JDCS

디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 18, No. 8, pp. 1509-1516, Dec. 2017

# 자동 팔 영역 분할과 배경 이미지 합성

김 동 현[1] · 박 세 훈[1] · 서 영 건[1*]
[1]경상대학교 컴퓨터과학과, 대학원 문화융복합학과

# Automatic Arm Region Segmentation and Background Image Composition

**Dong Hyun Kim[1] · Se Hun Park[1] · Yeong Geon Seo[1*]**
[1]Dept. of Computer Science and CCBM, Graduate School, Gyeongsang Nat'l University, Gyeongnam 52828, Korea

## [요 약]

일인칭 관점의 훈련 시스템에서, 사용자는 실제적인 경험을 필요로 하는데, 이런 실제적인 경험을 제공하기 위하여 가상의 이미지 또는 실제의 이미지를 동시에 제공해야 한다. 이를 위해 본 논문에서는 자동적으로 사람의 팔을 분할하는 것과 이미지 합성 방법을 제안한다. 제안 방법은 팔 분할 부분과 이미지 합성 부분으로 구성된다. 팔 분할은 임의의 이미지들을 입력으로 받아서 팔을 분할하고 알파 매트(alpha matte)를 출력한다. 이는 종단 간 학습이 가능한데 이 부분에서 우리는 FCN(Fully Convolutional Network)을 활용했기 때문이다. 이미지 합성부분은 팔 분할의 결과와 길과 건물 같은 다른 이미지와의 이미지 조합을 만들어 낸다. 팔 분할 부분에서 네트워크를 훈련시키기 위하여, 훈련 데이터는 전체 비디오 중에서 팔의 이미지를 잘라내어 사용하였다.

## [Abstract]

In first-person perspective training system, the users needs realistic experience. For providing this experience, the system should offer the users virtual and real images at the same time. We propose an automatic a persons's arm segmentation and image composition method. It consists of arm segmentation part and image composition part. Arm segmentation uses an arbitrary image as input and outputs arm segment or alpha matte. It enables end-to-end learning because we make use of FCN in this part. Image composition part conducts image combination between the result of arm segmentation and other image like road, building, etc. To train the network in arm segmentation, we used arm images through dividing the videos that we took ourselves for the training data.

색인어 : 팔 영역, 팔 분할, 배경 합성, FCN
**Key word :** Arm region, Arm segmentation, Background composition, FCN

**http://dx.doi.org/10.9728/dcs.2017.18.8.1509**

**Received** 01 November 2017; **Revised** 13 November 2017
**Accepted** 25 December 2017

***Corresponding Author; Yeong Geon Seo**

**Tel:** +82-55-772-1392
**E-mail:** young@gnu.ac.kr

## Ⅰ. Introduction

Training simulation needs to practice specific skills, such as fire fighter training platform[1], flight pilot training simulator[2], car driving simulator[3] and etc. Such these simulations are often operated on virtual reality based environment. These also have third-person perspective or first-person perspective system, but the effects of each perspective are different[4]. For obtaining immersive experience, a simulator should offer the users first-person perspective. And it will be even meaningful as a simulator also provides realism. Therefore, a simulator needs to provide virtual and realistic images at the same time as mixed reality[5]. With this need, we focused on image matting.

Image matting takes an image I as input and divide it into background B and foreground F assuming that I is composited linearly by B and F. The formulation of image matting can be expressed as

$$I = \alpha F + (1-\alpha)B, \alpha \in [0,1] \qquad (1)$$

where $\alpha$ is factor to decide the foreground opacity(alpha matte). However, the formulation is ill-posed because F and B are unknown in eq(1). Despite of the reason, conventional image matting methods approached as closed-form matting[6] and KNN matting[7], but those methods do not produce accurate results. Some works tried to overcome the difficulty as using deep learning[8, 9], though. On the other hand, there are applications for portrait image matting[10, 11]. Those perform image matting through getting semantic segment that it divide into F and B, such as segment and non-segment. Like this approach, we also perform image matting through decomposing cockpit and non-cockpit area to show real and virtual image at the same time.

In this paper, we propose a fully automatic segmentation method for application to aircraft pilot training simulation. It takes a cockpit image as input and makes a score map as output. This score map means the probability whether a pixel belongs on cockpit or not. So, it can be used as alpha matte and we can obtain cockpit area in image. For this task, we use recent convolutional neural networks(CNNs) that have encouraged to solve some visual recognition problems like image classification[12, 13], semantic segmentation[14, 15], object detection[16] and etc. After taking cockpit segment in image, we combine this segment and other images.

## Ⅱ. Related Studies

### 2-1 Image Matting

Conventional image matting methods have poor performance because these only use low-level features and weak high-level context. Closed-form matting[6] derives alpha matte in a closed form without explicit information whether a pixel is foreground or background area. As it considers local region, it is often called as local matting. In contrast, KNN matting[7] is similar with closed-form matting, but it solves the limitations of local matting by propagating alpha values across non-local neighbors as using K-nearest neighbors method in a high dimensional feature space. For this reason, this is called as non-local matting. Recently, some works have applied deep learning to image matting. Cho et al.[8] proposed deep convolutional networks for image matting. Their system takes RGB image, closed-form matting result, and KNN matting result as input. And then, the system predicts high-quality alpha mattes. Xu et al.[9] use encoder-decoder network to obtain alpha mattes and small convolutional network to refine the result of encoder-decoder network.

### 2-2 CNNs for Semantic Segmentation

CNN is initially proposed as LeNet[17] to recognize hand-written digits. After a long time, AlexNet[12] showed the best result in ImageNet Large-Scale Visual Recognition Challenge(ILSVRC) in 2012. Because of success of AlexNet, many studies have used CNN and VGG16[13], FCN[14], deconvolutional network[15] and WFSO[18] were released. Especially, FCN enabled end-to-end learning to conduct segmentation task because FCN replaced fully connected network to fully convolutional network. For this reason, many researches have made use of FCN and modified the architecture.

### 2-3 Application for Image Matting and Segmentation

There are some applications for specific purpose segmentation. ADAS(Advanced Driver Assistance System)[3] uses an encoder-decoder based convolutional neural network to show a segmented image that segments are composed with road, tree, car and etc. Shen et al.[10, 11] proposed that the application for human portrait stylization using image matting. They showed that the segmented image by person that is inferred by FCN can be used in portrait image matting.

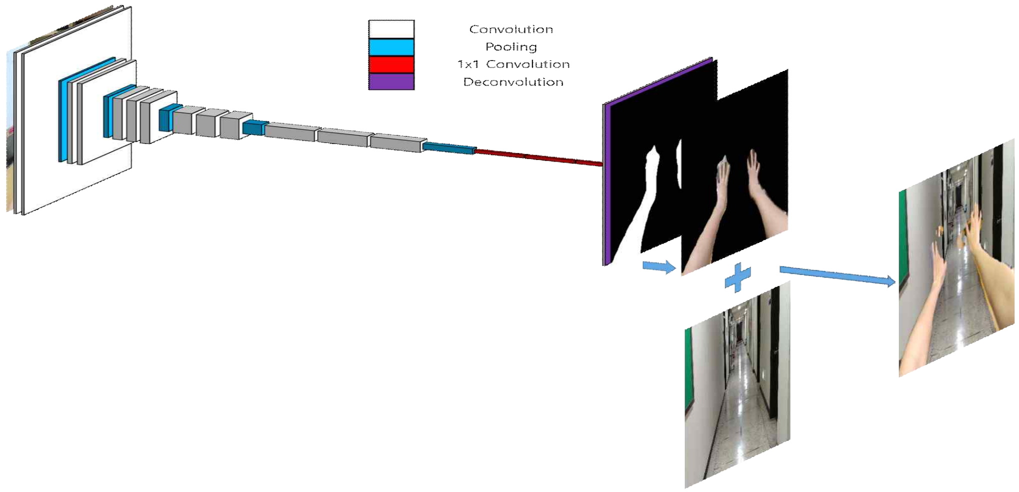

**Fig. 1.** Our system architecture

## Ⅲ. Our Approach

### 3-1 System Architecture

Our system architecture is illustrated in Fig. 1. It consists of FCN based on VGG16 and matting method. VGG16 network consists of 13 convolution layers, 5 pooling layers, and fully connected network. Each convolution layer performs 2d-convolution with 3x3 filters and the result of previous layer. And then, convolutional layer also performs ReLU function as a nonlinear activation function (ReLU is $f(x) = \max(0, x)$). Each pooling layer reduces the size of the input. Fully connected network is responsible for recognition on VGG16, but it can only fixed size.

### 3-2 FCN based on VGG16

FCN solved this problem as changing fully connected network to convolutional network with 1x1 filters. So FCN takes an arbitrary sized image as input. For representation of segment, FCN has deconvolutional layers that it can learn filters to upsample the previous layer. As passing through deconvolutional layers, FCN can infers the segment same sized with input. So this network can be learned end-to-end and does not need any interaction, whereas closed-form and KNN matting need further information like trimap. each pixel on inferred image has same value corresponding cockpit class.

### 3-3 Image Compositing

Next, we conduct conversion inferred image to alpha matte as eq. (2),

$$A(i,j) = \begin{cases} 255 \; (I(i,j) \geqq Class_j) \\ 0 \quad (I(i,j) < Class_j) \end{cases} \tag{2}$$

$Class$ is a set of value for each class. We specify the value of cockpit class as 21(e.g. $Class_0$ means none and $Class_{21}$ means a cockpit). So the values of an inferred image that contains cockpit pixel area are 21 or 0. The extracted image is composited other image like background. The compositing equation is expressed as eq. (3),

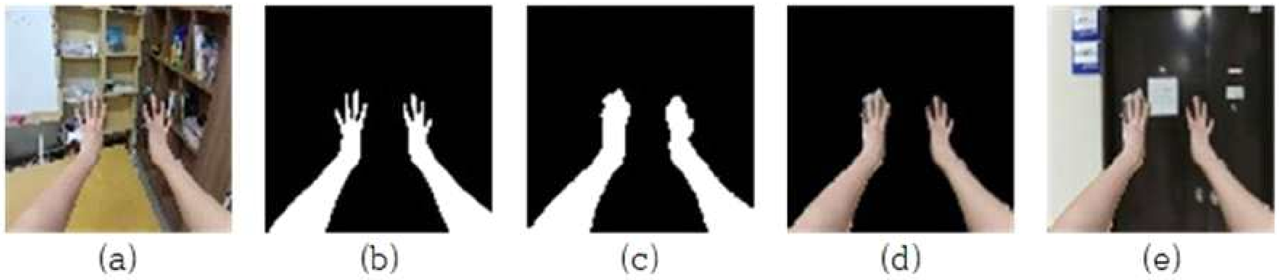

**Fig. 2.** (a) input (b) ground truth,(c) inference (d) foreground and (e) composited image. (d) is generated from (a) and (c), and (e) is composited (d) and another image.

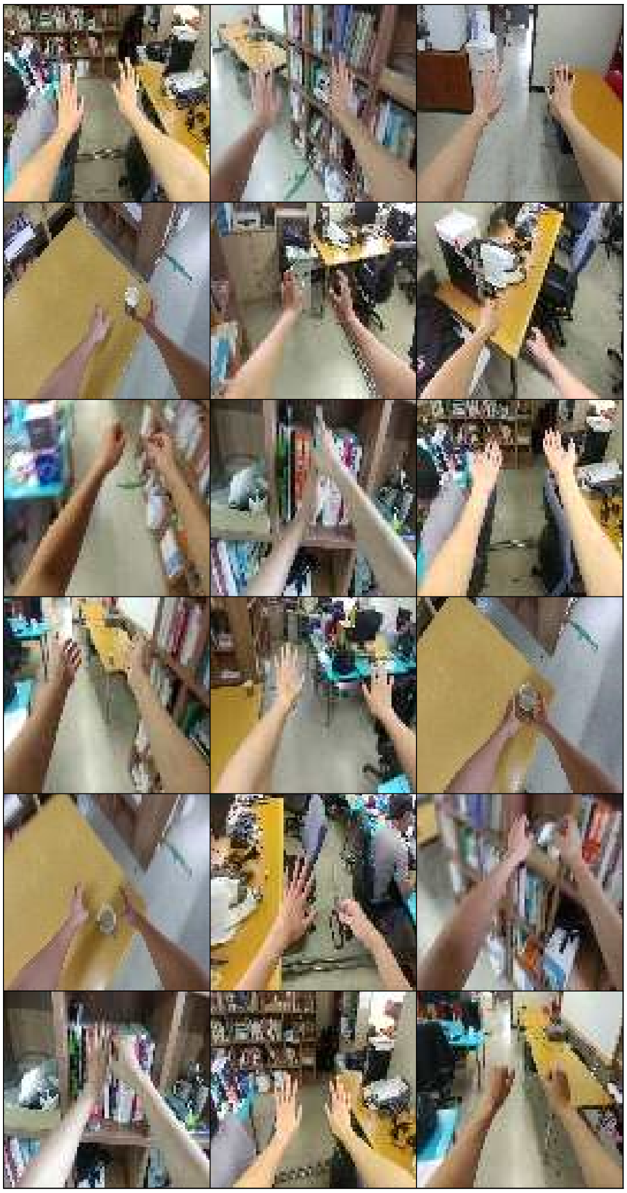

**Fig. 3.** The part of images of our arm dataset

$$Dst(i,j) = \begin{cases} Src(i,j) & (A(i,j) = 255) \\ Background(i,j) & (A(i,j) = 0) \end{cases} \qquad (3)$$

where $Dst(i,j)$ is the result of a composited image, $Src(i,j)$ is input image like an input image containing cockpit, $Background(i,j)$ is an image to be background. Fig. 2 shows the images that are the source images or some intermediate images, and final image.

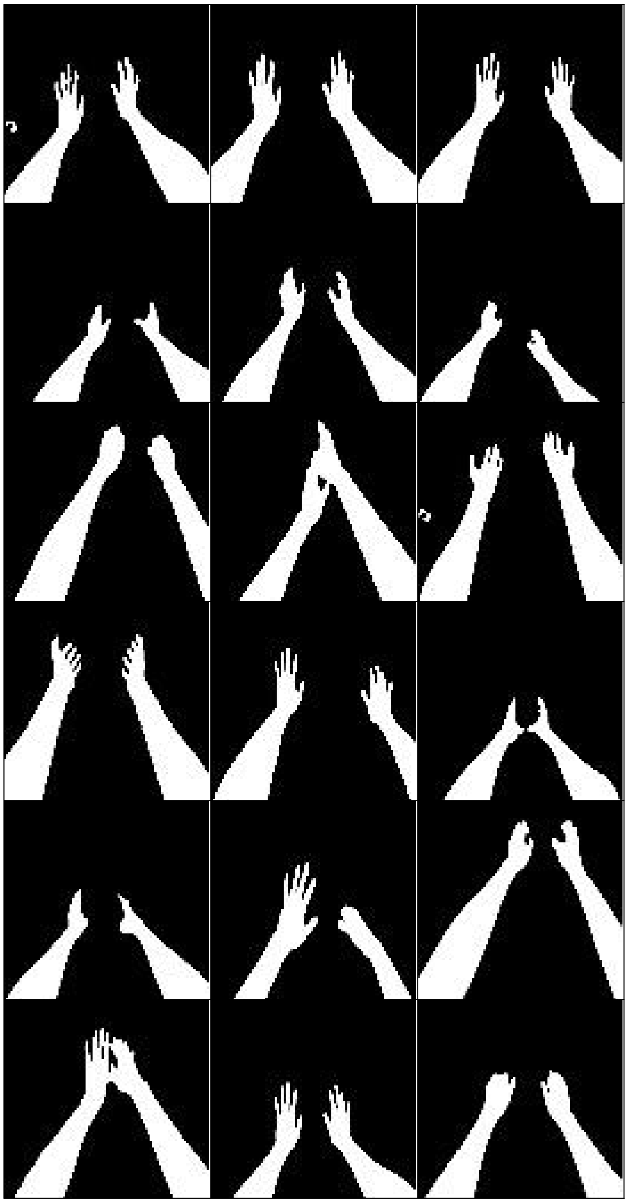

**Fig. 4.** The part of ground truth of our arm dataset

### 3-4 Data Preparation

To collect real arm images, we take a video shot, and divided each frame. We choose a few images in these frames because most frames are alike each other. Fig. 3 and Fig. 4. are illustrated as our images and labels of our dataset. The size of images is 1920x1080. When they put in our system, the size is changed 224x224 because of time due to the curse of the dimension. Other composited images are collected on another video shot.

## Ⅳ. Experimentation and Evaluation

### 4-1 Training and Testing

We use pre-trained VGG16 network on convolution layers of FCN. As training 168 images of our small dataset as much as 20 epochs, the network is intentionally overfitting. The reason of overfitting is to use our work on limited environment. Each image has 224x224 size when it is put in our system. We use softmax function as activation function on last layer and cross-entropy as loss function. Softmax function and cross-entropy can be expressed as

$$\overline{y}_{i,j} = \frac{e^{x}_{i,j}}{\sum_{i}^{n-1}\sum_{j}^{m-1} e^{x_{i,j}}} \tag{4}$$

$$-\frac{1}{n}\sum_{x}(y\ln\overline{y}+(1-y)\ln(1-\overline{y})) \tag{5}$$

where $n$ and $m$ are width and height of an image in eq. 4, and $n$ is a number of data, $y$ is inference from the network, $\overline{y}$ is ground truth in eq. 5. We also use Adam[19] as optimizer in learning rate 1e-6. Training task takes total 1141.201 sec. Testing images are 13 images of our dataset. Each test image has also same size with a train image. Testing task take total 5.709 sec, which take about 0.02 sec per each image(The process time of 1st data is 1.78 sec because of GPU memory allocation).

### 4-2 Experimental results

We evaluate our system on our testing task result. We report region IU(Iintersection over Union) evaluation. The evaluation is expressed as

$$IU_i = \frac{I_i}{U_i}, mean\,IU = \frac{1}{n}\sum_{i=1}^{n} IU_i \tag{6}$$

$$P_i(j,k) = \begin{cases} 0 \; (S_i(j,k) \neq GT_i(j,k)) \\ 1 \; (S_i(j,k) = GT_i(j,k)) \end{cases} \tag{7}$$

$$p_i = \frac{1}{mn}\sum_{j=0}^{m-1}\sum_{k=0}^{n-1} P_i(j,k), mean\,P = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{8}$$

where $I_i$ is intersection, $U_i$ is union between each inference and corresponding label in eq.. 6, $P_i$ is accuracy map, $S_i$ is a segment image, $GT_i$ is ground truth image in eq. 7, and $p_i$ is the accuracy that is from summation $P_i$ being divided by the size of $P_i$ in eq. 9. The result is on Table 1, Fig. 5. Our work is executed with Tensorflow[20] on a single NVIDIA GTX 1070 8GB, Intel i7-7820HK 2.90GHz CPU, 16GB RAM, and Windows 10 64-bit.

**Table 1.** The result of proposed system

| | $mean\,P$ | $mean\,IU$ |
|---|---|---|
| Closed-form matting[6] | 0.844 | 0.652 |
| KNN matting[7] | 0.916 | 0.776 |
| Proposed system | 0.913 | 0.916 |

## Ⅴ. Conclusion

We proposed automatic a person’s arm segmentation and image compositing system. It consists of segmentation part and image composition part. The segmentation part outputs arm segment or alpha matte after training end-to-end. The composition part composites the result of segmentation part and other images. The composite result of our system can be used in virtual-reality like FPP simulation.

