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 $10^5$ 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.

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 network 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...
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:

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

(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 ~ 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).

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 \textit{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

