Journal of Information Systems Engineering and Management
2025, 10(33s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Hand Gesture Recognition with RCNN Optimized Vision Transformer

Bhavana Sharma [1], Jeebananda Panda [2*]
[1, 2] *Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India*
**Corresponding Author Email: [2] bhavana_2k19phdec02@dtu.ac.in*

ARTICLE INFO

Received: 30 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

The deep learning models and convolutional neural networks have been comprehensively used for different applications of gesture recognition. The vanishing gradient problem, which is a vital problem for precise feature extraction, is the reason why these models do not produce high-quality results for hand gesture recognition because of long-range dependencies. In order to solve this issue and to handle the vision-based tasks for classification and segmentation, respectively, an efficient model is proposed that combines the Vision Transformer (ViT) block with RCNN. This model is trained with three-fold cross validation on two publically available datasets i.e., Ego hand dataset and VIVA dataset. In this study, 20% of the data from both datasets are utilized for testing while 80% of the data are used for training. The other parameters like batch size are 16 and learning rate of 0.001 over the course of 100 epochs. The test accuracy obtained for Ego hand dataset and VIVA dataset are 96.04% and 98.39%, respectively.

Keywords: Feature extraction, Gesture recognition, RCNN, Segmentation, Vision Transformer.

INTRODUCTION

In recent years, the convolutional neural networks (CNNs) have been used for segmentation and classification of gesture recognition. The main components of the recognition systems in the past were the conventional image segmentation algorithms, which incorporated threshold-based, edge detection-based, and region-based approaches [1]. Earlier for object detection R-CNN means Regions with CNN is designed with the combination of a variety of low-level visual attributes for segmentation, and high-level context for classification [2]. For dynamic dataset position sensitive is major issue, therefore a R-FCN without position-sensitivity is implemented where ResNet are designed by Region-based Fully Convolutional Networks [3]. For a variety of applications, numerous techniques are developed using various R-CNN models, including fast R-CNN, faster R-CNN, region-based fully convolutional networks, single shot detectors, deconvolution single shot detectors, R-CNN minus R, you only look once, and mask R-CNN [4]. A multi head R-CNN for multiple recognition tasks between classes and Auto-calibrated R-CNN for generation of pseudo-labels are used for dense pose recognition [5]. Hand pose estimation is main requirement for classification. The traditional segmentation methods have challenges for a complex background and occlusion in images and videos. Therefore, in order to recognize 3D hand gestures, Khan et al. combine regions with convolutional neural networks (R-CNN) and grass hopper optimization, where, five distinct elements make up the Mask-RCNN: the backbone, the region proposal network (RPN), the ROI alignment, the network head, and the loss function. [6]. Transformer, which include an encoder and a decoder through an attention model performs classification [7]. Initially, transformer works only for language processing only but now it is applied for vision-based application also by using direct sequence of image for classification task [8]. There so many visions transformer, which are designed for different purpose like real time applications and vision and video processing [9]. Due to high resolution in images and videos and major changes in visual structure, a Swin Transformer is designed, which is backbone of computer vision [10]. Therefore, due to complex background and occlusion in images and videos, practically vanishing gradient problem introduced. Therefore, there is some restriction with such segmentation algorithms. So, in this paper there is a fusion of vision transformer as classification and RCNN as segmentation have been done.

Journal of Information Systems Engineering and Management
2025, 10(33s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The three main contributions made by the works are as follows:

- Instead of learning for the full video sequence, only the main frames of arbitrary human hand characteristics are learned in the RCNN stream in order to increase the effectiveness of identification. By eliminating unnecessary information, it improves the framework's capacity for learning.
- The uniqueness of motion stream lies in its use of transfer learning to highlight the distinctive motion characteristics of each action performed by a human hand. Transfer learning is therefore utilized for classification in this paper.
- The three-fold cross validation has been done on two difficult publicly available datasets, such as the ego hand and viva hand datasets, are used in experiments to determine whether the suggested structure is effective. Then comparing the suggested framework to comparable state-of-the-arts reveals that it performs better.

PROPOSED METHOD

Fig. 1 provides a description of the suggested approach. For implementation, Ego hand dataset and VIVA dataset for hand gesture recognition are using. In pre-processing, the segmentation is done by using RCNN and for classification vision transformer is implemented for better recognition accuracy.

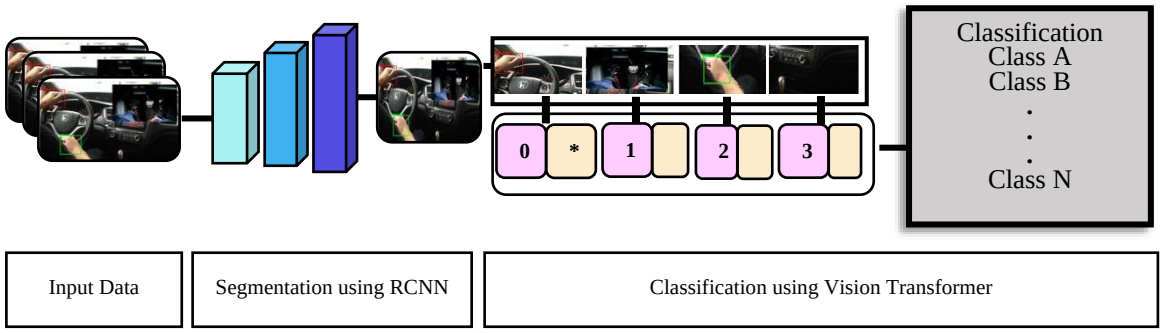


Fig. 1 Proposed Method

1. Input Data

The Ego hand dataset has complex background where a person has interaction with another person. Same 19 different dynamic hand gestures were made by 8 subjects inside a car in the multi-modal dynamic hand gesture dataset known as VIVA (Vision for Intelligent Vehicles and Applications) [11], [12], [13], [14]. Table 1 presents a thorough description of the datasets.

Table 1. The description of datasets

Dataset	Description	Source
Ego gesture dataset	Dynamic and RGB-D video Person's view gestures samples collected by Intel real sense SR 300	http: //www.nlpr.ia.ac.cn/iva/yfzhang/datasets/egogesture.html
VIVA dataset	Dynamic and RGB-D video Driver hand gesture samples collected by Microsoft Kinect	http://www.site.uottawa.ca/research/viva/projects/hand_detection

2. Regions with CNN

It is also known as R-CNN, and it is a model for object detection that uses high-capacity CNNs to produce bottom-up region recommendations in order to localize and segment items. It uses selective ROI search to find several boundary box object regions, then independently extract features for classification from each region. The isolated hand contour is discovered by ROI for hand gesture recognition. For calculation of minimal and maximal coordinates, the optical

Journal of Information Systems Engineering and Management
2025, 10(33s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

flow technique is used to detect the moving object with respect to time [15]. Segmentation is done by calculating center of mass of moving hand with elliptic least-squares fitting method.

3. Vision Transformer

A Transformer is a deep learning model which is differ from RNN because it takes on the self-attention mechanism, inherent non sequential input data and most important it uses positional embedding to reserve words' position in a sentence. The novel work for classification is explained in Fig 2, where vision transformer is used on images directly [8] and main three blocks are at encoder end. One is multi-head self-attentions (MSA), which is the main issue of high computational cost. Therefore, window-based MSA is used in proposed work. Second one is Multi-Layer Perceptron (MLP) and last one is layer norm which process the training images. The MLP head with two layers' classification system and Gaussian Error Linear Unit (GELU) is used as output of transformer. The transformer layers with respect to MSA and MLP can be explained in Equation 1 and 2:

