Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.032832
Article

Tech Science Press

A Hybrid Deep Learning Model for Real Time Hand Gestures Recognition

S. Gnanapriya[1,*] and K. Rahimunnisa[2]

[1]Department of Information Technology, Easwari Engineering College, Chennai, 600089, India
[2]Department of Electronics and Communication Engineering, Easwari Engineering College, Chennai, 600089, India
*Corresponding Author: S. Gnanapriya. Email: gnanapriya.s@eec.srmrmp.edu.in
Received: 31 May 2022; Accepted: 05 July 2022

Abstract: The performance of Hand Gesture Recognition (HGR) depends on the hand shape. Segmentation helps in the recognition of hand gestures for more accuracy and improves the overall performance compared to other existing deep neural networks. The crucial segmentation task is extremely complicated because of the background complexity, variation in illumination etc. The proposed modified UNET and ensemble model of Convolutional Neural Networks (CNN) undergoes a two stage process and results in proper hand gesture recognition. The first stage is segmenting the regions of the hand and the second stage is gesture identification. The modified UNET segmentation model is trained using resized images to generate a cost effective semantic segmentation model. The Central Processing Unit (CPU) utilization and training time taken by these models with respect to three public benchmark datasets are also analyzed. Recognition is carried out with the ensemble learning model consisting of EfficientNet B0, EfficientNet B4 and ResNet V2 152. Experimentation on NUS hand posture dataset-II, OUHANDS and HGRI benchmark datasets show that our architecture achieves a maximum recognition rate of 99.07% through semantic segmentation and the Ensemble learning model.

Keywords: Convolutional neural networks; EfficientNet; ensemble learning; ResNet; semantic segmentation; UNet

1 Introduction

Recognizing gestures in real time is complex as it involves other noises like background objects, self occlusion of hands and body parts, background objects with skin color, and the computational time involved in realtime segmentation of hands from the cluttered background. To overcome these difficulties a Solution is achieved using wearable sensors which detect the hand position accurately. The Kinect sensor product of Microsoft [1] is used for the recognition of hand gestures. The Kinect sensor scans hand objects in three dimensions and extracts the features. Convex hull and convexity depths are used to trace the contour in segmented hand images. The author [2] captures the depth image of the shown gestures, to detect hands from complex backgrounds or other body parts. The depth information is used to segment the hand from the rest of the objects in the scene, as the distance between the hand and the camera is usually less compared to other body parts and the background. The solution proposed by [1]

and [2] needs an additional device which may not be possessed by the needy at all times; therefore, vision based approaches are used.

An adaptive hand segmentation based on the skin color threshold is used to handle images with varying illumination and complex background. The image is processed in YCbCr space. To detect skin pixels the threshold of blue and red chrominance are fixed as $76 < Cb < 126$ and $132 < Cr < 173$. The chosen threshold segregated the skin and non-skin regions [3], but this approach does not differentiate the skin and non-skin regions when both the background and foreground objects are of the same color.

The probabilistic Gaussian Mixture Model (GMM) is used to extract the foreground and eliminate the background [4]. The Fuzzy Gaussian Mixture Model (FGMM) takes a little time to segregate gestures and non gestures in real time effectively [5]. The drawback of GMM and FGMM is, it results in noisy output images which leads to misclassification.

The fully connected conditional Random Field (CRF) algorithm with Gaussian Filter is used [6] in segmenting the foreground object by convolving the Gaussian filter over the entire image. The dense pixel connectivity leads to more accurate pixel level classification of objects. But the performance of the proposed approach is found to be less with respect to time.

HGR-Net a two stage cascaded CNN architecture overcomes illumination changes and segments hands from complex backgrounds and recognizes it. The author [7] proposed a two stage CNN, where the first stage segments the hand region, the second stage identifies the gesture; the system is found to be robust against illumination variation and complex backgrounds. The segmentation stage avoids color confusion even when the background color is similar to the skin color. The second stage of the network fuses feature representations from both the Red Green Blue (RGB) color image and its segmentation map before classification. The system performance is studied using the OUHANDS dataset. The limitation of this cascaded model is that it has to learn and fuse features from two CNN Models to recognize gestures.

The advancement in deep learning enables localization of the human hand as an object [8] using semantic segmentation. The result of semantic segmentation is then improved using the CRF. This give good results in less time and improved the quality of image segmentation. The hybrid network consisting of UNET architecture and CNN [9] is used to segment the hand region from a complex background and CNN architecture VGG16 to classify the gesture. The system performance is studied with benchmark datasets and found that the accuracy of the system is good, but the time for recognition is almost the same as that of the state- of-the- art approaches.

The proposed architecture combines both the segmentation model and the classification model. To reduce resource utilization in terms of training time, the resized input image is considered for training. The direct image magnification affects the quality of the image, and this has been addressed using the principle of bicubic interpolation algorithms [10]. The result of the analysis shows that bicubic is found to give sharp images with improved edges within an acceptable processing time. The important characteristics of the proposed system are, it uses the modified UNET;, unlike the traditional UNET it uses only a single convolutional Layer, and to recognize gestures the majority decisions of the CNN models are used. To study the system performance, both segmentation and classification models are analyzed with respect to the number of epochs and elapsed time they took for training and testing. The ensemble learning based classification model is evaluated using three readily available datasets in the public domain NUS hand posture dataset-II, OUHANDS and HGRI. The CPU utilization and the prediction time of the system are also evaluated with respect to the existing works.

