THESIS REVIEWER'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Visual Localization in Dynamic Environments**
Author's name:	**Martina Dubeňová**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Cybernetics
Thesis reviewer:	RNDr. Zuzana Kúkelová PhD.
Reviewer's department:	Department of Cybernetic

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

How demanding was the assigned project?

The topic of Martina Dubeňová's master thesis is visual localization in dynamic environments, i.e., estimating the 6-DoF pose of a query image in a scene containing moving objects. This is an interesting and challenging problem with many applications, e.g., self-driving cars, mobile robots, and augmented reality. In all these applications, the scene contains many dynamic objects and is permanently changing. While some smaller changes in the scene may not affect the quality of the pose estimate, larger changes or changes in environments where most of the features are on potentially moving objects (e.g., chairs) may significantly deteriorate the pose estimates. Therefore, it is important to develop visual localization methods that are robust to changes in the scene. The goal of the thesis is to investigate the influence of scene changes on the accuracy of localization and suggest improvements of the InLoc localization method that will provide robustness against moving objects. Moreover, the thesis is supposed to create a new dataset with dynamic objects moving in the environment and test the suggested improved InLoc method on this new dataset.
The thesis is part of the EU Horizon 2020 project SPRING [3]. The goal of SPRING is to develop socially assistive robots capable of moving, hearing, and communicating in complex and unstructured public places. The SPRING project will be tested at Broca Hospital (a gerontology hospital in Paris). Therefore the localization and datasets in the thesis focus on the home and medical environment.

Fulfilment of assignment **fulfilled with minor objections**

How well does the thesis fulfil the assigned task? Have the primary goals been achieved? Which assigned tasks have been incompletely covered, and which parts of the thesis are overextended? Justify your answer.

The thesis fulfills all four given tasks with minor objections.

1. Review the state-of-the-art in indoor visual localization.

   The student briefly reviewed state-of-the-art visual localization methods and described in more detail the InLoc visual localization pipeline [2] and the master thesis [7] that is extending [2] to localization of sequences of images and data acquired with Hololens.
   While the review of the state-of-the-art localization methods is sufficient, I would also appreciate a review of methods dealing with dynamic scenes.
   The suggested improved localization method that works on dynamic scenes is based on masking dynamic objects. In practice, this will require the detection and segmentation of such objects. There are many works dealing with this topic, e.g.

   [Zhou] Zhou et. al. Dynamic Objects Segmentation for Visual Localization in Urban Environments, IROS 2018 Workshop "From Freezing to Jostling Robots: Current Challenges and New Paradigms for Safe Robot Navigation in Dense Crowds", 2018

Moreover, there are many methods, e.g., in SLAM, that use a similar idea, i.e., detecting and filtering /masking dynamic/moving objects to improve the performance, e.g.

[Vincent] Vincent et. al, Dynamic Object Tracking and Masking for Visual SLAM, IROS 2020
[Wu] Wu et. al. OC-SLAM: Steadily Tracking and Mapping in Dynamic Environments, Front. Energy Res., 2021
[Sun] Y. Sun, M. Liu, and M. Q.-H. Meng, "Improving RGB-D SLAM in dynamic environments: A motion removal approach," Robotics and Autonomous Systems, vol. 89, pp. 110–122, 2017

2. Adjust method [2]/[7] to home/medical environments. Create a new data set for home/hospital environments from Matterport scans and simulate segmentation of dynamic objects moving in the environment.

   In general, this task was fulfilled.
   The student provided an implementation of an automatic method for creating localization datasets from Matterport scans that can be directly used by the InLoc localization pipeline [2, 7]. This method is explained in detail in the thesis. This part is maybe even overextended, with unnecessary technical details like the time after which the link for downloading images expires.
   For the segmentation of dynamic objects, the student decided to use manual segmentation instead of an automatic method such as [38]. Such manually segmented objects were loaded and placed in new locations in the scene. While manual segmentation is not a good solution for practical applications, for creating datasets, it is sufficient, and it allows to place objects present in the original scene in new positions without the need to do new scans. In this way, one "static" and two dynamic datasets of two scenes (the Hospital and the Living Lab scene) were created.
   Unfortunately, for some reason (that was not explained in the thesis), the texture maps for the segmented objects were lost. Therefore, the "dynamic objects" in the generated datasets did not have any color, and the default white material was applied. This negatively affected the rest of the thesis (testing of the suggested modified localization method) significantly. Moreover, the newly generated datasets are not useful for most of the applications because of the missing texture.

3. Investigate the influence of scene changes on the accuracy of localization of the method from 2) and suggest improvements providing robustness against moving objects.

   This task was fulfilled partially
   The influence of the scene changes on localization accuracy was tested only on the new datasets. It would have been good to also test this influence on standard datasets in which query images were taken in different times than the database (e.g., just by running detectors of chairs/cars, etc., and running standard and suggested localization pipelines with masked objects.)

   The suggested improvement of InLoc based on masking/filtering dynamic objects is reasonable and will most likely provide good robustness against moving objects (I am writing most likely since the proposed method was due to the missing texture not properly evaluated). While this improvement is not fully new, e.g., as mentioned above, it was used in SLAM methods and suggested for visual localization in [Zhou], to the best of my knowledge, it was not explicitly described and evaluated in the context of indoor visual localization. Therefore, the suggested solution would have been a good contribution if properly evaluated.

4. Demonstrate and evaluate the improved method on the new data set.

   This task was again fulfilled partially.
   The suggested method with filtering/masking of dynamic objects as well as the original InLoc method were evaluated on the new datasets. However, as already mentioned, due to the missing texture on dynamic objects, the evaluation was not very useful and was not really demonstrating an improved performance of the suggested method over the standard InLoc method. The student correctly identified the problem in the evaluation: The dynamic objects used in the new datasets were white and did not have any texture. First of all, such untextured objects will have a very low number of detected features. Second, in the original scene (3D map w.r.t. which the

localization is done), these objects have a texture, and therefore, almost no correspondences were found between white projections of dynamic objects in the query images and instances of these objects in the original 3D map. Therefore, there were "no correspondences to filter", and the original method without filtering was performing almost equally to the proposed method with filtering on "dynamic scenes". There was a very small difference on the "Broca_dataset_dynamic_1" dataset; however, this can be a result of some random correspondences on untextured objects or due to the randomness inside P3P Ransac in the InLoc pipeline.

Methodology **correct**

Comment on the correctness of the approach and/or the solution methods.

The thesis can be divided into three parts

1. Automatic generation of datasets for localization from a Matterport scanner

   In this case, the student provided an implementation of an automatic method for creating localization datasets from Matterport scans that can be directly used, e.g., in the InLoc localization pipeline.
   The methodology was selected properly, and the provided method seems to work. For this part, understanding different file formats, camera geometry, and transformations between different coordinate systems, as well as understanding the Matterport API and the AI Habitat framework, was necessary.

2. Developing a method for the creation of datasets with dynamic (moving) objects

   As mentioned above, the segmentation of dynamic objects was done manually instead of using an automatic method such as [38]. The manually segmented objects were loaded and placed in new locations in the scene. The query images with new objects placed in the scene were rendered using AI Habitat, which can also return segmentations and depth maps.
   While manual segmentation is not a good solution for practical applications, it is a reasonable choice for creating datasets. It allows placing objects present in the original scene in new positions without the need to do further scans.

   Unfortunately, in this case, the texture maps for the segmented objects were lost, and therefore the "dynamic objects" in the generated datasets do have just white material applied. This negatively affected the rest of the thesis (e.g., the evaluation of the proposed localization method). Moreover, the newly generated datasets are not really practically useful because of the missing texture.

   It is unclear what the reason for the missing texture was and whether it was somehow possible to solve it. I would assume that it was only a technical issue.
   If it was not possible to solve this issue, the student could have considered another way of generating such datasets. For example, one possibility is to place some existing textured 3D models, e.g., models of chairs, in the scene. Such chairs would not have been present in the original 3D map and database images. However, using a similar approach to generating query images, i.e., rendering images together with depth maps using AI Habitat, it would have been possible to generate new database images containing these "dynamic objects" in different places than in the query images.
   Another possibility was to acquire at least one more Matterport scan of the scene with some objects moved. However, I am not sure if this was technically possible. It is not clear if the student was present in the hospital during the scanning and had access to the Matterport scanner or if the scans were done by other partners in the EU project SPRING.

3. Modifying the InLoc localization pipeline to work in dynamic environments

   Even though the suggested improvement of the InLoc pipeline, which is based on masking/filtering dynamic objects is not fully novel (see above), it is a reasonable solution that will most likely provide good robustness

against moving objects. Unfortunately, due to missing textures on dynamic objects, this was not really demonstrated in the thesis.
A similar approach is, e.g., used in SLAM methods [Vincent]; however, to the best of my knowledge, in the context of indoor visual localization, it was not explicitly described and evaluated in the literature.

Technical level **C - good.**

Is the thesis technically sound? How well did the student employ expertise in the field of his/her field of study? Does the student explain clearly what he/she has done?

The thesis is technically sound. The student demonstrated a good understanding of different parts of localization pipelines, camera models, different file formats, and tools for processing data. On the other hand, even after figuring out the problem with untextured objects and their effect on the tested localization pipeline, the student did not suggest an alternative solution to generate datasets or to use some existing datasets on which the method could be properly evaluated.

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

Are formalisms and notations used properly? Is the thesis organized in a logical way? Is the thesis sufficiently extensive? Is the thesis well-presented? Is the language clear and understandable? Is the English satisfactory?

The level of English in the thesis is good. The thesis is organized in a logical way, and used notations and formalisms are used properly.
On the other hand, I think that the presentation of the thesis can be slightly improved.
Some parts of the thesis are unnecessarily detailed, e.g., there are a lot of technical and implementation details about processing the data (like the time after which the link for downloading images expires, the structure of the folders storing the data, etc.). I think some of these details can be a part of the code documentation and do not need to be mentioned in the thesis. On the other hand, some other parts lack details, e.g., experiments are not sufficiently described – e.g., it is not clear what "Correctly localized queries" means, i.e., how was this measured.

Selection of sources, citation correctness **C - good.**

Does the thesis make adequate reference to earlier work on the topic? Was the selection of sources adequate? Is the student's original work clearly distinguished from earlier work in the field? Do the bibliographic citations meet the standards?

The references are satisfactory. Even though the discussion of the state-of-the-art on visual localization is short, it covers the most important works. However, as mentioned above, I'm missing a discussion on methods for detecting and segmenting dynamic objects as well as methods that are using a similar method as the one suggested in the thesis, i.e., dynamic object filtering/masking to improve performance of, e.g., SLAM systems.

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS, SUGGESTED GRADE

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

THESIS REVIEWER'S REPORT

Summarize your opinion on the thesis and explain your final grading. Pose questions that should be answered during the presentation and defense of the student's work.

The thesis fulfills all its stated goals with minor objections. The student demonstrates a good understanding of different parts of localization pipelines, camera models, different file formats, and tools for processing data. Unfortunately, due to, most likely, technical issues with missing textures on dynamic objects, the main goals of the thesis, i.e., the generation of datasets with dynamic objects and the evaluation of the suggested improvements of the InLoc pipeline on dynamic scenes, were fulfilled only partial. The proposed datasets in the current version, as well as the results of the experiments with the modified InLoc method, are not very useful. After figuring out these technical problems, the student could have suggested and tried a different solution for generating dynamic datasets for testing. Another possibility was to at least test the method on existing datasets (e.g., the original InLoc dataset contains query images that were taken at a different time as database images, and as such they most likely contain many changes and dynamic object. Such objects could have been detected and filtered to test the proposed approach for localization). Still, the thesis presents a useful automatic method for generating datasets from Matterport scans that can be directly used, e.g., in the InLoc localization pipeline. Moreover, after solving technical issues with textures and properly evaluating the suggested localization method, the work can present some interesting contributions. In summary, the topic of the thesis is important to the field; the thesis goals were partially met, and a useful automatic method for generating datasets from Matterport scans was proposed. I recommend the thesis for defense and propose a grade of C (good).

Additional comments and questions:

1. What was the reason for the missing textures on the segmented objects?
2. In the Conclusion, you mentioned that you started working on a new method for comparing query images with synthesized images in the last step of InLoc (pose verification). Can you say more about this method?

The grade that I award for the thesis is **C - good.**

Date: **31.5.2022** Signature: