Displays 74 (2022) 102226

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

A lightweight hand gesture recognition in complex backgrounds☆

Weina Zhou *, Kun Chen

Shanghai Maritime University, Shanghai 201306, China

ARTICLE INFO

Keywords:
Hand segmentation
Hand gesture recognition
Semantic segmentation
Double-channel CNNs
Features fusion

ABSTRACT

Hand Gesture Recognition (HGR) is widely used in human–computer interaction due to its convenience. However, there are still some challenges in real-world scenarios, such as recognizing hand gestures in the complex backgrounds. To this end, the paper proposes a two-stage HGR system to solve the above issue. Specifically, the first stage performs accurate segmentation to segment the hand from the background. The segmentation network combines dilated residual network, atrous spatial pyramid pooling module and a simplified decoder. The segmentation network can effectively determine hand region even in challenging backgrounds. In the second stage, the double-channel Convolutional Neural Networks (CNNs) are presented to improve the recognition performance. The double-channel CNNs can learn features from the RGB input images and the segmented hand images separately. Experiment results show that the proposed method has an accuracy of 91.17% with the model size of 1.8 MB, both of which are better than other state-of-arts in hand gesture recognition. The method successfully constructed a lightweight model while keeping a high gesture recognition accuracy at final.

1. Introduction

In recent years, with the development of machine vision, human–computer interaction [1–4] has become closely related to daily lives such as action recognition [5–9] or gesture recognition. As a common way for people to communicate, gestures can provide intuitive interaction with machines, which has attracted many researchers [10,11]. Indeed, HGR has long been an important research area of machine vision and has extensive applications in intelligent driving, machine control, and virtual reality. HGR can be implemented through two methods: wearable equipment-based gesture recognition and vision-based gesture recognition. The first method makes the recognition system uncomfortable and inconvenient for users due to the requirement of additional devices. However, vision-based gesture recognition only needs a low-cost camera, making users communicate with computers more naturally. In this paper, a novel HGR system based on vision is proposed to overcome the difficulties of hand recognition in complex background. Our goal is to build a lightweight HGR system while ensuring its high recognition accuracy.

The proposed method contains two stages, i.e., hand segmentation and hand gesture recognition. At the stage of hand segmentation, a deep architecture based on fully convolutional residual network is used to segment the hand out of the image, which is a highly challenging task, due to the variability of the image background, skin colour differences, shadows, and other illumination variations. Thus, it integrates the advantage of Dilated Residual Network (DRN), an Atrous Spatial Pyramid Pooling (ASPP) module and a simple decoder hoping for a high accuracy. Then in the second stage of hand gesture recognition, a double-channel CNNs was built to learn features from the input RGB images and their segmented image separately. Features would then be fused to obtain for a better result afterwards. The contributions of our work are in four sides.

- First, an encoder-decoder structure is firstly proposed to segment hand gestures in complex backgrounds. As proved by experiments, the proposed structure can obtain a clearer hand edge than existing vision-based hand segmentation methods.
- Second. DRN and ASPP are combined as the encoder module, to efficiently solve the problem in segmenting different scales of hand regions.
- Third, only three layers of convolution are used in the proposed decoder module to reduce the complexity of the encoder-decoder network.
- Finally, Depthwise Separable Convolution (DSC) is used to both encoder module and decoder module, which makes the network less in computation amount and smaller in size (only 1.8 MB).

☆ This paper was recommended for publication by Prof G Guangtao Zhai.

* Corresponding author.
E-mail address: wnzhou@shmtu.edu.cn (W. Zhou).

https://doi.org/10.1016/j.displa.2022.102226
Received 14 September 2021; Received in revised form 30 March 2022; Accepted 22 April 2022
Available online 27 April 2022