2 Related Work

2.1 Segmentation

Segmentation Is the basic problem in all image classification and recognition systems. It can be performed by different methods in different dimensions. Segmenting an image with high accuracy is still a challenging task. Sometimes, after segmentation the image will not be clear and during training it fails to recognize the image. This is due to the image quality;, a robust system must recognize all kinds of images and segment them. The quality of segmentation is another important paradigm in the recognition system. The quality is measured by Segmentation Quality Assessment (SQA). It is measured in two stages; in the first stage the segment is based on a fixed parameter and in the second stage it tests the quality of the segmented image for accurate results using the grab cut method and Fully Convolutional Network (FCN) [11]. Automatic evaluation of segmentation is inaccurate as it is difficult to compare the performance with low quality segmented images. So adaptive segmentation is developed with multi labeled segmentation, which has a composed reference image [12]. A novel approach [13] for image segmentation is rigid detection and non-rigid detection using elastic regularization. Usually, rigid and non-rigid detection are performed separately and combined at last to classify, but elastic regularization combines both detections at the same stage. Yet another method to evaluate segmentation is by multiple reference segmentation using region-based exemplar reference [14]. It compares with other methods to test the quality of a segmented image for recognition. Segmentation based on the ratio in the image is to regulate the boundary of the region [15]. It imposes a curve into the boundary region to trace and then the images are segmented; this is considered as a parameter to classify the image. In image processing, objects are segmented in two stages. Instead of two-stage segmentation, it is possible to segment the images of the two-stage process at one shot using a FCN [16]. Evaluating segmentation with multiple ground truths [17] is an important aspect in the classification and recognition of images. A probabilistic measure is introduced to identify the similarity between the various methods and recommend the quality segmented image for recognition. In-the case of low-quality images, U-NET is used to segment, which restores the original image because it has a guidance image in the structure [18]. The authors [19] refined the U-Net architecture in three ways. The Refined double U-Net (RDU-Net) reduced one down sampling and concatenation to up sampling part, the Refined Res U-Net (RRU-Net) replace convolutional pairs with ResNet this give better accuracy but increase the training time. The Refined simple U-Net (RSU-Net) reduced the number of decoder layers and used cosine distance to find the distance between two encoder layers, all these refinement increases the inference speed and prediction accuracy. As UNET does good prediction based even on a few features, it is proposed to reduce a convolutional layer of traditional architecture. This reduces the number of training parameters and increases speed.

2.2 Gesture Recognition

The feature extraction descriptors like Histogram of Oriented Gradient (HOG) and Hu moments, extracts only the background features which are highly used to recognize the gestures shown. To reduce the background detail [20], a histogram of oriented gradient is combined with skin color segmentation which recognizes the hand gestures. The Size of the cells influences the recognition rate; so, combining the histogram of the oriented gradient in the skin color model of different cells extracted from local features is used. Considering the foreground features, CNN fails to perform well as it is not trained to extract all the morphological features. The Vision based approach of [21] combines the Hu invariant moment, the region of hand gesture and Fourier descriptors. The images are binarized and noises are removed by smoothing filters. Gesture is recognized by a multilayer perceptron which has a recognition rate of 97.4%. The hand gesture is recognized based on database driven threshold. The Hand is segmented based on the color code and the Otsu threshold method is applied to separate the hand from the background [22]. The Principal Component Analysis (PCA) is used for the recognition of hand gestures.

Atrous convolution for dense feature extraction [23] is used for the semantic segmentation of an image using Diffusion convolutional neural networks (DCNN). It extracts the textural features and morphological features into a deep network to recognize the hand gesture. For region-based methods [24], Hand gesture recognition using multiple learning models is proposed, which reduces the deep features and accurately identifies the American sign language. It is evaluated by conducting tests with a benchmark dataset. Alex Net features are pre-trained [25], the fully connected layers of CNN are used to extract the features. The supervised learning algorithm Support Vector Machine (SVM) is then applied to classify the postures and gestures of the hand. PCA reduces the deep feature dimension to improve the recognition rate. A two dimensional template is created for each gesture [26] which differentiates between the template gesture and original gesture images by the modified probable Longest Common Subsequence (LCS) algorithm. The result is based on an extracted sequence which differs in length and is used for interfacing systems. The input is identified and tested to see whether it recognizes accurately. A fuzzy rule-based system is used in hand gesture recognition [27]. Templates are generated for each hand shape using the IF-THEN rules extracted from the weights of trained hyper rectangular composite neural networks (HRCNN). Unknown gesture samples are tested in a fuzzy IF-THEN rule system and are classified based on similarity, is used in dynamic hand recognition systems.

The traditional convolutional neural network consists of a multilayer perceptron where the convolutional layer is used to extract the features and the fully connected layer classifies the labels accordingly. Compared to other feed forward neural networks, CNN is efficient and easier to train as it has many connection layers and parameters. CNN is used in many research problems that includes image classification and recognition, video analysis, natural language processing, time series etc. In earlier recognition systems, a feature is extracted from both still and moving images which are common, and then the hand gesture is recognized. Static feature is the processing of a single image using feature descriptors. Dynamic feature is processing the sequences of images which are complex in nature. A Feature descriptor is used to extract the image features. Various descriptors are used for the feature extraction including Gabor filters, gradient location and orientation histogram, scale invariant feature transformation, shape context, jet descriptors, Hu moments and Zernike moments. The Most commonly used descriptors are Gabor filters, Hu moments and Zernike moments that achieve greater accuracy compared to other methods. All these methods also take the morphological features other than the object and background, but their performance is low when data is overfitted. An Image can be segmented based on the color, shape and polygon approximation. The deep learning based semantic image segmentation algorithm UNET is used to segment the hand gestures from images with complex backgrounds [28,29].

The hardware resource utilizations of deep learning networks are analyzed [30] by throughput and CPU utilization using deep learning libraries and Frameworks on GoogLeNet, ResNet50 and squeezeNet. The proposed Ensemble model improves the recognition rate as it works based on majority voting and reduces the computation time as the image cropped by the bounding box is taken as the input, and the region of interest is prioritized. The recognition accuracy is also evaluated with respect to publically available NUS hand posture dataset-II, OUHANDS and HGRI benchmark datasets.

3 Proposed System

The proposed modified UNET reduces training time as the two convolutional layers of the traditional model is replaced with a single convolutional layer, as the number of features learned by the convolutional layers remain the same. During testing the object is already focused as the bounding box is used to capture the input hand gesture, which further reduces the processing time. The segmented output images are then split into training sets and test sets for building the feature extraction and classification

models. To reduce the probability of wrong prediction, Ensemble learning is used [31]. The architectural system flow of the proposed system is shown in Fig. 1

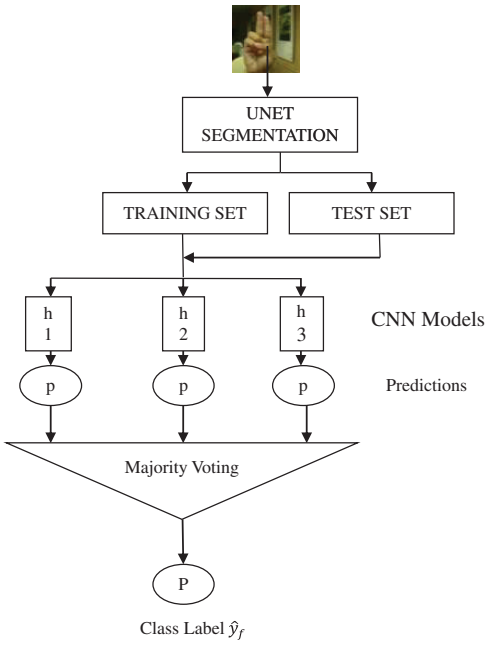


Figure 1: Architectural system flow of proposed model

3.1 UNET Segmentation

The proposed modified UNET architecture receives localized area in a complex image as input as the bounding box is used to capture it. This avoids much background and region of interest being focused. The original UNET developed for biomedical image segmentation uses two convolution layers where first layer learn the hand contour and the second one learn the timer details. The HGR system need only the segmented hand contour. This eliminates the need for two convolutional layers used by the traditional UNET architecture. The neural network model CNN is used to mine the object features in localized areas and classify them. UNET uses an alternate arrangement of convolution and max pooling layers during the contraction and expansion part of the network. The network parameters vary based on the chosen input. The objective of UNET is to produce an output image with the same size as that of the input; this is achieved by using padding equal to the same. The 2D convolution on down sampling starts with 64 feature filters of size 3×3 with a stride of 1. The kernel initialization is carried out with the normal. The system is trained with the ReLu activation function with a learning rate of 0.001using the Adam optimization algorithm. The dropout rate of 0.5 is considered to avoid overfitting during training. Convolution is followed by max pooling of size 2×2, and stride of 2, which reduce the dimension by half. The concatenation of the feature map from the encoder path during up sampling gives the localized information. The UNET configuration detail for the input image of size 64×64 is shown in Fig. 2.

3.2 Convolutional Neural Networks

The segmented hand outputs shown in Fig. 3 is used to train, validate and test the CNN models. The three scaled CNN imageNet architectures EfficientNet B0, EfficientNet B4 and ResNet V2 152 with average, high and low speed are used for image feature extraction and recognition. The scaling of ConvNets [32] MobileNets and ResNet improves accuracy and leads to better performance with reduced

size. From the family of EfficientNets and ResNets it is proposed to use three Neural Network models Bo, B4 form EfficientNets which are smaller and faster and ResNet 153 a deeper network. The default split of 80%, 10%, 10% is used for training, validation and testing the models. As UNET maintains the size of the output image as the same as input, all the three CNN architectures are tuned to process the segmented images of size 64×64 during training. These images are processed as a batch of 26 images, with the Adam optimizer, initialized with 100 epochs under a constrained learning rate of 0.000333 with the patience of 5. At the fully connected layer or dense layer, the Softmax activation function is used to scale the vector of numerical values to the vector of probabilities.

Resized Input Image			
Down sampling		Up sampling	
Conv2	32x32x128	Conv8	32x32x128
Max pool	16x16x128	Concat	32x32x256
		Conv8	32x32x128
		Upsample	32x32x256
Conv3	16x16x256	Conv7	16x16x256
Max pool	8x8x256	Concat	16x16x512
		Conv7	16x16x256
		Upsample	16x16x512
Conv4	8x8x512	Conv6	8x8x512
Dropout	0.5	Concat	8x8x1024
Max pool	4x4x512	Conv6	8x8x512
		Upsample	8x8x1024
Bottleneck			
Conv5	4x4x1024		
Dropout	4x4x1024(rate=0.5)		

Figure 2: Modified UNET configuration details

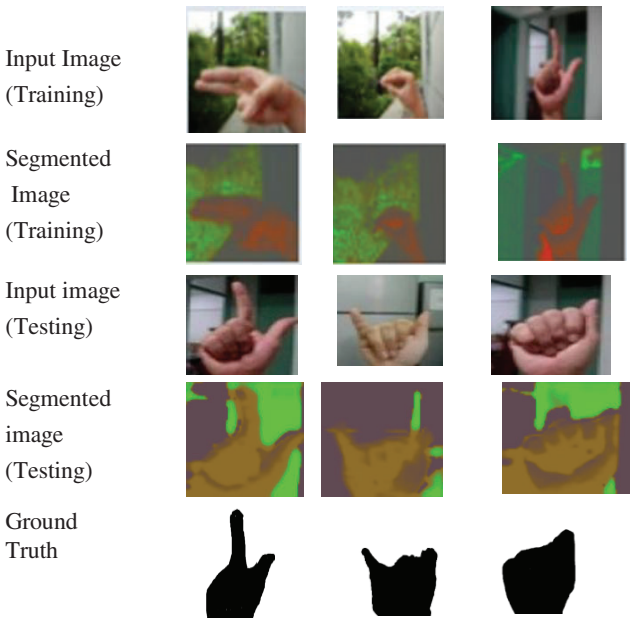


Figure 3: UNET segmentation outputs

The categorical cross entropy loss function is used to quantify the loss by measuring the difference between two discrete probability distributions,

$$Loss = -\sum_{i=1}^{n} y_i \, log\hat{y}_i \tag{1}$$

where $\hat{y}_i$-model output for ith scalar value

y_i- target value

n- output size

when loss sums to one, no surprise and exactly an event occur, and the result is 100% accurate prediction. If the loss is small it results in close prediction with many classes, or it results in poor prediction.

3.3 Ensemble Learning

Ensemble learning improves the prediction accuracy further by combining many base models to provide one optimal predictive model; this increases the stability and predictive power of the model. Ensemble prediction uses the majority vote to find the final class label $\hat{y}_f$

Final prediction

$$\hat{y}_f = mode\{h1(x),\ h2(x),\ h3(x)\} \tag{2}$$

where $h_i(x) = \hat{y}_i$

As n independent classifiers are used to predict, the base error rate (ε) which is better than random guessing. The probability of wrong prediction via ensemble is less, and this is given by the binomial coefficient

$$p(k) = \binom{n}{k} \varepsilon^k (1-\varepsilon)^{n-k} \tag{3}$$

where

n-is the number of classifiers

k-the no of combinations of picking particular classifier and $k > \lceil n/2 \rceil$

The probability of incorrect prediction as a result of using the ensemble of classification models, when k classifiers predict the same class label is given by

$$\varepsilon_{ens} = \sum_{k}^{n} \binom{n}{k} \varepsilon^k (1-\varepsilon)^{n-k} \tag{4}$$

If the base error of a classifier is less than 50%, then the ensemble error rate decreases; else the ensemble error rate increases.

4 Experimental Evaluation

The performance of both the segmentation and recognition process are considered for evaluation. The proposed model is evaluated on publicly available NUS Hand Posture dataset-II, OUHANDS and HGRI benchmark datasets, whose original image sizes are 160×120, 640×480, 640×480 respectively. The experiment is performed using Intel i5 64-bit 1.9 GHZ processor with 8 GB RAM.

4.1 Datasets

OUHANDS dataset has ground truth for segmentation which is used to compare and recognize the hand gesture. It has ten different hand gestures out of a total 2600 images which are divided for training and

testing. Each image is captured under a different atmosphere and situations with complexity in background, variation in illumination, occlusions with different shapes, sizes and colors. Differentiating this feature is a real challenge in the segmentation process. HGRI is another dataset with 1023 images with different gestures corresponding to numbers 1 to 5 and characters A to Z. Datasets with only a few classes do not help in evaluating the recognition of hand gestures. The segmentation performance of the system is done with the NUS hand posture dataset-II with both regular and resized images. This dataset has 11 classes and 200 samples under each class with varying backgrounds and illumination conditions. All the three datasets are used to compare the classification accuracy of the proposed work.

4.2 UNET Performance Evaluation

The UNET semantic segmentation networks training performance is evaluated with the original images and image resized to the dimension 64 × 64 using Bicubic interpolation. Bicubic interpolation retain image information, while it accurately scales the image and render it at real-time at a faster processing speed. The images are processed in a batch size of 32. The performance of the network is evaluated on the basis of the number of epochs and Elapsed time taken for training. The network parameters used to tune the semantic network model are the learning rate of 0.001, the default exponential decay rate for the first moment and second moment of the optimizer β_1 as 0.9 and β_2 as 0.999. Early stopping is used with the patience of 5, to terminate training when the loss no longer decreases after each epoch. The default split of 80%, 10% and 10% of the dataset is used for training, testing and validation respectively.

From Fig. 4 it is evident that the number of epochs required for training UNET with the resized image is less than that of the original image. Similarly, from Fig. 5 it is evident that the time required for training and validating the resized image is less than that of the original image that is also displayed in Tab. 1. Hence the rest of the procedure is carried out with the UNET segmented output obtained from resized input images.

The testing time of each image by the state-of-art-method is compared with the proposed approach and it is found that the current implementation on resized images with modified UNET outperform them. This is highlighted with bold font in Tab. 2, as it is used in further steps of processing.

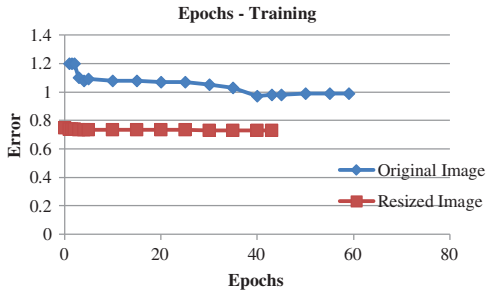


Figure 4: Epochs for training original and resized images

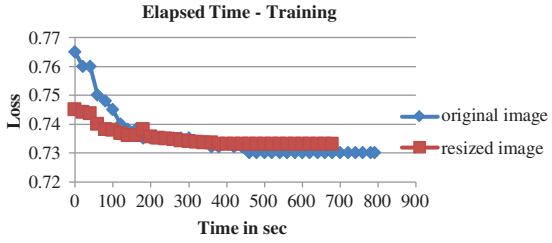


Figure 5: Elapsed time for training original and resized images

Table 1: UNET training time based on original and resized images

Input image	Best epoch	Elapsed time (Training)
Original image	59	791 s
Resized image	43	679 s
Best epoch, elapsed time	**43**	**679 s**

Table 2: Modified UNET performance with other semantic segmentation models

Method	Dataset	Input size	Testing time (ms)
FCN-8S [7]	OUHANDS	224 × 224	63 ms
PSPNET [7]	OUHANDS	224 × 224	50 ms
Deep lab V3 [7]	OUHANDS	224 × 224	43 ms
HGR Net (stage1 no ASPP) [7]	OUHANDS	320 × 320	20 ms
HGR Net (stage1) [7]	OUHANDS	320 × 320	21 ms
Modified UNET [proposed]	OUHANDS	64 × 64	13 ms
Modified UNET [proposed]	HGR1	64 × 64	10 ms
Modified UNET [proposed]	**NUS-II**	**64 × 64**	**9 ms**

4.3 CNN Models Evaluation

The performance of the CNN models is evaluated based on the number of epochs and time taken to train the individual models, this is shown in Tab. 3. The classification accuracy of CNN Architectures EfficientNet B0, EfficientNet B4, ResNet V2 152 is analyzed based on the confusion matrix of the predictive model and various validation and training metrics like accuracy, precision, recall, F1-score and loss. The confusion matrices of the models EfficientNet B0, EfficientNet B4, ResNet V2 152 are shown in Figs. 6–8 respectively.

Table 3: CNN models training performance

CNN Model	Best Epoch	Elapsed time(Training)
EfficientNet B0	33	25 min
EfficientNet B4	29	28 min
ResNet v2 152	28	40 min

The performance of the Ensemble model is better than that of the individual CNN models, since the majority opinions of these models are used to predict the resultant gesture by the bagging approach. These models are independent and run in parallel. The prediction accuracy of the model on the NUS Hand posture dataset-II is shown in Fig. 9.

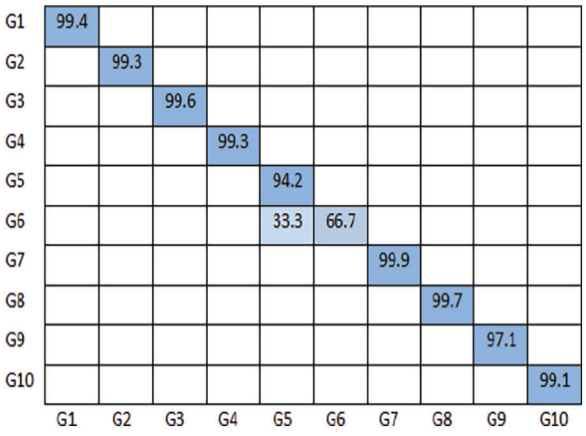


Figure 6: Confusion matrix of EfficientNet B0 model on NUS hand posture dataset-II

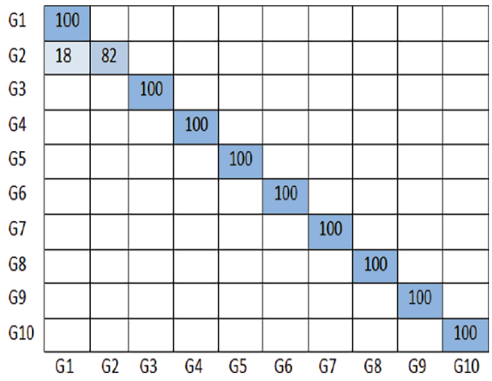


Figure 7: Confusion matrix of EfficientNet B4 model on NUS hand posture dataset-II

