, it is entirely expected to pre-train a ConvNet on an extremely enormous dataset (for example ImageNet, which contains 1.2 million pictures with 1000 classifications), and afterward utilize the ConvNet either as initialization or a fixed feature extractor for the interested task.

The proposed framework system can be visualized in two modules: Training and testing i.e. gesture recognition. The block diagram for training the framework is as appeared in Figure 3.

In training, a new strategy wherein three different modules corresponding to MobileNet and inception V3 architectures are trained. For this training, 10 different classes of images corresponding to 10 different hand signs are used. Each class of images contain data corresponding to same sign but in all possible poses. This helps in designing a pose invariant model for hand sign detection. Transfer learning is implemented due to availability of smaller dataset. The images of all 10 classes are trained using “MobileNet 0.5” pre-trained model to generate the training model 1. The training is also done with the same images using “MobileNet 1.0” pre-trained model to generate model 2. The model 2 is much deeper than model 1. Also two extra layers for re-training in model 2 are released as compared to model 1.

For some images edge detection is performed on the images to generate the edge map of the dataset. These edge maps are used as training data for another pre-trained model named “inception V3” to generate model 3. Generation of three models completes the process of training. We then use these trained models in the proposed system for hand sign detection. The above discussed system for hand sign detection is shown in Figure 4.

Region of interest in the test image is detected using the YOLO architecture. Authors then estimated the size of the ROI detected. The size of the ROI is then threshold to choose the further path in the system.

If size of ROI is above the threshold, path1 is followed as shown in Figure 2. The test image is given to model 1 and the probability of test image belonging to each of the classes is obtained. This probability is then fed to the hypothesis to decide the final class of the image.

While if size of ROI is below the threshold or if ROI is not detected, then the image has very small portion of hand or the image is blurry in nature. In such cases, testing is done through two stream implementation rather than one stream implementation. Also we propose to use deep neural networks rather than shallow neural networks for better accuracy.

If size of ROI is below the threshold or if the ROI is not detected, we choose path 2 as shown in Figure 2. Edge detection is performed on the test image to generate its corresponding edge map. The test image is supplied to Model 2 and the edge map to Model 3. The probabilities of test images belonging to each class from both the models are obtained. These probabilities are then combined to generate the joint probability which is sent to the hypothesis to decide the class of the test image.

The implementation details of the system are shown in Table I. The plots of weights after the convolution layer 1 and convolution layer 2 are presented in Figure 5 and 6. In actual CNN, each layer comprises various sub layers of Neurons operating in parallel on the previous layers. Examples are given in Figure 5 and 6 which shows the filters learned in layer 1 and layer 2. Each filter elaborates the three input color bands producing by convolution a corresponding to a feature map.

Table I: Implementation details of the system

Parameters	Inception V3	MobileNet
Bottlenecks	900	900
Cross entropy	0.0084	0.001122
Training Accuracy	1	1
Validation Accuracy	0.95	0.95
Test Accuracy	0.92	0.93

Configuration of CNN for transfer learning:

1. The number of layers left free for retraining in MobileNet and Inception V3: 2
2. Number of epochs used to converge:
   MobileNet -459
   Inception V3-487
3. Cross-entropy threshold for convergence: 0.000000001
4. Final cross entropy:
   Inception V3 (0.0084) and MobileNet (0.001122)

V. RESULTS AND DISCUSSIONS

The proposed system shows improvement in accuracy of categorization and estimation of pose invariant hand sign detection as compared to the state of the art systems.

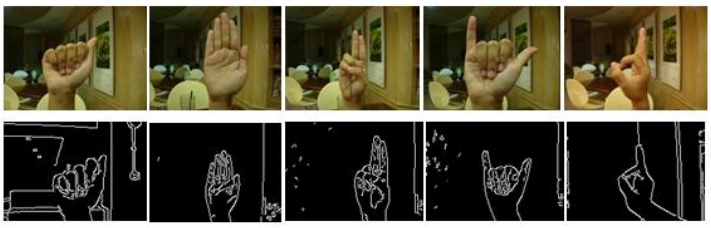


Fig. 7: Samples from the NUS dataset and respective Edge map

We consider 200 images of each class for testing from. NUS hand gesture dataset [24]. The results for 10 class classifications are shown in tables II, III, and IV. MobileNet shows an accuracy of 91.5%, inception V3 shows an accuracy of 90%.