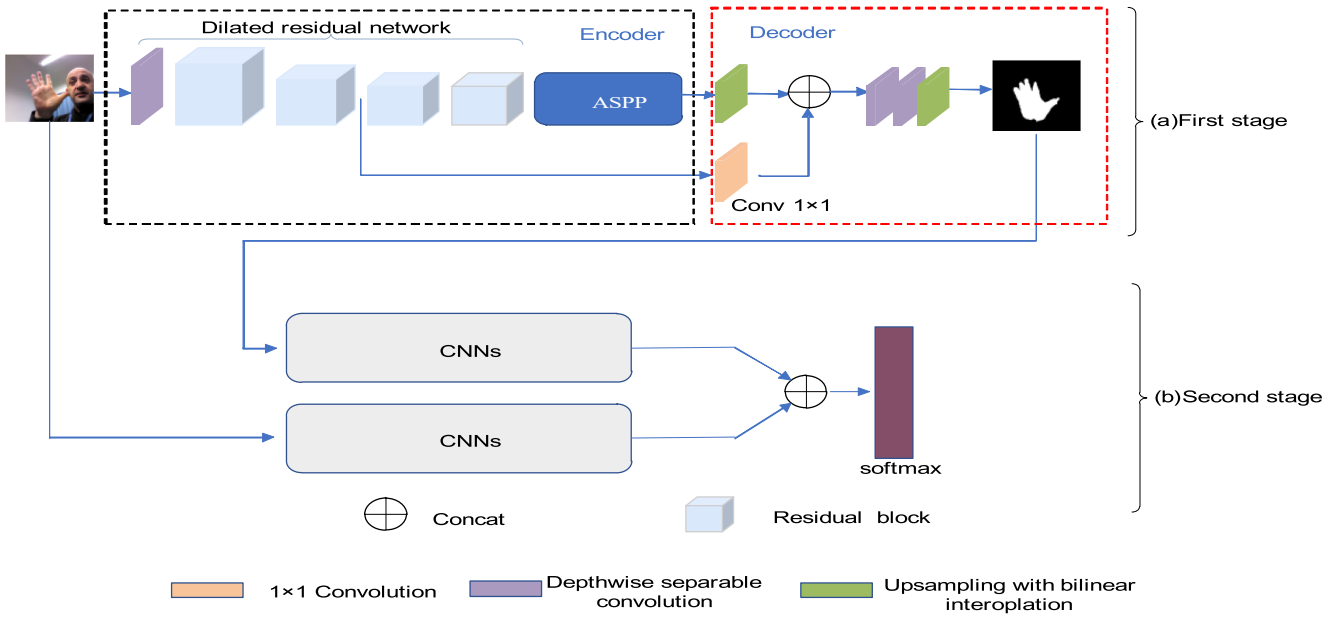

Fig. 1. The proposed HGR system with two stages.

2. Related work

In most vision-based HGR systems, hand segmentation plays an essential role for segmenting the hand from the background. Some researchers [12–14] tested their methods in a single background and obtained the hand region by simply pre-processing the original image. For example, Sun. et al. [15] established a gaussian mixture model considering of the particularity of hand skin. Sangi et al. [16] described gestures by extracting the histogram of the image's directional gradient. Therefore, depth information was used in gesture recognition by researchers [17–19]. Specifically, the method proposed by Duong. et al. [17] discriminated hand components and located fingertips in depth images. Liu. et al. [18] proposed a two-stream CNN architecture with color image and its pseudo depth image to improve the recognition performance. Kang et al. [19] proposed a hand segmentation method for hand-object interaction by using only a depth map. However, the depth images should be captured by the depth camera, which is not widespread applied, especially in outdoor scenarios. Recently, semantic segmentation methods have achieved great success based on deep convolutional neural networks [20–24]. Jiang et al improved Faster-RCNN algorithm, which can simultaneously realize the indoor scene semantic segmentation, target classification and detection multiple visual tasks [25]. All have a high-quality performance in solving many object segmentation problems. Thus, semantic segmentation network become the best choice to segment hands from complex backgrounds.

Hand gesture recognition, as another crucial part of the HGR system, aims to classify the extracted hand gestures features into a specific gesture category. Different kinds of features are designed in these methods [26–28]. Priyal et al. [26] presented a gesture recognition system, which used geometry-based normalizations and krawtchouk moment features for classifying static hand gestures. Pisharady et al. [27] proposed a feature based visual attention method to recognize multi-class hand postures against complex natural backgrounds. The system was implemented by using a combination of high level (shape, texture) and low level (color) image features. Avraam et al. [28] combined graph and appearance features to recognize static gestures. In recent years, researchers began to learn gesture features based on deep learning methods [29–31]. Liao et al [29] improve Mobilenet-SSD network to recognize occlusion gestures. Wu et al [30] presented a novel recognition algorithm based on a double-channel Convolutional Neural Network (CNNs) which selected the hand gesture images and the hand edge images as two input channels. Cheng et al combined the sEMG feature image and the Convolutional Neural Network (CNN) to recognize 52 gesture movements [31]. However, there is still a lack of a high-precision solution for hand gesture recognition in complex backgrounds.

In this paper, we propose a lightweight two-stage hand gesture recognition system that exploits the feature from the RGB image and the corresponding hand segmentation map. Our experiments show that the proposed HGR system has an excellent performance in complex environments with low computational costs.

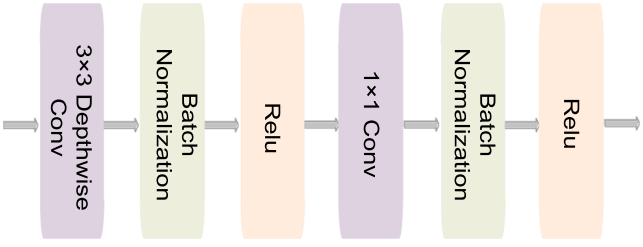

Fig. 2. Diagram of depthwise separable convolution.

3. Methods

The block diagram of the proposed HGR system is shown in Fig. 1. The system consists of two stages. In the first stage, an encoder-decoder framework is used to segment the hand regions out of the original image. The framework is comprised of the DRN, ASPP module, and a simplified decoder. In the second stage, a double-channel CNNs takes the segmented hand image and original RGB image as input, and then fuses the extracted features of the both images to obtain final results of HGR system. The detailed introduction of each module is as follows.

3.1. Hand segmentation

In our work, a novel encoder-decoder framework that integrates low-level detail features and high-level semantic features is presented to make the segmented hand areas complete. Constituted by an improved DRN module and ASPP module, the encoder part can efficiently segment different scales of hand regions. The decoder module is simplified and

Table 1
The algorithm process of batch normalization.