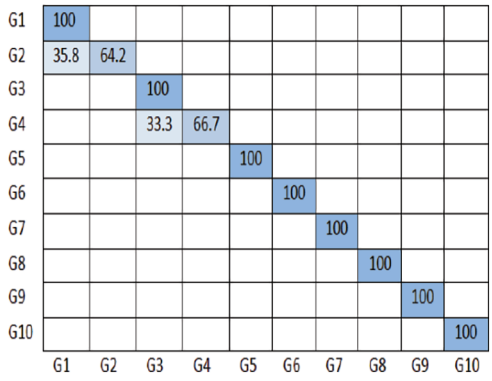


Figure 8: Confusion matrix of ResNet V2 152 model on NUS hand posture dataset-II

4.4 Experimental Setup

The performance measure F-score is used for both the individual models and ensemble models. This is shown in Tab. 4. It is observed that our proposed model performed well, in comparison with other existing deep learning based Neural Network models on benchmark datasets, as shown in Tab. 5. The challenges in segmentation such as background, variation in illumination, Training time and CPU utilization are dealt better and maximum accuracy is achieved.

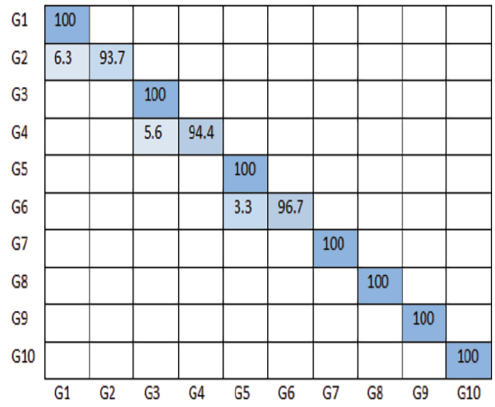


Figure 9: Confusion matrix of ensemble model on NUS hand posture dataset-II

Table 4: Training accuracy and validation accuracy of CNN models

CNN model	EfficientNet B0	EfficientNet B4	ResNet V2 152	EfficientNet B0	EfficientNet B4	ResNet V2 152
	Training accuracy			Validation accuracy		
Accuracy	1.000	0.996	1.000	0.962	1	0.962
Precision	1.000	0.997	1.000	0.867	0.900	0.880
Recall	1.000	0.996	1.000	0.867	0.900	0.867
F1-score	1.000	0.996	1.000	0.860	0.900	0.869
Loss	**0.140**	**0.270**	**0.302**	**0.184**	**0.282**	**0.386**

Table 5: Validation accuracy of ensemble models on benchmark datasets

Dataset	NUS-II	OUHANDS	HGR1
Accuracy	0.983	1	0.982
Precision	0.956	0.980	0.940
Recall	0.938	0.970	0.947
F1-score	0.947	0.978	0.944
Loss	**0.016**	**0.022**	**0.086**

The experimental result is validated on traditional classification methods by comparing with support vector machines, Artificial Neural Network (ANN), k-nearest neighbor (KNN) and CNN. The HGRI dataset is used to compare our proposed model against traditional methods in terms of training and testing time. Traditional methods failed to differentiate background pixels. Our proposed model highly removes skin like pixels in the background during the segmentation stage.

The scores of SVM, KNN, and ANN are presented along with the proposed CNN method in Tab. 6. It is inferred that the proposed ensemble based classifier shows better performance in terms of accuracy with respect to training, testing and validation. The Tab. 6 illustrates the accuracy of the OUHANDS, HGRI

and NUS-II datasets, with and without segmentation. The proposed ensemble model achieves the maximum accuracy on both datasets compared to other methods.

Table 6: Comparison of recognition accuracy with different classifiers on benchmark datasets

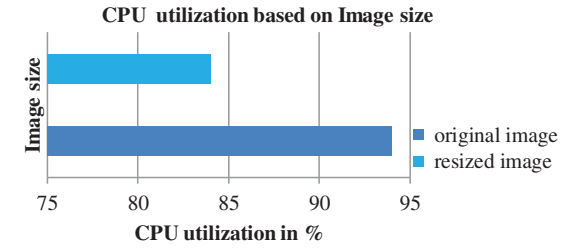
Classifier	Without segmentation	With segmentation	Without segmentation	With segmentation	Without segmentation	With segmentation
	OUHANDS dataset		**HGR1 dataset**		**NUS-II dataset**	
SVM	88.35	-	89.11	-	89.73	-
ANN	90.58	-	86.23	-	91.56	-
KNN	87.89	91.14	93.29	93.60	94.89	92.42
CNN	93.67	96.87	95.08	96.37	93.76	96.78
Proposed ensemble model	**96.05**	**98.98**	**96.25**	**98.76**	**96.67**	**99.07**

4.5 Resource Utilization Inference

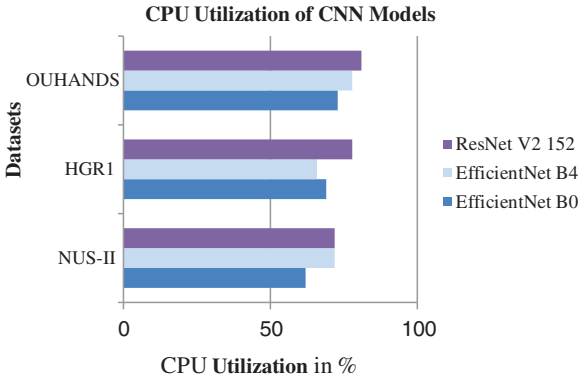
The experiment demonstrates that, as the size of the image to be classified increases, execution time on CPU increases. The execution time depends on processing the image, image splits and distribution to the classifier. The CPU utilization of both segmentation and Classification models are examined using Python cross-platform library psutil with parameter cpu_percent for the duration of 6 min, and accounted for analysis. During the semantic segmentation, the CPU utilization of the original image and resize images with respect to the public bench mark NUS Hand posture dataset-II for 6 min duration is considered for resource utilization inference, and plotted in Fig. 10a. From the figure it is evident that the amount of resource utilized by the original image is greater than the resized images. The resource utilization of the classification models with respect to benchmark datasets is plotted in Fig. 10b. The average CPU utilization of the proposed network models with respect to the benchmark dataset is tabulated in Tab. 7. The average CPU utilizations of the state-of-the-art [30] CNN network models and proposed CNN networks based on Tensor flow library are tabulated in Tab. 8. Comparing the ResNet models of the proposed and state-of-the-art models, the CPU utilization of the proposed methods is comparatively reduced.

5 Future Work

In future the work could be extended for dynamic gestures related to sentences. The dependency path between the hand gestures on different video frames could be combined to form a dependency tree [33,34] and Neural Network Model could be trained with the dependency tree instead of video frames. This could further reduce the time and cost involved in constructing and training the model.



(a) CPU Utilization of UNET with resized and original images of NUS-II Dataset



(b) CPU Utilization of CNN Models on benchmark datasets

Figure 10: (a) CPU utilization of UNET with resized and original images of NUS-II dataset **(b)** CPU utilization of CNN models on benchmark datasets

Table 7: Average CPU utilization of CNN models based on benchmark datasets

CNN architecture	NUS-II %	HGR1%	OUHANDS %
EfficientNet B0	62	69	73
EfficientNet B4	72	66	78
ResNet V2 152	**74**	**78**	**81**

Table 8: Comparison of average CPU utilization of state- of-art [30] with proposed CNN models

State of art [30] (6 min interval)	Proposed (6 min interval)	CPU utilization %
GoogleNet	-	78
ResNet 50	-	82
SqueezeNet	-	61
-	EfficientNet B0	68
-	EfficientNet B4	72
-	ResNet V2 152	78

6 Conclusion

Gestures, which are the non-verbal communication of emotions, play an important role in social interaction, as they convey specific messages effectively. The field of computer vision enables systems to learn human hand gestures using algorithms by providing accuracy and stability. Segmentation of images results in improving the accuracy in recognizing hand gestures. The main issues like variation in image quality during scaling, background complexity, CPU utilization and training time are overcome by the proposed method. Deep learning methods have resulted in improving the accuracy of the segmentation process. The proposed modified UNET model eliminates the background and segments it to recognize the hand gestures in an acceptable time. The reduced convolutional layers which increases the speed, and the ensemble model which improves the prediction accuracy are considered as the main advantages of the system. The results are compared with the existing system with and without the segmentation process. Our experimental results achieved an accuracy of 98.98%, 98.76% and 99.07% with OUHANDS, HGR1 and NUS hand posture-II datasets respectively.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Shukla and A. Dwivedi, "A method for hand gesture recognition," in *Proc. of CSNT*, Karnataka, India, pp. 919–923, 2014.

[2] J. Gangrade and J. Bharti, "Vision-based hand gesture recognition for Indian sign language using convolution neural network," *IETE Journal of Research*, vol. 66, no. 1, pp. 1–10, 2020.

[3] R. A. Elsayed, M. S. Sayed and M. L. Abdalla, "Skin-based adaptive background subtraction for hand gesture segmentation," in *Proc. of ICECS*, Glasgow, UK, pp. 33–36, 2015.

[4] A. Nurhadiyatna, W. Jatmiko, B. Hardjono, A. Wibisono, I. Sina *et al.,* "Background subtraction using Gaussian mixture model enhanced by hole filling algorithm," in *Proc. of ICSMC*, Himachal Pradesh, India, pp. 4006–4011, 2013.

[5] T. Zhang, H. Lin, Z. Ju and C. Yang, "Hand gesture recognition in complex background based on convolutional pose machine and fuzzy Gaussian mixture models," *International Journal of Fuzzy Systems*, vol. 22, no. 4, pp. 1330–1341, 2020.

[6] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with Gaussian edge potentials," *Advances in Neural Information Processing Systems*, vol. 24, no. 10, pp. 109–117, 2011.

[7] A. Dadashzadeh, A. T. Targhi, M. Tahmasbi and M. Mirmehdi, "HGR-Net: A fusion network for hand gesture segmentation and recognition," *IET Computer Vision*, vol. 3, no. 8, pp. 700–707, 2019.

[8] S. Y. Kazdorf, Z. S. Pershina and A. B. Kolker, "Development and research of hand segmentation algorithms on the image based on convolutional neural networks," in *Proc. of INTELS*, St. Petersburg, Russia, pp. 450–454, 2019.

