20 January, 2020
Québec

Ing. Vladimír Kubelka, Ph.D.

Master's thesis review / Posudek oponenta

Author of the reviewed thesis: **Varun Burde**
Thesis title: **Deep neural network for city mapping using Google Street View data**

The aim of the master's thesis of Varun Burde was to design and test a new software pipeline that would exploit state-of-the-art image classifiers to analyze Google Street View images from a given area selected by the user. The detected objects found in the images would be later inserted into a digital map as additional points of interest. The motivation for this work is the automatic enhancement of existing maps by the means of machine learning algorithms.

I appreciate the effort put into re-implementing various components necessary for the whole pipeline, especially the state-of-the-art deep neural networks (NN) utilized for the image analysis. The thesis contains a brief overview of available NN architectures both for object detection and classification and for the depth reconstruction. The author selected the best-performing ones. These main building blocks were integrated with the Street View and other map components to fulfill the thesis goal. The user interface was implemented using the Google Colab environment.

However, regarding the technical part, I see several issues that need clarification (these are also my questions). The author reports the problem of obtaining detections of a single object from multiple views. That is actually expected, the difficulty is to match these together into a single point added to the map in the end. Clustering using the k-means algorithm was not satisfactory and it was not used in the final implementation. How was the problem resolved then? Are all the detections just added to the map including the duplicates? There is a possible resolution proposed in the future work which would utilize a tool from Google to be able to download images in sequence along a street that would allow searching for the related detections. Is that necessary? The downloaded images contain information about their location, why it is not possible just to sample the area as proposed in the thesis and then process the images in sequence according to their coordinates? Since the depth estimation is not always perfect, this would also allow better localization instead of using fixed thresholds.

My other comment is about the way the location (longitude and latitude) of each detection is computed from the pixel coordinates. The proposed formula (6.2) is very approximative, the left-right displacement w.r.t. the camera depends on the distance, not only the pixel

coordinates. How does the author justify this approximation? Is the allowed distance so constrained that the dependency does not matter in the formula (6.2)? There was a proposal in Sec. 6.14 to resolve this problem by finding an explicit mapping between the pixel coordinates and the world coordinates, but this was later rejected as unsatisfactory. How large the position errors actually are? I am missing a clear comparison or an experiment.

The presentation of the work done is, unfortunately, the weaker part of the thesis. The structure of the thesis is fine up to the not-so-logical split of the pipeline description between chapters 5 and 6, these two parts could have been just one chapter, perhaps with a dedicated section for discussion of the encountered problems. The text is however difficult to read, with whole sentences that do not make sense. For example, the abstract contains statements as: “A bar graph to visualize the number of detection per class.” That is not a sentence. Sentences “*The author projected a general framework for classifying the practicality of individual buildings.*” or “*Network Architecture: Mask R-CNN have multiple architectures) Convolution backbone architecture used for feature extraction over an entire image) network head for bounding box recognition (classification and regression).*”(sic) also do not help the reader to understand the message. Several figures lack an explanation of the units used (Figs. 3.8, 3.9: what is the x-axis?, Fig. 6.4: what is the indicated depth, meters?). Abbreviations should be explained when used for the first time in the text, e.g. “CNN” and “GSV” in Sec. 1.1. What is “Overpass API”? It was used in the work but not explained properly. There are many more issues like these and they degrade the thesis. I do acknowledge that the thesis was written in English but I propose to the author to use more proofreading next time to avoid this unnecessary problem.

In conclusion, the goal of the thesis was generally satisfied except for the last point from the assignment guidelines (the comparison of the whole pipeline to the state-of-the-art approaches is missing). The pipeline accepts the input from the user, downloads and analyses the related Google Street View images, and finally puts obtained detections into a digital map. The presented solution would benefit from a better way of dealing with multiple detections of a single object and with a failing depth estimation. The formal part of the thesis is the weakest point. As the final mark for the thesis, I propose “**GOOD**” (**C**).

In Québec, 20 January 2020

Vladimír Kubelka