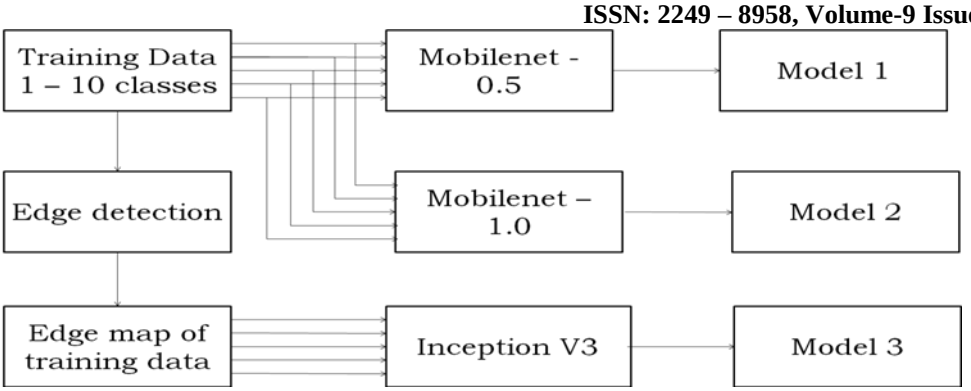


Fig. 3: Training for the system

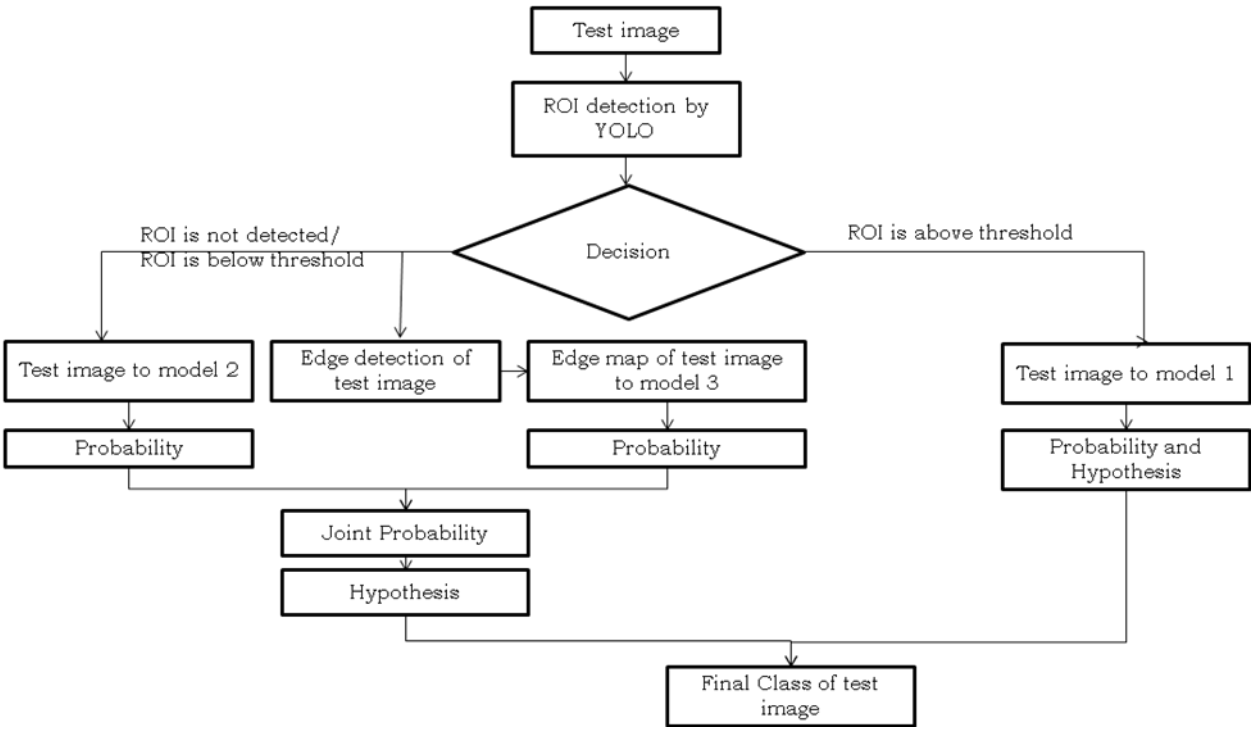


Fig. 4: Hand Sign Recognition System

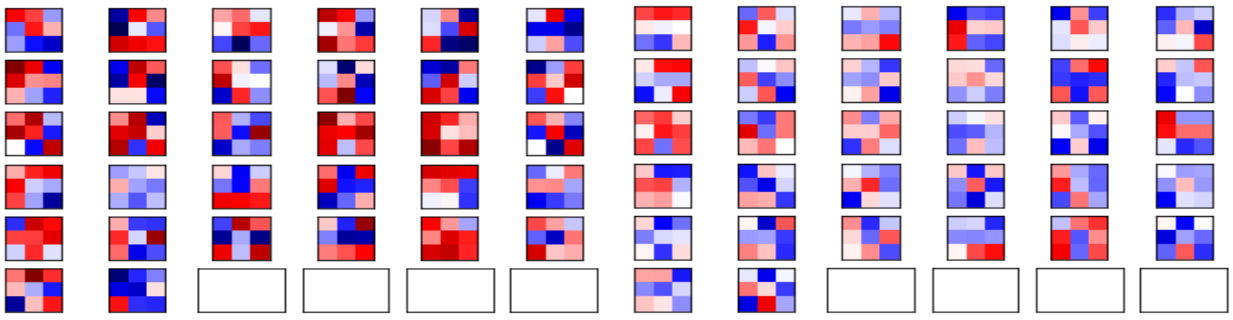


Fig. 5: Weights after convolution layer 1. **Fig. 6: Weights after convolutional layer 2.**

Table II: Confusion matrix generated using the proposed algorithm.

G	1	2	3	4	5	6	7	8	9	10
1	197	0	0	0	0	0	1	2	0	0
2	0	198	0	0	0	0	0	0	0	2
3	1	0	192	2	0	1	1	3	0	0
4	0	0	0	196	2	0	0	0	1	1
5	0	0	0	0	200	0	0	0	0	0
6	0	1	0	1	0	195	0	3	0	0
7	0	1	0	0	0	3	196	0	0	0
