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 poses 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 keypoints detector for bouldering, such that it makes use of temporal smoothing for infering the position of the keypoints. 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 detecting the keypoints of a human in two dimensions of a given video recording. This model does not make use of temporal smoothing, but instead just process the frames 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 processing 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. These pose estimation methods for video sequences may use a pose estimation methods for still images, for getting an initial estimation of the poses within each frame of a video sequence. Then later use temporal-inclusive methods for smoothing out these still image pose estimation [3].

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 regressionbased, 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], which consists of unary terms that model body part appearances and pairwise terms between adjacent body parts and/or joints, capturing their preferred spatial arrangement [14, 1, 9, 22].

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 earlies 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 [18, 3]. In 2015 Tompson *Et al.* intrdouced 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 [17]. 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 [19] 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 [21]. In 2022 ViTPose was introduced, which instead used a vision transformer for the pose estimation [20].

The incorporation of temporal information video sequences was firstly done by Simonyan *Et al.* who used a convolutional neural network for capturing this temporal information for action recognition [16]. 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 [13, 8]. Girdhar *Et al.* expanded on this idea by using a 3-dimensional convolutional layers to capture the temporal information [6].

Further, with the introduction of convolutional LSTM networks [15] in 2015, these types of networks were also experimented on articulated human pose estimation by Luo *Et al.*, who in 2018 achieved state-of-the-art results with a unidirectional convolutional LSTM [12]. This idea was expanded by Artacho and Savakis in 2020, who showed that their pose estimator for still images could deliver state-of-the-art results in video sequences by extending it with a unidirectional convolutional LSTM [2].

Like in the case of the still images, transformer-based networks have also started to been 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 [23].

1.2 Choice of Models

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 convolutional layer 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 a machine learning method.

For the 3-dimensional convolutional layer based approach, we will be implementing a very simple model, consisting of only a 3-dimensional convolutional layer followed by an activation function. We picked this model, as we needed something simple to be our baseline model and as we find it interesting whether a more complex model is even needed.

For the convolutional LSTM we let us inspire by the work of Artacho *Et al.* as they had a similar problem, where they aimed at extending a pose estimator, such that that it included temporal information, which ended up yielding state-of-the-art results [2]. However, we will be modifying their convolutional LSTM, as we see some potential shortcomings of their model.

Finally, for the transformer-based approach we will be implementing DeciWatch by Zeng *Et al.* [23]. We picked this model, as this model provides state-of-the-art results as well as it being very fitting for our project, as it is being described as being a model build on top of a keypoint detector for human pose estimators, such that it includes temporal information.

1.3 Reading Guide

The following section, section **??**, covers the most relevant machine learning theory for this project. This is then followed by section **??**, where we go into details with the various models that we will be developing. Section **??** then covers the used datasets. In section **??** and section **??** we describe the experiments we perform, as well as the preprocessing of the datasets. Finally, in section **??** we discuss the obtain results, as well as some of our shortcomings. Lastly, we conclude our results in section **??**.