THESIS REVIEWER'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Václav Vávra**
Author's name:	**Matching pohledů na scény s planárními povrchy**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Computer Science
Thesis reviewer:	Torsten Sattler
Reviewer's department:	Czech Institute of Informatics, Robotics and Cybernetics

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

How demanding was the assigned project?

Matching local features between two images taken from very different viewpoints is an important component in many 3D computer vision applications, including 3D reconstruction, Structure-from-Motion, and visual localization. An established approach is to detect planar regions and unwarp the corresponding image regions before feature detection to remove the effect of perspective distortions. Recently, Toft et al. proposed one such approach that uses single-view depth predictions to identify planar regions. The goals of this project were to re-implement the approach by Toft et al. (as not all parts are publicly available), analyze its components, develop an improved version of the method, and to evaluate the re-implementation and its improved version through detailed experiments. Given the complexity of replicating published results, the challenge of understanding an existing approach to the level of detail that it becomes possible to extend it, and the challenge of developing such an extension, the assigned project is suitably challenging for a master thesis.

Fulfilment of assignment **fulfilled**

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 fulfils all individual parts of the assigned task: the thesis presents the re-implemented version of the approach by Toft et al., discussing the main design choices and validating them through experimental results on the dataset published by Toft et al. The thesis proposes improvements of the original approach on two levels: the first level consists of changing smaller components of the method (using weighted instead of unweighted SVD for normal estimation, normal estimation based on binning the space of surface normals rather than using k-means clustering, finding connected components) and introducing filtering steps (filtering pixels based on the planarity of the corresponding regions, filtering out normal clusters). For these smaller modifications, the thesis discusses multiple choices for the components and selects the best-performing combination based on detailed experiments. The second level replaces the way planar regions are unwarped used by Toft et al. (based on homographies) with an approach by Rodriguez et al. based on affine transformations given by a convolutional neural network for detecting affine features. The thesis further presents detailed experiments using the re-implementation and the proposed variations with multiple types of local features, showing improvements over the method by Toft et al. (even though not every modification helps to improve performance).

Methodology **correct**

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

The chosen methodology is appropriate for the project and correct: detailed ablation studies, i.e., measuring the impact of individual components and design choices on the overall performance of an algorithm, is a standard approach in the computer vision and machine learning literature. The thesis does a very good job of validating the individual components and especially the proposed modifications through extensive experiments, including parameter tuning. I especially appreciate that the thesis evaluates different types of local features, thus demonstrating that the re-implementation with the proposed modifications is widely applicable. The proposed modifications are technically sound and combining the methods from Toft et al. and Rodriguez et al. is an interesting idea. It is quite interesting to see that using affine transformations rather than homographies improves results, even under strong viewpoint changes. This is to some degree

surprising as affine transformations are a subgroup of homographies and not able to completely remove perspective distortion.
My main point of criticism is the evaluation measure used in Chapter 2. Since the end-goal is to improve feature matching, why not evaluate the modification and components under this aspect rather than using a metric that measures how well the dominant planes are recovered? Given that the thesis gives one example where the performance on one task does not translate to performance on the other, it seems that evaluating matching quality (as done in Chapter 3) would have been more appropriate to ensure that the choice of parameters is indeed the best choice.

Technical level **B - very 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. It is clear that the student is very familiar with the techniques employed in the thesis (local features, benchmarking algorithms, detecting planes from depth maps, removing perspective distortion). As such, the student employed his expertise in the field very well. There are some minor issues, e.g., not all combinations are tested in Figure 4.3, which makes it hard to draw conclusions about design choices, but these do not take away from the contributions of the thesis.

Formal and language level, scope of thesis **C - 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?

Overall, the organization and structure of the thesis can be improved:

- Key concepts are not defined: what exactly is a patch (the word is used quite often, but it is unclear to me what it means: is it a region around a keypoint, a connected component of pixels belonging to the same plane, …)? What are r-balls? What is meant by sparse and dense covering? What are affine normalizing maps? What is a naïve covering?
- There are inconsistencies in the mathematical definitions, e.g., the variable psi does not occur in the set in Eq. 3.7.
- Evaluating individual components as part of the method description in Chapter 2 rather than in a separate section is a good idea that helps the reader to see which components and which modifications are important. Unfortunately, it is very hard to understand the results as the evaluation metric and experimental setup is only described towards the end of Chapter 2 while the dataset used for evaluation is described in Chapter 4.
- The meaning of the y-axis of the plots in Chapter 2 is not fully clear to me: evaluation is performed over many pairs of images, but how are the individual results aggregated into the results shown in the figures? Is a simple mean used?
- While Chapter 2 clearly describes what was done, I found it much harder to follow Chapter 3. The chapter lacks a higher-level overview over the approach from Rodriguez et al. and instead directly presents the method in great detail. Having a higher-level overview typically helps to follow technical details more easily.
- Chapter 4.5 seems incomplete: there is no reference for the EDV dataset and results are never discussed (there is a comment to "Explain" them in Tab. 4.2 though).
- It would have been good to motivate some of the design choices in the thesis, e.g.: why not use the MonoDepth2 network used by Toft et al.? How were the hyperparameters used in the thesis chosen? Why use a uniform kernel and not a Gaussian or some other kernel that gives less weight for far away data for mean shift clustering? Why use only a single scene for training and not multiple ones?
- There is no discussion of potential future research directions based on the results of the thesis.

There are multiple minor points such as typos, inconsistent notation, variables that are not introduced when they are used for the first time, and missing references for RANSAC, BRISK, RootSIFT, and mean shift clustering. I can provide details directly to the author.
The level of English is satisfactory.

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?
As detailed in Toft et al., there is quite some classical work on removing perspective distortion before feature extraction. The thesis briefly mentions some of that work in a few sentences without providing references. Discussing the thesis in the context of this prior work would help the reader to better understand the contributions of the thesis and how they relate to existing methods. This could have been done either in the introduction or a separate related work section. Otherwise, the thesis makes adequate reference to prior work. The bibliographic citations meet the standards with minor issues (citing non-peer-reviewed arXiv versions instead of peer-reviewed conference versions, missing venues for citations).

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

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.

Overall, this is a good thesis. I very much like the detailed ablation studies presented in the thesis as they help to understand which parts of the method of Toft et al. are crucial for good performance. The proposed modifications are technically sound and sensible and the experimental results show that they can help to improve performance. The result that affine transformations can lead to better results than homographies is interesting and a bit surprising at first glance. I can imagine that this insight will inspire future work to use simpler rectification pipelines. I would have suggested a better grade if not for the quality of the report as I found the thesis in parts very hard to read due to missing information.

Questions for the presentation and the defense:

1. How is a patch defined?
2. Is there an intuitive explanation for why affine warps give similar or better results than homographies?
3. Based on your experience with your method and your results, what are promising directions for future research?

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

Date: **19.1.2022** Signature: