

1 Introduction

Video analysis in sports have throughout the last decade become more and more common, as these recordings contain a lot of important information. By analyzing such a video recording of people engaging in sports, we can for instance help a referee make the correct calls or help the people engaged in the sport improve their technique. However, most of these analyses requires the system to know where the relevant people are in the video recordings. The models that perform such a task have already been developed for the most popular sports, such as football or basketball, and tend to deliver very accurate results. On the other hand, for the less popular sports, such as bouldering, such models are not as common

To perform video analysis, machine learning methods are often used, where they can for instance be used for estimating the pose of the people in the video. Often however, these models process the frames of an input video independently of each other, leading to suboptimal results. As the individual frames of a video contain some temporal information that correlates across the frames, one may incorporate this temporal information to improve the performance of the model that otherwise processes the frames independently of each other. Incorporating this temporal information is however not always straight forward and can sometimes require a lot of data, which does not align with the less popular sports, such as bouldering, where annotated data does not come in large quantities. Further, the poses and movements of the people in some of these sports are often very unlike most large public datasets, making these datasets unapplicable for these less popular sports.

Thus, the aim of this thesis is to implement various methods for extending an already developed pose estimator for bouldering, such that it makes use of temporal smoothing. This will be done by developing and testing various machine learning methods through multiple different experiments, such that we end up with the most optimal results.

The thesis is done in collaboration with ClimbAlong at NorthTech ApS. ClimbAlong provides an annotated dataset of video recordings of people climbing a bouldering wall. This dataset is not a synthetic dataset but instead a real dataset, making the developed results with this dataset be more realistic. Secondly, ClimbAlong provides an already developed and trained machine learning model for estimating pose of the humans in 2 dimensions of a given video recording. This model does not make use of temporal smoothing, but instead just processes each frame independently of each other.

1.1 Related Work

2-dimensional articulated human pose estimation can be divided into two groups. The first group focuses on still images, where the pose estimation of multiple images is done independently of each other. The second group focuses on video sequences, where the temporal information across the frames of a video sequence may be used for performing the pose estimation. Further, these pose estimation methods for video sequences may use the pose estimation methods for still images for getting an initial pose estimation of the frames of a video sequence and later use temporal-inclusive methods for smoothing out these still-image pose estimation.

Whether a human pose estimation method is for still images or a video sequence, the method can be categorized in various ways. First off, they can be categorized as being a generative method, which is model based, or a discriminative method, which is not model based. Secondly, they can be categorized as being a top-down method or a bottom-up method, based on which level they start the processing. Thirdly, they can be categorized as being regression-based, where they directly map from input image to the body joint position, or as being

detection-based, where they generate intermediate image patches or heatmaps of joint location. Lastly, they can be categorized as being one-stage, where end-to-end training is used, or multi-stage, where the human pose is estimated in multiple stages and are accompanied by intermediate supervision [3].

The early methods for articulated human pose estimation for still images focused on using the pictorial structures model [5] and consist of unary terms that models body part appearances and pairwise terms between adjacent body parts and/or joints capturing their preferred spatial arrangement [13, 1, 9, 21].

More recent methods for articulated human pose estimation for still images are based on deep learning. Especially convolutional neural networks [11] tend to be very popular. AlexNet [10] was one of the earliest deep learning based methods for human pose estimation. In 2014 Toshev and Szegedy introduced DeepPose, which was an AlexNet-like deep neural network that learned joint coordinates from full images in a very straightforward manner without using any body model or part detectors [17, 3]. In 2015 Tompson *Et al.* introduced an architecture that included a 'pose refinement' model that estimates the joint offset location within a small region of an image, which was trained in combination with a convolutional neural network [16]. In 2017 He *Et al.* introduced the Mask R-CNN, which could perform both articulated pose estimation and segmentation simultaneously [7].

With the recent introduction of transformers [18] and vision transformers [4], recent still image methods have started to incorporate these techniques. For instance, in 2020 TransPose was introduced which used multiple transformer encoder layers to perform the pose estimation [20]. In 2022 ViTPose was introduced, which instead used a vision transformer for the pose estimation [19].

The incorporation of temporal information in video sequences was firstly done by Simonyan *Et al.* who used a convolutional neural network for capturing this temporal information for action recognition [15]. This was rather quickly adapted for human pose estimation by Jain *Et al.* who also used a convolutional neural network for capturing the temporal information across the frames of a video sequence [12, 8]. Girdhar *Et al.* expanded on this idea by using 3-dimensional convolutional layers to capture the temporal information [6].

Further, with the introduction of convolutional LSTM networks [14] in 2015, these type of networks was also experimented on articulated human pose estimation by Girdhar *Et al.* [6] and Artacho *Et al.* [2].

Like in the case of the still images, transformer-based networks have also started to be used to capture the temporal information for human pose estimation. One example of this is DeciWatch, introduced by Zeng *Et al.* in 2022, which efficiently delivers state-of-the-art results [22].

1.2 Choice of Model

- Vi vælger både en 3d-conv, en convlstm og en transformer da vi så repræsenterer alle 3 metoder for at implementere temporal smoothing
- 3d-conv: simpel
- Unipose-lstm: De havde et lignende problem, så dem lader vi os inspirere af
- DeciWatch: Transformer giver lovende resultater, så det vil vi også eksperimentere med

As seen in section [1.1](#), there are generally three different approaches for incorporating the temporal information of the video into the pose estimator: (1) a 3-dimensional convolution-based approach, (2) an approach based on a convolutional LSTM, and (3) a transformer-based approach. As we find all three approaches interesting and see potential in all of them, we will for each of the approaches be implementing machine learning method.

1.3 Reading Guide