[9] S. Sharma, H. P. J. Dutta, M. K. Bhuyan and R. H. Laskar, "Hand gesture localization and classification by deep neural network for online text entry," in *Proc. of ASPCON*, Kolkata, India, pp. 298–302, 2020.

[10] D. Han, "Comparison of commonly used image interpolation methods," in *Proc. of ICCSEE*, Hangzhou, China, pp. 1556–1559, 2013.

[11] F. Meng, L. Guo, Q. Wu and H. Li, "A new deep segmentation quality assessment network for refining bounding box based segmentation," *IEEE Access*, vol. 7, no. 5, pp. 59514–59523, 2019.

[12] B. Peng, L. Zhang, X. Mou and M. H. Yang, "Evaluation of segmentation quality via adaptive composition of reference segmentations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 10, pp. 1929–1941, 2016.

[13] J. C. Nascimento and G. Carneiro, "One shot segmentation: Unifying rigid detection and non-rigid segmentation using elastic regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 12, pp. 3054–3070, 2019.

[14] B. Peng, X. Wang and Y. Yang, "Region based exemplar references for image segmentation evaluation," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 459–462, 2016.

[15] T. Schoenemann, S. Masnou and D. Cremers, "The elastic ratio: Introducing curvature into ratio-based image segmentation," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2565–2581, 2011.

[16] K. K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixé *et al.,* "Video object segmentation without temporal information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 6, pp. 1515–1530, 2018.

[17] B. Peng. and T. Li, "A probabilistic measure for quantitative evaluation of image segmentation," *IEEE Signal Processing Letters*, vol. 20, no. 7, pp. 689–692, 2013.

[18] P. Yin, R. Yuan, Y. Cheng and Q. Wu, "Deep guidance network for biomedical image segmentation," *IEEE Access*, vol. 8, no. 6, pp. 116106–116116, 2020.

[19] T. H. Tsai and S. A. Huang, "Refined U-net: A new semantic technique on hand segmentation," *Neuro Computing*, vol. 495, no. 7, pp. 1–10, 2022.

[20] L. Tiantian, S. Jinyuan, L. Runjie and G. Yingying, "Hand gesture recognition based on improved histograms of oriented gradients," in *Proc. of CCDC*, Qingdao, China, pp. 4211–4215, 2015.

[21] C. Yu, X. Wang, H. Huang, J. Shen and K. Wu, "Vision-based hand gesture recognition using combinational features," in *Proc. of IIHMSP*, Adelaide, SA, Australia, pp. 543–546, 2010.

[22] M. K. Ahuja and A. Singh, "Static vision based hand gesture recognition using principal component analysis," in *Proc. of MITE*, Amritsar, Punjab, India, pp. 402–406, 2015.

[23] L. C. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv Preprint, vol. 3, no. 6, pp. 1–14, 2017.

[24] R. Girshick, J. Donahue, T. Darrell and J. Malik "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of CVPR*, Columbus, OH, USA, pp. 580–587, 2014.

[25] J. P. Sahoo, S. Ari and S. K. Patra, "Hand gesture recognition using PCA based deep CNN reduced features and SVM classifier*,"* in *Proc. of iSES*, Rourkela, India, pp. 221–224, 2019.

[26] D. Frolova, H. Stern and S. Berman, "Most probable longest common subsequence for recognition of gesture character input," *IEEE Transactions on Cybernetics*, vol. 43, no. 3, pp. 871–880, 2013.

[27] M. C. Su, "A fuzzy rule-based approach to spatio-temporal hand gesture recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part C Applications and Reviews*, vol. 30, no. 2, pp. 276–281, 2000.

[28] A. Sreekumar and M. Geetha, "Hand segmentation in complex background using UNet," in *Proc. of ICIRCA*, Coimbatore, India, pp. 440–445, 2020.

[29] H. P. J. Dutta, D. Sarma, M. K. Bhuyan and R. H. Laskar, "Semantic segmentation based hand gesture recognition using deep neural networks," in *Proc. of NCC*, Kharagpur, India, pp. 1–6, 2020.

[30] D. Velasco-Montero, J. Femández-Bemi, R. Carmona-Gálán. and A. Rodríguez-Vázquez, "On the correlation of CNN performance and hardware metrics for visual inference on a low-cost CPU-based platform," in *Proc. of IWSSIP*, Niteroi, Brazil, pp. 249–254, 2019.

[31] M. Andrew, S. Bhattiprolu, D. Butnaru and J. Correa, "The usage of modern data science in segmentation and classification: Machine learning and microscopy," *Microscopy and Microanalysis*, vol. 23, no. 1, pp. 156–157, 2017.

[32] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. of PMLR*, Vancouver, BC, Canada, pp. 6105–6114, 2019.

[33] H. Sun and R. Grishman, "Lexicalized dependency paths based supervised learning for relation extraction," *Computer Systems Science and Engineering*, vol. 43, no. 3, pp. 861–870, 2022.

[34] H. Sun and R. Grishman, "Employing lexicalized dependency paths for active learning of relation extraction," *Intelligent Automation & Soft Computing*, vol. 34, no. 3, pp. 1415–1423, 2022.