Input:Values of $\times$ over a mini-batch:
$B = \{x_{1\cdots m}\}$;
Parameters to be learned:γ, β
Output:$\{y_i = BN_{\gamma,\beta}(x_i)\}$
$\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$ //mini-batch mean
$\delta_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2$ //mini-batch variance
$\widehat{x}_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\delta_B^2 + \epsilon}}$ //normalize
$y_i \leftarrow \gamma\widehat{x}_i + \beta$ //scale and shift

Table 2
The proposed dilated residual network.

Layer name	Output size	Layers
Conv1	320 × 320 × 16	3 × 3, 16
ResBlock_1	320 × 320 × 32	$\begin{bmatrix} 1\times1, 8 \\ 3\times3, 8 \\ 1\times1, 32 \end{bmatrix}$ × 3, stride = 1
ResBlock_2	160 × 160 × 64	$\begin{bmatrix} 1\times1, 16 \\ 3\times3, 16 \\ 1\times1, 64 \end{bmatrix}$ × 3, stride = 2
ResBlock_3	80 × 80 × 128	$\begin{bmatrix} 1\times1, 32 \\ 3\times3, 32 \\ 1\times1, 128 \end{bmatrix}$ × 3, stride = 2
ResBlock_4	80 × 80 × 128	$\begin{bmatrix} 1\times1, 32 \\ 3\times3, 32 \\ 1\times1, 128 \end{bmatrix}$ × 3, rate= [2,4,2]

cascaded with the ASPP module to obtain sharper hand boundaries with a small module size. All of them adopt DSC into their architectures to save the total computing resources and reduce the model size.

3.1.1. Depthwise separable convolution (DSC)

DSC is proposed by Howard et al. [32], which is a form of factorized convolutions. Fig. 2 is the diagram of DSC module. It factorizes a standard convolution into a depthwise convolution and a 1 × 1 convolution called pointwise convolution. The depthwise convolution and the pointwise convolution are used for filtering each input channel and combining the outputs of the depthwise convolution, respectively. This factorization helps reduce computational complexity and model size. Together with depthwise convolution and 1 × 1 convolution, batch normalization (BN) and Relu are also used in DSC to prevent overfitting of the hand segmentation network. The processing procedure of the BN algorithm is shown in Table 1, where m is the number of features in each batch.

3.1.2. Dilated residual network (DRN)

Deep networks are needed for many visual recognition tasks [33,34]. However, a deep learning network cannot achieve good performance by simply stacking more layers. Many problems, such as the disappearance and explosion of gradients, would appear in practice. Although the bottleneck residual structure [35] solved the above problems, and accelerated the network training and convergence speed, the resolution of output feature maps can be gradually reduced until retaining little spatial information. Thus, in hand segmentation task, dilation convolution is finally adopted based on bottleneck residual structure to increase neurons' receptive field [36], reducing the loss of resolution.

In that case, the proposed Dilated Residual Network (DRN) adopts bottleneck residual unit to construct deeper network and dilation convolution to reduce the loss of resolution. As shown in Table 2, the proposed DRN consists of one 3 × 3 convolution and four residual blocks, where each block contains three bottleneck residual units. The DRN contains only 12 bottleneck residual units which could be lightweight when obtaining rich semantic information. It is known that the down-sampling operation is beneficial to expand the receptive field of neurons, and a wider receptive field helps capture larger-scale contextual information [37], the stride of the last 1x1 convolutional layer in the second and third residual blocks is determined to set as 2. In the last residual block, different dilation rates are adopted for the three cascaded bottleneck residual units. The proposed strategy enlarges the receptive field of neurons while keeping the resolution of the output feature map.

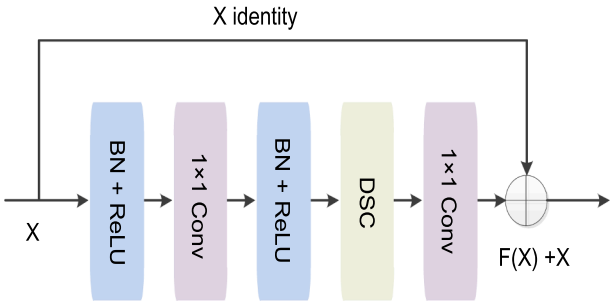

Fig. 3. The proposed bottleneck residual unit with DSC.

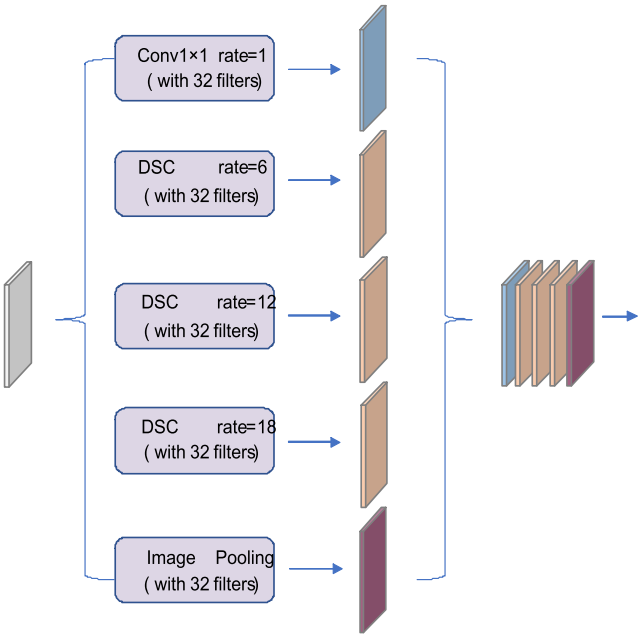

Fig. 4. The proposed ASPP module.

Fig. 3 shows the architecture of bottleneck residual unit, which is the core of DRN. In the unit, skip connection is used for identity mapping. Two 1 × 1 convolutions are used to reduce and then restore the dimension of the input feature map, which is beneficial to decrease parameters and computation costs. DSC is adopted to reduce the computational complexity of the unit. Formally, each unit is defined as:

$$y = \mathrm{F}(\mathrm{x}) + \mathrm{x} \tag{1}$$

where x and y are the input and output vectors of the bottleneck residual unit. The function F represents the learned residual mapping and must have the same dimensions as x.

3.1.3. Atrous spatial pyramid pooling (ASPP)

Multiple scale is the main factor in improving the accuracy of hand segmentation. Due to the poor performance of DRN on segmenting objects at different scales, ASPP module is introduced to capture multi-scale contextual information, improving the segmentation results' quality with various hand size. As shown in Fig. 4, the proposed ASPP

Table 3
The architecture of DCNNs.

Layer name	Output size	Layer type
Conv2d_1	320 × 320 × 16	convolution
Pooling_2d_1	106 × 106 × 16	max-pooling
Conv2d_2	106 × 106 × 32	convolution
Pooling2d_2	35 × 35 × 32	max-pooling
Conv2d_3	35 × 35 × 64	convolution
Pooling2d_3	11 × 11 × 64	max-pooling
Conv2d_4	11 × 11 × 128	convolution
Pooling2d_4	128	global average pooling
Dense_1	64	fully connected
Dense_2	64	fully connected

consists of an image pooling operation, one 1 × 1 convolution, and three depthwise separable convolutions with the dilation rate of 6,12 and 18, respectively. The global average pooling is applied to output feature map of the DRN, which implements image pooling operation. Then the obtained image-level features are fed to a 1 × 1 convolution. Finally, we fuse the features of five branches as the encoder output feature maps by concatenation operation. The proposed ASPP module can capture context information of different scales by changing the dilation rate of convolution. Additionally, the module obtains global contextual information by image pooling operation.

3.1.4. Proposed decoder

Different from most encoder-decoder architectures based on semantic segmentation network, our proposed decoder module is more lightweight, and simultaneously fuses shallow detail features and deep semantic features to improve the model segmentation performance well. In the paper, the decoder is the last module of hand segmentation network. Generally, low-level features have rich detail information, while high-level features contain semantic information of targets. In recent years, the idea of decoder [38–40] is proposed to recover the spatial detail information to obtain sharper segmentation. Therefore, to segment the hand with sharp boundaries, we present a simplified decoder that integrates low-level detail features to the high-level semantic features. The proposed decoder can be seen in the top-right corner of Fig. 1. The output features of the encoder are upsampled by bilinear interpolation. The number of channels of low-level features is reduced through 1 × 1 convolution to avoid outweighing the importance of encoder features. By experiments, the channel of low-level features is set to 16 to achieve better performance. In addition, two cascaded depthwise separable convolutions (all with 160 filters) is applied to refine integrated features. Another bilinear upsampling is used to obtain the final hand segmentation map, and the sigmoid function is used as the last activation function in the architecture.

3.1.5. Loss function in first stage

In this stage, RGB image is converted into a binary image, which contains hand area and non-hand area. We use the binary cross-entropy function to achieve the pixel-level binary classification of the image, which is given as:

$$\mathrm{L}(t,z) = -\frac{1}{N}\sum_{i=1}^{N}[t_i log(z_i) + (1-t_i) log(1-z_i)], \tag{2}$$

where N, $t_i \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 1}$, and $z_i \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 1}$ are the number of samples, the segmentation masks, and the predicted segmented map respectively. H and W represent the height and width of the image. The sigmoid function is adapted to produce the predicted segmented map, which is defined as.

$$z_i = sigmoid\left(w^T x_i + b\right), \tag{3}$$

where $x_i \in \mathbb{R}^{h\times w\times 3}$ is the RGB image, w and b are the learned parameters of the model.

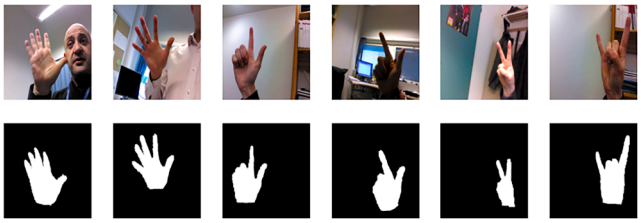

Fig. 5. Some hand gesture images and their segmented binary masks from the OUHANDS dataset.

3.2. Hand gesture recognition

Double-channel Convolutional Neural Networks (DCNNs) [18] is beneficial to improve the accuracy of gesture recognition. Therefore, the simple DCNNs is proposed to recognize hand gestures, which combines the features of RGB image and the segmented hand image. As shown in Table 3, the architecture of the CNNs consists of four convolution layers, four pooling layers, and two fully connected layers. The convolution layer uses a 3*3 convolution kernel, and the downsampling step size of the pooling layer is set to 3. Since global average pooling can reduce the network overfitting [41] and is more robust to the spatial transformation of the input at the same time, we use the global average pooling as the fourth pooling layer. The feature fusion operation can cascade the output features of each CNNs, and the fused features (with 128 channels) are feed into the softmax layer to achieve the final gesture classification. Formally, let $v_1 \in \mathbb{R}^m$ and $v_2 \in \mathbb{R}^n$ be the output feature of each CNNs, respectively. Note that, m and n present the number of CNNs' output features. The output of the feature fusion operation is $= [v_1, v_2] \in \mathbb{R}^{m+n}$.

