Bachelor's Thesis Review
Implementation of a Computer Vision Algorithm for Onboard Detection of Unmanned Aircraft
submitted by *Lukáš Bauer*

The bachelor's thesis of Lukáš Bauer is concerned with vision-based algorithms for the detection and localization of UAVs in images from an on-board UAV camera. This is an interesting problem with several applications. The first set of goals included studying different methods for the real-time detection and localization of UAVs, selecting a method for implementation from the candidates, and then integrating and optimizing the selected method as a Robot Operating System (ROS), which operates on the embedded computer of the UAV. The second set of goals included testing the method on simulated and real-word data, evaluating the precision and the computational speed of the implementation on the embedded device, and preparing the system for integration into a formation-control algorithm.

The goals of the thesis were only partially fulfilled. The thesis includes a study of the literature and describes several approaches for object detection. Lukáš Bauer chose the YOLOv2 visual detector and implemented this detector as a ROS node operating on-board of a UAV. The YOLO detector is a state-of-the-art convolutional neural network that is tuned for real-time object detection. The configuration file and the pre-trained weight file for the YOLO detector are available online. For the given problem it was sufficient to make only very small changes to the configuration of the detector in order to run this detector for the given problem on standard PC. Slightly more effort was necessary to port the YOLO detector to the embedded UAV computer. The thesis presents a simple method for filtering false positives from detections. In addition a method for computing the relative position of the detected UAV with respect to the UAV with the camera from a detected bounding box is introduced.

The YOLO network was trained on 2500 annotated images; however, the images were taken from the same recording, which may not provide enough variation to properly train the network. The annotated images should have been drawn from different recordings to improve the performance of the detector. The lack of diversity in training data seems to be confirmed by the conclusion of the thesis, which states that the current version of the detector is functional only under very specific conditions and, even then, many false positives and imprecise bounding boxes are produced.

The main weakness of the thesis is in the proposed experimental evaluation. At the time of evaluation, the relative position estimator was not implemented and the latency of the detector was 7s. Therefore, in the first "leader-follower" experiment, the results do

not show the functionality of the proposed relative pose estimator, and these results are significantly affected by detection latency. The second experiment, where the precision of relative pose estimator was tested, was performed on the desktop PC instead of the embedded computer. The errors of the relative UAV position estimated from the bounding boxes received from the detector are huge - varying from 1 to 10 meters. Moreover, as mentioned by the student, the YOLO detector produces less accurate bounding boxes on the embedded GPU; therefore, these errors would likely increase in real applications. Since the position estimated from the bounding box obtained from the detector follows a similar trajectory to the position from ground truth, but with a constant offset, it's more likely that the error is caused by a "bug" in the proposed method, e.g., an error in the camera calibration. However, I think that simple unit-tests could be designed to cover all parts of the proposed method, which would likely reveal the cause of the error. For example, it's easy to measure the error of the camera calibration, and most calibration toolboxes report this error. Other parts of the proposed pipeline could be replaced by ground-truth data and, in this way, tested. Unfortunately, the two proposed experiments do not really show any strength of the work. I believe that the same detector can achieve significantly better results.

The text of the thesis could be improved. Many technical details are missing, or they are not clear and the proposed equations are wrong. Listed below are a few examples:

1. Equation 4.3 doesn't correspond to Fig 4.1 and in this sense it is not correct. The correct equation should contain the projection matrix of the camera or, equivalently, the camera rotation and translation. Moreover, it is not clear what the weighted position of the object is. The equations 4.3 and 4.4 have slightly different meaning than what is described in the thesis. E.g, from Equation 4.4 it follows that Z = 1, which is not true if Z is the position of the the object in the camera coordinate system.

2. The used radial distortion model is not described. Fig 4.3 is redundant. A more interesting and useful image would be an undistorted image; i.e. an image showing a result after calibration and undistortion of the calibration chessboard.

3. Fig 4.1 does not show a situation from the proposed problem (this figure was downloaded from the Internet). A more useful figure would be a figure showing the used coordinate systems, e.g. in the text the orientation of the coordinate system is changing and therefore it's not completely clear what are the X,Y,Z axes in the experiments.

4. Coordinate systems should be correctly defined at the beginning, and the coordinates w.r.t. the different coordinate systems should be clearly distinguished (e.g. using subscripts). Moreover the student should use consistent notation, e.g., Equation 4.12 contains homogeneous coordinates $[x', y', z', 1]$ and Equation 4.3 inhomogenous coordinates $[X, Y, Z]$ for the coordinates of the object in the camera coordinate system.

5. It is not clear why the equations for $v_X(k)$ and $v_Z(k)$ are the same in Equation 5.2.

6. The method for filtering out false positives from detections is not sufficiently described. Moreover, it would be useful to see some results of this method. It is not

clear what the description in Figure 5.7. means - ”these should be also filtered out by the described filter”. Are these detections filtered out or not? Moreover, by using the proposed filtering method, it is not guaranteed that the best detection will remain and will be used for relative pose estimation. Since the proposed relative pose estimation algorithm may be very sensitive to the precision of bounding box detections, a method that uses all possible candidate bounding boxes (e.g. with some ”trust” weights) for the relative pose computation would be more robust.

7. It is not clear why the YOLO detector produces less accurate bounding boxes on the embedded GPU than on a standard PC. I would expect a decrease in detection speed but but a similar detection accuracy. Is the decreased accuracy caused by the detector latency?

In summary, the topic of the thesis is of importance in the field; however, the goals of the thesis were met only partially. I recommend the thesis for defense and propose the grade of **D (satisfactory)**.

3. 6. 2018

RNDr. Zuzana Kúkelová, PhD
Czech Technical University in Prague,
Faculty of Electrical Engineering