Annex B to the Bachelor Thesis

Real-Time Egocentric Segmentation of Local Reality

Jorge Calvar Seco

Universidad Pontificia Comillas (ICAI)

Director: Ester González-Sosa

Nokia eXtended Reality Lab

November 15, 2022

Contents

1 Introduction **2**

2 State of the art **3**

3 Motivation and objectives **5**

4 Methodology and timeline **5**

5 Resources to use **6**

1 Introduction

Virtual Reality is a rapidly growing field that tries to simulate an environment and make it appear real to the user. To do this, a headset is used with a mounted display, so that the user can only see what is shown on the screen, and therefore has the feeling of being in an alternate world. To make this experience more authentic, motion sensors (e.g., accelerometers, gyroscopes, eye tracking devices, ...) are used, and the display adapts accordingly when you move your head, giving the sensation of actually being in another world. Additionally, to allow user interaction with this VR, a set of hand controllers allow to track the position and orientation of the user's hands, as well as having some buttons for the user to click.

Similarly, the field of Extended Reality (XR) tries to combine the virtual environment with parts of the real world. The most common way to implement this is with an egocentric camera on the VR headset, which captures what the user would see if it did not have the headset covering his eyes. Depending on the percentage of Local and Virtual Reality being combined, it can also be called Augmented Reality (AR) or Augmented Virtuality (AV).

The applications of these technologies are limitless, and encompass a wide variety of sectors. It can already be used in some games for a more real first person experience, or to fly FPV drones, for example. Meta (previously Facebook) is a company that has heavily invested in XR, what they refer to as the Metaverse.

To be able to immerse the person into an XR environment, it is necessary to create an algorithm that decides which parts of the real world will be kept and which ones will be replaced by the virtual world. This algorithm must be able to identify each of the objects or parts of an image, and their borders. This task is known in Machine Learning as Semantic Segmentation. Its mission is to classify each of the pixels of an image according to the semantic meaning of the object to which that pixel belongs.

Finally, an important part of this problem is that it is intended to be used in real-time,

i.e., the user needs to have the appearance that he is living in the virtual world, and that will not be possible if there is a lag. Therefore, the time it takes to segment an image must be, at most, 1 divided by the desired frame rate.

2 State of the art

It is usual in the field of semantic segmentation to use the following metrics to measure the quality of the results:

- **IoU** (Intersection over Union): for each class that needs to be segmented, we compute the pixels that are positive for both the prediction and the ground truth, and then calculate both the intersection and the union of these regions. Finally, the metric is obtained by dividing the number of pixels in each of the resulting regions.
- **PA** (Pixel Accuracy): the percentage of pixels that have been classified correctly.

There has already been work trying to solve the problem of real-time egocentric segmentation. Gonzalez-Sosa et al. [1] [2] have already presented significant results. The architecture they have used is based on Thundernet [4], which is especially suitable for this task because it manages to comply with the required compatational time restrictions.

Additionally, to train the model they have created their own dataset from several sources. One of them is the THU-READ dataset [5], which was created by researchers from Tsinghua University to work on solving the task of egocentric action recognition. The other datasets used are EgoHuman, and EgoOffices [3].

The EgoHuman dataset was created in a semi-synthetic manner, by capturing human arms from an egocentric view with and then, using a chroma-key, changing the background to a realistic one. Therefore, obtaining the label was straightforward. However, the labeling of the THU-READ and EgoOffices datasets was done manually through *Amazon Mechanical*

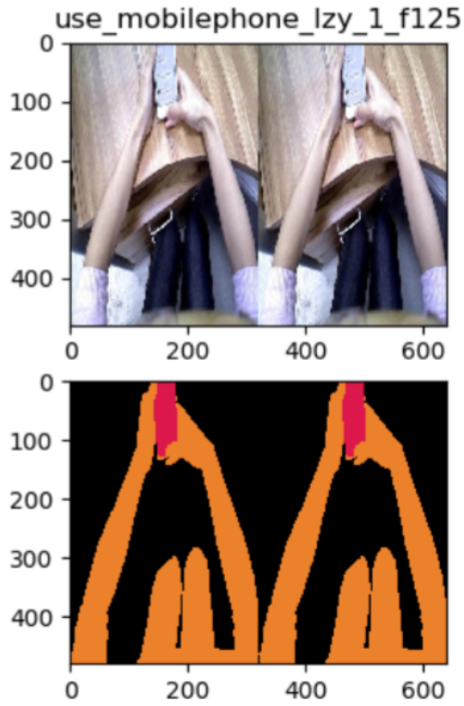

Figure 1: An example of a labeled image

Turk. This service allows to define a simple task, which is then solved by the thousands of workers that Amazon puts at your disposal. The researchers manually checked the correctness of the results that were returned by AMT.

In figure 1, we observe one of the images from the dataset and its label. We can see how the human body was segmented in orange, and the phone in red. We also appreciate that the image is repeated twice. This is because the algorithm is being trained to work with a stereo camera, which are 2 cameras separated just a few centimeters.

3 Motivation and objectives

In this Bachelor thesis, we will aim to improve the current algorithms of Semantic Segmentation for the application of eXtended Reality, i.e., taking into account that the algorithm will be performed in real-time and that the input is egocentric.

We believe this is an important topic to work on because this field may expand significantly in the following years, especially since the world has become more virtual after the Covid-19 pandemic hit. New mainstream applications of this technology beyond the ones mentioned before will also appear. A common one is for virtual calls, like Zoom but instead using a VR headset, which will give a feeling of almost being in person.

4 Methodology and timeline

The main part of the thesis will be spent working on researching ways in which to improve the algorithms for real-time egocentric semantic segmentation. This will be done in an iterative manner: define the experiment, try it out, and analyze the results. During this process, we will record all the experiments performed as well as our findings.

The things that will be tried out may be grouped in the following categories:

- **Algorithm architecture:** research the literature on model architectures used for Semantic Segmentation, including the optimizer and the loss function. Then, we will try them out or make our own architecture.
- **Tuning hyperparameters:** try out different values for the parameters used to train a model, such as the number of epochs, the learning rate, or the regularization parameter.
- **Datasets:** find new datasets and ways in which they can help us, as well as feature or data engineering of the current datasets to produce better results.

Finally, around March 2023, we will start crafting our overall conclusions with the contributions we have made. The thesis we will be defended around June 2023.

5 Resources to use

To solve the task at hand will use the following tools:

- **Hardware:** a computer equipped with an NVIDIA graphics card will be used to train the models within a reasonable time frame.
- **Software**: we will write the code using Python. The most important libraries that we will use are Tensorflow and its high-level interface, keras, for defining and training models, and OpenCV to manipulate images, including recording, visualization, and transformation.

References

[1] Gonzalez-Sosa, E., Gajic, A., Gonzalez-Morin, D., Robledo, G., Perez, P., & Villegas, A. (2022, July 4). Real time egocentric segmentation for video-self avatar in mixed reality. arXiv.org. Retrieved November 15, 2022, from `https://arxiv.org/abs/2207.01296`

[2] E. Gonzalez-Sosa, G. Robledo, D. Gonzalez-Morin, P. Perez-Garcia & A. Villegas, "Real Time Egocentric Object Segmentation for Mixed Reality: THU-READ Labeling and Benchmarking Results," 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), 2022, pp. 195-202, doi: 10.1109/VRW55335.2022.00048.

[3] E. Gonzalez-Sosa, P. Pérez, R. Tolosana, R. Kachach and A. Villegas, "Enhanced Self-Perception in Mixed Reality: Egocentric Arm Segmentation and Database With Automatic Labeling," in IEEE Access, vol. 8, pp. 146887-146900, 2020, doi: 10.1109/ACCESS.2020.3013016.

[4] Qin, Z., Li, Z., Zhang, Z., Bao, Y., Yu, G., Peng, Y., & Sun, J. (2022, March 28). ThunderNet: Towards real-time generic object detection. arXiv.org. Retrieved November 15, 2022, from `https://arxiv.org/abs/1903.11752`

[5] Tsinghua University RGB-D egocentric action dataset. THU-READ. (n.d.). Retrieved November 15, 2022, from `http://ivg.au.tsinghua.edu.cn/dataset/THU_READ.php`