ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

# REVIEWER'S OPINION OF FINAL THESIS

## I. IDENTIFICATION DATA

| | |
|---|---|
| **Thesis name:** | **Visual Localization with HoloLens** |
| **Author's name:** | **Pavel Lučivňák** |
| **Type of thesis :** | master |
| **Faculty/Institute:** | Faculty of Electrical Engineering (FEE) |
| **Department:** | Department of Computer Science |
| **Thesis reviewer:** | RNDr. Zuzana Kúkelová PhD. |
| **Reviewer's department:** | Department of Cybernetics |

## II. EVALUATION OF INDIVIDUAL CRITERIA

**Assignment** **challenging**

*Evaluation of thesis difficulty of assignment.*

The topic of Pavel Lučivňák thesis is indoor visual localization using HoloLens and sequences of images. This is an interesting research topic that has attracted a lot of attention in recent years [1,2]. Indoor visual localization is a challenging task with many open problems and a large space for improvement as the majority of prior work has focussed on outdoor settings, resulting in pipelines optimized for outdoor but not indoor scenarios.

The goal of the thesis was to review state of the art in indoor visual localization, especially [1,2], adjust the method [1] to a new local environment, and perform image acquisition using HoloLens. For this purpose, Pavel was supposed to create a new 3D dataset for the local environment and evaluate the accuracy of the localization w.r.t. a ground truth in that environment.

The next goal was to evaluate the InLoc method [1] on the newly created dataset and then investigate the possibility of using multiple images (in the form of image sequences) and the HoloLens data for improving the localization. Finally, the performance of the newly developed algorithm was supposed to be evaluated.

**Satisfaction of assignment** **fulfilled with minor objections**

*Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.*

The thesis fulfills all given tasks with mostly minor objections.

1) Pavel Lučivňák reviewed state of the art for indoor visual localization methods.
2) Pavel created a new InLocCIIRC dataset consisting of two rooms scanned at CIIRC using a Matterport scanner. The dataset contains two sets of query images taken by a smartphone camera (Samsung Galaxy S10) and HoloLens. To generate the reference pose for each query image, a Vicon system was used.
   The newly generated dataset captures interesting scenes for indoor localization and Pavel Lučivňák did a large amount of work to acquire the data and estimate reference poses. However, there are still some issues with this dataset. These issues are mostly connected to the precision of the computed reference poses and calibrations. To obtain precise reference poses, cameras have to be precisely calibrated, several coordinate systems have to be aligned and Vicon and HoloLens have to be synchronized. In the thesis, most of these issues were solved manually or using brute-force search, seemingly ignoring prior practices. Therefore, it is not fully clear how good these reference poses and calibrations are. Moreover, the used approach might not be easily applicable to new data.
3) Pavel adapted the InLoc approach [1] to the newly recorded dataset.
4) He suggested two approaches to handle sequential data and evaluated them in the adapted InLoc pipeline on the new dataset. While using sequential information shows promising results, the proposed approaches are currently limited to very short sequences due to sub-optimal design decisions. The thesis discussed potential modifications that could improve the results, but does not evaluate them experimentally, likely due to a lack of time.

**Method of conception** **partially applicable**

*Assess that student has chosen correct approach or solution methods.*

Overall, the methodology used in the thesis is adequate: the approach used to generate the new dataset handles the main challenges, i.e., camera calibration, synchronization between the query images and the Vicon poses, and the alignment between the query / Vicon poses and the coordinate system used by the Matterport data. The two approaches proposed to handle image sequences are technically sound and complementary in the sense that the first one uses additional poses for verification while the second one uses multiple images for pose estimation in the first place. Modeling image sequences as generalized cameras, as done in the second approach, is a good choice.
However, some choices seem rather ad-hoc and not principled:

- Intrinsic camera parameters are estimated through manual inspection and it is unclear whether radial distortion is modeled (which typically has an impact on the pose accuracy). It is common practice to use existing calibration toolboxes, e.g., the Matlab calibration toolbox or Kalibr, to obtain accurate calibrations.
- Rather than brute-force search over a discretized set of parameters, one could use non-linear optimization, e.g., via the Levenberg-Marquardt algorithm, to optimize over a continuous range of parameters. In particular, one could optimize over all parameters (coordinate system alignments and synchronization), e.g., following the approach described in the supplementary material of Schops et al., CVPR 2019.
- An alternative to manually annotating matches is to align the Matterport cutouts and the query images using Structure-from-Motion. Here, using all query images for the alignment typically stabilizes the process and helps recover poses for images that cannot be directly matched against the Matterport data. One such approach is described in Schops et al., CVPR 2017 (in the context of highly accurate registration of images against laser scans). This approach would naturally benefit from the HoloLens poses.
- The thesis states that the MultiCameraPose approach has to deal with a combinatorial explosion (each query image has 10 retrieved images, resulting in $10^k$ potential combinations for a sequence of length k). However, this ignores that not all combinations are plausible, e.g., if one query image had retrieved a cutout from scan position A, then using a cutout from another scan position far away from A for the next image in the sequence is physically implausible. In addition, many of the retrieved cutouts could come from the same scan position due to the small size of the dataset, in which case matches can be combined. As such, the valid number of combinations should be significantly lower than $10^k$. An alternative would be to use all 2D-3D matches from all retrieved images for the pose estimation stage to circumvent the problem. None of these approaches is discussed in the thesis.

**Technical level** **C - good.**

*Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.*

Pavel Lučivňák demonstrates an understanding of different steps and issues connected to dataset generation, e.g., necessary transformations between different coordinate systems, camera calibration, synchronization between the query images and the Vicon poses etc. However, some parts of the dataset generation process could have been solved better using established methods from the literature, especially the intrinsic calibration step and the use of continuous optimization of all parameters together (coordinate systems and synchronization), and a more rigorous mathematical formulation of the alignment process. The technical description of the coordinate systems used and the relation between the coordinate systems could have been explained more clearly and in more detail.
The thesis clearly reflects the knowledge gained by Pavel through his work on the topic. This is especially apparent in the parts of the thesis that discuss current shortcomings and propose alternatives, which he was mostly due to the lack of time not able to test. However, some of these alternatives seem easy to evaluate, e.g., scoring the camera poses for multi-image queries based on median or maximum values or integrating a P3P solver in the software used for the MultiCameraPose approach to handle the case where all matches are found for a single image. While the InLoc pipeline is not efficient, ablation studies could have been performed on a subset of the queries to reduce experimentation runtime.

**Formal and language level, scope of thesis** **C - good.**

*Assess correctness of usage of formal notation. Assess typographical and language arrangement of thesis.*

The thesis is structured into six chapters that correspond to the different goals of the thesis (Literature overview, Dataset, Implementation, and Experiments). The literature overview is divided into unnecessarily many short sections corresponding to individual papers and it contains some sections that are not satisfactory motivated or connected to other sections, e.g. 2.6 Procrustes analysis.

Sections 3 and 4 contain some details that are not particularly important, e.g., Section 3.2.1 Usage. On the other hand, these sections would benefit from a better and more detailed explanation of the proposed solutions and issues. The clarity could be improved by summarizing the proposed sequential methods and the dataset creation, e.g., listing all steps of the data reference pose estimation process in one section, before going into technical details. In general, the clarity of the thesis could be improved by focusing more on the motivation behind the design choice and the underlying technical and mathematical concepts and less on implementation details.
The English used in this thesis is satisfactory.

**Selection of sources, citation correctness** **B - very good.**

*Present your opinion to student's activity when obtaining and using study materials for thesis creation. Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.*

The references are satisfactory. However, I would appreciate a more detailed description of the state-of-the-art methods. Some sections in the Literature review chapter are not really a literature review and are more Technical Background – e.g., 2.4. Camera coordinate system or 2.6 Procrustes analysis. Furthermore, a sub-section discussing previous work on dataset generation for visual localization is missing, e.g., the work by Sattler et al., CVPR 2018 on building outdoor localization benchmarks, and the references listed above.

**Additional commentary and evaluation**

*Present your opinion to achieved primary goals of thesis, e.g. level of theoretical results, level and functionality of technical or software conception, publication performance, experimental dexterity etc.*

Additional questions:

1) Is radial distortion taken into account when calibrating the cameras? This seems to be especially important for the wide-angle camera of the smartphone.
2) For the initialization of the query poses in the Matterport coordinate system, why not use correspondences obtained via feature matching between the queries and the panoramas or by solving a Structure-from-Motion problem over all queries and cutouts?
3) Why not also capture sequence data for the smartphone?
4) Why not calibrate the S10 and HoloLens cameras using an existing calibration toolbox instead of manual calibration via a tripod and ruler?
5) Any suggestions on how to automate the synchronization process, e.g., by automatically finding the synchronization constant?
6) Why is it necessary to pad the query images? The difference in the aspect ratio is caused by different fields of view but does not lead to deformations of the image content. At the same time, the neural network is fully convolutional and should be able to deal with arbitrary resolutions.

Additional comments:

1) Section 2.4 explains camera calibration, but not really the local camera coordinate system.

**III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION**

*Summarize thesis aspects that swayed your final evaluation. Please present apt questions which student should answer during defense.*

In summary, the thesis fulfills the stated goals with mostly minor objections. The topic of the thesis is of importance to the field; the goals of the thesis were met. A new indoor localization dataset and two approaches to handle image sequences in localization were proposed. However, some of the choices used in the dataset creation process and for the proposed localization methods seem rather ad-hoc and not principled, and there is still a large space for improvement. The text of the thesis could also be improved (see the review). I recommend the thesis for defense and propose the grade of C (good).

I evaluate handed thesis with classification grade **C - good.**

Date: **29.8.2020** Signature