Indian Sign Language Recognition using Vision Transformer based Convolutional Neural Network

Sunil G. Deshmukh[1], Shekhar M. Jagade[2]

[1]*Department of Electronics and Computer Engineering, Maharashtra Institute of Technology, Aurangabad 431010, Maharashtra, India*
[2]*Department of Electronics and Telecommunication, Sri N B Navale Sinhgad College of Engineering, Solapur-413255, Maharashtra, India*

---***---

Abstract - *Hand gesture recognition (HGR) is a popular issue in the areas of learning algorithms and visual recognition. Certain Human-Computer Interactions technologies also need HGR. Traditional machine learning techniques and intricate convolutional neural networks (CNN) have been employed to HGR up till now. Despite the fact that these approaches work adequately well on HGR, we employed a more modern model, vision transformer, in this research. Vision Transformer (ViT) is created to enhance CNN's performance. ViT has a strong similarity to CNN, but its classification work utilizes distinct layers. The ViT was performed to gesture datasets via learning algorithms. In trials, a testing dataset having crossed test strategy is evaluated, and classification accuracy is employed as productivity metric. According to test findings, the presented approach framework attains an achieved accuracy of 99.88% on the image database used, which is considerably higher than the state-of-the-art. The ablation study also supports the claim that the convolutional encoding increases accuracy on HGR.*

Key Words: **Convolutional neural network, Vision transformer (ViT), Transfer learning, Training of images, Accuracy, Hand gesture, Human-computer interaction (HCI).**

1. INTRODUCTION

Direct contact is becoming the most popular way for users and machines to communicate. People connect with one another naturally and intuitively through contactless techniques including sound as well as body actions. The versatility and efficacy of these contactless communication techniques have motivated several researchers to adopting them to further HCI. Gestures are a significant part of human language and an essential non-contact communication technique. Wearable data gloves were frequently used in the past to grab the positions and angles of every user's joints as they moved.

The complexity and price of a worn sensor have limited the extensive application of such a technology. Gesture recognition refers to a computer's ability to understand gestures and execute specified instructions in response to such motions. The main goal of HGR is to establish a system that can identify, analyze, and communicate information based on specific motions [1].

Techniques for recognizing gestures on the basis of contactless visual inspection are now prominent. This is because their accessibility and price. Hand gestures are an expressive communication technique employed in the healthcare, entertainment, and educational sectors of the economy, as well as to assist people with special needs and the elderly. For the purpose of identifying hand gestures, hand tracking which combines a number of computer vision operations such as hand segmentation, detection, and tracking is crucial. HGR are used to convey information or emotions in sign language to those who have hearing loss. The major problem is that the typical person may easily misinterpret the message. AI and computer vision developments may be utilized to identify and comprehend sign language.

With the assistance of modern technology, the typical person may learn to recognize sign language. This article introduces a deep learning-based technique for hand gesture identification. In this regard, operating the system remotely necessitates the use of gestures. The devices record human motions and recognize them as the ones that are used to control them. The movements employ a variety of modes, including static and dynamic modes. The static gestures are maintained while the dynamic gestures shift to various areas when the machine is being controlled. So, rather than using static gestures, it is essential to identify or recognize dynamic motions. The camera that is connected to the apparatus initially captures people's movements [2].

When the backdrop of any motions that were identified is removed, the gesture's foreground is gathered. To locate and eliminate the noises in the foreground gesture, filtering techniques are applied. These noise-removed gestures are compared to pre-stored and taught movements in order to verify the meaning of the gestures. The automotive and consumer electronics sectors utilize a gesture-based machine operating system that doesn't require any human input. Static and dynamic gestures, as well as online and offline actions, can all be classified as human gestures. The machine's icons can be changed using offline gestures, but the system or menu where the items are shown cannot be altered. The movements of the internet cause the machine's symbols to shift or tilt in a variety of ways.

Online gestures are significantly more useful than offline gestures in real-time machine operating systems. These strategies required a large number of training samples and could not handle large training datasets. This flaw is fixed by suggesting ViT-based CNN in this study. The complexity of this technique is low, and it does not require a big amount of data to train. The incorporation of a deep learning algorithm and a cutting-edge segmentation approach into a HGR system is an original component of the proposed study. Enhancing users to communicate by programming software more receptive to user demands is the fundamental aim of human-computer interaction (HCI). Gestures are intentional, suggestive body movements that include the physical action of the hands, wrists, face, shoulders, and face with the intention of engaging with the surroundings or conveying the relevant information. The creation of human-centric interfaces is made possible by hand gesture recognition (HGR), a crucial job in HCI and a highly helpful method for computers to comprehend human behavior.

A vital part of our everyday lives is being able to communicate effectively. People who are deaf or dumb find it challenging to engage with people due to their incapacity to talk and listen. One of the most effective and well-known techniques is arguably the use of hand gestures, sometimes referred to as sign language. Programs that can identify sign language motions and movements must be developed in order for deaf and dumb people to communicate more easily with those who do not understand sign idioms. The purpose of this study is to employ sign languages as a first step in removing the barrier to communication between hearing-impaired and deaf people. When employed to solve picture recognition issues in computer vision, modern (CNN) yield successful performance. The most effective way to build a complete CNN network is via the use of transfer learning. In a broad range of fields, including automation, machine learning is used. Aspects of artificial intelligence include techniques for measuring, recording, detecting, monitoring, identifying, or diagnosing physical phenomena. Moreover, a variety of sensors, such as vision-based sensors, capacitive sensors, and motion-based sensors can record hand movements [3].

Data Glove Techniques: Glove-based sensors are the only technologies that can handle the complex needs of hand-based input for HCI. In order to determine the hand postures, this approach uses mechanical or optical sensors that are mounted to a glove and translate finger flexions into electrical impulses. The following disadvantages render this innovation less well-liked: The user is required to carry a burden of wires that are linked to the processor and it also needs calibrating and setup processes, thus contact with the computer-controlled surroundings lacks its ease and naturalness. Techniques that are based on vision Approaches based on system vision have the ability to obtain more non-intrusive, natural solutions since they are based on how people interpret information about their environment. While it is challenging to create a vision-based interface for general use, it is possible to do so for a controlled environment without facing many difficulties, such as accuracy and processing speed [4].

Heap et al. [5] developed a deformable 3D hand model, and they surface-mapped the whole hand using PCA from training instances. Real-time tracking is made possible by locating the most likely to be distorted model that matches the picture. While more processing is necessary to extract valuable higher metadata, such as pointing direction, it has been shown that such a representation is quite successful at precisely locating and tracking the hand in pictures. Nevertheless, the approach is not scale and rotation invariant and cannot tackle the obstruction issue. Compositional methods are used in to identify hand posture. Configurations of parts serve as the foundation for a hand posture depiction. Characteristics are classified in accordance with the perceptual principles of grouping to provide a list of potential candidate compositions. Based on overlapped sub-domains, these subgroups provide a sparse picture of the hand position.

A novel method for identifying static motions in challenging circumstances based on wristband-based contour features was proposed by Lee et al. [6]. Pair of black wristbands is applied to both hands to properly divide the hand area. By the use of a matching algorithm, the gesture class is detected. When the background colour reappears, the system fails to effectively segregate the hand region. For the purpose of detecting static hand gestures, Chevtchenko et al. [7] optimized a coalitation of characteristics and dimensions using a multi-objective evolutionary algorithm. Up to 97.63 percent identification accuracy was reached on 36 gesture postures from the MU dataset. The aforementioned accuracy was attained utilizing a holdout cross-validation test and the combined Gabor filter and Zernike moment features. A variety of geometric characteristics including angle, distance, and curvature features were also produced by the contour of the hand motion, and these aspects were classified as local descriptors. Following that, local descriptors were improved using the Fisher vector, and gesture recognition was accomplished using a support vector machine classifier. Deep features from AlexNet's fully connected layer and VGG 16 were used to recognize sign language. The obtained attributes were classified using the SVM classifier. The recognition performance was estimated to be 70% using a conventional dataset and the leave-one-subject-out cross-validation test. The results show that for RGB input photos, detection performance is constrained by background variation, human noise, and high inter-class similarity in ASL gesture postures.

Parvathy et al. [8], the authors deploy a vision-based HGR system to extract key characteristics from radar-based pictures before categorizing them with a 96.5% accuracy SVM classifier. Deep learning models based on (CNN) architecture have attracted more interest in human computation and object detection. Convolutional filters are used by CNN to extract the key details of an image's subject the object classification convolution processes that cover important attributes should therefore be minimized. The authors Zhan et al. [9] have suggested a CNN model for the real-time HGR. This model's accuracy was 98.76% using a dataset of 500 photos and 9 distinct hand motions. They also suggested a double dimensional (2D) CNN model for dynamic HGR that includes of max and min resolution sub-networks, which was later demonstrated to have attained an accuracy of 98.2% on the same dataset.

For dynamic HGR that incorporates actions as well as motion sensing, the suggested works are not appropriate. So, it is necessary to investigate the selection of reliable classifiers and suitable hyper-parameters for the same. Adithya and Rajesh [10] suggest a deep CNN architecture for HGR. The suggested architecture avoids detecting and segmenting hands in collected pictures. The authors tested the suggested architecture on the NUS hand posture and American Finger spelling A datasets, achieving accuracy of 94.26% and 99.96%, respectively. The authors Islam et al. [11] developed a CNN model with data augmentation for static HGR. In this case, 8000 photographs from the dataset are utilized for training, whereas 1600 images are used for testing after being divided into 10 classes. The dataset has been augmented using various data augmentations such as zooming, re-scaling, rotation, shearing, height, and width shifting.

Neethu et al. [12] used a cutting-edge deep CNN model to perform hand motion recognition and HGR. Here, the hand and finger photos are separated into separate mask images, which are subsequently sent to CNN for classification into several groups. The results demonstrate that the suggested CNN model produced a 90.7% recognition accuracy. Wu [13], suggested a double channel CNN (DC-CNN) model for HGR. The hand gesture photos and hand edge images are created here by pre-processing the source photographs. The CNN channels then categories these pictures using two feeds. These pictures are then sent into two CNN channels for classification. The suggested DC-CNN demonstrated accuracy of 98.02% and 97.29% on the Jochen-Triesch Database and the NAO Camera hand posture Database, respectively. The suggested DC-CNN approach has not been tested with complicated backdrop pictures. Further features are being added to make the model more versatile. The model's applicability to dynamic HGR must be investigated.

Lai and Yanushkevich [14] proposed a mix of CNN and recurrent neural network (RNN) for autonomous HGR based on depth and skeleton data. RNN is utilized to process the skeleton data in this case, whereas CNN is used to analyze the depth data. Using the dynamic hand gesture-14/28 dataset, the suggested model attained an accuracy of 85.46%. The presented approach may be expanded to recognize human activities, which is an intriguing subject to investigate. Bao et al. [15] suggested a deep CNN model for small HGR directly from pictures without segmentation or pre-processing. The suggested HGR network can distinguish between up to seven groups. The suggested model attained accuracy of 97.1% for simple backdrop and 85.3% for complicated background, respectively. It has been said that the proposed solution has the benefit of being able to be implemented in real-time

A realistic and effective way for creating the model's contours which are then evaluated with the picture data is produced by using quadrics to create the 3D model. An Unscented Kalman filter (UKF), which reduces the geometrical uncertainty between the profiles and edges retrieved from the pictures, is used to predict the posture of the hand model. Although giving more accuracy than the enhanced Kalman filter, the UKF allows for faster frame rates than more complex estimating techniques like particle filtering. The hand gesture recognition job may depend on a variety of input data types and combinations, much like other computer vision-related tasks. As a result, methodologies described in the literature may generally be divided into two categories: monotonic and multifunctional. To address the action and sign language recognition challenges, the authors of suggest transformer-based methods that are comparable to ours. An input translator and recognizing paradigm that resembles the Faster R-structure CNN's uses a slightly adapted variant of the transformer architecture. A data extractor and a transformer-like framework are combined to recognize actions in real-time. The temporal linkages are not explicitly modelled since it uses 1D convolutional layers between consecutive decoder blocks rather than any kind of positional encoding. On the contrary side, in our method, positional encodings is used to encapsulate the temporal features on the frames order (PE) [16].

The CNNs is the state-of-the-art in machine learning and extensively employed for various image identification applications, have a strong competitor in the shape of Vision Transformer (ViT). In terms of accuracy and computational efficiency, ViT models performed approximately four times better than the most advanced CNNs currently available. New vision transformer frameworks have also exhibited surprising skills, reaching equivalent and even superior performance than CNNs on several computer vision applications, despite convolutional neural networks having long since dominating the area of machine learning.

Here, it has been researched and developed a Transformer-based framework to recognize hand gestures in this article. In a sense, we take use of the Transformer architecture's recent ground-breaking work in solving many complicated ML issues, as well as its enormous potential to use more input parallel processing with attentive mechanisms. To solve the issues mentioned above, Vision Transformers (ViT) might be regarded as an acceptable architecture. The presented Vision Transformer-based Hand Gesture Recognizing (ViTHGR) framework may address issues with assessment accuracy and training time that are often encountered when using other networks like the traditional ML techniques. As a result, the primary signal is divided into smaller sections using a certain window opening in the recommended ViT-HGR topology, and one of these segments is then sent to the ViT for additional assessment.

2. VISION TRANSFORMER (ViT)

The advent of Vision Transformer (ViT) puts CNN, the state-of-the-art in computer vision that is so often used in a variety of image identification applications, into intense competition. Convolutional neural networks (CNNs) cannot compete with the ViT models in terms of processing power, effectiveness, and accuracy. Modern performance criteria for transformer topologies in the area of natural language processing.

In certain computer vision tasks, the transformer outperforms existing convolution approaches [10]. A vision transformer is a transformer used specifically for a computer vision job (ViT). ViTs perform impressively and with great promise in computer vision applications. ViT offers two key advantages: 1) Self-attention method, in which the model interprets a wide variety of input seeds (tokens) in an all-encompassing context. 2) The capacity to practice challenging activities. The ViT concept is shown in Fig. 1, where initial photos are patched together in line with the prototype model. Patches are then sent immediately into the layer of the straight projections. The procedure of patch embedding is carried out in the second step. The sequencing of the inserted patches now includes the class token. As a result, patches grow in size by one. Positional embedding additionally adds embedded patches to the database spatial sequencing of patches. Lastly, the encoder layer serves as the first transformer layer and is given patch embedding and positioning of encoding with a class token. The most crucial part of a transformer, notably in ViT, is encoding. The feeding forward layer receives the output of the attentive layer, which then produces the encoder's final outcome [17].

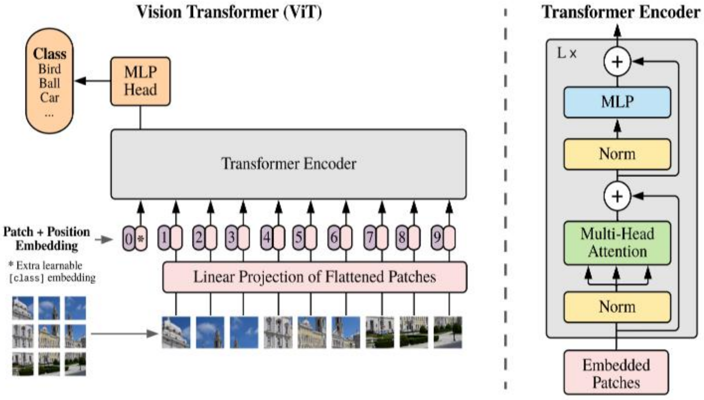

Fig-1: Sketch of overall ViT concept

The ViT encoder has many stages, and each component consists of three significant processing components:

1. **Layer Norm**
2. **Multi-head Attention Network (MSP)**
3. **Multi-Layer Perceptions (MLP)**

1. Layer Norm helps the model adjust to the differences between the training pictures while keeping the training phase on track.

2. A network called the Multi-head Attention Network (MSP) creates attention mappings from the integrated visual signals that are provided. These concentration maps assist the network in concentrating on the image's most crucial areas, such as object (s). The idea of attention mapping is similar to that explored in the research on conventional computer recognition (e.g., saliency maps and alpha-matting).

3. The Gaussian Errors Linear Unit (GELU) is the last layer of the two-layer classification network known as MLP. The final MLP block, commonly known as the MLP head, serves as the transformer's output. Data augmentation may be used on this output to provide categorization labels.

The transformer's architectural has been preserved, and the majority of the alterations relate to how the photos are altered before being sent to the transformer. Whereas Vision Transformer (ViT) uses pictures as its input, NLP-based models take a sequence of words as input. Images are split into 2D patches of a given size to account for this disparity (usually 16x16). Prior to the patches being supplied to the Transformer as a flattened 1D sequence, recommends utilizing linear projections on flattened 2D patches and positional embedding to get patch embedding while retaining positional information (just like tokens). The Transformer Encoder, which is composed of an attention layer and a Multi-Layer-Perceptron, receives a series of patches as input in the basic model's architecture (MLP). This is then sent to a MLP that is not connected to the encoder, and this MLP eventually produces a probability distribution of all the classes. The picture is initially run through a featured extruder, which identifies crucial characteristics that describe the image and normalizes images across RGB channels while maintaining mean and standard deviations. The ViT model is then applied to the picture, with each layer teaching the model that something new before producing the result [17].

Vaswani et al. [18] served as an inspiration for the Transformers encoding image. A picture is split into small patches, each of which is linearly embedded, positional embeddings are added, and the resultant sequences of matrices are sent to a typical Transformer encoding. It has been used the conventional method for adding an additional trainable "classifying token" to the sequences as a way to do classification.

3. RESULTS AND DISCUSSION

In this article, the interpretation of predefined Indian sign languages is examined. The suggested strategy may aid disabled individuals in conversing with normal individuals. In the perspective of the Indian language, the dataset utilized for research is a static ISL database with digit and alphabetic signals. Functioning with the transformer encoding does not need any data preparation. ViT assists the model in enhancing its performance. The performance of the suggested technique is shown in Table 1. Recall is the total proportion of entries that were mistakenly categorized as negative, calculated by (“equation 2”) dividing the number of real positives by the summation of true positives and false negatives. Precision and F-1 score are calculated by “equation 1” and “equation 3”, respectively.

Mathematically, the report can be represented in a similar manner as;

$$precision = \frac{True\ positive}{True\ positive + False\ positive} \qquad (1)$$

$$\mathrm{Re}\,call = \frac{True\ positive}{True\ positive + Negative} \qquad (2)$$

$$F1\text{-}Score = 2 * \frac{precision * \mathrm{Re}\,call}{precision + \mathrm{Re}\,call} \qquad (3)$$

Table -1: Performance of vit methodology

Data set	**Accuracy**	**Precision**	**F1-Score**	**Classification Error**
ISL	99.88%	0.99	0.99	0.72

As seen in Fig. 2. the proposed model demonstrates great convergence throughout the training and validation phases. The samples were evaluated by means of the ViT model. These are the testing accuracy results for ten epochs. The total average accuracy was 99.88%; as shown in Fig. 2. out of the 48281 images utilized for testing, 3531 were correctly categorized, while 165 were misclassified. It is seen that the ViT model performed having good index of accuracy level. From Fig. 3 it can be observed that the convergence time and losses are minimized. Table 2 displays the training loss, percentage accuracy, validation accuracy, and validation loss as they occurred with the response time.

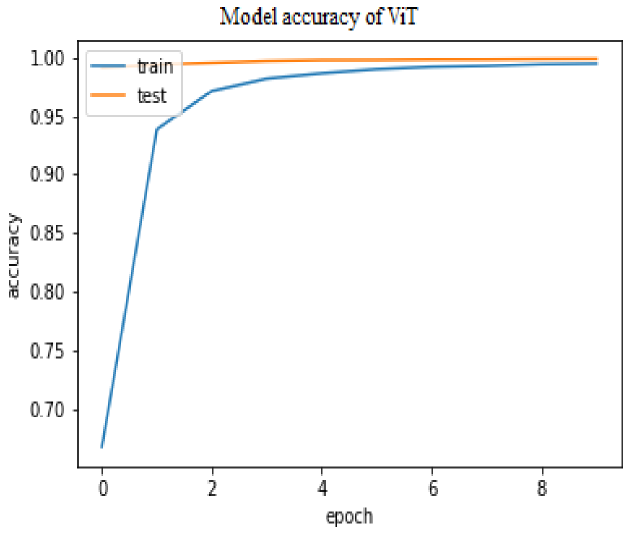

Fig-2: Model accuracy of ViT model

A confusion matrix is also plotted in Fig. 4. A confusion matrix displays the real positives and negatives identified during the execution of the algorithm. This illustrates where the algorithm went wrong as well as the number of accurate exclusions and false positives. The genuine labeling is shown on the y axis and the anticipated labeling is displayed on the x axis. The confusion matrix helps to determine whether or not the bulk of the model's predictions are true.

