Towards Self-Perception in Augmented Virtuality: Hand Segmentation with Fully Convolutional Networks

E. Gonzalez-Sosa & P. Perez & R. Kachach & J. J. Ruiz & A. Villegas

Nokia Bell Labs

Abstract

In this work, we propose the use of deep learning techniques to segment items of interest from the local region to increase self-presence in Virtual Reality (VR) scenarios. Our goal is to segment hand images from the perspective of a user wearing a VR headset. We create the VR Hand Dataset, composed of more than 10.000 *images, including variations of hand position, scenario, outfits, sleeve and people. We also describe the procedure followed to automatically generate groundtruth images and create synthetic images. Preliminary results look promising.*

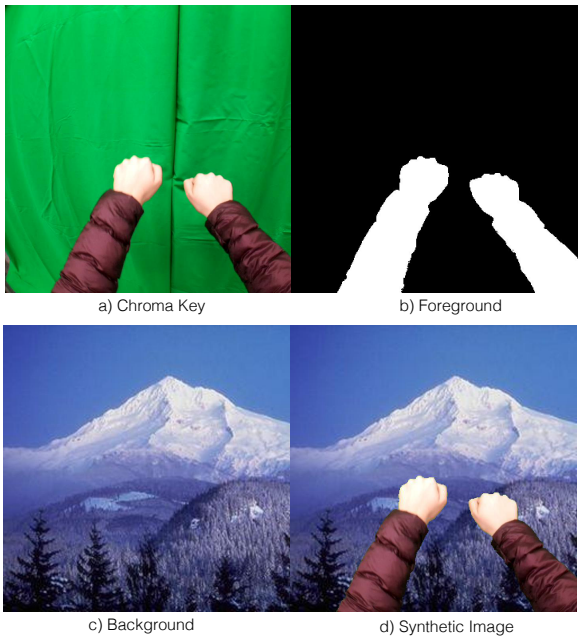

Figure 1: Pipeline of VR Hand Dataset creation.

1. Introduction

Augmented virtuality is a subcategory of mixed reality technologies in which real world objects are merged into the virtual world with different purposes: *i*) integrate self-presence or awareness of other people to prevent isolation, or *ii*) ease interaction with local items such as keyboards, smartphones, coffee cups, etc. [MBMSB15]. Concerning enhancing self-presence in Virtual Reality (VR), researchers have focused on integrating full body or some body parts (e.g. hands). Approaches followed to date have been mainly based on leveraging colour and depth information from Kinect devices, virtual hand tracking user movements, or using some chroma key to select particular items. In this work we propose to use deep learning to smartly segment items of interest from the local region. This cutting edge technology will be useful to circumvent problems derived from former methods.

2. Fully Convolutional Networks

Fully Convolutional Networks (FCN) [LSD15] were proposed to adapt Convolutional Neural Networks (CNN) from the classification to the segmentation task, where output are images instead of text labels. FCN architecture is composed of two subnetworks: a CNN and a deconvolutional network. The CNN subnetwork is used to obtain a feature representation of the input image. To prevent training from scratch, VGG-16 pre-trained model has been used as the CNN subnetwork [SZ14]. Then, since CNN gradually loses spatial information, a deconvolutional netwok is stacked after the CNN to recover that spatial information while preserving the information needed for distinguishing between classes. We have adapted a FCN network proposed by Long *et. al* [LSD15] to a binary segmentation problem, considering only two classes: 1) hand and 2) background.

3. VR Hand Dataset

The goal is to be able to segment hand images from the point of view of a subject wearing a VR headset. As there are not enough

DOI: 10.2312/egp.20181012

Table 1: VR Hand Dataset. It has been acquired with different configurations: people, position, scenario, outfit, and sleeve.

People	11 male and 1 female
Position	close hand, open palm, open dorsum, left hand, right hand
Scenario	outdoors, indoors
Outfit	outfit1, outfit2
Sleeve	Long sleeve, short sleeve

images of this nature, we have created our own database—VR Hand Dataset—to train the deep learning architecture presented in Section 2. In order to prevent manual labelling of thousands of images, a chroma key was used with a double purpose: *i*) automatically obtain groundtruth images, and *ii*) generate synthetic images for the training stage.