3.2.1. Loss function in second stage

In this stage, a double-channel CNNs is applied to classify the learned features into specific gesture categories. We adopt the categorical cross-entropy function to the multi-classification tasks. Suppose that N training samples and K hand gesture classes are existed, the loss is given as.

$$L(y,p) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(y_{ik}) log(p_{ik}), \tag{4}$$

where $y \in \mathbb{R}^N$ are the true labels of hand gestures and $p \in \mathbb{R}^N$ are the predicted values of hand gestures.

4. Experimental results

4.1. Datasets

We evaluate the proposed HGR system on the OUHANDS dataset [42]. Among all the public gesture recognition datasets, the OUHANDS dataset is one of the few databases with a complex background and hand segmentation ground truth at the same time. The dataset consists of ten different gesture classes from 23 subjects with background disturbed by illumination, complex scenes, face-hand collusions, etc. Fig. 5 shows some hand gesture images and their segmented binary masks. There are totally 3000 images are in the dataset, which is divided into 1920, 480, and 600 images for training, validation, and testing, respectively. Note that, the testing set is independent from the training set and the validation set.

4.2. Data augmentation

To prevent overfitting and enhance the generalization of the HGR system, an effective offline data augmentation is proposed. The training set is expanded to 2-fold by random zoom (10–15%) and horizontal/

Table 4
Effectiveness comparison with different modules.

BaseNet	DC	ASPP	Decoder	MIoU
√				0.8676
√	√			0.8968
√	√	√		0.9108
√	√	√	√	**0.9217**

Table 5
Segmentation results with different frameworks.

Method	MIoU(%)	Model Size(MB)	FLOPs(1e5)
SegNet	90.49	34.6	181.2
PSPNet	90.81	251.1	1313.9
DeepLabv3+	92.02	157.3	821.2
HGR-Net	90.40	1.1	5.5
Ours	**92.17**	**1.0**	**5.1**

vertical translation (10–15%) to increase the diversity of data.

4.3. Implementation details

The proposed framework is implemented based on Tensorflow. All images of the dataset are uniformly resized to 320 × 320. The experiments are performed on GeForce RTX 3080 GPU, and the parameters of Adam optimizer [39] are set to default (initial learning rate lr = 0.001, betas= (0.9,0.999)). The Adam optimizer is used to adjust the learning rate in training, which could avoid the decay of initial learning rate with the number of training iterations. Considering the convergence of the model in training, the number of iterations is set to 100. Besides, the batch size is set as 8, which comprehensively considering the training time, convex problems and memory resource of GPU.

4.4. Experiments on hand segmentation

Mean Intersection over Union (MIoU) is adopted to evaluate our hand segmentation model. The criterion is used to represent the correlation between the ground truth and the predicted segmentation map. The higher the value of MIoU, the greater the correlation. MIoU is defined as:

$$\mathrm{MIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij} + \sum_{j=0}^{k}p_{ji} - p_{ii}}, \tag{15}$$

where $k+1$ presents the total number of categories, including background, p_{ij} represents the number of pixels where i is predicted to be j. p_{ji} represents the number of pixels where j is predicted to be i, and p_{ii} represents the number of pixels where i is predicted to be i.

To demonstrate the effectiveness of each module, we conduct a comparison between nets with different modules. As shown by the results in Table 4, the evaluation criteria have greatly improved after applying dilation convolution (DC) to BaseNet(the proposed residual network without dilation convolution). The performance can further be improved by successively adding ASPP module and the decoder module. The experimental results show that all the proposed modules are beneficial to hand segmentation. Our method is really effective in segmenting hands in complex backgrounds.

We also compare the proposed framework with three classical semantic segmentation architectures and the HGR-Net(stage1) proposed by Dadashzadeh. et al. [43]. The three classical architectures of hand segmentation, include SegNet [18], PSPNet [21] (using Resnet-101[35] as the backbone), and Deeplabv3+ [22] (using Xception[44] as the backbone). All the networks are trained by the same dataset for fair comparisons. MIOU, model size, and FLOPs are adopted to evaluate these frameworks in Table 5. FLOPs is short for floating-point operations, which is used to measure the complexity of a model. The results show that the MIoU value of our proposed method is 92.17%, which is