8	0	0	3	0	0	1	2	194	0	0
9	2	1	1	0	1	1	0	0	194	0
10	1	0	0	1	0	0	0	0	0	198

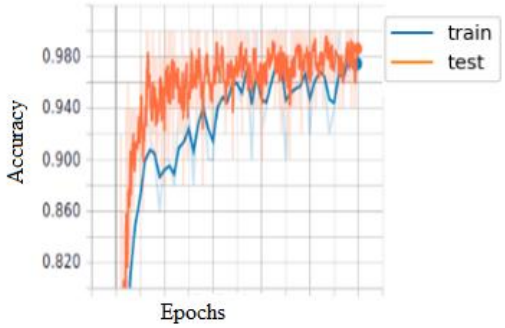
Table III: Confusion matrix generated using MobileNet architecture

G	1	2	3	4	5	6	7	8	9	10
1	176	0	4	0	12	0	1	4	2	1
2	0	184	1	3	2	3	2	0	3	2
3	5	2	190	0	0	1	0	2	0	0
4	4	2	0	181	3	5	1	1	3	0
5	2	7	1	1	176	2	3	2	1	5
6	2	0	3	1	0	190	0	1	3	0
7	7	0	0	3	2	1	183	0	0	4
8	2	1	0	3	4	2	1	185	1	1
9	10	3	2	0	2	0	0	3	178	2
10	0	0	2	1	2	0	2	1	5	187

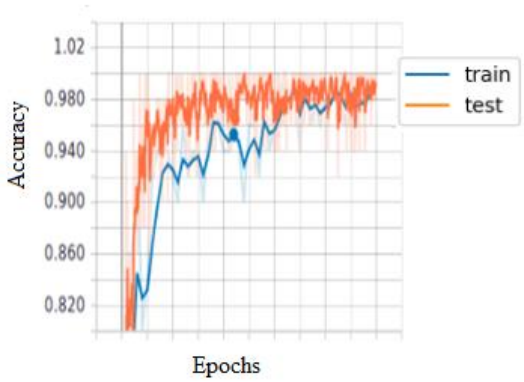
Table IV: Confusion matrix generated using Inception V3 architecture