3.1. Configuration

Our purpose is to acquire a dataset with a wide range of variations that could help the FCN algorithm to maximize their generalization capabilities. Table 1 presents the VR Hand Dataset configuration. As can be seen, 5 different hand positions are considered. Different scenarios, outfits and sleeve are also considered.

3.2. Chroma Key and Groundtruth Generation

1) Acquisition. An Android application is developed in order to record 30 fps videos from the Samsung-S8 frontal camera while the subject is wearing the Gear VR Samsung headset with the smartphone, in front of a chroma key backdrop. Each session is designed to record videos with a particular *{people,scenario,outfit,sleeve}*-configuration. A voice assistant ensures that, at each session, videos from the 5 different hand positions are recorded. Fig.1a shows a particular chroma key frame.

2) HSV Filtering. With the recorded chroma key videos, a HSV-based filter is applied to obtain the foreground images (values are in the range $0-1$), as follows:

$$f(x) = \begin{cases} 1 & \text{if } H(x,y) \leq 0.22 \wedge H(x,y) \geq 0.45 \wedge S(x,y) \geq 0.20 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

3) Frame Selection and Pre Processing. To prevent high similarity, final images are selected every 5 frames. Then, some morphological operations are applied to delete noisy areas. As a result, the VR Hand Dataset contains more than 10000 images. A resulting frame could look like Fig.1b.

3.3. Synthetic Images

Synthetic images are created combining background with chroma key images masked with foreground images. Background images are obtained from the MIT Scene Parsing Bechmark. Given a chroma key (Fig.1a), a background image (Fig.1c) is randomly chosen from a $\sim$ 4000-set. Then, pixels from the chroma key image belonging to the foreground (Fig.1b) are overlapped to the background to conform the final synthetic image (Fig.1d).

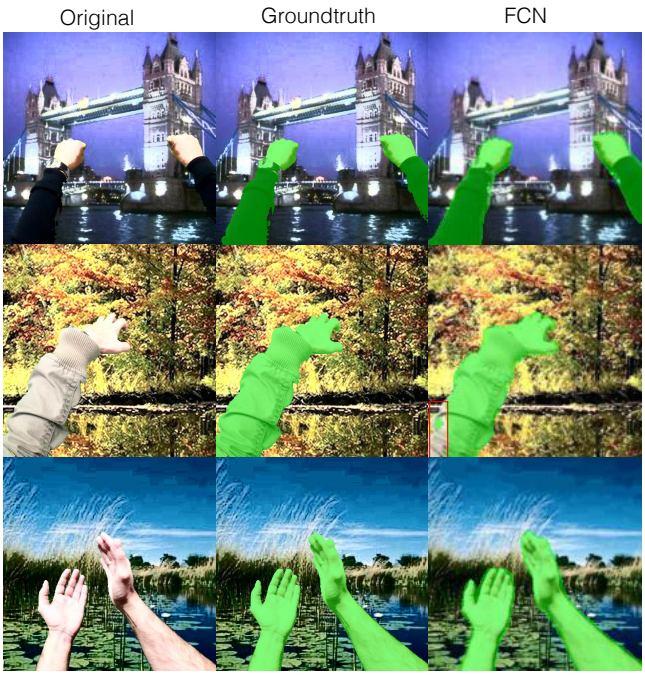

Figure 2: Segmentation Samples. Left) Original Images; Center) Groundtruth Images; Right) Output from the FCN semantic segmentation algorithm. Green areas are overlapped to original images to indicate which pixels are labeled as hands.

4. Results and Conclusions

Fig.2 shows some preliminary visual results of our algorithm. The second row example presents a particular case in which the algorithm is not accurately segmenting the hand (red box area); we hypothesize it has to do with the similarity between clothes and background colour.

Our plan for future work is to explore more in depth this semantic segmentation problem, giving more quantitative results in terms of *Intersection over Union*, test the generalization capabilities of the FCN algorithm with real images, and deploy the algorithm in the smartphone device to be used in our Augmented Virtuality scenario.

References

[LSD15] LONG J., SHELHAMER E., DARRELL T.: Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3431–3440. 1

[MBMSB15] MCGILL M., BOLAND D., MURRAY-SMITH R., BREWSTER S.: A dose of reality: Overcoming usability challenges in vr head-mounted displays. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM, pp. 2143–2152. 1

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2014). 1