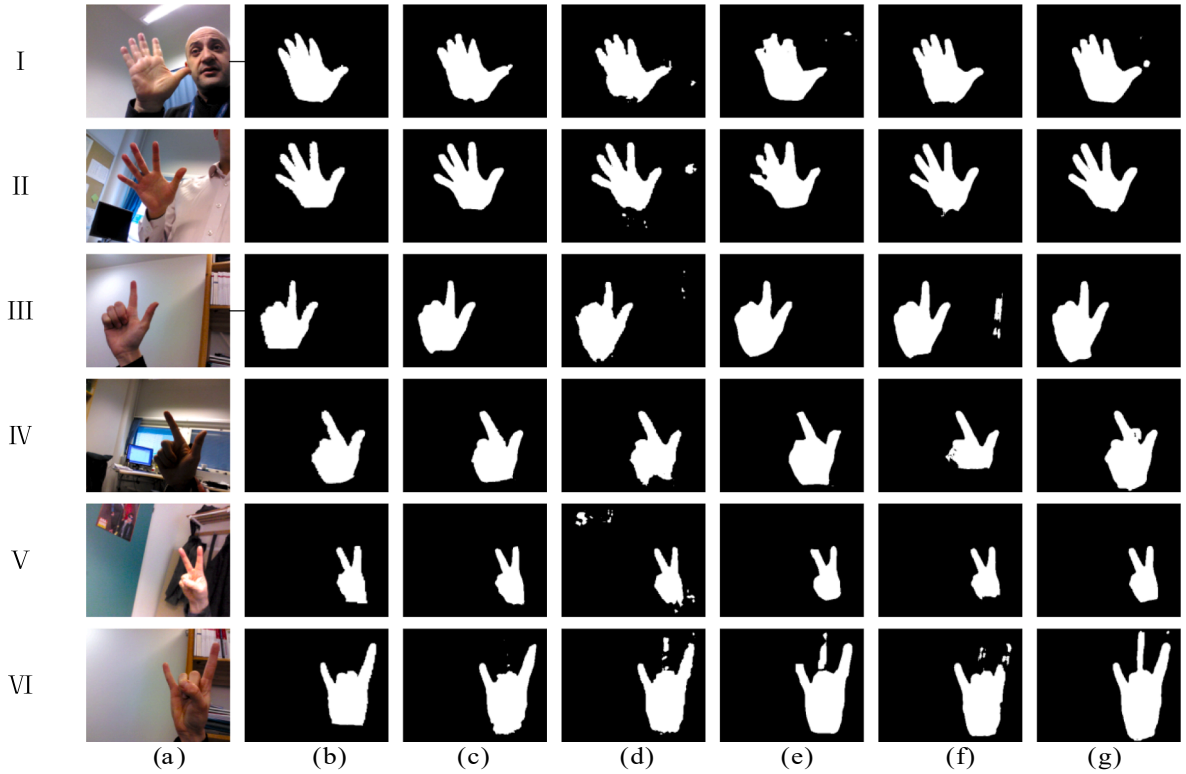

Fig. 6. Visual comparisons of different methods on OUHANDS dataset. (a)original image, (b)ground truth, (c)the proposed framework, (d)HGR-Net(stage1), (e) SegNet, (f)PSPNet, (g)DeepLabv3+.

Table 6
Evaluation of the proposed systems.

Method	Accuracy	Macro-F1
CNNs	0.8017	0.8031
Proposed system (without feature fusion)	0.8767	0.8772
Proposed system (with feature fusion)	**0.9117**	**0.9114**

Table 7
Recognition results with different methods.

Method	Accuracy (%)	Macro-F1 (%)	Model Size (M)	FLOPS (1e5)
ResNet-101	83.33	83.75	162.8	850.4
ShuffleNetV2	86.17	86.12	7.4	38.2
MobileNetV3	87.52	87.58	11.6	60.6
HGR-Net	87.13	88.10	1.9	9.9
Ours	**91.17**	**91.14**	**1.8**	**9.5**

higher than other state-of-the-arts algorithms. The experimental results show that the proposed framework has better performance on hand segmentation while requiring less hardware resources.

The hand segmentation results of our framework and other frameworks are visualized in Fig. 6. The images in column (a) and (b) of Fig. 6 show six different HGR situations and its corresponding ground truth. Row (I) and (II) show the results of segmenting hands with face noise in the background. Row (III) and (VI) show the results obtained in complex backgrounds containing skin-colored objects. Row (IV) and (V) show the results when we deal with much severer illumination changes. As shown in Fig. 6, the proposed framework not only perform better than other methods but also are closer to the ground truth.

4.5. Experiments on hand gesture recognition

In this paper, Accuracy and Macro-F1 are used to evaluate the hand gesture classifier. Accuracy is defined as:

$$\text{Accuracy} = \frac{\text{CN}}{\text{TN}} \tag{6}$$

where CN is the number of correctly classified samples, and TN is the total number of samples. Macro-F1 represents the average value of F1-Scores for the whole categories, and Macro-F1 is defined as:

$$\textit{Macro}\text{-F1} = \frac{1}{C}\sum_{i=1}^{C}\text{F1-Score}_i \tag{7}$$

where C is the total number of categories of the hand gestures. F1-Score is the weighted average of precision and recall, which is defined as:

$$F1 - Score \;=\; 2 \times \frac{P \times R}{P + R} \tag{8}$$

where P and R represent the precision and recall, respectively.

The evaluated results of systems with different gesture recognition architecture are shown in Table 6. In Table 6, Row 1 shows the result of recognizing gestures by only using CNNs. Row 2 and row 3 are the results of proposed HGR system with or without the feature fusion operation. As can been seen from the result, the proposed HGR system performs better by using feature fusion operation in the second stage.