## References

[1] M. Cha et al, "A virtual reality based fire training simulator integrated with fire dynamics data", *Fire Safety Journal*, Vol. 50, pp. 12-24, 2012.

[2] Q. Kennedy et al, "Age and Expertise Effects in Aviation Decision Making and Flight Control in a Flight Simulator", *Aviation, Space, and Environmental Medicine,* Vol. 81, No. 5, pp. 489-497, 2010.

[3] P. Backlund et al, "Games for traffic education: An experimental study of a game-based driving simulator", *Simulation & Gaming*, Vol. 41, No. 2, pp. 145-169, 2010.

[4] P. Salamin et al, "Quantifying effects of exposure to the third and first-person perspectives in virtual-reality-based training", *IEEE Transactions on Learning Technologies*, Vol. 3, No. 3, pp. 272-276, 2010.

[5] F. S. Dean, P. Garrity and C. B. Stapleton, "Mixed reality: A tool for integrating live, virtual and constructive domains to support training transformation", *The Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, 2004.

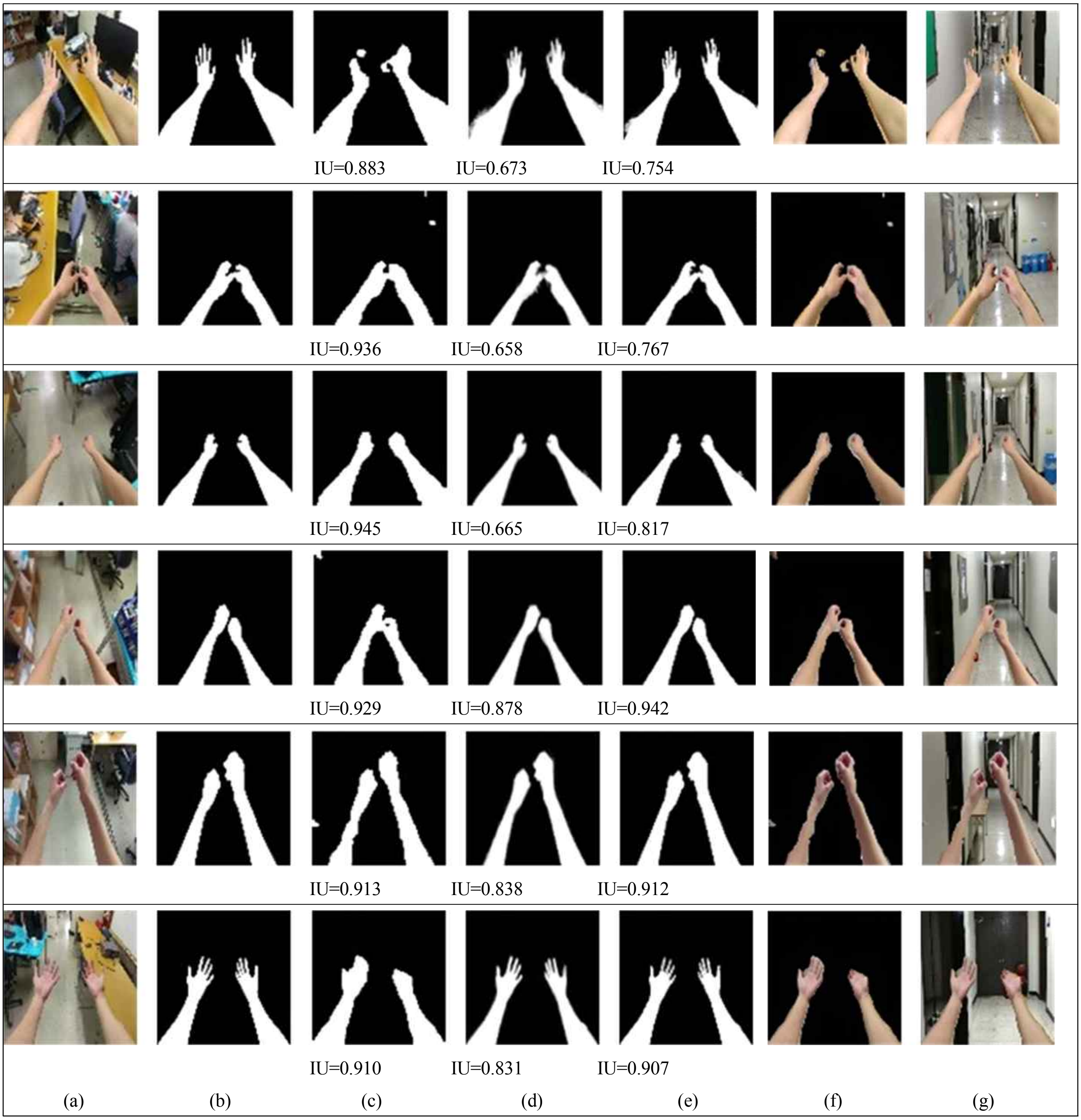

**Fig. 5.** (a) input (b) ground truth (c) inference (d) the result of closed-form matting (e).KNN matting (f) foreground and (g) composited image

[6] A. Levin, D. Lischinski and Y. Weiss, "A closed-form solution to natural image matting", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 30, No. 2, pp. 228-242, 2008.

[7] Q. Chen, D. Li and C. Tang, "KNN matting", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 35, No. 9, pp. 2175-2188, 2013.

[8] D. Cho, Y. Tai and I. Kweon, "Natural image matting using deep convolutional neural networks", *European Conference on Computer Vision,* pp. 626-643, 2016.

[9] N. Xu et al, "Deep Image Matting", Available: http://arxiv.org/abs/1703.03872, 2017.

[10] X. Shen et al, "Automatic Portrait Segmentation for Image Stylization", *Computer Graphics Forum,* Vol. 35, No. 2, pp.

93-102, 2016.

[11] X. Shen et al, "Deep automatic portrait matting," *European Conference on Computer Vision,* pp. 92-107, 2016.

[12] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012.

[13] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", 2014.

[14] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431-3440, 2015.

[15] H. Noh, S. Hong and B. Han, "Learning deconvolution network for semantic segmentation", *Proceedings of the IEEE International Conference on Computer Vision,* pp. 1520-1528, 2015.

[16] J. Redmon et al, "You only look once: Unified, real-time object detection", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 779-788, 2016.

[17] Y. Lecun et al, "Gradient-based learning applied to document recognition", *Jproc*, Vol. 86, No. 11, pp. 2278-2324, 1998.

[18] S. Jang and H. Jang, “Training Artificial Neural Networks and Convolutional Neural Networks Using WFSO Algorithm”, *Journal of Digital Contents Society*, Vol. 18, No. 5, pp. 969-976, 2017.

[19] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", 2014.

[20] M. Abadi et al, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems", 2016.

Dong Hyun Kim

2016. 2 : Department of Computer Science of Gyeongsang National University (B.S Degree)

2018. 2 Department of Computer Science, Graduate School of Gyeongsang National University (M.S Degree Candidate)
※ Research Interests : AI, Deep Learning, Image Segmentation

Se Hun Park

2016~now : Department of Computer Science of Gyeongsang National University

※ Research Interests : AI, Deep Learning, Virtual Reality

Yeong Geon Seo

1987.2 : Department of Computer Science of Gyeongsang National University (B.S Degree)
1997.2 : Department of Computer of Soongsil University (Ph.D Degree)

1989~1992 : Trigem Computer Inc.
1997~now : Department of Computer Science of Gyeongsang National University, Professor
2014~now : Department of CCBM, Graduate School of Gyeongsang National University, Professor
※ Research Interests : AI, Deep Learning, Medical Imaging, IT Convergence, Computer Network