$$TMSA = MSA(f) + TMSA_{n-1} \quad (1)$$

Where, function (f) = $LN(TMLP)_{n-1}$

$$TMLP = MLP(f) + TMLP_{n-1} \quad (2)$$

Where, function (f) = $LN(TMSA)_{n-1}$ and LN is layer normalization

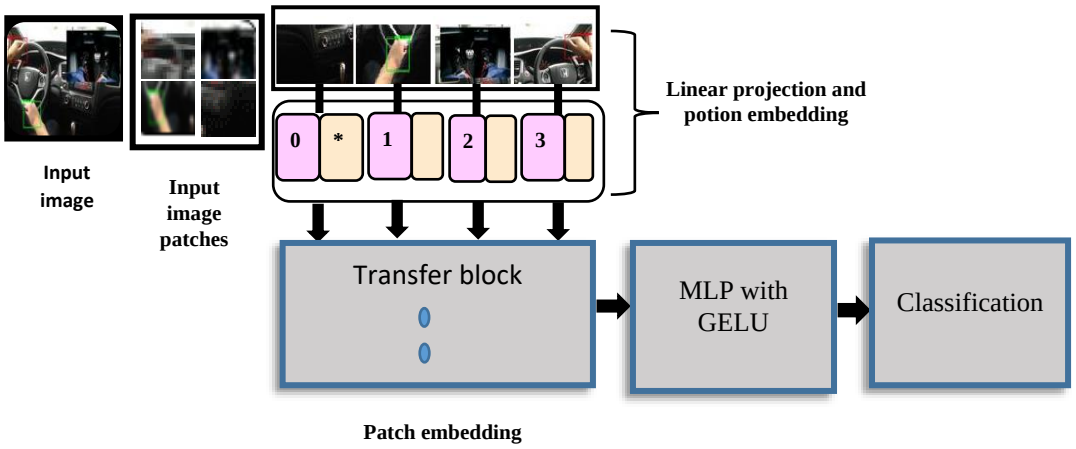


Fig 2: Processing steps of vision transformer

The value associated with each key is multiplied by the attention weight calculated by the layer normalization and attention mechanism between the query token and key token. Then the formation of projection matrix of each head of multi-head attention mechanism are used to calculate attention weight [8], [9]. The steps of working of proposed system is explained in the following algorithm.

Algorithm

Input: I(mxn); image with m = no. of rows and n = no. of columns

Output: Ic3

Begin Procedure

Initialize the image I (mxn)

Apply RCNN segmentation

I (mxn)= (I1, I2, ..., Iy, ..., Iz)

Update the image patches obtained through segmentation

While (X< RCNNiteration) do,

Journal of Information Systems Engineering and Management
2025, 10(33s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

For each path do

Update

End for

End while

Apply vision transformer for classification,

For I = 1 ton (n= no. of columns in image)

V_T(I)= classify (I),

IC1= patch embedding + positional embedding

IC2= IC1+ Multi head attention

IC3= Ioutput

Define class of IC3

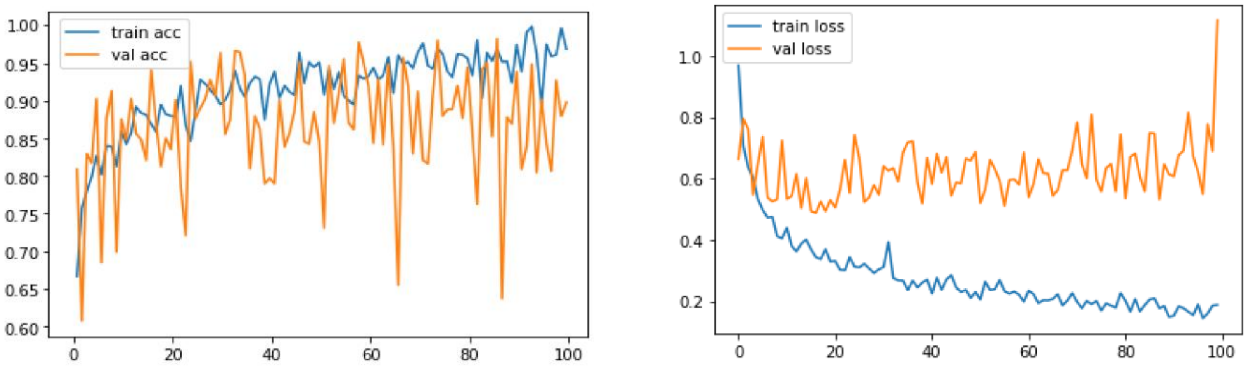
End For

Return archive

End procedure

RESULTS AND DISCUSSION

The two distinct datasets utilized to train and evaluate the suggested model are the Ego gesture dataset and the VIVA hand dataset. The transfer learning approach makes use of 80% of each dataset's data for training and 20% of each dataset's data for testing. Using 16 images batch size and a 0.001 learning rate, the trained model was run 100 times. Adam is used to optimize this model. After 100 iterations, the validation parameters applied to the testing data produce the best results for classification. The Fig. 3 and Fig. 4 show the output of Vision transformer as a classifier, where we can see the results of accuracy and losses of both datasets at specific parameters. For the validation of proposed model, we can see in this graph that overall classification accuracy of ego hand (dataset 1) in Fig. 3(A and B) and VIVA hand (dataset 2) achieves best performance in Fig. 4(A and B). Ego hand dataset achieves 0.9604 classification accuracy during training and 0.9100 accuracy during testing. While for VIVA hand dataset classification accuracy is 0.9839 during training and 0.9490 accuracy and during testing.

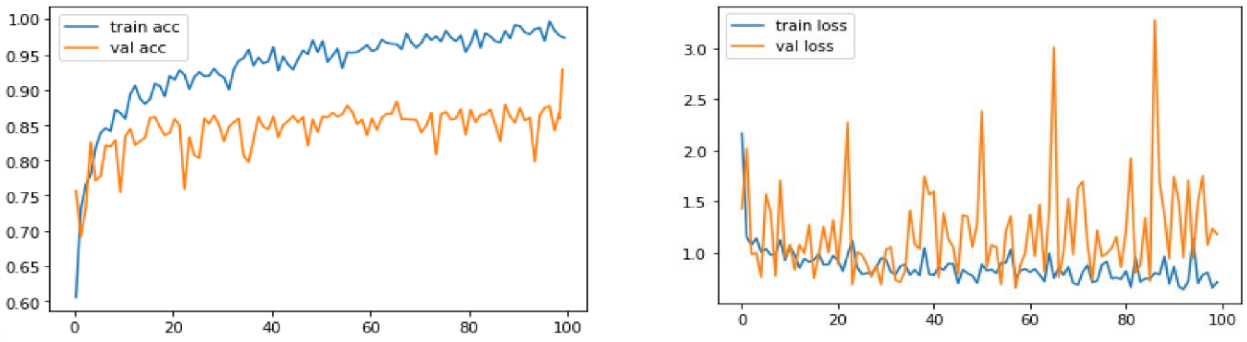


3(a) 3(b)

Fig. 3 Accuracy and loss of training and testing dataset of Ego hand (Dataset 1)

Journal of Information Systems Engineering and Management
2025, 10(33s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



4(a) 4(b)

Fig. 4 Accuracy and loss of training and testing dataset of VIVA hand (Dataset 2)

Table 2 presents the test and train results for both datasets. The results of the test and train clearly show minimal fluctuation, demonstrating that the suggested methodology is not overfitting.

Table 2. Summary of obtained accuracy of train and test for both datasets

	Epoch	Accuracy%	
		Ego hand	VIVA hand
Train	25	90.00	92.00
	50	94.20	95.21
	100	96.04	98.39
Test	25	95.00	85.82
	50	89.90	89.02
	100	91.00	94.90

It is abundantly clear from the results that the model performed best for the VIVA hand dataset and only slightly down for the Ego hand dataset. The classifiers are trained and evaluated using a 3-fold cross-validation in which 70% of the dataset is utilized for training and 30% is used for testing in order to provide robust validation of the proposed work. For the 3-fold cross validation testing, the dataset samples were randomly divided into three divisions of equal size. As shown in Table 3, the suggested method yielded the best results in terms of precision, sensitivity, specificity, and F1-score, with train scores of 96.51, 92.98, 93.03, 95.02 and test scores of 92.89, 90.72, 89.67, 92.08 respectively for the EGO hand dataset. Similar, to that for the VIVA dataset, the train accuracy, sensitivity, specificity, and F1-score are 98.83, 93.61, 93.93, 92.81 respectively, and for the test 96.99, 91.83, 90.07, 90.62.

Table 3. A brief of the performance evaluation of 3-fold cross validation

Parameters	Datasets	Fold I		Fold II		Fold III	
		Train	Test	Train	Test	Train	Test
Accuracy%	Ego hand	91.26	88.04	96.51	92.89	94.72	93.83
	VIVA hand	96.83	95.09	95.09	93.61	98.83	96.99
Sensitivity %	Ego hand	91.99	90.27	92.98	90.72	92.23	90.12
	VIVA hand	93.23	91.81	90.89	88.98	93.61	91.83
Specificity %	Ego hand	92.99	90.72	93.03	89.67	91.62	90.27
	VIVA hand	90.72	88.62	92.88	91.89	93.93	90.07

Journal of Information Systems Engineering and Management
2025, 10(33s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

F 1-score%	Ego hand	94.97	92.89	95.02	92.08	90.89	88.79
	VIVA hand	81.89	80.61	88.78	86.90	92.81	90.62

The proposed work is contrasted with recent work on the ego hand and VIVA hand datasets in Table 4. When compared to previous work on the ego hand and VIVA hand datasets, our model produced the best results for dynamic activity recognition.

Table 4. Comparison with recent works on hand gesture recognition

Technique with Reference	Dataset	Accuracy%	Sensitivity%	Specificity%
SSD [16]	VIVA hand	82.49	NR	NR
SSD [16]	Ego hand	77.96	NR	NR
CCNN model [12]	VIVA hand	92.17	NR	NR
CCNN model [12]	Ego hand	88.81	NR	NR
MFS-CCNN model [12]	VIVA hand	91.31	NR	NR
MFS-CCNN model [12]	Ego hand	91.76	NR	NR
Segmentation based Embedding + LSTM [13]	Ego hand	96.9	NR	NR
RCNN + ViT (Proposed Model)	Ego hand	96.04	92.14	93.08
RCNN + ViT (Proposed Model)	VIVA hand	98.39	93.34	93.20

NR*- Not reported

CONCLUSION

We proposed a model that combines the vision transformer with RCNN for the purpose of hand gesture detection. In addition, we put forth a powerful model for dynamic gesture recognition that was inspired by numerous multimodal feature fusion studies. According to the experimental findings on Datasets 1 and 2, the proposed model had a good accuracy score on Dataset 2 but a slightly lower score on Dataset 1. It demonstrates that three-fold cross validation with transformer-based networks are capable of delivering excellent performance for any recognition applications. In addition to the transformer for the recognition system, we intend to test out some self-supervised new learning techniques in the future.

REFERENCES

[1] Sharma, Bhavana, and Jeebananda Panda. "A Review of State of Art Techniques for 3D Human Activity Recognition System." *Modern Electronics Devices and Communication Systems: Select Proceedings of MEDCOM 2021* (2023): 1-9.
[2] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.
[3] Dai, Jifeng, et al. "R-fcn: Object detection via region-based fully convolutional networks." *Advances in neural information processing systems* 29 (2016).
[4] Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey (2016)
[5] Sanakoyeu, Artsiom, et al. "Transferring dense pose to proximal animal classes." *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 2020.
[6] 3D Hand Gestures Segmentation and Optimized Classification Using Deep Learning (2021)
[7] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
[8] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

Research Article

[9] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." *Proceedings of the IEEE/CVF international conference on computer vision* (2021).
[10] Han, Kai, et al. "A survey on vision transformer." *IEEE transactions on pattern analysis and machine intelligence* 45.1 (2022): 87-110.
[11] Sarma, Debajit, and Manas Kamal Bhuyan. "Methods, databases and recent advancement of vision-based hand gesture recognition for hci systems: A review." *SN Computer Science* 2.6 (2021): 436.
[12] Wang, Qiangyu, Guoying Zhang, and Shu Yu. "2D hand detection using multi-feature skin model supervised cascaded CNN." *Journal of Signal Processing Systems* 91 (2019): 1105-1113.
[13] Chalasani, Tejo, and Aljosa Smolic. "Simultaneous segmentation and recognition: Towards more accurate ego gesture recognition." *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops 2019*.
[14] Boughnim, Nabil, et al. "Hand posture recognition using jointly optical flow and dimensionality reduction." *EURASIP Journal on Advances in Signal Processing* 2013.1 (2013): 1-22.
[15] Zhou, Tian, Preeti J. Pillai, and Veera Ganesh Yalla. "Hierarchical context-aware hand detection algorithm for naturalistic driving." *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2016.
[16] Liu, Wei, et al. "Ssd: Single shot multibox detector." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer International Publishing, 2016.