G	1	2	3	4	5	6	7	8	9	10
1	180	2	0	1	5	2	0	10	0	0
2	3	177	6	2	0	1	5	0	0	6
3	4	0	176	1	5	3	1	3	2	5
4	0	5	1	185	2	1	3	2	0	1
5	3	0	0	0	180	5	2	10	0	0
6	0	5	2	5	0	183	0	0	5	0
7	0	0	0	2	0	0	178	0	18	2
8	6	0	0	0	7	0	3	184	0	0
9	0	18	1	0	1	1	3	0	175	1
10	1	0	0	1	0	1	4	2	9	182

Figure 8 summarizes the learning curves of the model, showing accuracy for both without and with decision model on both the trained (blue) and tested (orange) datasets at the end of each training epoch. In this case, we can see that the proposed model learned the problem reasonably quickly and well, and remaining reasonably stable. It can be seen from the graph that training and testing errors are reasonably smoother for the reported system with decisioning.



a)



b)

Fig. 8: Learning curve a) without decisioning b) with decisoning

To prove the superiority of machine learning, the experiment of hand gesture recognition was also performed with SVM learning. The results of both approaches are compared and are presented in Table V below. It is observed that the proposed two stream transfer learning architecture provides 98% accuracy which is 30 % more than SVM learning.

Table V: Comparison with SVM

Parameters	Proposed architecture	SVM learning
Training Accuracy	100.00%	98%
Validation Accuracy	100%	82%
Validation Loss	0.351	NA
Time elapsed (Training)	00:16:26	00:05:15
Accuracy on Test-Set	98% (940 / 960)	67.7%(650/960)

VI. CONCLUSIONS AND FUTURE SCOPE

This paper investigates the deep learning based techniques utilized for robust hand sign recognition invariant to scale, rotation, translation, view, and illumination. Transfer learning based CNN has been discussed in this paper. The technique was made robust by avoiding a strategic distance from skin color based segmentation, blob detection, skin region cropping and shape feature extraction for hand gestures.

The model is tested on NUS hand gesture dataset which incorporates 10 gesture vocabularies with noteworthy accuracy of 98%. The accuracy is improved using joint probability of three models namely MobileNet 0.5, MobileNet 1 and Inception v3. It is now evident that the project can find its applications in many fields where accurate recognition is expected.

The reported system also opens the door to multiple pathways. The reported algorithm can be extended to detect multiple signs in a single image and also to identify two-handed gestures. The techniques such as auto-encoders can be used for classification.

