SSEC: Semantic Segmentation and Ensemble Classification Framework for Static Hand Gesture Recognition using RGB-D Data

Dayananda Kumar NC[1], K.V Suresh[2], Chandrasekhar V[3], Dinesh R[4]
Dept. of Electronics and Communication Engineering[1,2,3]
Siddaganga Institute of Technology, Tumkur, India[1,2]
IIIT Sricity, India[3]
Dept. of Information Science and Engineering, Jain University, Bangalore, India[4]

***Abstract*—Hand Gesture Recognition (HGR) refers to identifying various hand postures used in Sign Language Recognition (SLR) and Human Computer Interaction (HCI) applications. Complex background in uncontrolled environmental condition is the major challenging issue which impacts the recognition accuracy of HGR system. This can be effectively addressed by discarding the background using suitable semantic segmentation method, where it predicts the hand region pixels into foreground and rest of the pixels into background. In this paper, we have analyzed and evaluated well known semantic segmentation architectures for hand region segmentation using both RGB and depth data. Further, ensemble of segmented RGB and depth stream is used for hand gesture classification through probability score fusion. Experimental results shows that the proposed novel framework of Semantic Segmentation and Ensemble Classification (SSEC) is suitable for static hand gesture recognition and achieved F1-score of 88.91% on OUHANDS test dataset.**

***Keywords*—Hand gesture recognition; semantic segmentation; ensemble classification; score fusion**

I. Introduction

Hand gestures plays significant role in many real time applications like robotics control, gaming, 3D modeling, virtual environment etc,. Various methods are used to detect and recognize the hand gestures depending on the data acquisition and processing system [1] [2]. Hand gesture recognition systems can be broadly classified into sensor based and vision based systems.

In sensor based systems, data is captured using the sensor modules connected to the hand which converts the hand movements into varying time series signal. These devices captures the hand movement data very precisely but not ease for usage as it impose the constraint to wear the device and not supporting the contact less operation [3] [4].

In vision based systems, RGB cameras are widely used to capture the hand pose data as color images. Vision based systems involve hand detection or segmentation as one of the important pre-processing step involved in gesture recognition pipeline to localize the hand region and discard the background in the image [5]. Feature extraction methods are used on the segmented hand data to obtain its characteristic representation which are used in classification stage to effectively recognize the various hand gestures. Current state-of-the-art recognition systems based on the color image data face many challenges in hand segmentation and recognition due to the complex background and varying illumination conditions. Gestures performed by various subjects differing in their hand size and color is difficult to identify due to large intra-class and inter-class variations. Addressing these issues is difficult by using the color modality data. Hence recent systems are developed on dual modality using RGB and Depth data known as RGB-D data. In these systems, data is simultaneously captured using both color and depth sensor to obtain the RGB-D aligned data pair registered on the same view of camera coordinate system [6].

Depth map can be obtained using various techniques like stereo, time-of-flight of IR etc., where it provides the distance information between depth sensor and the object scene [7]. Kinect device sensor uses time-of-flight between the emitted IR light and the reflected light on projector to provide raw depth map in which each pixel location represents the distance in millimeter. Depth modality can be effectively used to discard the far away background based on depth distance range. It helps in effective hand segmentation to group hand region pixels into foreground and rest of the image pixels into background [8]. Also in case of low light scenarios RGB sensors fails to capture the data, this issue can be resolved using Kinect depth sensor as it captures the data using IR light.

Various image processing and computer vision algorithms are discussed in literature for hand region segmentation. According to recent studies, CNN architectures are widely used to address various real time problems and segmentation is one of the majorly studied area. CNN based semantic segmentation networks performs pixel level localization of region of interest where it classifies each pixel into its corresponding segmentation class. It provides the fine boundary of each distinct region mapped to the unique segmentation class.

In this paper, we analyzed various segmentation methods on RGB and depth data and provided comparative analysis to identify the best suitable method for hand segmentation.

The organization of the paper is as follows. In Section II, a brief review of different methods that exist in vision based hand gesture recognition is presented. In Section III, problem statement and the proposed method is discussed. In Section IV, detailed experimental results is presented. Experimental outcomes are briefly discussed in Section V. Finally, conclusion

is drawn in Section VI mentioning the limitation of current work and scope of future work.

II. Literature Review

In this section, we discuss on the state-of-the-art methods for hand region segmentation and gesture classification along with their advantages and limitations.

Earlier approaches of hand segmentation in color image were based on color intensity thresholding in RGB, HSV, YCbCr and other color spaces [9]. In these methods suitable color range was identified based on the experimentation to segment the skin region. Limitation of this approach is difficulty in selecting the threshold range to segment all variation of skin color and the lighting variations which significantly affect the segmentation accuracy.

Hand region segmentation based on human skin tones was proposed in [10] using an MLP network to learn the skin color tones and classify the pixels of image which belongs to the skin color sets.

User independent recognition system using low-cost Microsoft Kinect depth sensor was proposed in [11] to overcome illumination and background variations issue in color-based sign language recognition. Here hand region was segmented by using a pre-processing algorithm on depth image. Features are extracted from hand segmented data using CNN based unsupervised Principal Component Analysis Network (PCANet) and classified using Support Vector Machine (SVM) classifier.

Real-time hand gesture recognition method was put forth in [12] using light-weight semantic segmentation method (FASSD-Net) to produce hand segmentation masks which are combined with RGB frames in gesture classification using Temporal Segment Networks (TSN) and Temporal Shift Modules (TSM) tested on IPN Hand dataset.

Various interactive methods like Graph cut, Random walker, geodesic star convexity etc., were analyzed in [13] for hand region segmentation. Five distinct types of hand motions in various backdrops were tested using the Expectation Maximum technique to learn the parameters of the Gaussian Mixture Model and the Gibbs random field to image segmentation by minimising the Gibbs Energy using the Min-cut theorem. According to experimental findings, utilising manually segmented photos improves recognition accuracy when compared to unsegmented images.

