Hand Localization Using YOLO on Depth Data

Maikel Both

Master of Electronics and ICT Engineering Technology

Introduction

The Human Interface Mate (HIM) developed by Arkite is an augmented reality technology that assists production operators in real time. The HIM currently detects when any volume enters a predefined area on the workbench to proceed to the next step in the process. However, it is not checked whether this is a person's hand, which indicates a certain intention, or something else, their elbow for example (accidentally). That could result in the HIM going to the next step in the process too quickly and skipping important steps. Detecting the location of the hand of the operator is a solution to this problem, ensuring that no steps in the process are skipped and to avoid accidental (false) triggers. The hand detection can be done on RGB images, but this could possibly raise a privacy concern. That concern can be avoided when only using depth and/or IR images to detect the hands.

Objective: Hand Localization

The goal of this project is to research various existing techniques related to hand detection and choose the most viable one for implementation based on the requirements of Arkite. The method selection is visualized on figure 1. The requirements of Arkite are:

- The method must locate the hands from top-view perspective.
- The method must locate the hand in each frame.
- The method works in real-time at a preferred rate of 30 frames per second.
- The method uses Depth and/or IR images.

The viable methods remaining after checking these requirements are YOLO, OpenPose, Voxel-to-Voxel and Anchor-to-Joint. Among those, YOLO is the only method that can do hand localization and use depth data. The method gets trained on depth data to localize hands.

