Review of Master's Thesis

Student: Konečný Daniel, Bc.
Title: Self-Supervised Learning for Recognition of Sports Poses in Image (id 24543)
Reviewer: Beran Vítězslav, Ing., Ph.D., DCGM FIT BUT

1. Assignment complexity — **average assignment**

The task requires familiarity with more advanced machine learning methods and, in particular, the preparation of the dataset itself, which is a time-consuming and experience-demanding process.

2. Completeness of assignment requirements — **assignment fulfilled**

Quality training data is a key to properly learn models. The author's focus in iterative refinement is on exploring and refining the selection process of training data. This shows a good understanding of the problem and leads to usable models.

3. Length of technical report — **in usual extent**

4. Presentation level of technical report — **85 p. (B)**

The technical report is logically structured, with all key information and areas of expertise well presented, balanced in scope and presented in a clear, professional and understandable manner. Apart from the information that the "self-supervised approach requires significantly less training data", which is repeated somewhat frequently in the text, the text contains no unnecessary information or ballast. The author, probably inspired by scientific publications, does not provide many implementation details (even in the appendix). It is beneficial for the reader that the technical text does not contain dozens of pages of unnecessary descriptions of the author's scripts and functions; on the other hand, a little more information about the implementation of the solution would improve the reader's idea of the complexity and scope of the technical solution. Minor errors, such as the discrepancy between the figures on page 11 of the cited publication and those in Table 2.2, are isolated.

5. Formal aspects of technical report — **95 p. (A)**

The thesis is written in English. The reviewer is not a native speaker, but the language quality can be assessed as excellent, the text is written in good technical English, without errors and incomprehensible sentence constructions. The typographical level is also excellent.

6. Literature usage — **95 p. (A)**

The selection of literature is more modest in scope, but very relevant. The selection includes both book publications and scientific articles. It is evident from the text that the author has a good understanding of the subject under study and draws appropriately from the literature.

7. Implementation results — **90 p. (A)**

The core implementation results are a set of support tools for efficient dataset creation in a given context, the dataset with two sets of annotations, and two trained models.
The tools for dataset creation exploit configuration information at the time of acquisition (capturing multiple cameras from different viewpoints) and temporal continuity, the so-called time-contrastive approach. They actually contain a proposed algorithm for temporal synchronization of videos, semi-automatic cropping of the object of interest, annotation of images, etc.
Two datasets were acquired as part of the solution (hand pose and upper-body pose). The second one is the key one, the size is almost 4 thousand images of upper-body poses and contains two sets of annotations dividing the samples into 4, resp. 16 classes. The final model is based on the ResNet-50 architecture and the implementation is based on the TensorFlow library. Model training includes visualization and discussion of the latent vector distribution of the trained encoder, also a procedure for selecting the appropriate latent vector length, and finally a comparison of the two resulting models learned by the two approaches: supervised and self-supervised.
The source files include a header regarding authorship, but lack source code comments as well as an overall description and explanation of the contents of the source files and their parts.

8. Utilizability of results

The key results are useful for further development and research in the field, as well as for solving similar tasks in other domains. More work is needed to train a model for sports pose classification that would be applicable in a production software. Especially a larger and more diversified dataset is needed.

9. Questions for defence

- Explain better the findings that the visualization of latent vectors has brought.

- What else specifically needs to be done with your result to make it usable in a real user application?

10. Total assessment **90 p. excellent (A)**

Mr. Konecny created a basic dataset for training a statistical model to recognize upper-body poses in images. He experimented with a self-supervised approach for training. At the end of his work, he compares this approach with the classical supervised approach when the training data is small. He presents the procedure and results expertly in his technical report, which is carefully structured and written in very good English. An important part of the solution is the semi-automated tools using video processing methods that can be used for efficient creation of training datasets in similar tasks based on time-contrastive approach. The high professional quality of the solution demonstrates a good knowledge of the studied materials and the given problem.

In Brno 16 August 2022

Beran Vítězslav, Ing., Ph.D.
reviewer