Bin et. al [14] proposed a fine-tuned Inception V3 RGB-D static gesture recognition method. This framework eliminates the gesture segmentation and feature extraction steps in traditional algorithms. The proposed framework consists of a CNN architecture in which feature concatenate layer concatenates the features of RGB and depth images. Compared with general CNN, the Inception V3 based gesture recognition resulted in improved accuracy.

D Kumar et. al [15] proposed a two stage approach for static hand gesture recognition using RGB-D data. In first stage k-means clustering algorithm is applied on the depth image to cluster the foreground and background depth pixels based on the distance. Depth threshold is computed as the mean of cluster centers and using this dynamic threshold background is discarded. In classification stage, segmented RGB-D data is stacked to form the input to data layer of custom CNN network.

Coarse to fine segmentation approach using depth map was proposed in [8] where pre-trained YOLO-v3 model was used to detect and localize the hand region at coarse level. The hand detected bounding region was used to initialize the foreground in graph cut segmentation algorithm which refines the hand region boundary and discards the background. Hand segmented RGB-D data was further used in classification stage to recognize the hand gestures.

The hand region in the depth map was segmented using the depth thresholding approach in [16]. Additionally, a two stream network with AlexNet and VGG16 was employed using score-level fusion technique to recognise the static hand gestures from the datasets from Massey University (MU) and HUST American Sign Language (HUST-ASL) with accuracy of 98.14 % and 64.55 % respectively.

Hand Gesture Recognition Approach called HGRA on RGB data using two stream was proposed by [17], in first branch U-Net combined with Multi-Scale Attention module is used to segment the hand region and extracting shape features. In second branch, Multi-Scale Fusion (MSF) and Light-Weight Multi-Scale (LWMS) modules are used to extract multi-scale appearance and color features. This method was evaluated on OUHANDS and HGR1 datasets and achieved the accuracy of 90.9% and 83.8% respectively.

Three stage spatial attention-based neural network was proposed in [18]. First two stages include generation of feature vector and attention map with the feature extraction architecture and self-attention technique. Final feature is generated after multiplying the features and attention map and feed to classification module in third stage to predict the label of hand gesture. This model achieved 99.75%, 99.46% and 99.67% accuracy in Kinematic, NTU and senz3D datasets respectively.

Dual-stream dense residual fusion network(DeReFNet) was proposed in [19] which utilizes the strength of global features and spatial information from the residual stream and other stream. Both the streams are fused using the feature concatenation module. Subject-independent cross-validation technique is used to validate DeReFNet four publicly available benchmark datasets.

Kinect sensor device is used to capture hand gesture depth images. Serial binary image extraction is used in [20] to eliminate the undesired shadow region in depth image and improve the recognition accuracy using VGG-type CNN. Emergence of industry 4.0 with need of natural human-robot interaction in manufacturing using vision-based and wearable-based approaches for gesture-based interaction is discussed in [21]. Position data from Microsoft Kinect RGB-D cameras and acceleration data from inertial measurement units (IMUs) is compared to evaluate the recognition accuracy.

Based on the brief literature review it can be observed that most of the recent research in hand gesture recognition use RGB-D data. Early methods of hand region segmentation used skin color based segmentation in different color space, later CNN based semantic segmentation methods gained much attention due to its efficiency and robustness even in complex

scene. Depth modality can be used both in segmentation and classification, hence active research is being carried out in state-of-the art methods to evaluate various ensembling and fusion techniques of RGB and Depth modalities [22]. In further section, we discuss about the details of proposed method and experimental analysis.

III. Proposed Method

Based on the literature review, it is evident that hand gesture recognition is still an active area of research trying to solve the challenges of gesture recognition is real scenarios with complex background scene and varying lighting conditions. Current research methods have also showed that multi-modal RGB and depth stream data is effective than uni-modal RGB data for hand gesture recognition.

In this paper, we analyze various semantic segmentation methods to effectively segment the hand region using RGB and Depth stream data. Hand segmented RGB and Depth data are further used to train custom CNN model for gesture classification. Suitable approach for fusing the probability scores from both the models are analyzed and proposed ensemble classification framework for static hand gesture recognition.

The main contributions of this paper are:

1) Analysis of semantic segmentation model accuracy using RGB data, depth data and combined RGB-D data.
2) Proposed the ensemble approach of score fusion for static HGR classification on RGB and Depth data.

A. Semantic Segmentation

Semantic segmentation is a pixel-based classification in which each pixel of an image is classified to its corresponding class. Here the class labels of all the pixels of image are predicted, hence segmentation is also termed as dense prediction. Hand region segmentation is a binary case of segmentation which has two output class and provides the pixel level mapping into required foreground and background regions as in Fig. 1. In this work, we evaluate various CNN architectures like UNet, ResUNet and DeeplabV3-Plus for semantic segmentation of hand region on both RGB and Depth data.

B. UNet

U-Net architecture [23] adopts auto-encoder framework which consists of two components known as encoder and decoder as in Fig. 1. Encoder generates the compressed feature representation of the image using down sampling and strided convolution, these features contribute in classifying the pixels into its corresponding segmentation class. The encoder and decoder layers are symmetrical to each other. Decoder includes up-sampling and transpose convolution which generates the output segmentation map which has the same resolution as the input image to segmentation model. The least squares reconstruction error is back propagated from the decoder to encoder using which the weights are updated to obtain optimal feature representation.

