ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

REVIEWER'S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Accuracy of 3D body pose and shape estimation of infants from RGB and RGB-D data**
Author's name:	**Vojtěch Volprecht**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Computer Science
Thesis reviewer:	Dr.-Ing. Nikolas Hesse
Reviewer's department:	Swiss Children's Rehab, University Children's Hospital Zurich

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **extraordinarily challenging**

Evaluation of thesis difficulty of assignment.

The thesis assignment includes many different aspects, like hardware setup (different RGB-D cameras, motion capture system, laboratory), different experiment setups (objects, doll, real infants), and the implementation of pipelines to evaluate multiple state-of-the-art methods for human pose estimation. This requires many different skills and poses significant challenges.

Satisfaction of assignment **fulfilled**

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 majority of assignment items have been very well fulfilled and extensive experiments were carried out.
Only for item 6, the estimation of complete body meshes with respect to point clouds or with respect to pose estimation performance could have been more extensive.

Method of conception **correct**

Assess that student has chosen correct approach or solution methods.

Big part of the thesis was to set up a highly complex recording system (including multiple RGB-D cameras and a motion capture system) and a pipeline for the evaluation of pose estimation methods. This has been very well executed.

However, there are some shortcomings in the evaluation of the pose estimation methods. The most important question that this thesis should have answered (in my opinion), is if these cameras and methods can capture infant movements at a quality so that they can be used for further (clinical) motion analysis, and which of the methods and cameras are best suited for that task, and what their shortcomings are.

The evaluation metrics do not reflect this very well. The Euclidean distance to the origin, either using only XY or Z (depth) coordinates, does not really show if the movement characteristics of infants over time can be captured. The plots could have provided more information, but are showing mostly static joints like hips and shoulders. Joint-wise results are only given as average values over all data. (Also, the 3D joint and the 3D mesh estimation methods should have been analyzed in more detail, since these are the most promising candidates for high quality motion capture. But due to the extensive content of the thesis, I understand that there a time constraints.)

The results indicate what kind of noise levels to expect, but they don't give any insights on how important these are for infant motion analysis – the choice of experiments is good, and they are very extensive, but the evaluation metrics should have been chosen differently.

Technical level **A - excellent.**

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

The assignment required knowledge in many different areas. The successful completion of many different experiments

shows that the student was able to fulfill all the requirements.

Formal and language level, scope of thesis **A - excellent.**

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

The language level is good, and the thesis is easy to read.

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 selection of sources is good. The structure of the related work section could be improved, as the mentioned literature seems to be in a relatively arbitrary order. The current work is not set in relation to the literature. Some citations are incomplete ([13], [20], [22], [38], [39]).

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.

First of all, I would like to acknowledge the large amount of work that has obviously been put into this thesis. Following are some more detailed comments and evaluations.

The motivation of the thesis is the (automated) diagnosis of movement disorders like cerebral palsy. The most important aspect of such a system is the accurate capture of an infant's movement characteristics, e.g., the movement complexity and variation over time. Therefore, this should be the focus of the experiments – how well the captured motions correspond the the actual movements of the infants.
A big part of the experiments concentrates on depth accuracy – the first experiments evaluates if measured depth values correspond to "real" values. The second part evaluates the depth accuracy for a synthetic doll – how close a photogrammetric scan can match point clouds captured with the different cameras. It is good to evaluate the technical aspects of the hardware, but I feel like the relevance to the main topic is relatively small.
The third, and most important, part focuses on methods for 2D and 3D keypoint estimation. 2D keypoints are projected to the depth images to compute 3D values. These are then compared to the output of a motion capture system.
The analysis is, however, divided into XY plane and depth (Z) values. Instead of computing the distance of estimated joint positions to the ground truth (Mocap) positions, the Euclidean distance to the origin is presented. The plots show results for (mostly) static joints (hips, shoulders), and only tables contain averaged values for all joints. This does not provide many insights on how well movements over time (or motion patterns) can be captured with the tested systems. In addition to evaluation 3D joint position error, the correlation between estimated movements and ground truth (Mocap) movements over time would have been much more meaningful for the task of analyzing motions of infants.

