1 Dataset

To perform the pose estimation in Section **??**, we need some data on which to train, validate and test our models. The following section describes the datasets that will be used, as well as the preprocessing of these datasets.

Keypoint label	ClimbAlong	BRACE	Penn Action
Head	No	No	Yes
Nose	Yes	Yes	No
Left ear	Yes	Yes	No
Right ear	Yes	Yes	No
Left eye	No	Yes	No
Right eye	No	Yes	No
Left shoulder	Yes	Yes	Yes
Right shoulder	Yes	Yes	Yes
Left elbow	Yes	Yes	Yes
Right elbow	Yes	Yes	Yes
Left wrist	Yes	Yes	Yes
Right wrist	Yes	Yes	Yes
Left pinky	Yes	No	No
Right pinky	Yes	No	No
Left index	Yes	No	No
Right index	Yes	No	No
Left thumb	Yes	No	No
Right thumb	Yes	No	No
Left hip	Yes	Yes	Yes
Right hip	Yes	Yes	Yes
Left knee	Yes	Yes	Yes
Right knee	Yes	Yes	Yes
Left ankle	Yes	Yes	Yes
Right ankle	Yes	Yes	Yes
Left heel	Yes	No	No
Right heel	Yes	No	No
Left toes	Yes	No	No
Right toes	Yes	No	No

Table 1: Overview of the annotated keypoints of the three used datasets

1.1 The ClimbAlong Dataset

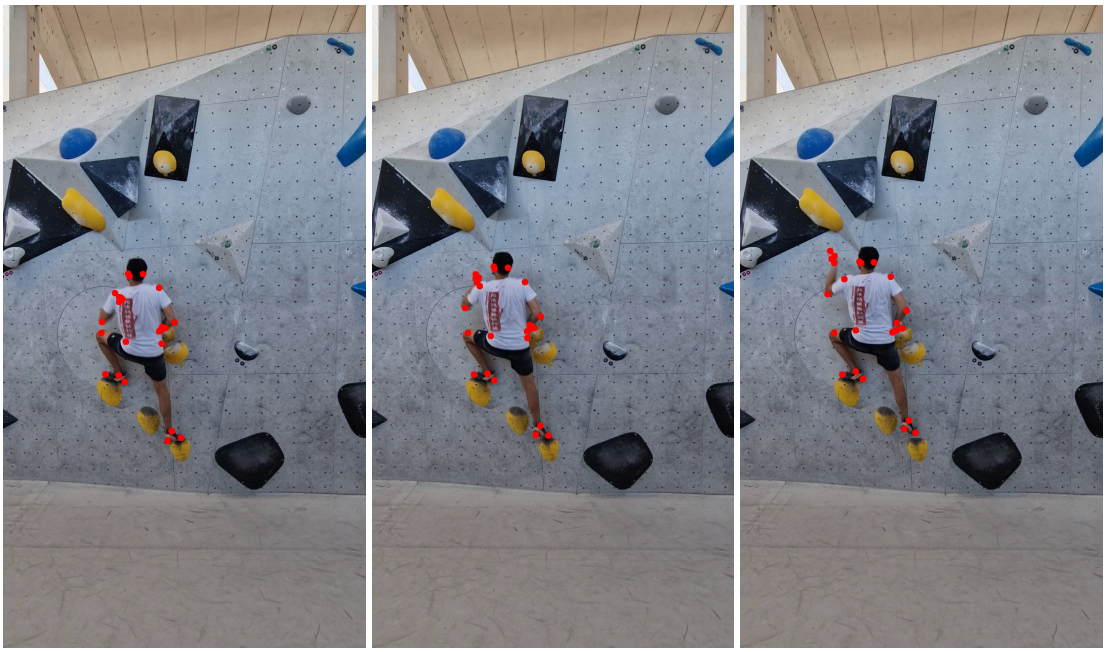
As the aim of our models is to perform well on climbers, we will be using some annotated data of climbers. For this, ClimbAlong ApS has developed a dataset that we will be using. The dataset consists of videos of various climbers on bouldering walls, where each video contains just a single climber. Figure 1 and 2 illustrates two windows of five consecutive frames of a single video from the ClimbAlong dataset. As shown in the figures, the videos in the dataset contains both static positions, where the climber holds a position for a while, as well as quick movements.

The dataset consists of 30 fully annotated videos and a total of $10,310$ fully annotated frames, where each annotation consists of 25 keypoints. Table 1 gives an overview of which key-

(a) Frame 17 (b) Frame 18 (c) Frame 19

(d) Frame 20 (e) Frame 21

Figure 1: Example of five consecutive frames of a video from the ClimbAlong dataset with the corresponding groundtruth keypoints, where the actor holds his position for a while.



(a) Frame 31 (b) Frame 32 (c) Frame 33

(d) Frame 34 (e) Frame 35

Figure 2: Example of five consecutive frames of a video from the ClimbAlong dataset with the corresponding groundtruth keypoints, where the actor performs a quick movement.

baseball_pitch	baseball_swing	bench_press
bowling	clean_and_jerk	golf_swing
jumping_jacks	jump_rope	pull_ups
push_ups	sit_ups	squats
strumming_guitar	tennis_forehand	tennis_serve

Table 2: The original 15 action-types in the Penn Action dataset.

points are annotated in the dataset. Each videos is filmed in portrait mode with a resolution of 1080×1920 and 30 frames per second.

1.2 The BRACE Dataset

The second dataset we will be using is the *BRACE* dataset [1]. This dataset consists of $1,352$ video sequences and a total of $334,538$ frames with keypoints annotations of breakdancers. The frames of the video sequences have a resolution of 1920×1080 [1].

We chose to use this dataset as breakdancers tend to swap between static and quick poses, as well as containing some acrobatic poses, similarly to the ones seen in the ClimbAlong dataset. Generally, the movements of the BRACE dataset are quicker than the movements of the ClimbAlong dataset. The static poses of the BRACE dataset tend to occur less frequently than the static poses of the the ClimbAlong dataset, as well as the quick movements tend to be quicker than the quick movements of the ClimbAlong dataset. However, as both the actors of both datasets swap between static and quick poses, as well as both datasets containing acrobatic poses, we found the BRACE dataset relevant for our experiments in Section **??**. Figure 3 and 4 contains two consecutive sequences, each of five frames, that illustrates these two cases.

The frames of the video sequences have been annotated by initially using state-of-the-art human pose estimators to extract automatic poses. This was then followed by manually annotating bad keypoints, corresponding to difficult poses, as well as pose outliers. Finally, the automatic and manual annotations were merged by using interpolating. Each frame-annotation consists of 17 keypoints, following the COCO-format, as illustrated in Table 1 [1].

1.3 The Penn Action Dataset

The final dataset we will be using is the *Penn Action* dataset [2]. This dataset consists of $2,326$ video sequences of 15 different action-types. Table 2 lists these 15 action-types [2].

Each sequence has been manually annotated with human joint annotation, consisting of 13 joints as well as a corresponding binary visibility-flag for each joint. The frames have a resolution within the size of 640×480 [2].

Unlike the BRACE dataset, most of the poses in the Penn Action dataset are not very unusual and thus are not very relevant for the poses of climbers. For that reason, we have decided to focus on the action-types that may contain more unusual poses. Thus, we only keep the sequences that have `baseball_pitch`, `bench_press` or `sit_ups` as their corresponding action-type [2]. Further, the movements of the Penn Action dataset tend to be of a more similar pace to the ClimbAlong dataset than the BRACE dataset, making the Penn Action dataset relevant for our task.

In total, we use 307 video sequences from the Penn Action dataset, consisting of a total of

(a) Frame 2450

(b) Frame 2451

(c) Frame 2452

(d) Frame 2453

(e) Frame 2454

Figure 3: Example of five consecutive frames of a video from the BRACE dataset with the corresponding groundtruth keypoints, where the actor holds his position for a while.

(a) Frame 1148

(b) Frame 1149

(c) Frame 1150

(d) Frame 1151

(e) Frame 1152

Figure 4: Example of five consecutive frames of a video from the BRACE dataset with the corresponding groundtruth keypoints, where the actor performs a quick movement.

(a) Frame 1148

(b) Frame 1149

(c) Frame 1150

(d) Frame 1151

(e) Frame 1152

Figure 5: Example of five consecutive frames of a video from the Penn Action dataset with the corresponding groundtruth keypoints.

$26,036$ frames. Figure 5 illustrates five consecutive frames with its groundtruth annotations for one of these video sequences.

1.4 Preprocessing of the Data

In the following subsections we describe our preprocessing of the various datasets.

1.4.1 BRACE and Penn Action

As our models take a sequence of estimated poses as input, we will not be using the images of the frames, hence why we discard the images of all frames from BRACE and Penn Action, such that we only keep the annotated poses.

We start by extracting the bounding-box of the annotated pose in each frame by using the annotated keypoints. Further, we expand each side by 10%, such that no keypoint lies on any of the boundaries of the bounding-box. To ensure that the aspect ratio of the pose is kept later on, we transform the bounding-box into a square by extending the shorter sides, such that they have the same length as the longer sides. Next, we discard everything outside the bounding-box and rescale the bounding-box to have a sidelength of 56, such that it has the same size as the output of the already developed pose-estimator.

Next, we transform each frame into twenty five heatmaps. This is done by creating twenty five 56×56 zero-matrices for each frame, such that each zero-matrix represents a single keypoint of a single frame. Further, for each keypoint we insert a fixed value $c \in \mathbb{R}$ at the position of the keypoint in its corresponding zero-matrix and apply a Gaussian filter with mean $\mu = 0$ and standard deviation $\sigma = 1$ to smear out each heatmap. For missing keypoints, we do not place the value c in the corresponding heatmap, making the heatmap consist of only zeros. Further, as Penn Action is the only dataset with the position of the head annotated, as well as the only dataset missing a annotation for the nose, we treat the head-annotation of Penn Action as if it was a nose-annotation, as the position of the two annotation would be very close to each other.

The heatmaps that we produce by following the above description will be used as the groundtruth output of our models. However, as we will be pretraining our models detached from the already developed pose-estimator, we will also need some data as input. We acquire this data by adding some noise to the data, such that they become similar to the output of the already developed pose-estimator, essentially simulating the output of the already developed pose-estimator. The noise is introduced by randomly shifting each keypoint of each sample and by smearing our each keypoint of each sample by using a Gaussian filter, where the standard deviation is randomly chosen. For the shift-value, we use $x \cdot k$, where $k > 0$ is some fixed positive number and x is equal to 20% of the mean torso-diameter. We clip the position of the shifted keypoints between 0 and 55, such that they cannot be outside of their corresponding heatmaps. For the random standard deviation we sample uniformly at random from the set $\{1, 1.5, 2, 2.5, 3\}$.

1.4.2 ClimbAlong

For the ClimbAlong dataset we perform only minor preprocessing. First, the preprocessing of each video is done by having the already developed pose-estimator process the video, such that we have the output heatmaps of the pose-estimator, containing all of the pose-estimations of each video. Next, we preprocess the heatmaps by setting all negative values to 0 and normalizing each heatmap, such that each heatmap sums up to the fixed value $c \in \mathbb{R}$ that we used

when preprocessing the BRACE and Penn Action datasets, essentially making the heatmaps more similar to the preprocessed heatmaps of BRACE and Penn Action. These heatmaps will then be used as the input for our models.

For the groundtruth heatmaps we create twenty five heatmaps of each frame, similarly to how we did it for the BRACE and Penn Action datasets, however, in this case we use the predicted bounding-box of the pose-estimator as our bounding-box. In cases where the groundtruth keypoint is placed outside of the bounding-box, we place the groundtruth at the border of the bounding-box.