Attention Based Residual Network for Micro-Gesture Recognition

Min Peng*, Chongyang Wang†, Tong Chen‡
* Intelligent Media Technique Research Center, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China, pengmin@cigit.ac.cn
† UCL Interaction Centre, University College London, London, United Kingdom, chongyang.wang.17@ucl.ac.uk
‡ College of Electronic and Information Engineering, Southwest University, Chongqing, China, c_tong@swu.edu.cn

***Abstract*—Finger micro-gesture recognition is increasingly become an important part of human-computer interaction (HCI) in applications of augmented reality (AR) and virtual reality (VR) technologies. To push the boundary of micro-gesture recognition, a novel Holoscopic 3D Micro-Gesture Database (HoMG) was established for research purpose. HoMG has an image subset and a video subset. This paper is to demonstrate the result achieved on the image subset for Holoscopic Micro-Gesture Recognition Challenge 2018 (HoMGR 2018). The proposed method utilized the state-of-the-art residual network with an attention-involved design. In every block of the network, an attention branch is added to the output of the last convolution layer. The attention branch is designed to spotlight the finger micro-gesture and reduce the noise introduced from the wrist and background. With an extensive analysis on HoMG, the proposed model achieved a recognition accuracy of 80.5% on the validation set and 82.1% on the testing set.**

Keywords-finger micro-gesture;residual network;attention;

I. INTRODUCTION

As a popular role of HCI in applications of AR and VR, body and hand gesture showed many advantages in device controlling and gaming interaction. According to [1], finger gestures can express more information than body gestures in many aspects. Generally, gesture recognition has attracted much attention in the past few years. For finger gesture recognition, due to unavailability of data, there is little research conducted on finger micro-gesture recognition. In order to push the research on finger micro-gesture, a holoscopic 3D micro-gesture database (HoMG) [2] was collected with a holoscopic 3D (H3D) imaging system. Comparing with previous datasets collected with Kinect and RGB camera, a H3D imaging system is able to capture more accurate finger motion [3].

Residual network [4] has achieved great success in image recognition, which is mainly constructed with a stack of residual block. While for image-based object recognition, the attention mechanism [5] has seen many advantages on improving the performance of respective models. For micro-gesture recognition, as the existence of wrist and background can influence the recognition result, we propose to combine the idea of residual network and attention mechanism to design the recognition model. Details of the proposed model can be seen in Section 3.

HoMG has an image subset and a video subset, while a three-class labelling was created as "Button, Dial and Slider" according to the most commonly used gestures in AR and VR applications. Fig. 1 show the three types of gesture included in the database. Based on HoMG, the HoMGR 2018 is organized. This paper aims to demonstrate our result on the image subset using the proposed residual network with an attention-involved design. With an extensive analysis on HoMG, the proposed model achieved a recognition accuracy of 80.5% on the validation set and 82.1% on the testing set.

II. RELATED WORK

Due to some drawbacks of RGB-D camera and Kinect sensor, e.g. they are unable to provide accurate result with wide view coverage [6] [7], they are not suitable to collect finger micro-gesture data. While recent research [8] [9] have shown the advantage of 3D imaging in panoramic capturing and some other potential areas. [3] proposed that using a H3D imaging system can capture more accurate finger micro-gesture. To push the research of finger micro-gesture recognition, [2] collected the HoMG database with a H3D imaging system. The data is provided with an image subset and a video subset which is further divided into 3 classes: "Button, Dial and Slider" according to the gestures commonly used in AR and VR applications. The distance between the camera and subject was also considered, so there are data with close or long distance in each subset.

Figure 1. Three types of finger micro-gesture descriped in the HoMG database.

For the recognition of finger micro-gesture, [2] also provided some results acquired with traditional feature-based methods. The best recognition accuracy on image subset was 50.9 using Local Binary Pattern (LBP) and k-Nearest Neighbor (k-NN), while a better recognition accuracy of 86.8 was achieved on the video subset using Local Phase Quantization from Three Orthogonal Planes (LPQTOP) and Support Vector Machine (SVM). The higher recognition accuracy in video subset may be due to the motion information that video data can provide. However, as [2] reasoned, to build a video-based finger micro-gesture recognition model would need a large volume of data, and the size of the video subset of HoMG is limited, comparing with the image subset (Totally 960 video clips and 30635 images were collected from 40 subjects). Moreover, the recognition accuracy on the image subset using traditional method is quiet low. Therefore, we

978-1-5386-2335-0/18/$31.00 ©2018 IEEE

chose to design a model basing on the image subset and try to acquire higher recognition result.

The residual network [4] has achieved great success in image-based recognition tasks, which uses a stack of residual block to form the network. Fig. 2 shows the architecture of a classic residual block. In each block, a shortcut connection would be applied to do element-wise summarization of the input and output of the block. This operation would accelerate the training process and improve the performance of the model without introducing more parameter and computation. By stacking such block, the model can also reduce the degenerating problem.

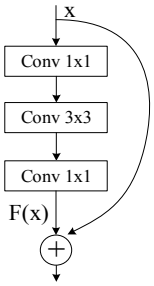

Figure 2. The classic residual block.

For image-based object recognition, the attention mechanism [5] [10] [11] is used to find the region of interest of the subject and highlight the representation of the respective location. In a recent work [10], the attention mechanism is used with the residual network. Specifically, attention modules are designed to fit into the residual block. By stacking such attention enabled residual block, better result is achieved in their experiments, comparing with former residual network. A similar attention mechanism was used in SeNet [11], which added an attention branch to the output of the last convolution layer in the block to learn the channel weight. The SeNet also achieved the best result in a recent ImageNet challenge.

In order to combine the residual network and attention mechanism for finger micro-gesture recognition, we add a compact attention branch to the output of the last convolution in the block which can compute spatial and channel attention value. By doing so, the network can learn to put focus on the gesture-related parts in an image and reduce the noise introduced by the wrist and background.

III. METHOD

In this work, we utilized the state-of-the-art residual network with an attention-involved design. Specifically, a modification has been made to each residual module in the network, the architecture shown in Fig. 3 is the proposed attention-based residual block where the attention branch is shown in the dotted-line area.

Within the attention branch, we used a bottom-up top-down structure [12] [13] [14] to learn the weight feature for the interested spatial area and different channels. Let the size of the feature output F(x) of the last 1×1 convolution in Fig. 3 to be N×C×W×H (N denotes the number of the input sample, C denotes the number of convolutional channels, W and H denote the width and height of the input image, respectively), the size of the output of the following max pooling layer should be N×C×W/2×H/2. Therefore, the following convolution layer can get a better spatial receptive field. For a convolutional neural network, the activation value for different parts of the image show the different importance of them. The following 3×3 convolution layer, as a result, is used also to show the non-linear characteristic between the activation value and the pattern of finger gestures. Then, after an upsampling and another 3×3 convolution, the spatial attention mask A(x) is acquired. The size of A(x) would be the same with F(x) and further normalized to range 0-1 by a sigmoid activation function. Higher value of A(x) represent higher importance of the area. The 1×1 convolution in the branch is used to reduce dimensionality where the number of output feature map would be 1/16 of the output from the former layer. For the original output F(x) of the 1×1 convolution, the output of the proposed block as shown in Fig. 3 is

$$F(x) + F(x) \bullet A(x) + x \tag{1}$$

where $\bullet$ means dot multiplication. Comparing with the original block, the proposed branch only bring in a F(x)·A(x) section which denote the computation of the attention value. And the A(x) will approximately approaching 0 when the attention is not needed for F(x).

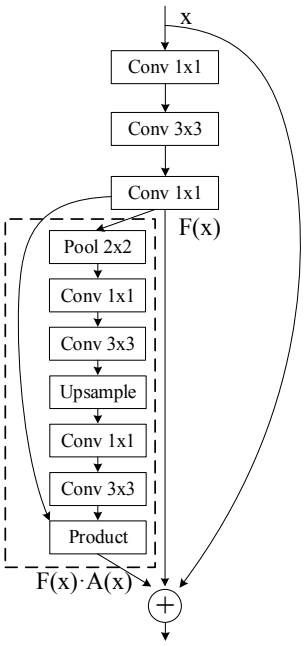

Figure 3. Attention-based residual block.