Encoder generally have the following sub layers, Convolution layer, Relu activation layer and pooling layer. The input

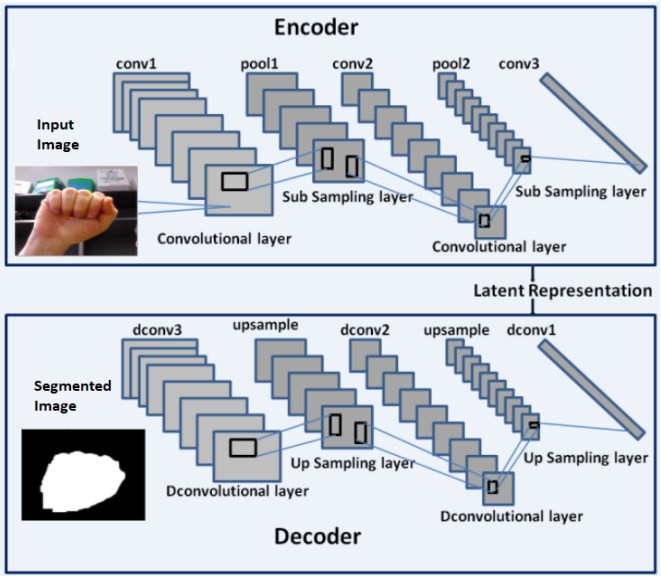

Fig. 1. Auto-encoder based segmentation network.

image is fed in to the data layer followed by convolution layer which consists of filter of size 3x3 followed by Relu activation to add non-linearity. In the subsequent layers the number of kernels is doubled with constant kernel size and the max pooling layer is used to reduce or down sample the feature map and to maintain the local dominant features in the image patch. Here we have modified original architecture by removing last block of convolution layers and used only three blocks which consists of two convolution layers in each block for model convergence.

Decoder is used to reconstruct the input image using the reduced representation from encoder layer. Encoded input images are decoded by a series of up sampling and de-convolution block. The up-sampling operation of the decoder layers use the max-pooling indices of the corresponding encoder layers. The decoder architecture follows certain pattern based on its encoder design, where the decoder is mirror replica of encoder. The decoded image is evaluated against the input image while self learning the feature representation.

C. ResUNet

Residual U-Net [24] is a semantic segmentation network in which the residual blocks are used in encoder and decoder block of U-Net architecture. This residual learning helps to improve the U-Net results and only with fewer parameters. Fig. 3 shows the basic unit blocks of U-Net (a) and ResUnet (b). Each residual unit can be mathematically shown as in Eq.1.

$$\begin{aligned} y_l &= h(X_l) + F(X_l, W_l), \\ X_l + 1 &= f(y_l) \end{aligned} \tag{1}$$

where X_l and $X_l + 1$ are the input and output of the l^{th} residual unit, F is the residual function, $f(y_l)$ is activation function and $h(X_l)$ is a identity mapping function, where $h(X_l) = X_l$.

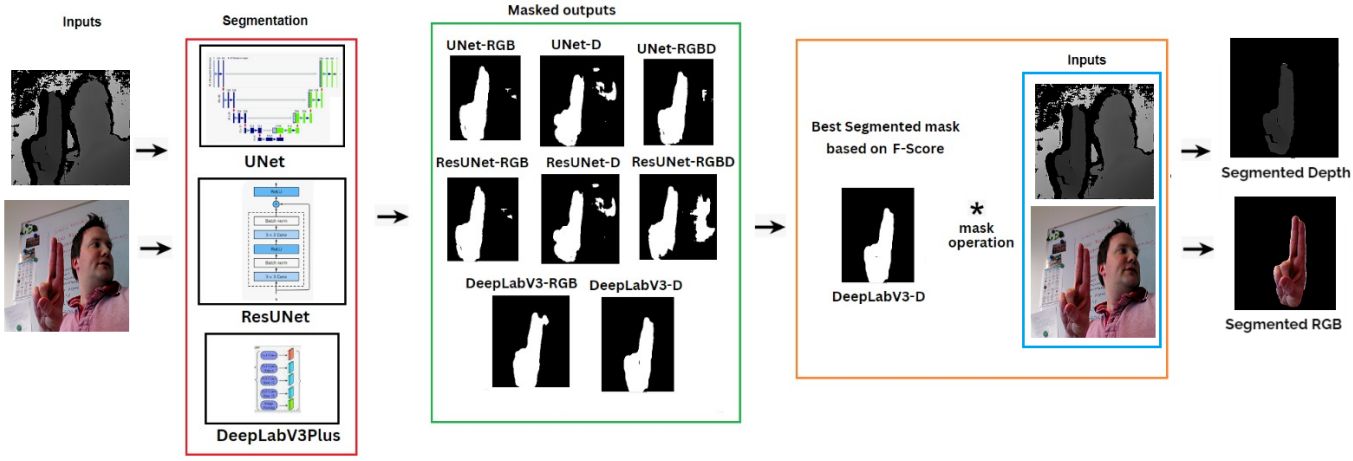

Fig. 2. Block diagram of RGB-D semantic segmentation analysis.

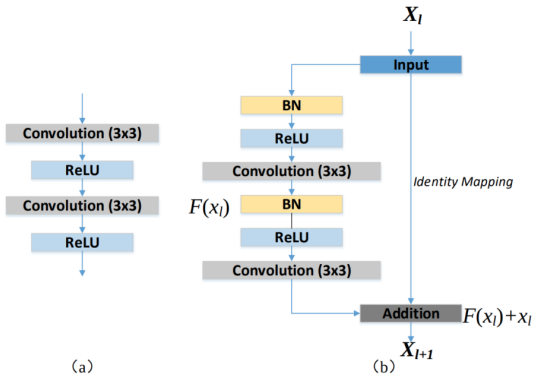

Fig. 3. (a) U-Net unit block and (b) ResUnet unit block.