We also compare recognition performances of proposed architecture with ResNet-101 [35], ShuffleNetV2 [45], MobileNetV3 [46] and HGR-Net [43] in Table 7. They are not only the state-of-the-art methods but also have their characteristics. ShffleNetV2 and MobileNetV3 are two well-known lightweight networks. Their accuracy is relatively high and model size is much smaller than typical networks for classification such as ResNet-101. And HGR-Net is a network which has a great competitiveness in performing hand recognition task as far as we know. From the results in Table 7, we can see that that the accuracy of our proposed

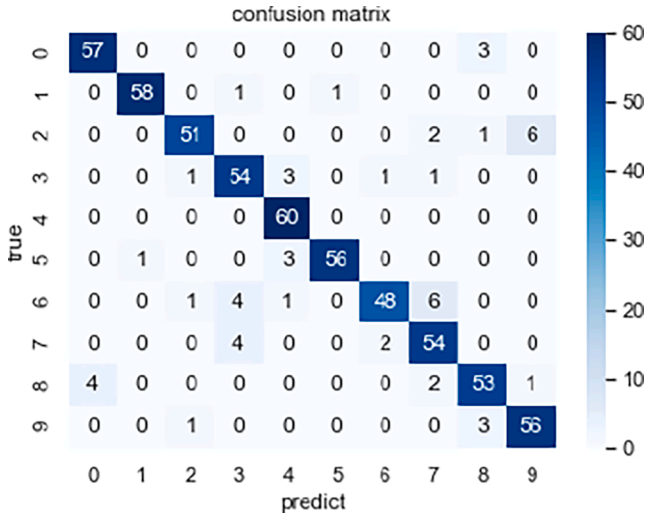

Fig. 7. Confusion matrix of the proposed HGR system.

method is 91.17%, which is 3.6% higher than MobileNetV3 that has the second highest accuracy. And the model size and FLOPS of our proposed method are also lower than other models. This experiment proves that our proposed HGR system not only outperforms other methods in accuracy and computation amount, but also is the most lightweight network.

To further visualize the hand gesture recognition results of our proposed HGR system, confusion matrix is introduced and shown in Fig. 7, the value on the main diagonal indicates the number of correct recognitions for each category. From the matrix, we can find out that our architecture performs very well in hand recognition.

5. Conclusion

In this paper, we propose a two-stage HGR system for hand gesture recognition in complex backgrounds. To improve hand gesture recognition performance, we present a hand segmentation network based on a novel encoder-decoder architecture without using depth information. The encoder consists of a DRN and an ASPP module. They can capture deep semantic features and multi-scale contextual information. Then, a simplified decoder is built to integrate high-level semantic features and low-level detail features to segment hands completely. After that, double-channel CNNs are designed to learn features from the RGB images and the corresponding hand segmentation maps, which can perform much better than single-channel convolutional neural networks. In all, our HGR system performs better on HGR in different complex backgrounds. What's more, HGR is also a lightweight network and can recognize hand gestures within limited resources at the same time.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant No. 52071200, 61404083) and State Key Laboratory of ASIC & System (2021KF010).

References

[1] X. Zhang, J. Liu, Q. Gao, et al., Adaptive robust decoupling control of multi-arm space robots using time-delay estimation technique, Nonlinear Dyn. 100 (3) (2000) 2449–2467.

[2] X. Zhang, J. Liu, J. Feng, et al., Effective capture of nongraspable objects for space robots using geometric cage pairs, IEEE/ASME Trans. Mechatron. 25 (1) (2019) 95–107.

[3] A. Singla, P.P. Roy, D.P. Dogra, Visual rendering of shapes on 2D display devices guided by hand gestures, Displays 57 (2019) 18–33.

[4] F. Başçiftçi, A. Eldem, An interactive and multi-functional refreshable Braille device for the visually impaired, Displays 41 (2016) 33–41.

[5] D.K. Vishwakarma, R. Kapoor, Integrated approach for human action recognition using edge spatial distribution, direction pixel and-transform, Adv. Robot. 29 (23) (2015) 1553–1562.

[6] D.K. Vishwakarma, R. Kapoor, R. Maheshwari, et al., Recognition of abnormal human activity using the changes in orientation of silhouette in key frames, in: 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 336–341.

[7] D.K. Vishwakarma, A two-fold transformation model for human action recognition using decisive pose, Cogn. Syst. Res. 61 (2020) 1–13.

[8] D.K. Vishwakarma, T. Singh, A visual cognizance based multi-resolution descriptor for human action recognition using key pose, AEU-Int. J. Electron. Commun. 107 (2019) 157–169.

[9] C. Dhiman, D.K. Vishwakarma, A Robust Framework for Abnormal Human Action Recognition Using R-Transform and Zernike Moments in Depth Videos, IEEE Sens. J. 19 (13) (2019) 5195–5203.

[10] S.S. Rautaray, A. Agrawal, Vision based hand gesture recognition for human computer interaction: a survey, Artif. Intell. Rev. 43 (1) (2015) 1–54.

[11] O.K. Oyedotun, A. Khashman, Deep learning in vision-based static hand gesture recognition, Neural Comput. Appl. 28 (12) (2017) 3941–3951.

[12] D.K. Vishwakarma, R. Kapoor, Simple and intelligent system to recognize the expression of speech-disabled person[C], in: 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI), 2012, pp. 1–6.

