Computer Vision Analysis of Speed Climbing Human Pose Estimation

Manny Cassar
Queen's School of Computing
Kingston, Canada
m.cassar@queensu.ca

Kieran Green
Queen's School of Computing
Kingston, Canada
21kg38@queensu.ca

Mike Stefan
Queen's School of Computing
Kingston, Canada
21mgs11@queensu.com

I. Introduction

The sport of speed climbing made its debut on the world stage in the 2020 Tokyo Olympic Games due to its remarkable surge in popularity in recent years. This rise from a niche discipline to a mainstream competitive activity has underscored the need for sophisticated tools to analyze climbers' movements that will analyze and optimize training methodologies.

The goal of a speed climber is to navigate up a 15-meter vertical wall containing commonly placed climbing holds with precise and swift body positions in the shortest time possible. The primary challenge for coaches in speed climbing lies in separating and grading specific movements and sequences in a rapid and dynamic climbing run. This fast-paced format makes conventional pose estimation techniques less effective, necessitating a specially trained system capable of processing data swiftly to provide immediate feedback. Addressing this gap is critical for advancing training methodologies in speed climbing and can have broader implications for motion analysis in other high-velocity sports and activities.

This project aims to develop a deep neural network model tailored explicitly for speed climbing capable of detecting and overlaying skeletal coordinates onto live video feeds of climbers. The neural network will process individual video frames as input and output the coordinate positions of critical joints, including feet, hips, hands, elbows, and knees. For each frame of the climbing run, a complimentary Python script will then use these coordinates to produce a skeletal overlay, connecting the joints with lines to visualize the climber's real-time posture and movement. The video will then be reconstructed to produce a live skeleton overlay on the climber.

II. Objective

There is a lack of an accurate quick real-time model of a climber's body during their speed climb. We plan to build a model that will capture the skeletal coordinates of the climber and overlay them in a real-time video using a deep neural network. This will allow coaches and climbers to more quickly spot mistakes made during a climb without the need to wait in between climbing sections. It is our goal to develop this revolutionary tool using a deep convolution neural network model which will allow our model to be greatly optimised allowing for real-time viewing and application.

III. Related Work

Overlaying skeletons onto human bodies is a widely studied field with a blind spot for algorithms specialized for climbing. One study looks at climbers' body position and motions to spot errors made by the climber while they are bouldering.[1] This is done by mapping a skeleton over the climber, mapping the joints, and then following the motions to determine when the climber makes mistakes. Well, this model takes too long to be used for speed climbing; the training information used in this model will significantly improve our accuracy in recognizing climbing-specific positioning.

In their research, Pieprzycki et al. [2] investigated methods to analyze speed climbers' runs through video recordings. They developed a system that captures spatial and temporal parameters of climbers' movements without requiring intrusive sensors utilizing high-frame-rate cameras and visual markers placed near the climber's center of mass for effective tracking. Their approach employed algorithms such as the Kanade-Lucas-Tomasi (KLT) tracker and the OpenPose convolutional neural network for keypoint detection. This methodology allowed for the extraction of various kinematic parameters, including velocity, acceleration, and movement trajectories, providing valuable insights into climbers' performance. While their work showcased the potential of video analysis in evaluating climbers, its dependence on physical markers and post-processing limits its practicality for real-time applications. There is a clear need for a noninvasive, efficient system capable of real-time pose estimation that can handle the rapid and complex movements characteristic of speed climbing. Our project seeks to address this gap by developing a deep neural network model tailored explicitly for speed climbing. This model aims to enable real-time skeletal overlays on live video feeds without requiring markers, thereby enhancing the applicability and scalability of pose estimation in the sport. Our model will build upon the foundation established by Pieprzycki et al., moving towards a more practical and immediate analysis tool for athletes and coaches alike.

Our application requires our model to be extremely lightweight and able to analyze high-resolution video on standard hardware. Two common lightweight human pose estimation approaches are shuffle blocks [3] and HRNet [4]. Shuffle blocks improve the algorithm's performance by separating the convolutions into a linear combination of depthwise convolutions and 2 other convolutions, drastically reducing the compute time since these convolutions are more computationally efficient than the standard convolutional step. The HRNet architecture starts with high-resolution convolutions and adds high-to-low-resolution streams connected in parallel with their output, eventually being fused. With both of these approaches having drawbacks, Lite-HRNet [5] proposes a novel combination of these two algorithms, which replaces the costly high-resolution convolutions found in the HRNet architecture with the split stream approach derived in shuffle blocks. This approach also provides novel optimizations to the shuffle block algorithm that reduces the number of 1X1 convolutions, an extremely costly operation on video feeds.

IV. Methods

Our approach is modeled after the architecture proposed in Lite-HRNet [5], where our images are passed through the high-resolution network and fused with high-to-low streams. This entire process is a subset of the deep convolution neural network model. The approach taken in Lite-HRNet proposes novel optimizations that greatly benefit our real-time application. Being significantly lighter while maintaining the accuracy presented by significantly larger and slower models, this architecture meets all of our research objectives. We also look to include the climbing position-specific training information used in models trained for bouldering [1]. This should improve our model accuracy in recognizing climbing-specific positioning, which is typically not included in standard HPE training sets, including COCO [6].

V. Data Collection and Preprocessing

This project will use high-quality competition data from multiple World Cup competitions between 2018 and 2020 [7]. This data is split up into two different parts. The first part is a list of 362 individual runs with a link to the online YouTube video with the video metadata. The second part of the data is a folder of files named with the run_id found in the videos. The files contain the 16-joint XY-coordinate positions of the climber for each video frame. This data will be quickly collected by running a script which will go through all the video links and capture the portions provided with the time stamps. Afterwards, they will be reviewed to ensure all videos are correctly downloaded.

VI. Experimental Design

To rigorously evaluate the performance and effectiveness of our real-time model, we will employ the following quantitative benchmarking techniques, which will provide metrics of accuracy, efficiency, and robustness.

Joint Localization Accuracy

- Mean Absolute Error (MAE): We will calculate the MAE between the predicted joint coordinates and the ground truth annotations for each key joint—feet, hips, hands, elbows, and knees. This metric provides an average measure of the key joint's prediction error in pixels.
- Percentage of Correct Keypoints (PCK): The PCK metric assesses the percentage of correctly predicted joints within a specified threshold distance from the ground truth. We will use thresholds based on a fixed pixel value.

Model Efficiency and Real-Time Performance

- Inference Speed (Frames Per Second): We will measure the model's inference speed to ensure real-time performance when overlaying skeletons onto live video feeds. The target benchmark is a minimum of 30 frames per second (FPS) on a conventional laptop brought to the climbing event.

Robustness and Generalization

- Cross-Condition Evaluation: To ensure robustness, the model's performance will be fed a custom testing set collected under diverse conditions, with the lighting variation, image capture angle, climber attire, and background complexities controlled.

VII. Expected Results and Analysis

We anticipate that our deep neural network model will achieve high accuracy in estimating the positions of key joints in speed climbing videos and will operate efficiently enough for real-time applications. The outline below shows our expected outcomes and the methods we will use to analyze and interpret the results.

High Accuracy in Joint Localization

- Expected Outcomes: We expect the model to achieve a Mean Absolute Error (MAE) of less than 5 pixels on average and a Percentage of Correct Keypoints (PCK) exceeding 85%.
- Analysis Plan: We will perform per-joint and overall evaluations to identify any joints with consistently higher errors.

Real-Time Processing Capability

- Expected Outcomes: The model should maintain an inference speed of at least 30 FPS on a standard GPU-equipped laptop system.
- Analysis Plan: We will profile the model's computational performance using the psutil Python package and NVIDIA's Nsight Systems to identify bottlenecks. We will optimize the model through techniques such as model pruning or quantization if necessary.

Statistical and Computational Techniques

- Error Distribution Analysis: We will examine the distribution of errors across joints using histograms and box plots to assess model bias and variance.
- Cross-Validation: We will employ k-fold cross-validation to ensure that the model's performance is generalizable and not a result of overfitting the training data.

Comparison with Baseline Models

- Expected Outcomes: When evaluated on the same input data, our model's accuracy and efficiency should outperform or be competitive with existing models.
- Analysis Plan: We will perform paired statistical tests, such as the paired t-test, to compare our model's performance with baseline models, ensuring that any improvements are statistically significant.

VIII. Timeline

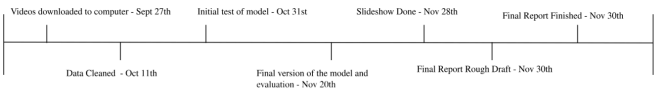

Figure 1: Our Timeline

IX. Potential contributions

Using a strictly single-camera-based algorithm for climber pose estimation for speed climbing is not a well-researched application. From our literature review, all existing, skeletal-based pose estimations of climbers require many camera views [8] or require special sensors placed in the wall to track the forces the climbers are applying to the wall [9]. The other case is targeted specifically towards bouldering and requires a fixed camera at a static and known location, which is less useful when applied to speed climbing, especially for use in international competition, where a permanent camera setup is only sometimes possible [1]. Our proposal will help improve these existing systems by allowing for dynamic, accessible, inconsistent setups. This will help make CV-assisted analysis and coaching tools more applicable and valuable for speed climbers. This proposal also expands on how athletes can gain insights into their climbing. Speed climbing is a unique climbing discipline where each minuscule action significantly impacts the final time. With the entire sport taking < 5 seconds at the highest level, athletes must move flawlessly every run, and our system will help highlight mistakes in their training runs.

References

[1] R. Beltrán, J. Richter, G. Köstermeyer, and U. Heinkel, "Climbing Technique Evaluation by Means of Skeleton Video Stream Analysis," *Sensors*, vol. 23, p. 8216–8217, 2023, doi: 10.3390/s23198216.

[2] A. Pieprzycki, T. Mazur, M. Krawczyk, D. Król, M. Witek, and R. Rokowski, "Computer-Aided Methods for Analysing Run of Speed Climbers," *Preprints*, Feb. 2023, doi: 10.20944/preprints202302.0166.v1.

[3] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018, pp. 122–138.

[4] J. Wang *et al.*, "Deep High-Resolution Representation Learning for Visual Recognition," 2020, [Online]. Available: https://arxiv.org/abs/1908.07919

[5] C. Yu *et al.*, "Lite-HRNet: A Lightweight High-Resolution Network." [Online]. Available: https://arxiv.org/abs/2104.06403

[6] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context." [Online]. Available: https://arxiv.org/abs/1405.0312

[7] P. Elias, V. Skvarlova, and P. Zezula, "SPEED21: Speed Climbing Motion Dataset," 2021, pp. 43–50. doi: 10.1145/3475722.3482795.

[8] F. Q. P. L. Lionel Reveret Sylvain Chapelle, "3D Visualization of Body Motion in Speed Climbing," *Front Psychol.*, 2020.

[9] J. Richter, R. B. Beltrán, G. Köstermeyer, and U. Heinkel, "Human Climbing and Bouldering Motion Analysis: A Survey on Sensors, Motion Capture, Analysis Algorithms, Recent Advances and Applications," in *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, SciTePress, 2020, pp. 751–758. doi: 10.5220/0008867307510758.