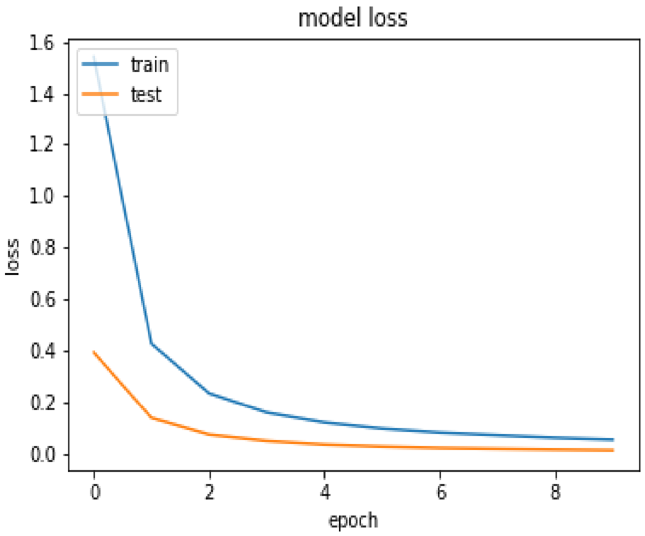

Fig-3: Model loss of ViT model

Table-2: Percentage accuracy, validation accuracy and validation loss of vit model

Epoch	Res. Time (S/step)	T-Loss	% Accu.	Val. Loss	Val. Accu.
1/10	871s 18m	1.5406	0.6672	0.3923	0.9912
2/10	911s 19m	0.4262	0.9386	0.1380	0.9939
3/10	891s 18m	0.2336	0.9715	0.0729	0.9955
4/10	864s 18m	0.1587	0.9820	0.0483	0.9972
5/10	854s 18m	0.1200	0.9865	0.0341	0.9980
6/10	858s 18m	0.0962	0.9899	0.0263	0.9981
7/10	852s 18m	0.0804	0.9920	0.0208	0.9985
8/10	853s 18m	0.0701	0.9929	0.0177	0.9939
9/10	892s 18m	0.0604	0.9942	0.0147	0.9985
10/10	908s 19m	0.0531	0.9948	0.0122	0.9988
Res. Time = Response time, T-Loss: Training loss, % Accu.: Percentage Accuracy, Val. Loss: Validation Loss, Val. Accu.: Validation accuracy.					

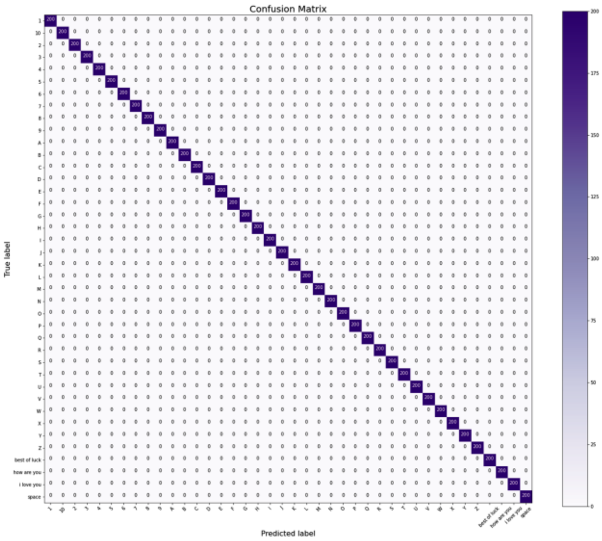

Fig-4: Confusion matrix of ViT model

4. CONCLUSION

In this research, the viability of the ViT approach for the categorization of the HGR is investigated. The shortcomings of the Convolutional networks were addressed using the newly introduced strategy known as ViT. The suggested ViT-HGR system can categorize a significant variety of hand gestures correctly from start without the necessity of data augmenting and/or transfer learning, hence overcoming the retraining time issues with recurrent platforms. According to test findings, the presented approach framework attains an achieved accuracy of 99.88% on the image database used, which is considerably higher than the state-of-the-art. The ablation study also supports the claim that the convolutional encoding increases accuracy on HGR. In next research, more pre-trained ViT algorithms will be utilized for the purpose of improving the accuracy of hand gesture identification. Furthermore, additional hand gesture datasets as well as the MLP mixing models will be employed in future investigations.

REFERENCES

[1] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," IEEE transactions on multimedia, vol. 15, no. 5, 2013, pp. 1110–1120.

[2] S. Ahmed, K. D. Kallu, S. Ahmed, and S. H. Cho, "Hand gestures recognition using radar sensors for human-computer-interaction: A review," Remote Sens., vol. 13,no. 3, Feb. 2021.

[3] Van den Bergh, M., & Van Gool, L, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction,"

In 2011 IEEE workshop on applications of computer vision (WACV), January 2011, pp. 66-72,

[4] Karbasi, M.; Bhatti, Z.; Nooralishahi, P.; Shah, A.; Mazloomnezhad, S.M.R, "Real-time hands detection in depth image by using distance with Kinect camera," Int. J. Internet Things, Vol. *4*, 2015, pp. 1–6.

[5] A. J. Heap, D. C.Hogg, "Towards 3-D hand tracking using adeformable model," In 2nd International Face and Gesture Recognition Conference (1996), pp 140–45.

[6] D. Lee and W. You, "Recognition of complex static hand gestures by using the wristband-based contour features," IET Image Processing, vol. 12, no. 1, 2018, pp. 80–87.

[7] S. F. Chevtchenko, R. F. Vale, and V. Macario, "Multi-objective optimization for hand posture recognition," Expert Systems with Applications, vol. 92, 2018, pp. 170–181.

[8] Parvathy P, Subramaniam K, Prasanna Venkatesan G, Karthikaikumar P, Varghese J, Jayasankar T, "Development of hand gesture recognition system using machine learning," J Ambient Intell Humaniz Comput 12(6): 2021,6793–6800.

[9] Zhan F, "Hand gesture recognition with convolution neural networks," In: 2019 IEEE 20th international conference on information reuse and integration for data science (IRI), 2019, pp 295–298.

[10] Adithya V, Rajesh R, "A deep convolutional neural network approach for static hand gesture recognition," Procedia Computer Science third international conference on computing and network communications (CoCoNet'19) 171, 2020, pp. 2353–2361.

[11] Islam MZ, Hossain MS, ul Islam R, Andersson K, "Static hand gesture recognition using convolutional neural network with data augmentation,". In: 2019 joint 8th international conference on informatics, electronics vision (ICIEV) and 2019 3rd international conference on imaging, vision pattern recognition (icIVPR), 2019, pp. 324–329.

[12] Neethu PS, Suguna R, Sathish D, "An efficient method for human hand gesture detection and recognition using deep learning convolutional neural networks," Soft Comput, vol. 24(20), 2020, pp. 15239–15248,.

[13] Wu XY, "A hand gesture recognition algorithm based on dc-cnn", Multimed Tools Appl, vol. 79(13), 2020, pp. 9193–9205.

[14] Lai K, Yanushkevich SN, "Cnn+rnn depth and skeleton based dynamic hand gesture recognition," In: 2018 24th international conference on pattern recognition (ICPR), 2018, pp 3451–3456.

[15] Bao P, Maqueda AI, del-Blanco CR, García N, "Tiny hand gesture recognition without localization via a deep convolutional network,". IEEE Trans Consumer Electron, vol. 63(3), 2017, pp. 251–257.

[16] Song, L.; Hu, R.M.; Zhang, H.; Xiao, Y.L.; Gong, L.Y, "Real-time 3d hand gesture detection from depth images", Adv. Mater. Res., Vol. 756, 2013, pp. 4138–4142.

[17] Cheng, P. M., and Malhi, H. S. "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images," Journal of digital imaging, Springer*,* Vol. 30(2), 2017, pp.234-243.

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. "Attention is all you need", In: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017, pp. 1-15.

BIOGRAPHIES

Sunil G. Deshmukh acquired the B.E. degree in Electronics Engineering from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad in 1991 and M.E. degree in Electronics Engineering from Dr. Babasaheb Ambedkar Marathwada University (BAMU), Aurangabad in 2007. Currently, he is pursuing Ph.D. from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India. He is the professional member of ISTE, His areas of interests include image processing, Human-computer interaction.

Shekhar M. Jagade secured the B.E. degree in Electronics and Telecommunication Engineering from Government Engineering College, Aurangabad in 1990 and M.E. degree in Electronics and Communication Engineering from Swami Ramanand Teerth Marathwada University, Nanded in 1999 and Ph.D from Swami Ramanand Teerth Marathwada University, Nanded in 2008. He has published nine paper in international journals and three papers in international conferences. He is working as a vice-principal in the N. B. Navale Sinhgad College of Engineering, Kegaon-Solapur, Maharashtra, India. He is the professional member of ISTE, Computer Society of India, and Institution of Engineers India. His field of interests includes image processing, micro-electronics, Human-computer interaction.