ResUnet comprises of three parts built with residual units, encoding, bridge and decoding. The first part encodes the input image into latent feature representation. Second part forms a bridge connecting the encoding and decoding paths. Third part of decoding provides the semantic labels to each pixel by pixel-wise classification. Each residual units consists identity mapping and two 3 × 3 convolution blocks with BN layer and ReLU activation layer. The identity mapping connects the inputs and outputs of the unit. Decoding path consists of three residual units and concatenated with the feature maps from the corresponding encoding path. After the last level of decoding path, a 1 × 1 convolution and a sigmoid activation to obtain desired segmentation.

D. DeepLab-V3+

DeepLabv3+ [25] is the extended version of DeepLabv3 segmentation architecture. It follows encoder-decoder structure with Atrous Spatial Pyramid Pooling (ASPP) module in the encoder block. Hence encoder module processes multi-scale contextual information at multiple rates and multiple effective fields-of-view by applying dilated convolution at multiple scales. The decoder module with depthwise separable convolution refines the segmentation results along object boundaries by gradually recovering the spatial information.

Fig. 2 shows the block diagram of segmentation architecture analysis. UNet and ResUnet models are trained on RGB, Depth and stacked RGB-D data, DeepLabv3+ is trained on RGB and Depth data. All these models are trained separately and results are analyzed using mean IoU (Intersection over Union), average F1-score metrics. DeepLabv3+ trained on depth data gave better accuracy with comparatively less parameters, hence this model is selected as the best model for hand region segmentation. Predicted binary segmentation mask from this model is combined with RGB and depth data to obtain segmented RGB and segmented Depth data, this data is further used as input to the classification model.

E. Ensemble Classification

Classification block diagram is depicted in Fig. 4, where two classification models using segmented RGB and segmented depth data stream are trained independently using the custom CNN network as shown in Table I. Further, the classification probability of segmented RGB and depth model are analyzed and fused using max and average operator to select the best classification model.

Custom CNN-Net architecture in Table I consists of four groups with two layers of convolution CONV2D and RELU activation, Batch normalization and Max pooling. The number of kernels is increased as [16, 32, 64, 128] in subsequent groups. Global average pooling layer is used to get final feature map, two fully connected layers are used followed by Softmax activation to get the probability output of each class.

Model is trained using Adam optimizer with learning rate of 0.001 and categorical cross-entropy is used as loss function. Batch normalization and drop out layers are used avoid the model from over-fitting. RGB and depth data from OUHANDS dataset is resized to 320 x 320 and segmented using the binary mask obtained from Deeplabv3+ depth segmentation model.

Two classification models are trained using the segmented RGB and depth data, these models are evaluated on the test data and the miss-classified images are analyzed. It is observed that some of the images that were wrongly classified in the RGB stream are detected properly in the depth stream and

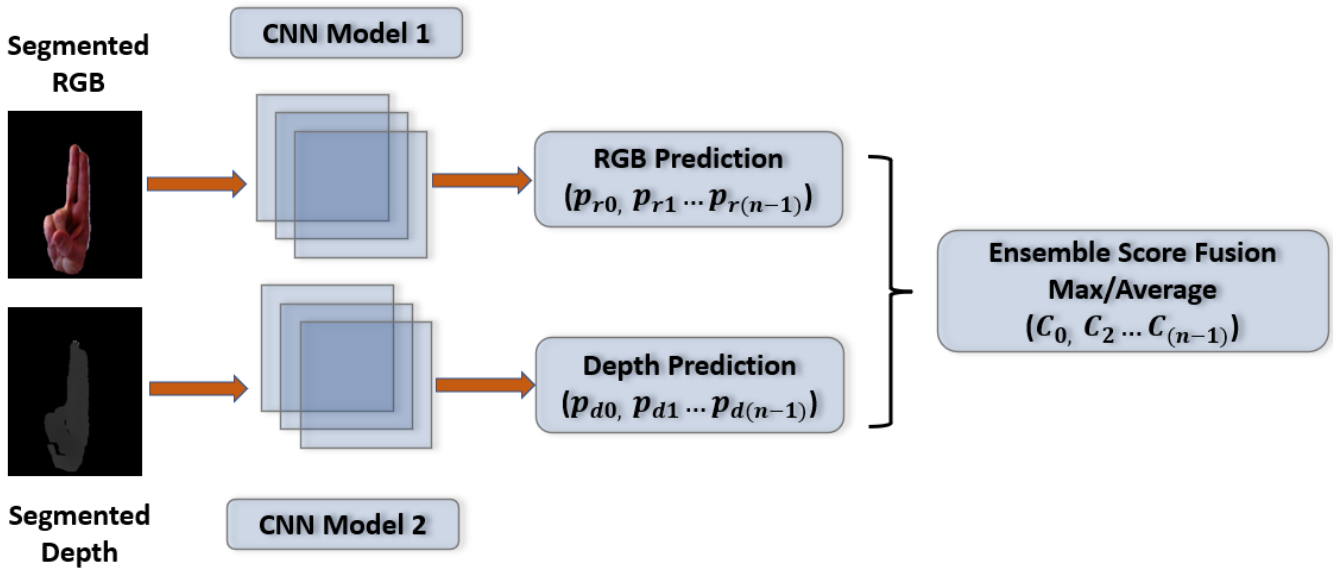

Fig. 4. Block diagram of ensemble score fusion for hand gesture classification.