I'm a bit surprised to see the small amount of usable Mocap data (and that this was termed a "success"). Capturing 3 infants for 20 minutes each gives 60 minutes = 3600 s, of which 146 seconds were usable, which corresponds to 4 % of the data. It would have been interesting to read about the biggest issues that happened and if there are ways to solve them. Nevertheless, this stresses further that other, markerless systems are needed.

The thesis didn't mention if there was any interference between the different systems or if anything was done to avoid it. The Realsense cameras and the motion capture system use an projected pattern in infrared to compute the depth image. It is highly likely that the lights of the Mocap system as well as the passive markers (often overexposed in IR images, since they are highly reflective) interfere with the Realsense cameras. This could have been tested by examining the infrared images, or compare point clouds with and without the marker-based system.

The discussion section mostly summarized the results, and repeated the "best" camera for each experiment, but did not provide more insights as to what the results mean for the analyzing infant motion.

Further comments:

4.1.3 YCB Object Dataset
What's the purpose of this experiment? I assume it is to evaluate if the point clouds captured with the RGB-D cameras correspond to the ground truth shapes. There are multiple variables that influence the quantitative evaluation. Cutting the object's GT point cloud in half does not necessarily reflect the exact same object the camera is seeing due to its perspective. As shown in Fig. 4.11 – 4.13, even though the point clouds are well aligned, there are many GT points that are unmatched (left side and top of the object. Particularly the mustard bottle's shape seems to differ a lot from the GT shape.

Table 4.7 and 4.8: table 4.8 shows that results are highly variable with respect to sensor and model. Merging them and declaring a "winner" doesn't seem to make much sense.

It is reported that Mediapipe "Full" performed better than "Heavy" – where does the difference come from? Giving some insights here would have been valuable.

The Mediapipe 3D results might need some further processing before comparing them to the motion capture results. I assume Mediapipe 3D isn't using any camera calibration, so it cannot be expected that the 3D positions correspond to real metric values, particularly since it was trained on adults, which is why the output is probably more of the size of an adult. This could be checked by computing a rough size from both systems (e.g., head keypoing to ankle keypoint distance), and an alignment of the whole body joints to the mocap joints could have been done.

I would have liked to see an evaluation of 4DHumans results. Of course, proportions are not good, but the positions of body parts/keypoints look very good in the example. (3D keypoints can be extracted from the meshes, and they should be fine (again, not in the same reference coordinate system as the mocap)). This would have been very interesting, and has the potential to be the best of all methods. If the 3D positions of the body are inconsistent, one might consider merging the results with the point cloud data, e.g., by registering the estimated mesh and the point cloud.

In the accuracy discussion it is mentioned that landmarks are often occluded. That's the reason why projecting 2D keypoints onto the depth image is generally not a good idea. The discussion should not only point out problems, but also provide (possible) solutions. Methods like the original Kinect tracking estimate the keypoints inside the body from depth images (Shotton et al. 2011) . Or estimate the body surface (mesh) from the point clouds and regress the body joint locations from that [4].

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.

The topic is important and relevant, the assignment tasks were fulfilled, the experiments were extensive and described in detail. The student seems to have a very good understanding of the assignment.

There still were some aspects that could be improved, mostly the evaluation and interpretation/discussion of results.

Overall, the thesis was well executed.

Questions:

- Why was only such a small portion of Mocap data usable? What were the most common problems? Are there ways to overcome these?
- Did the calibration between Mocap system and RGB-D cameras only include translation or also rotation? If it didn't include rotation, why not? Do you think this could have improved the plane estimation results?
- Why was the distance to the origin chosen as main evaluation metric? Could you think of other metrics that are better suited to determine if a system/method is usable for automated motion analysis (and why)?
- Which system/method is best suited as a basis for an automated system for detecting cerebral palsy from infant movements, and why?

I evaluate handed thesis with classification grade **B - very good.**

Date: **11.6.2024** Signature: