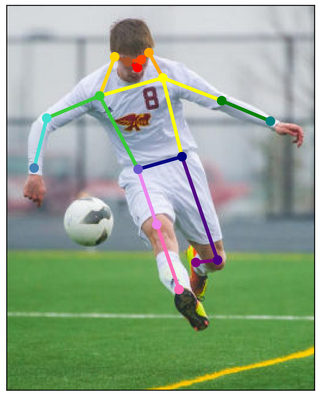
Automatic Human Pose Estimation using Neural Network

Author: Ing. Jakub Straka
Supervisor: Ing. Ivan Gruber, Ph.D. et Ph.D.

FACULTY
OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

What is human pose estimation?

Human pose estimation is a computer vision task that aims to estimate significant keypoints on the human body from image or video. Keypoints are often body joints and significant features in the face. A pose estimation system should be able to robustly estimate all poses in an image and accurately estimate the position of individual keypoints. The estimation should also be made for poses that are only partially in the image.



Motivation

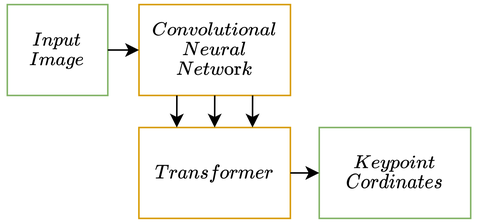
Human pose estimation can be applied in various fields. One of the obvious fields is a sport. Pose estimation can be used for quantification and movement analysis. This can be used for guidance during training. Similarly, pose estimation can be used in healthcare for the diagnosis and quantification of diseases affecting the musculoskeletal system. Another usage is in the entertainment industry. The movement of characters in games and animated films is mostly recorded using motion capture systems which are very expensive and difficult to set up. In some cases, pose estimation can effectively replace these systems.

Thesis goals

- Choose a suitable model for human pose estimation
- Train the model on an appropriate benchmarking dataset and compare results with state-of-the-art models

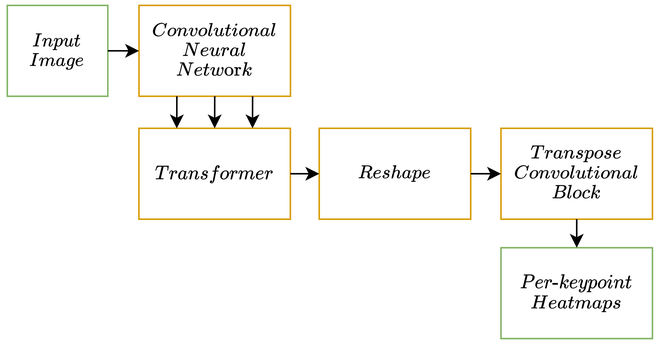
Deformable Pose Transformer

Deformable Pose Transformer (DePOTR) [2] is a model for 3D hand pose estimation. The model is based on transformer architecture and consists of a convolutional neural network followed by a transformer composed of an encoder and decoder stack.



For training the model, the MSCOCO [1] dataset was used, which is baseline for comparing models in computer vision. In order to be able to train the model, it was necessary to adjust the processing of the input images appropriately. Since the original model predicted coordinates in 3D, it was also necessary to reduce the dimension of the coordinates at the output. The output of the DePOTR model are the coordinates of the keypoints. This type of model belongs to the group of regression-based models. The second group are models based on heatmaps. These models produce a matrix for each keypoint, the value on each position in the matrix indicates the probability that a keypoint lies at a given position. Based on this approach, the second variation of the DePOTR model was made.

In order to produce heatmaps at the output additional structure was added to the output of the transformer. First, the output of the transformer is reshaped into the matrix for each keypoint and then transposed convolution is used to increase the spatial dimension of the matrices resulting in heatmaps at the output.



Results

Both variations of the model were trained with different hyperparameters. Among the tuned hyperparameters were input image size, depth of the convolutional network, optimizer, learning rate scheduler and data augmentation. The second variant of the model is labeled DePOTR-HM and the best results of the trained models are compared in the following table with state-of-the-art methods for pose estimation.

Model	AP
SimpleBaseline	73.7
TransPose	**75.0**
DePOTR	54.5
DePOTR-HM	65.0

Conclusion

The thesis focuses on training and modifications of the DePOTR model which was originally designed for 3D hand pose estimation. The model was modified so that it could be used for human pose estimation on the selected dataset. Additionally, the second variation of the model was created. Modifications of this variation were inspired by other successful models that were used to solve the pose estimation task. These modifications of the model led to a significant improvement in the result. All source files are available at: `https://github.com/strakaj/Automatic-Human-Pose-Estimation-using-Neural-Network`

References

[1] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context." In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[2] Hrúz Marek et al. *DePOTR: Deformable Transformer for Hand Pose Estimation*. 2022. URL: `https://github.com/mhruz/POTR`.