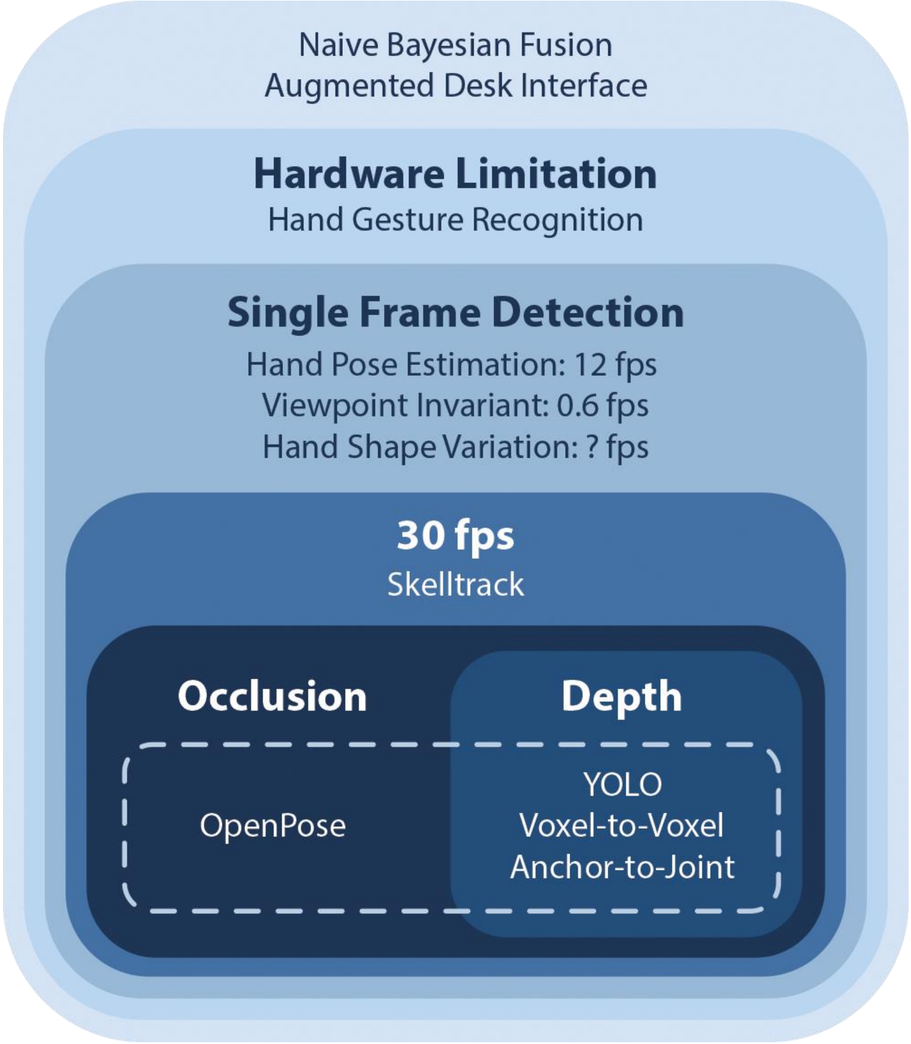

Figure 1: Methods and deciding factors

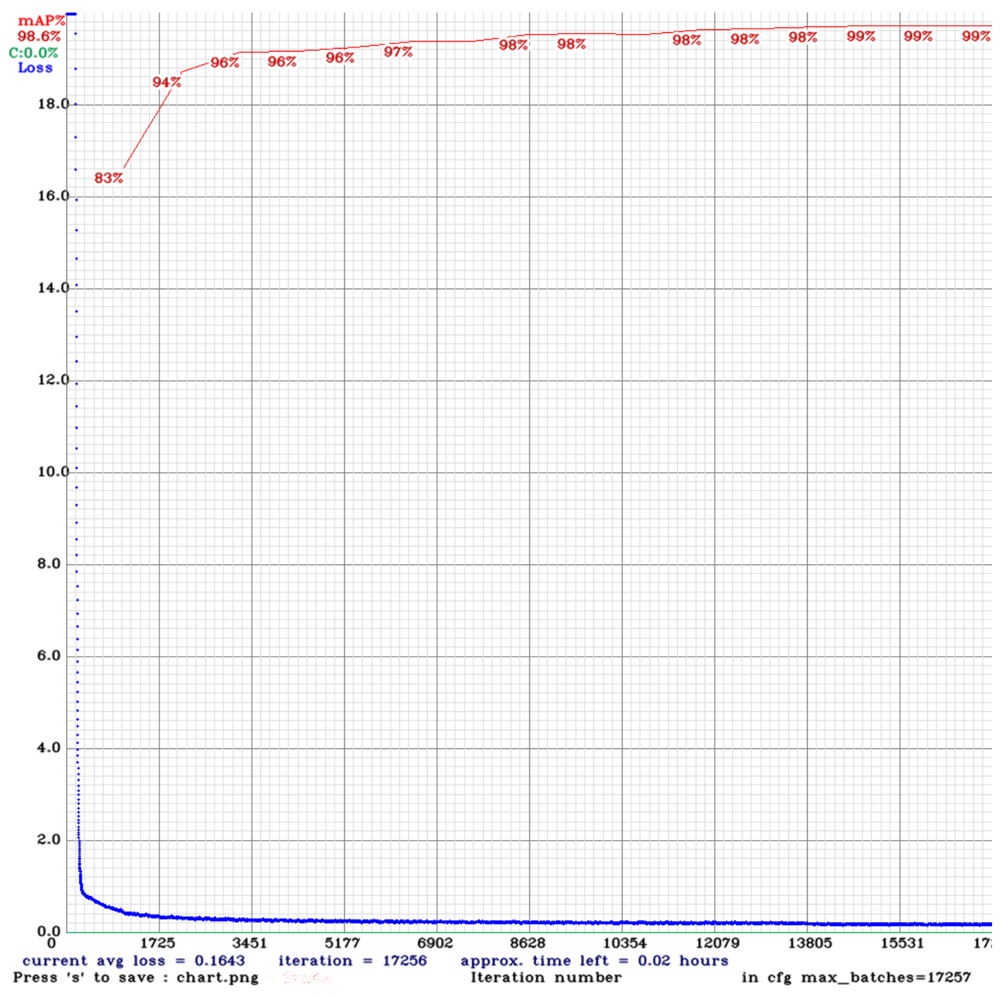

Figure 2: YOLO learning curve on training with ITOP dataset

Datasets

The ITOP dataset contains over 50,000 depth images of people from a top-view perspective where each body part is annotated. The images are saved in a float format where each value is the distance to the sensor in meters.

The Arkite dataset is limited in size of about 1,200 images. These images are not annotated yet and are saved in signed int16 format. Negative numbers are used in case the sensor fails to read the data. Each value is the distance to the sensor in millimeters. This is converted to meters to be compatible with the ITOP dataset.

Results

Training YOLO on the ITOP dataset results in a mean average precision (mAP) of 98,67% on ITOP dataset and 7,35% on Arkite dataset. The learning rate is adjusted upon adding the Arkite dataset through transfer learning. Depending on the learning rate, the mAP results on the ITOP and Arkite dataset differ greatly.

Table 1 shows that an equal learning rate to the one used for the ITOP training results in a drop of over 50% on the ITOP dataset and an increase of almost 15% on the Arkite mAP. Using a third of that learning rate shows the best results, with only a 0,07% drop in ITOP mAP and an increase of over 5% on the Arkite mAP. Using an even lower learning rate will fail to improve the Arkite mAP at a good rate. Figure 3 shows the bounding boxes detected by the new model on an image of Arkite.

