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.

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

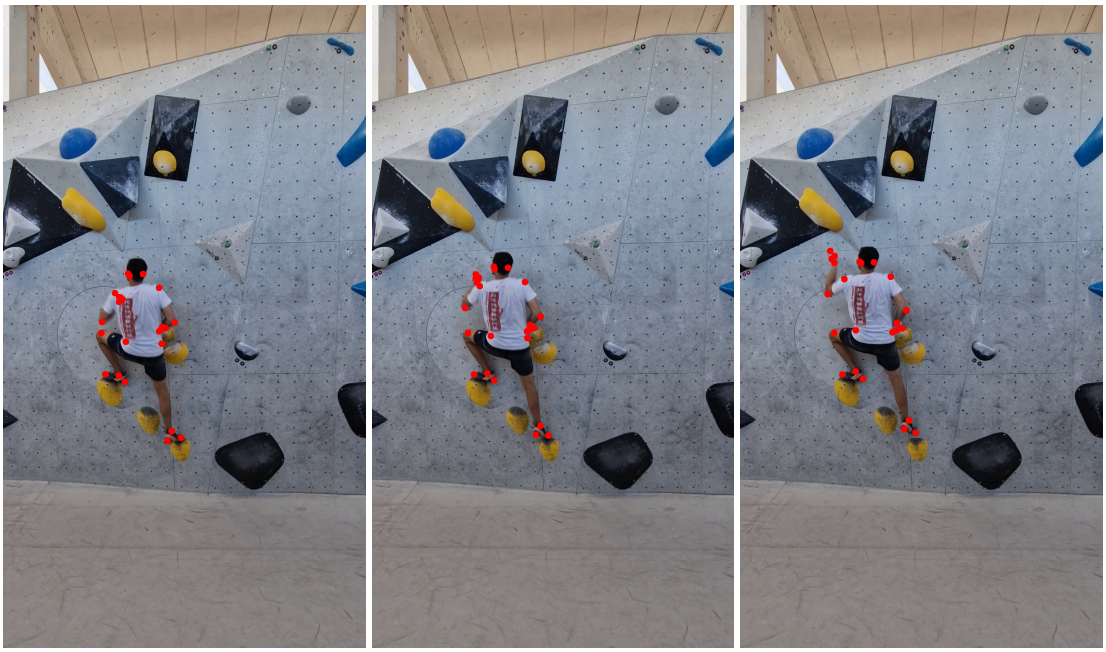
As the aim of our models is to perform well on climbers, we will be using some annotated data of climbers. For this, ClimbAlong ApS has developed a dataset that we will be using. The dataset consists of videos of various climbers on bouldering walls, where each video contains just a single climber. Figure 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 29 fully annotated videos and a total of $9,535$ fully annotated frames, where each annotation consists of 25 keypoints. Table 1 gives an overview of which keypoints are annotated in the dataset. Each videos is filmed in portrait mode with a resolution of

(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.



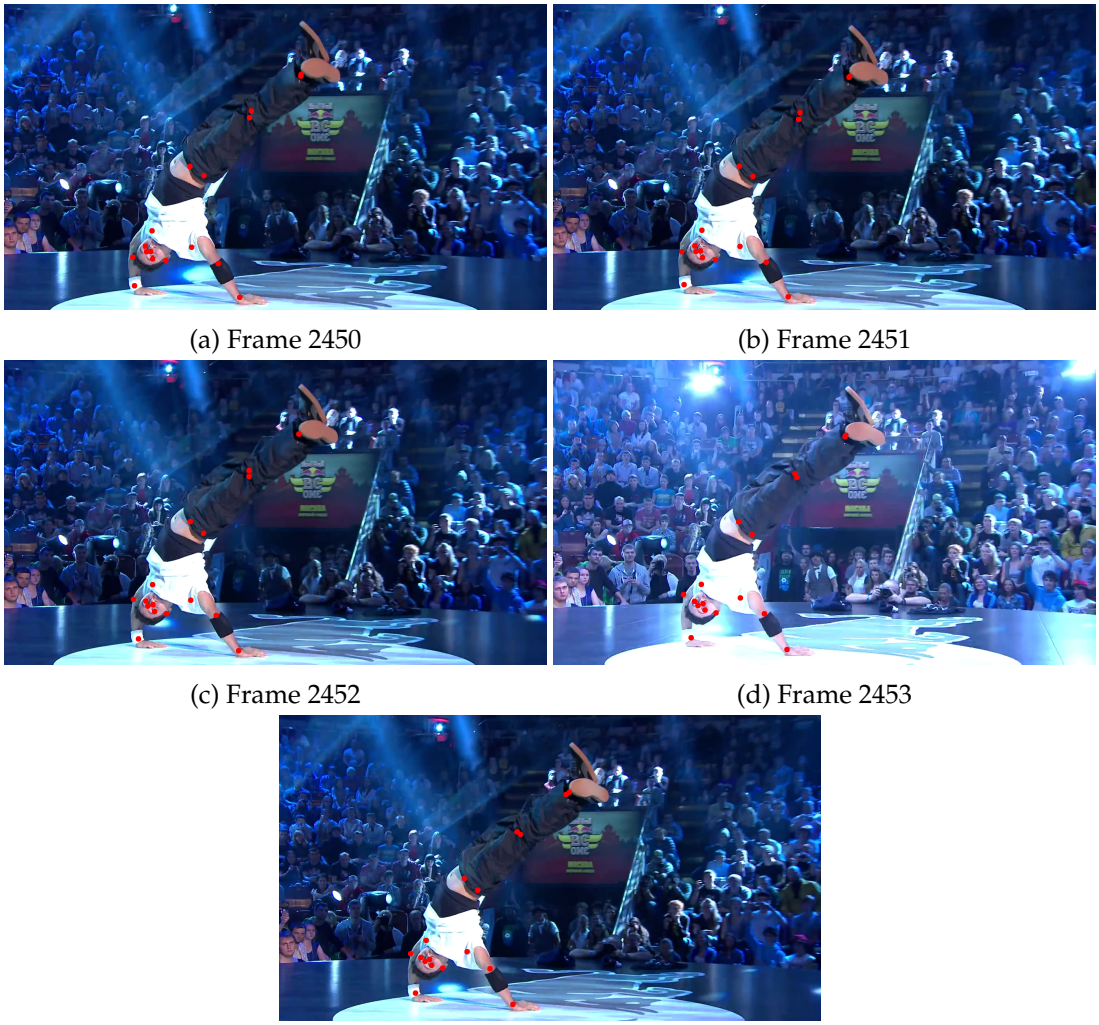
(a) Frame 31 (b) Frame 32 (c) Frame 33

(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.

1080×1920 and 30 frames per second.

1.2 The BRACE Dataset



(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.

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.

(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.

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.

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 frame have a resolution within the size of 640×480 [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 placement of the invincible keypoints is very inconsistent, as some of the positions have been estimated, whereas others have been placed near one of the corners the image.

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

(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.