TABLE I. CLASSIFICATION MODEL ARCHITECTURE

Layer Type	Layer Name	Output Shape
Data	RGB input	(320, 320, 3)
CONV2D + RELU	conv2d_relu_1	(320, 320, 16)
CONV2D + RELU	conv2d_relu_2	(320, 320, 16)
BatchNormalization	batch_norm_1	(320, 320, 16)
MaxPooling	max_pooling2d_1	(106, 106, 16)
CONV2D + RELU	conv2d_relu_3	(106, 106, 32)
CONV2D + RELU	conv2d_relu_4	(106, 106, 32)
BatchNormalization	batch_norm_2	(106, 106, 32)
MaxPooling	max_pooling2d_2	(35, 35, 32)
CONV2D + RELU	conv2d_relu_5	(35, 35, 64)
CONV2D + RELU	conv2d_relu_6	(35, 35, 64)
BatchNormalization	batch_norm_3	(35, 35, 64)
MaxPooling	max_pooling2d_3	(11, 11, 64)
CONV2D + RELU	conv2d_relu_7	(11, 11, 128)
CONV2D + RELU	conv2d_relu_8	(11, 11, 128)
BatchNormalization	batch_norm_4	(11, 11, 128)
Global average pooling	gap2d_1	(128)
Dense	dense_1	(64)
RELU	activation_1	(64)
Dropout (0.5)	dropout_1	(64)
Dense	dense_2	(10)
Softmax Activation	activation_2	(10)

vice versa, hence this forms the basis to build a score fusion ensemble model with both RGB and depth stream which gives better accuracy as compared to the uni-modal results.

Let $P_R = (p_{r0}, p_{r1}, p_{r2} \ldots p_{r(n-1)})$ and $P_D = (p_{d0}, p_{d1}, p_{d2} \ldots p_{d(n-1)})$ be the probability vectors obtained from RGB stream and depth stream respectively, where $n = 10$ represents the number of class.

$$P_{E_{(max)}} = (max(p_{r0}, p_{d0}), max(p_{r1}, p_{d1}), max(p_{r2}, p_{d2}) \ldots . max(p_{r(n-1)}, p_{d(n-1)})) \quad (2)$$

$$P_{E_{(avg)}} = 0.5 * ((p_{r0} + p_{d0}), (p_{r1} + p_{d1}), (p_{r2} + p_{d2}) \ldots . (p_{r(n-1)} + p_{d(n-1)})) \quad (3)$$

Probability score fusion of max and average methods is mathematically shown in Eq. 2 and Eq. 3. In max fusion, maximum value of RGB and depth probability is considered for each class, where in average fusion mean probability is taken. Further, max of these fused probability is taken to decide the class label of ensemble classification model. From the experiments, it is found the average fusion gives better results as compared to max fusion.

F. Evaluation Metrics

Most commonly used principal measures to evaluate semantic segmentation and classification performance are briefly explained below.

Intersection over Union (IoU) - It is computed as intersection of the pixels from a given class in the predicted results with the ground truth divided by their union. IoU is computed class wise in case of multi-class segmentation. In our work, it is binary case of foreground hand region and the background hence only the class of pixels belonging to foreground is considered.

$$IoU = \frac{T_p}{T_p + F_p + F_n} = \frac{c_{jj}}{c_{ij} + c_{ji} + c_{jj}} i \neq j \quad (4)$$

where, c_{jj} = T_p is the number of pixels which are labeled as class j in ground truth and also predicted as class j, c_{ij} = F_p is the number of pixels which are labeled as class i, but classified as class j that is False Positives class for class j. Similarly, c_{ji} = F_n , the total number of pixels labeled as class j, but classified as class i are the False Negatives (misses) for class j.

Mean Intersection over Union (mIoU): mIoU is the class-averaged IoU across all the images, where k is the number of class.

$$mIoU = \frac{1}{k} \sum_{j=1}^{k} \frac{c_{jj}}{c_{ij} + c_{ji} + c_{jj}} \quad (5)$$

Precision - It is the ratio of hits over summation of hits and false alarms. It indicates total positive cases predicted correctly, over all the predicted positive cases.

$$Precision = \frac{T_p}{T_p + F_p} = \frac{c_{jj}}{c_{ij} + c_{jj}} i \neq j \tag{6}$$

Recall - It is the ratio of hits over summation of hits and misses. It indicates total positive cases predicted correctly, over all the actual positive cases.

$$Recall = \frac{T_p}{T_p + F_n} = \frac{c_{jj}}{c_{ji} + c_{jj}} i \neq j \tag{7}$$

F1score - This measure also known as the dice coefficient, computed as harmonic mean of the precision and recall.

$$F1_{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

In this work precision, recall and F1-score are computed for both segmentation and classification models. In segmentation predicted pixel class is considered whereas in classification predicted image class label. Additionally IoU metrics are used to evaluate segmentation models.

IV. Experimental Results

The appraise the proposed semantic segmentation and ensemble classification framework, experiments are conducted on widely used benchmark OUHANDS [26] dataset for static hand gesture recognition. The experiments are performed using TensorFlow2.0 Keras deep learning library in Google Colab environment with NVIDIA GPU.

OUHANDS dataset [26] - contains RGB, raw depth data and segmentation ground truth images. It consists of 10 unique gestures captured in complex backgrounds and lighting changes from 23 subjects with different hand gesture sizes and shapes. Training dataset consists of 2000 images split into 1600 for training, 400 for validation. Test dataset contains 1000 images of the unseen individuals in the training set. All images in the training and test datasets are resized to 320 x 320 image resolution and used in segmentation and classification.

As depicted in Fig. 2, We have chosen three well known segmentation networks namely: UNet, ResUNet and DeepLabV3+. RGB and Depth images are used as input into these networks, which outcomes the hand region segmented mask. We trained these networks with various type of input data like RGB, Depth and the stacked RGBD data. Out of these 3 combination Depth data based model gave better accuracy with fine segmented hand region mask.

As shown in Table II, it can be observed that DeepLabV3+ segmentation results are better than UNet and ResUNet models. Also it can be observed that, the Depth data based based models provide better F1-score as compared to RGB data and stacked RGBD data. The DeepLabV3+ depth segmentation model resulted in the highest F1 score of 0.9235 compared to other networks.

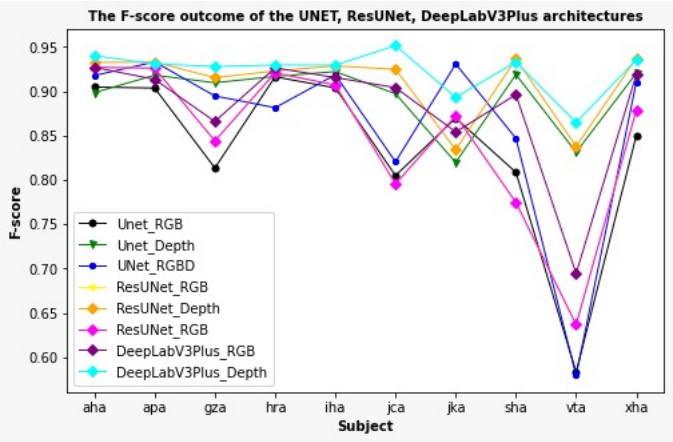

Fig. 5. Analysis of segmentation F1-score outcome by various networks specific to each subject.

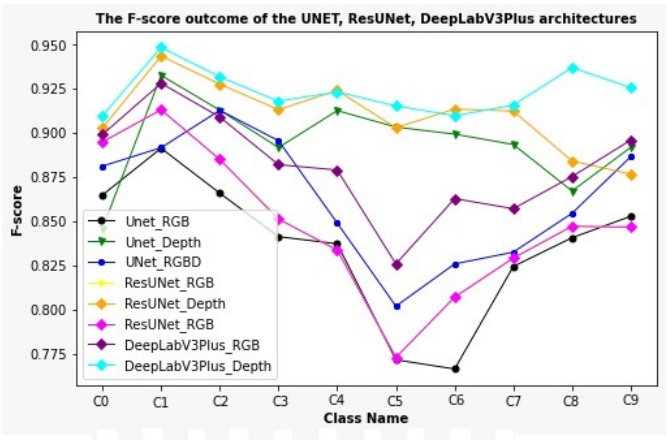

Fig. 6. Analysis of segmentation F1-score outcome by various networks specific to each class.

We have analyzed the segmentation accuracy for each subject performing various gestures under different lighting conditions as shown in Fig. 5. As it can be observed, Deeplabv3+ depth segmentation model gave better accuracy for all the subjects.

Segmentation results were analyzed over each gesture class as shown in Fig. 6. From the plot it is evident that Deeplabv3+ model gives better segmentation accuracy over UNet and ResUNet models. Hence we can conclude that DeepLabV3+ segmentation model trained with Depth data is suitable for efficiently segmenting the hand region in complex scenarios.

Based on experimental results DeepLabV3+ is selected as best segmentation model, predicted binary segmentation mask is used for masking to discard the background by bitwise AND operation on the input RGB and depth image. This results in fine segmented foreground hand region in RGB and Depth image constituting to segmented RGB and Segmented Depth images.

Table III shows the classification results on OUHANDS test dataset from the models trained on segmented RGB and Segmented Depth training data using custom CNN network

TABLE II. COMPARISON OF HAND REGION SEGMENTATION ACCURACY ON OUHANDS TEST DATASET

No	Method	Data Type	Mean IOU	Average Precision	Average Recall	Average F1 score	Num. of Parameters	Model Size
1	UNet	RGB	0.7404	0.8655	0.8303	0.8358	7.861 M	90.1 MB
2		Depth	0.8207	0.8898	0.9161	0.8952	7.861 M	90.1 MB
3		RGBD	0.7878	0.8905	0.8747	0.8633	7.861 M	90.1 MB
4	ResUNet	RGB	0.7577	0.8725	0.8503	0.8482	4.680 M	53.9 MB
5		Depth	0.8477	0.9223	0.9136	0.9102	4.680 M	53.9 MB
6		RGBD	0.7803	0.9250	0.8298	0.8542	4.680 M	53.9 MB
7	DeepLabV3 Plus	RGB	0.8020	0.8717	0.9055	0.8815	17.830 M	204.7 MB
8		Depth	0.8638	0.9022	0.9534	0.9235	17.830 M	204.7 MB

TABLE III. COMPARISON OF PRECISION, RECALL AND F1-SCORE OF OUHANDS TEST DATASET CLASSIFICATION USING SEGMENTED RGB, DEPTH, MAX AND AVERAGE ENSEMBLE DATA

Class	Segmented RGB			Segmented Depth			Max Ensemble			Average Ensemble		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
C0	0.9091	0.9000	0.9045	0.8704	0.9400	0.9038	0.8455	0.9490	0.8942	0.8440	0.9583	0.8976
C1	0.8879	0.9694	0.9268	0.9700	0.9898	0.9798	0.9897	0.9796	0.9846	0.9896	0.9794	0.9845
C2	0.8588	0.7374	0.7935	0.9146	0.7576	0.8287	0.8941	0.7677	0.8261	0.8929	0.7732	0.8287
C3	0.9545	0.8400	0.8936	0.8557	0.8300	0.8426	0.8864	0.7879	0.8342	0.8876	0.7980	0.8404
C4	0.9029	0.9300	0.9163	0.8911	0.9000	0.8955	0.8205	0.9600	0.8848	0.8288	0.9684	0.8932
C5	0.9524	0.8247	0.8840	0.9659	0.8763	0.9189	0.9778	0.8889	0.9312	0.9773	0.8866	0.9297
C6	0.8776	0.8687	0.8731	0.8800	0.8889	0.8844	0.8980	0.8800	0.8889	0.9053	0.8776	0.8912
C7	0.8990	0.8900	0.8945	0.8500	0.8500	0.8500	0.8947	0.8500	0.8718	0.9043	0.8763	0.8901
C8	0.7870	0.8500	0.8173	0.8654	0.9000	0.8824	0.8505	0.9100	0.8792	0.8660	0.8936	0.8796
C9	0.7521	0.9192	0.8273	0.7768	0.8788	0.8246	0.8190	0.8687	0.8431	0.8286	0.8878	0.8571
Average	**0.8780**	**0.8730**	**0.8731**	**0.8836**	**0.8810**	**0.8809**	**0.8875**	**0.8841**	**0.8837**	**0.8926**	**0.8895**	**0.8891**

TABLE IV. COMPARISON OF HAND GESTURE RECOGNITION ACCURACY ON OUHANDS TEST DATASET

No	Method	Input data	F1-score	Input Size	# Parameters	Model Size
1	ResNet-50 [27]	RGB	0.8138	224×224	23.60 M	99 MB
2	DenseNet-121 [28]	RGB	0.8281	224×224	7.04 M	33 MB
3	Two stream CNN [29]	RGB & sMask	0.8621	256×256	-	-
4	MobileNet [30]	RGB	0.8650	224×224	3.22 M	16 MB
5	RGB-D Early fusion [15]	sRGB & sDepth	0.8757	320×320	0.3035 M	3.6 MB
6	HGR-Net [31]	RGB & sMask	0.8810	320×320	0.499 M	2.4 MB
7	Proposed SSEC	sRGB & sDepth	**0.8891**	320×320	sRGB = 0.3034 M sDepth = 0.3034 M	3.6 MB 3.6 MB

sRGB = Segmented RGB
sDepth = Segmented Depth
sMask = Segmented Binary Mask

shown in Table I. Accuracy of these two models are analyzed and found that few miss-classified images are detected complementary, hence score fusion is performed using max and average operations to get ensembled gesture classification results. It can be observed that average ensemble gives the best result of 88.91%. It is also evident from Fig. 7, average ensemble provides best accuracy over all models for all the gesture class.

Proposed framework of semantic segmentation and ensemble classification (SSEC), is compared with the state of the art methods as shown in Table IV. Deeplabv3+ depth segmentation model followed by classification using average score fusion gives the best F1 score accuracy on OUHANDS test dataset.

V. DISCUSSION

Comprehensive experiments using RGB and Depth data stream are conducted in both segmentation and classification stage. As in Fig 2, hand region segmentation is performed using three segmentation networks UNet, ResUNet and Deeplab V3+ with RGB, Depth and stacked RGBD data. Corresponding experimental result is shown in Table II which indicates Deeplab V3+ with Depth data gives the better segmentation accuracy. This is also evident from the plots in Fig. 5 and Fig. 6, where segmentation accuracy is analyzed for each subject and each class respectively (both in aqua color plot).

Segmented RGB and Depth data is used to train the classification model as depicted in Fig. 4. Corresponding experimental results in Table III and plot in Fig. 7. shows that average ensemble (blue color plot) gives the better classification accuracy.

Proposed SSEC framework of Deeplab V3+ based semantic segmentation and average score ensemble classification is evaluated on OUHANDS benchmark dataset and compared the accuracy with existing methods as in Table IV. Experimental results shows that proposed methods gives the highest F1-score of 0.8891 and proved to be better than state-of-the-art methods.

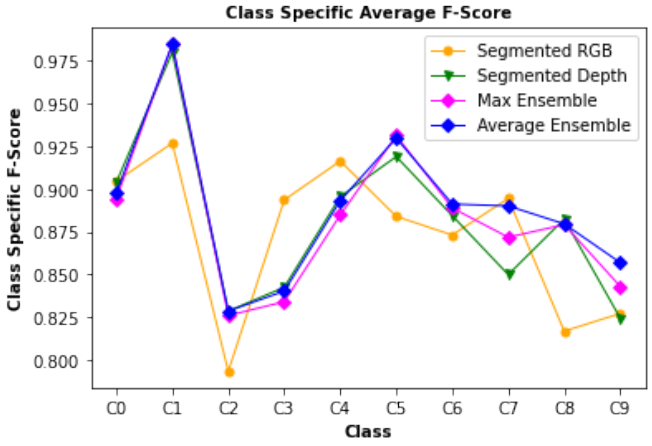

Fig. 7. Analysis of classification F1-score results of each class using segmented RGB, depth data models and various score fusion techniques.

VI. Conclusion

In this paper, we introduce SSEC, a novel semantic segmentation and ensemble classification framework on RGB-D Data for Static Hand Gesture Recognition. Specifically, we have analyzed three segmentation networks UNet, ResUNet and Deeplab V3+ with RGB, Depth and stacked RGBD data. Deeplab V3+ with Depth data gave higher F1 Score, the prediction outcome of this model is used as segmentation mask to discard background in RGB and depth data. In classification stage, custom CNN network was trained with segmented depth and segmented RGB data individually. Experimental results shows that, average score ensemble of these models can give better accuracy as compared to individual models. Hence it is inferred that RGB-D is with score fusion model ensembling is suitable for hand gesture classification as compare to the stat-of-the art methods.

Limitation of current approach is the usage of same network in both the RGB-D stream which may not give the diverse features, this can be further improved by using different CNN architectures in both streams to extract complimentary features.

In future work, we intend to develop a framework for dynamic hand gesture recognition considering the temporal sequence of RGB-D data for word and sentence level classification.

References

[1] F. Al Farid, N. Hashim, J. Abdullah, M. R. Bhuiyan, W. N. Shahida Mohd Isa, J. Uddin, M. A. Haque, and M. N. Husen, "A structured and methodological review on vision-based hand gesture recognition system," *Journal of Imaging*, vol. 8, no. 6, p. 153, 2022.