Methods

The configuration for the YOLO model is adjusted to be able to train on depth images. The datasets used are split in 80% for training and 20% for testing. The mean average precision (mAP) values are calculated using the testing subset.

The ITOP dataset is used to train YOLO on initially. The training process can be stopped earlier as it reaches a point where more training does not improve the result further. This is shown on the graph on figure 2 where the red line does not increase any further past iteration 14,000.

Once YOLO is trained on the ITOP dataset, it is used to train on the Arkite dataset by applying transfer learning. Transfer learning takes the model result of the ITOP training and uses it as a starting point for training on the Arkite dataset. The learning rate of the YOLO model is decreased for this phase as the Arkite dataset is much smaller than ITOP dataset. Three different learning rates are used to compare results.

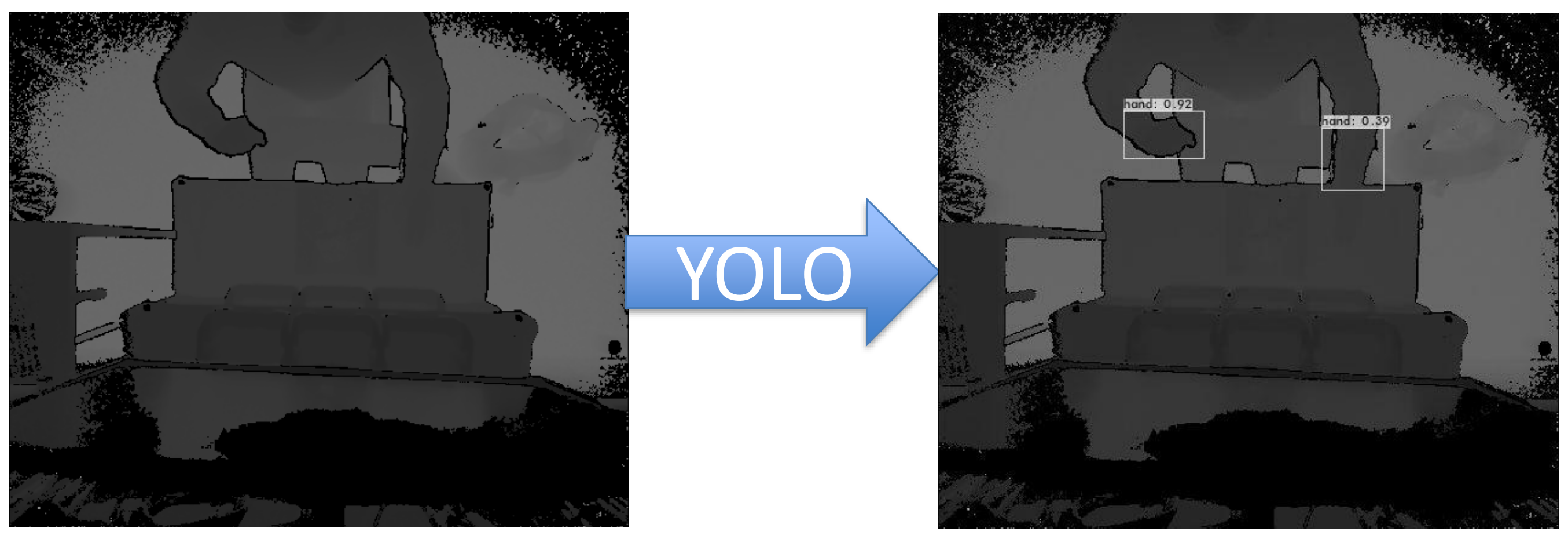

Figure 3: YOLO detection shown on Arkite data

Learning Rate	ITOP mAP	Arkite mAP
0.00261	42,79%	22,05%
0.00087	98,60%	13,59%
0.000261	98,66%	9,72%
0	98,67%	7,35%

Table 1: Results on Arkite and ITOP dataset using different learning rates for transfer learning.

Conclusion

YOLO initially retrained on the ITOP dataset resulted in a mean average precision (mAP) of 98,67%. Using the method of transfer learning has shown that the learning rate should be a third of the rate used for the ITOP dataset. It improved the mAP on Arkite dataset up to 13,59% while decreasing the ITOP mAP by only 0,07%. The YOLO model mAP can be improved by adding more Arkite data.

Supervisors / Co-supervisors / Advisors

Prof. dr. ir. Eric Demeester
Stijn Debruyckere
Yanming Wu