[13] D.K. Vishwakarma, R. Kapoor, An efficient interpretation of hand gestures to control smart interactive television, Int. J. Comput. Vis. Robot. 7 (4) (2017) 454–471.

[14] D.K. Vishwakarma, R. Maheshwari, R. Kapoor, An efficient approach for the recognition of hand gestures from very low resolution images, in: 2015 Fifth International Conference on Communication Systems and Network Technologies, 2015, pp. 467–471.

[15] J.H. Sun, T.T. Ji, S.B. Zhang, et al., Research on the hand gesture recognition based on deep learning, in: 2018 12th International symposium on antennas, propagation and EM theory (ISAPE), IEEE, 2018, pp. 1–4.

[16] P. Sangi, M. Matilainen, O. Silvén, Rotation tolerant hand pose recognition using aggregation of gradient orientations, in: International Conference on Image Analysis and Recognition, Springer, Cham, 2016, pp. 257–267.

[17] D.H. Nguyen, T.N. Do, I.S. Na, et al., Hand segmentation and fingertip tracking from depth camera images using deep convolutional neural network and multi-task signet, Multi-scale context aggregation by dilated convolutions, 1901.03465, 2019.

[18] J. Liu, K. Furusawa, T. Tateyama, et al., An improved hand gesture recognition with two-stage convolution neural networks using a hand color image and its pseudo-depth image, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 375–379.

[19] B. Kang, K.H. Tan, N. Jiang, et al., Hand segmentation for hand-object interaction from depth map, in: 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE, 2017, pp. 259–263.

[20] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.

[21] H. Zhao, J. Shi, X. Qi, et al., Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[22] L.C. Chen, Y. Zhu, G. Papandreou, et al., Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.

[23] W.N. Zhou, Y. Zhou, An attention nested U-Structure suitable for salient ship detection in complex maritime environment, IEICE Trans. Electron. E105-D (6) (2022) 1–7.

[24] X. Yan, S. Hou, A. Karim, et al., RAFNet: RGB-D attention feature fusion network for indoor semantic segmentation, Displays 70 (2021) 102082.

[25] D. Jiang, G. Li, C. Tan, et al., Semantic segmentation for multiscale target based on object recognition using the improved Faster-RCNN model, Future Gen. Comput. Syst. 123 (2021) 94–104.

[26] S.P. Priyal, P.K. Bora, A robust static hand gesture recognition system using geometry based normalizations and Krawtchouk moments, Pattern Recogn. 46 (8) (2013) 2202–2219.

[27] P.K. Pisharady, P. Vadakkepat, A.P. Loh, Attention based detection and recognition of hand postures against complex backgrounds, Int. J. Comput. Vision 101 (3) (2013) 403–419.

[28] M. Avraam, Static gesture recognition combining graph and appearance features, Int. J. Adv. Res. Artif. Intell. (IJARAI) 3 (2) (2014).

[29] S. Liao, G. Li, H. Wu, et al., Occlusion gesture recognition based on improved SSD, Concurr. Comput.: Pract. Exp. 33 (6) (2021) e6063.

[30] X.Y. Wu, A hand gesture recognition algorithm based on DC-CNN, Multimedia Tools Appl. 79 (13) (2020) 9193–9205.

[31] Y. Cheng, G. Li, M. Yu, et al., Gesture recognition based on surface electromyography-feature image, Concurr. Comput.: Pract. Exp. 33 (6) (2021) e6051.

[32] A.G. Howard, M. Zhu, B. Chen, et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications, Comput. Vis. Pattern Recogn. (cs.CV) 1704.04861 (2017).

[33] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Vis. Patt. Recogn. (cs.CV) 1409.1556 (2014).

[34] C. Szegedy, W. Liu, Y. Jia, et al., Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[35] S. Jian, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision & Pattern Recognition, 2016, pp. 770–778.

[36] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, Multi-scale context aggregation by dilated convolutions, 1511.07122, 2015.

[37] B. Zhou, A. Khosla, A. Lapedriza, et al., Object detectors emerge in deep scene cnns. Multi-scale context aggregation by dilated convolutions, 1412.6856, 2014.

[38] Z. Zhang, X. Zhang, C. Peng, et al., Exfuse: Enhancing feature fusion for semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 269–284.

[39] J. Fu, J. Liu, Y. Wang, et al., Stacked deconvolutional network for semantic segmentation, IEEE Trans. Image Process. (2019).

[40] Z. Wojna, V. Ferrari, S. Guadarrama, et al., The devil is in the decoder, in: British Machine Vision Conference 2017, BMVC 2017, BMVA Press, 2017, pp. 1–13.

[41] M. Lin, Q. Chen, S. Yan, Network in network, Multi-scale context aggregation by dilated convolutions, 1312.4400, 2013.

[42] M. Matilainen, P. Sangi, J. Holappa, et al., OUHANDS database for hand detection and pose recognition, in: 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE, 2016, pp. 1–5.

[43] A. Dadashzadeh, A.T. Targhi, M. Tahmasbi, et al., HGR-Net: a fusion network for hand gesture segmentation and recognition, IET Comput. Vis. 13 (8) (2019) 700–707.

[44] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[45] N. Ma, X. Zhang, H.T. Zheng, et al., Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 116–131.

[46] A. Howard, M. Sandler, G. Chu, et al., Searching for mobilenetv3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1314–1324.