IV. EXPERIMENT

A. Data augmentation and pre-processing

HoMG [2] has two subsets, a video subset with 960 video clips was collected from 40 subjects using a 3D holoscopic camera and an image subset with 30614 images was selected from the frames in the video subset. The challenge task is to recognize three types of gesture, namely Button, Dial and slider, from those data. The image subset was further divided into a training set with 16763 images from 20 participants, a validation set with 6560 images from another 10 participants and a testing set with 7291 images from the rest 10 participants (The testing data released for the challenge is selected from the testing

set described in [2]). For the challenge purpose, the testing set is not open for use until the final model-testing stage. The experiment was conducted on the training and validation set, while the submitted result was gained on the unlabeled testing set after it has been released.

In order to fit the training of the proposed deep neural network, data augmentation and pre-processing was used.

The original image in the image subset of HoMG with size of 1080×1920, as shown in Fig. 4, is divided into several blocks with size of 3×3. To form one augmented image from an original image, one pixel from each block are sequentially used so that a respective augmented set with 9 images can be created after each pixel in every block is visited. As a result, the size of the image the augmented dataset is 360×640. On the other hand, the augmented dataset is 10 times bigger than the original dataset.

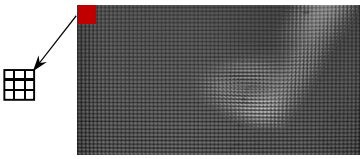

Figure 4. Extract one block from the original image.

The augmented data is first normalized to size of 224x224. In order to enhance the representation of the gesture area and spotlight the region of interest from the background in the original image, a gradient image g(x, y) is then computed from the normalized image f(x, y) according to (2). The computed gradient image can be seen in Fig. 5. For data in the testing set, without any augmentation method, they are directly resized to 224×224 and then transformed to the gradient images.

$$g(x,y)=\sqrt{[\frac{\partial f(x,y)}{\partial x}]^2+[\frac{\partial f(x,y)}{\partial y}]^2} \tag{2}$$

where

$$\begin{aligned}\frac{\partial f(x,y)}{\partial x}&=f(x,y)-f(x+1,y)\\ \frac{\partial f(x,y)}{\partial y}&=f(x,y)-f(x,y+1)\end{aligned} \tag{3}$$

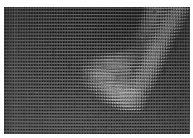

Figure 5. Pre-processed image in a gradient representation after done a size normalization.

B. *Implementation details*

Based on Resnet-50 [4], the original residual block was replaced with the proposed attention-based residual block as shown in Fig. 3. The detailed architecture of our Attention Based Residual Network is given in Table 1. In the table, conv2_x, conv3_x, conv4_x, and conv5_x denote attention-based residual blocks.

TABLE I. DETAIL FOR THE PROPOSED NETWORK

Layer name	***Output size***	***Kernel parameters***
Conv1	112×112	7×7, 64, stride 2
Conv2_x	56×56	3×3 max pool, stride 2 { 1×1,64 3×3,64 1×1,256 *2×2 max pool, stride 2 *1×1,16 *3×3,16 *Upsample layer, stride 2 *1×1,16 *3×3,256 } ×3
Conv3_x	28×28	{ 1×1,128, stride 2 3×3,128 1×1,512 *2×2 max pool, stride 2 *1×1,32 *3×3,32 *Upsample layer, stride 2 *1×1,32 *3×3,512 } ×4
Conv4_x	14×14	{ 1×1,256, stride 2 3×3,256 1×1,1024 *2×2 max pool, stride 2 *1×1,64 *3×3,64 *Upsample layer, stride 2 *1×1,64 *3×3,1024 } ×6
Conv5_x	7×7	{ 1×1,512, stride 2 3×3,512 1×1,2048 *2×2 max pool, stride 2 *1×1,128 *3×3,128 *Upsample layer, stride 2 *1×1,128 *3×3,2048 } ×3
Pool5	1×1	average pool

The parameter in each block is shown in the brackets, where * denotes the parameters is used in the attention branch, and the Upsample layer with stride 2 denotes that the spatial size of an image would be upsampled to two times bigger in this layer. The upsampling method that we used is a linear interpolation method.

For training process, the model was initialized according to Resnet-50 [4] that acquired a state-of-the-art result on ImageNet challenge. The pixel average of the training set has been subtracted from the input gray-scale image and the input image was further divided by the variance of the training set. The dropout ratio of the Pool5 layer was set to 0.5 with the minimum batch size set to 9. We also employed stochastic batch gradient descent algorithm with momentum of 0.9 and weight decay of 0.0005. The initial learning rate was set to 0.001, while the learning rate was divided by 10 after every 10 epochs. Our model were implemented on Caffe framework [15] with an Nvidia Titan X GPU, and the training time is around one day.

C. *Results for HoMGR 2018*

Table 2 shows the result achieved with our proposed model.

TABLE II. RECOGNITION ACCURACY (%) FOR IMAGE BASED MICRO-GESTURE RECOGNITION ON VALIDATION SET AND TESTING SET

Methods	*Validation Set*	*Testing Set*
LBP [2]	41.0	48.9
LPQ [2]	51.6	50.9
Our method	**80.5**	**82.1**

Where we can see the proposed method gained an improvement around 30% of recognition accuracy than the baseline [2]. Moreover, the high recognition accuracy acquired on both validation set and testing set may prove the robustness of our prosed model.

D. Data augmentation and gradient representation influence analysis

Table 3 shows the impact on the performance of proposed model introduced by using the proposed data augmentation method and the gradient representation.

TABLE III. THE INFLUENCE OF THE DATA AUGMENTATION AND GRADIENT RPRESENTATION ON THE RECOGNITION PERFORMANCE

Methods	*Validation Set*
Our method without data augmentation	78.6
Our method without gradient computation	78.9
Our method	80.5

Where we can see that, by using the data augmentation, the proposed model achieved an accuracy improvement of 2% approximately. Same for applying the gradient representation, the accuracy was also improved. Therefore, the data augmentation and gradient representation we proposed is validated to be effective.

E. Attention branch influence analysis

In order to show the impact of the proposed attention branch, we compare the proposed model with the original Resnet-50 [4]. The architecture of the Resnet-50 used here is the same with [4], while the training strategy and hyper-parameter setting is the same with our model. The result is shown in Table 4.

TABLE IV. THE INFLUENCE OF THE ATTENTION BRANCH ON THE RECOGNITION PERFORMANCE

Method	*Validation Set*	*Params*[a]
Resnet-50	77.5	37M
Resnet-101	79.4	68M
Our method	**80.5**	**39M**

a. The total number of parameters learned in each convolutional layer, which can be computed by multiplying the size of convolution kernels with the number of channels. Let the input and output channel number of a convolutional layer to be the same as C, the kernel size to be 3×3, then the Params of this layer is 3^2C^2.

where we can see an improvement of 3% in recognition accuracy is achieved by using proposed attention branch, while the parameters increased about only 2M. In order to show that the improvement gained with our model is not only at the expense of larger parameter size, we used Resnet-101 for comparison, which share a same architecture of ResNet [4], with the same training strategy and hyperparameters of our model. As we can also see from the table, the parameter of Resnet-101 is nearly 1.7 times larger than the proposed model but produced a lower recognition accuracy.

F. Feature visualization

In this section, an feature visualization method is utilized to demonstrate the attentional characteristics expressed in the proposed attention based residual network. As the architecture shown in Table 1, the proposed network used forward propagation to drive the image data flow through the model, while the feature map acquired in different convolutional layers can be seen in Fig. 6. For low-level layers, the outline information is mainly learned. After the attention branch added, in the middle-level layer of the network (Fig. 6 (c), Fig. 6 (f)), the model started to take more attention on the hand gesture area while denoising the edge information at four corners of the image. Moreover, compared with the layer shown in Fig. 6 (c), the layer at the same stage with attention branch added put more attention on the gesture area than the arm area. As a result, in higher level layers, most of the less-interested areas, such as arm and so on, are neglected.

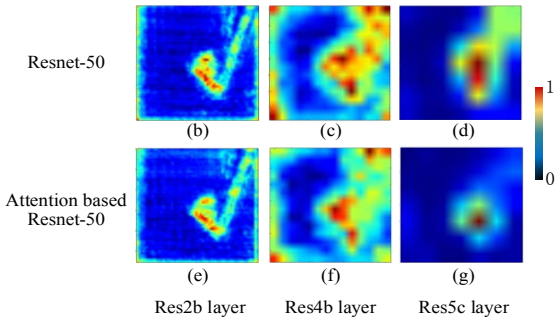

Figure 6. The influence of the attention branch on feature learning. First row and second row represent the feature maps learned by original Resnet-50 and proposed attention based Resnet-50, respectively at res2b layer (low-level) , res4b layer (middle-level) and res5c layer (high level). Higher value in the color bar on the right side denotes higher level of attention

V. DISCUSSION

This paper proposed to use Residual Network with Attention Mechanism for finger micro-gesture recognition. Specially, we proposed to add attention branches to each residual block within the residual network. With an extensive experiment conducted on the HoMG image subset, the proposed model achieved a recognition accuracy of 80.5% on the validation set and 82.1% on the testing set. Specifically, the proposed attention-based residual block employed the residual learning mechanism, where network is able to choose using the attention learning or not in our case. As the attention mechanism is used in a bottom-up top-down convolution structure, the network can learn to focus on the region of interest, e.g. finger gesture area. Generally, the compact and sufficient attention-based residual block designing leads to better performance achieved on HoMG. After this work, we would try to find better representation for the holoscopic 3D image as little research has been done for this purpose.

REFERENCES

[1] Renate Hauslschmid, Benjamin Menrad, and Andreas Butz, “Freehand vs. Micro Gestures in the Car: Driving Performance and User Experience,” vol. 76, 2015, pp.171–180.

[2] Yi Liu, Hongying Meng, Mohammad Rafiq Swash, Yona Falinie A Gaus, and Rui Qin, “Holoscopic 3D Micro-Gesture Database for Wearable Device Interaction”, arXiv preprint, arXiv:1712.05570, 2018.

[3] M. R. Swash, O. Abdulfatah, E. Alazawi, T. Kalganova, and J. Cosmas, “Adopting multiview pixel mapping for enhancing quality of holoscopic 3D scene in parallax barriers based holoscopic 3D displays," IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB, 2014, pp.1–4.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[5] L. Itti, C Koch, “Computational Modelling of Visual Attention,” Nature Reviews Neuroscience, 2001.

[6] Djamila Aouada, Bjorn Ottersten, Bruno Mirbach, Frederic Garcia, and Thomas Solignac, “Real-time depth enhancement by fusion for RGB-D cameras,” IET Computer Vision , vol.7, 2013, pp.335–345.

[7] Rodrigo Ibanez, Alvaro Soria, Alfredo Teyseyre, and Marcelo Campo, “Easy gesture recognition for Kinect,” Advances in Engineering Software , Vol. 76, 2014, pp.171–180.

[8] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, Ivan Poupyrev, and Google Atap, “Soli: Ubiquitous Gesture Sensing with Millimeter Wave Radar,” ACM Trans. Graph. Article , Vol.35, 2016.

[9] Giulio Marin, Fabio Dominio, and Pietro Zanuttigh, “Hand gesture recognition with jointly calibrated Leap Motion and depth sensor,” Multimedia Tools and Applications, 2016, pp.14991–15015.

[10] Wang, Fei, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. "Residual Attention Network for Image Classification," arXiv preprint arXiv:1704.06904, 2017.

[11] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." arXiv preprint arXiv:1709.01507, 2017.

[12] Long, Jonathan, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[13] Newell, Alejandro, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation." European Conference on Computer Vision. Springer International Publishing, 2016.

[14] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." arXiv preprint, arXiv:1511.00561, 2015.

[15] Jia, Yangqing, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. "Caffe: Convolutional architecture for fast feature embedding." In Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 675–678.