REFERENCES

1. Tolentino, Roselito, " Application of Template Matching Algorithm for Dynamic Gesture Recognition of American Sign Language Finger Spelling and Hand Gesture," Asia Pacific Journal of Multidisciplinary Research.2014
2. M. Elmezain, A. Al-Hamadi, J. Appenrodt and B. Michaelis, "A Hidden Markov Model-based continuous gesture recognition system for hand motion trajectory," 2008 19th International Conference on Pattern Recognition, Tampa, FL,2008, pp.1-4 doi:10.1109/ICPR.2008.4761080.
3. Stergiopoulou E., Papamarkos N., Atsalakis A. (2005), " Hand Gesture Recognition Via a New Self-organized Neural Network," In Sanfeliu A., Cortés M.L. (eds) Progress in Pattern Recognition, Image Analysis, and Applications. CIARP 2005. Lecture Notes in Computer Science, vol. 3773. Springer, Berlin, Heidelberg
4. P. Hong, M. Turk, and T. S. Huang, "Gesture modeling and recognition using finite state machines," in Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recogn., Grenoble, France, Mar. 2000, pp. 410–415.
5. M. Su., "A fuzzy rule-based approach to Spatio-temporal hand gesture recognition," IEEE Transactions on Systems, Man, and Cybernetics Part C, 30(2):276-281, 2000.
6. H, L. Liu and L. Shao, "Synthesis of Spatio-Temporal Descriptors for Dynamic Hand Gesture Recognition Using Genetic Programming," Published in Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops
7. Nasser H. Dardas and Nicolas D. Georganas, "Real-Time Hand Gesture Detection and Recognition using Bag-of-Features and Support Vector Machine Techniques," IEEE Transactions on Instrumentation and Measurement, vol.60, no.11, November 2011.
8. Z.Ni,J.Chen,N.Sang, C Gao, L.Liu, "Light YOLO for High-Speed Gesture Recognition", 2018 25th IEEE conference on Image Processing ICIP, Athens, pp 3099-3103.
9. X. Han and Q. Du, "Research on Face Recognition based on Deep Learning," 6th International Conference on Digital Information, Networking, and Wireless Communications (DINWC), *Beirut, 2018, pp. 53-58.*
10. C.Szegedy, W.Liu, Y. Jia, P.Sermanet, S.Reed, D.Anguelov, D.Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015
11. H. Shin et al., "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," in IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1285-1298, May 2016.
12. F. Flórez, J. M. García, J. García, and A. Hernández, "Hand Gesture Recognition Following the Dynamics of a Topology-Preserving Network," in Automatic Face and Gesture Recognition, Proceedings. 5th IEEE International Conference, 2002, pp. 318–323.
13. H.J. Kim, J. S. Lee, and J.H. Park, "Dynamic Hand Gesture Recognition using a CNN Model with 3D Receptive Fields," International Conference on Neural Networks and Signal Processing, 2008.
14. Nagi, J., Ducatelle, F., Caro, G.D., et al., "Max-Pooling Convolutional Neural Networks for Vision-Based Hand Gesture Recognition," in Proceedings of the 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA).
15. P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand Gesture Recognition with 3D Convolutional Neural Networks," 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2015.
16. V. John, A. Boyali, S. Mita, M. Imanishi, and N. Sanma, "Deep Learning-Based Fast Hand Gesture Recognition Using Representative Frames," International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2016.
17. X. Yingxin, L. Jinghua, W. Lichun, and K. Dehui, "A Robust Hand Gesture Recognition Method via Convolutional Neural Network," 6th International Conference on Digital Home (ICDH), 2016.
18. I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional Neural Network Architecture for Geometric Matching," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
19. Y. Kim and B. Toomajian, ''Hand gesture recognition using micro-Doppler signatures with convolutional neural network,'' IEEE Access, vol. 4, pp. 7125–7130, 2016.
20. Oyeb ade K. Oye dotun et al, "Deep learning in vision-based static hand gesture recognition", Neural Computing and Applications 2016
21. Strezoski, G., Stojanovski, D., Dimitrovski, I. and Madjarov, G., "Hand Gesture Recognition Using Deep Convolutional Neural Networks," in International Conference on ICT Innovations, Springer, Cham, pp. 49-58, 2016.
22. S. Hussain, R. Saxena, X. Han, J. A. Khan, and H. Shin, "Hand Gesture Recognition using Deep Learning," *International SoC Design Conference (ISOCC)*, Seoul, 2017, pp. 48-49.doi: 10.1109/ISOCC.2017.8368821
23. Ali A. Alani, Georgina Cosma et al., "Hand Gesture Recognition Using an Adapted Convolutional Neural Network with Data Augmentation," 4th IEEE International Conference on Information Management, 2018.
24. Pramod Kumar Pisharady, Prahlad Vadakkepat, Ai Poh Loh, "Attention Based Detection and Recognition of Hand Postures Against Complex Backgrounds", International Journal of Computer Vision, vol.101, no.3, pp.403-419, February, 2013
25. Anjali R.Patil, Dr. S. Subbaraman, 'Illumination Invariant Hand Gesture Classification against Complex Background using Combinational Features', International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No. 3, March 2018

AUTHORS PROFILE

Mrs. Anjali R Patil has accomplished BE (2002) and ME (2009) from Shivaji University, Kolhapur. Presently she is research Ph.D. student in Electronics Engineering at Shivaji University Kolhapur She has published 12 papers in reputed journals and conferences. Her research interest includes image processing, Pattern recognition, Soft computing etc.

Dr. Mrs. Shaila Subbaraman, Ph. D. from I.I.T., Bombay (1999) and M. Tech. from I.I.Sc., Bangalore (1975) has a huge involvement in industry in the limit of R and D architect and Quality Assurance Manager in the field of assembling semiconductor devices and ICs. She additionally has over 27 years of academic experience at both UG and PG level for the courses in Electronics Engineering. Her specialization is in Micro-hardware and VLSI Design. She has in excess of fifty publications in her credit. She retired as Dean Academics of autonomous Walchand College of Engineering, Sangli and as of now she is working as Professor (PG) in a similar college. Moreover she fills in as a NBA expert for engineering programs in India as per Washington Accord. Recently she has been felicitated by "Pillars of Hindustani Society" award instituted by Chamber of Commerce, Mumbai for contribution to Higher Education in Western Maharashtra