[2] M. Oudah, A. Al-Naji, and J. Chahl, "Hand gesture recognition based on computer vision: a review of techniques," *journal of Imaging*, vol. 6, no. 8, p. 73, 2020.

[3] S. Yuying, C. Sujie, L. Ming, L. Siying, P. Yisen, and G. Xiaojun, "Flexible strain sensors for wearable hand gesture recognition: From devices to systems," *Computers and Electrical Engineering*, vol. 1002, no. 170, pp. 1–17, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/epdf/10.1002/aisy.202100046

[4] S. Jaya Prakash, S. Suraj Prakash, A. Samit, and P. Sarat Kumar, "Rbi-2rcnn: Residual block intensity feature using a two-stage residual convolutional neural network for static hand gesture recognition," *Signal, Image and Video Processing*, vol. 1007, no. 170, pp. 1–17, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s11760-022-02163-w

[5] B. Gopa, V. Monu, C. Mahesh, and V. Santosh Kumar, "Hyfinet: Hybrid feature attention network for hand gesture recognition," *Multimedia Tools and Applications*, vol. 1007, no. 170, pp. 1–17, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s11042-021-11623-3

[6] H. Xu, G. Chen, Z. Wang, L. Sun, and F. Su, "Rgb-d-based pose estimation of workpieces with semantic segmentation and point cloud registration," *Sensors*, vol. 19, no. 8, p. 1873, 2019.

[7] M. Poggi, G. Agresti, F. Tosi, P. Zanuttigh, and S. Mattoccia, "Confidence estimation for tof and stereo sensors and its application to depth data fusion," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1411–1421, 2019.

[8] N. Dayananda Kumar, K. Suresh, and R. Dinesh, "Depth based static hand gesture segmentation and recognition," in *Cognition and Recognition: 8th International Conference, ICCR 2021*. Springer, 2021, pp. 125–138.

[9] C. N. Aithal, P. Ishwarya, S. S, Y. C. N, D. Kumar, and K. V. Suresh, "Dynamic hand segmentation," in *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 2021, pp. 1–6.

[10] R. F. Pinto, C. D. Borges, A. M. Almeida, and I. C. Paula, "Static hand gesture recognition based on convolutional neural networks," *Journal of Electrical and Computer Engineering*, vol. 2019, pp. 1–12, 2019.

[11] W. Aly, S. Aly, and S. Almotairi, "User-independent american sign language alphabet recognition based on depth image and pcanet features," *IEEE Access*, vol. 7, pp. 123 138–123 150, 2019.

[12] G. Benitez-Garcia, L. Prudente-Tixteco, L. C. Castro-Madrid, R. Toscano-Medina, J. Olivares-Mercado, G. Sanchez-Perez, and L. J. G. Villalba, "Improving real-time hand gesture recognition with semantic segmentation," *Sensors*, vol. 21, no. 2, p. 356, 2021.

[13] D. Chen, G. Li, Y. Sun, J. Kong, G. Jiang, H. Tang, Z. Ju, H. Yu, and H. Liu, "An interactive image segmentation method in hand gesture recognition," *Sensors*, vol. 17, no. 2, p. 253, 2017.

[14] B. Xie, X. He, and Y. Li, "Rgb-d static gesture recognition based on convolutional neural network," *The Journal of Engineering*, vol. 2018, no. 16, pp. 1515–1520, 2018.

[15] N. D. Kumar, K. Suresh, and R. Dinesh, "Cnn based static hand gesture recognition using rgb-d data," in *2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)*. IEEE, 2022, pp. 1–6.

[16] J. P. Sahoo, A. J. Prakash, P. Pławiak, and S. Samantray, "Real-time hand gesture recognition using fine-tuned convolutional neural network," *Sensors*, vol. 22, no. 3, p. 706, 2022.

[17] S. Wang, S. Zhang, X. Zhang, and Q. Geng, "A two-branch hand gesture recognition approach combining atrous convolution and attention mechanism," *The Visual Computer*, pp. 1–14, 2022.

[18] A. S. M. Miah, M. A. M. Hasan, J. Shin, Y. Okuyama, and Y. Tomioka, "Multistage spatial attention-based neural network for hand gesture recognition," *Computers*, vol. 12, no. 1, p. 13, 2023.

[19] J. P. Sahoo, S. P. Sahoo, S. Ari, and S. K. Patra, "Derefnet: Dual-stream dense residual fusion network for static hand gesture recognition," *Displays*, p. 102388, 2023.

[20] J. Ding and N.-W. Zheng, "Rgb-d depth-sensor-based hand gesture recognition using deep learning of depth images with shadow effect removal for smart gesture communication," *Sensors and Materials*, vol. 34, no. 1, pp. 203–216, 2022.

[21] L. Roda-Sanchez, C. Garrido-Hidalgo, A. S. García, T. Olivares, and A. Fernández-Caballero, "Comparison of rgb-d and imu-based gesture recognition for human-robot interaction in remanufacturing," *The International Journal of Advanced Manufacturing Technology*, vol. 124, no. 9, pp. 3099–3111, 2023.

[22] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal

fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1407–1417.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[24] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[26] M. Matilainen, P. Sangi, J. Holappa, and O. Silvén, "Ouhands database for hand detection and pose recognition," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2016, pp. 1–5.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[28] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[29] G. Jianchun, G. Jiannuan, and W. Lili, "Gesture recognition method based on attention mechanism for complex background," *Journal of Physics: Conference Series*, vol. 1873, no. 1, p. 012009, apr 2021.

[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *ArXiv*, vol. abs/1704.04861, 2017.

[31] D. Amirhossein, T. A. Tavakoli, M. Tahmasbi, and M. Mirmehdi, "Hgr-net: a fusion network for hand gesture segmentation and recognition," *IET Computer Vision*, vol. 13, no. 700-707, 2019.