Rock-Climber Pose Estimation Using a Pictorial Structures Framework with SIFT-detected Features and ESRGAN-Super-resolved Images

Hamza Muhammad Anwar
Computer Science, *Irving K. Barber Faculty of Science*
University of British Columbia, Okanagan
Kelowna, Canada
hanwar02@student.ubc.ca

Bridgette Hunt
Computer Science, Irving K. Barber Faculty of Science
University of British Columbia, Okanagan
Kelowna, Canada
bhunt02@student.ubc.ca

Jeena Javahar
Computer Science, Irving K. Barber Faculty of Science
University of British Columbia, Okanagan
Kelowna, Canada
jeena01@student.ubc.ca

Aadil Shaji
Computer Science, Irving K. Barber Faculty of Science
University of British Columbia, Okanagan
Kelowna, Canada
as0809@student.ubc.ca

***Abstract*—Human pose estimation (HPE) is a critical task in computer vision, enabling applications in animation, augmented reality, security, and sports analysis. Traditional HPE methods rely on probabilistic and statistical models, such as the Pictorial Structures framework, but lack generalizability and efficiency. Modern deep-learning techniques have improved accuracy and robustness but introduce complexity and high computational costs. We propose a specialized HPE pipeline tailored for both indoor and outdoor rock-climbing analysis. Our method enhances 2D joint detections using pre-processing techniques intended to improve accuracy in the rock climbing domain. Key characteristics of the system include assuming a single-person scenario to simplify detection and optimizing foreground extraction by isolating the climber against the image background. We also add pre-processing steps such as image super-resolution that aimed to improve the performance of the pose estimator. After the images are super-resolved, they are passed to a SIFT feature extractor to find relevant human features, which are used to draw a bounding box around the person. The results of the proposed human pose estimation pipeline are unsatisfactory, as the pose estimator is unable to accurately delineate the body parts of the climbers. This may have been because the available body models were meant for images with forward-facing and upright people, a requirement that most rock-climbing images do not satisfy.**

Keywords—human pose estimation, pictorial structures, image super-resolution, ESRGAN, SIFT, feature extraction

I. Introduction

Computer vision is a research area in computer and data science that aims to utilize computers to perform tasks in image processing similar to that of human visual capabilities. This involves developing systems, algorithms, and techniques that enable machines to mimic the functions of the human visual system and, in some cases, to attempt to create models of human vision. The various problems that research in computer vision aims to solve have a wide variety of real-world applications, such as in the development of autonomous vehicles, security, surveillance, manufacturing, health, retail, etc. A relevant problem in computer vision is human pose estimation, which involves creating methods for the isolation of person(s) in a given image or images.

Human pose estimation (HPE) is a technique in computer vision meant to locate, isolate, and highlight the locations and shapes of the human body from a two- and three-dimensional perspective. It is a specialization of techniques related to general object pose estimation in digital imagery. In this case, a pose is the configuration of a human body's joints and objects (body parts). Human pose estimation aims to approximate the position of these objects and joints in relation to the rest of the body.

As a long-standing problem in Computer Vision, pose estimation has numerous real-world applications. Computer animation, where tracking the location of extremities and the general body shape has been utilized to construct more realistic character animation, is a relevant use case for HPE. In more recent years, the increased interest in augmented reality (AR) techniques has made human pose estimation and general pose estimation an invaluable asset to curators of these experiences, especially in cases where multiple users must exist in the same digital world simultaneously. Other such uses, such as security and surveillance, often utilize HPE to locate human bodies in their application domain.

Over years of research, human pose estimation has evolved beyond traditional methods, with deep-learning methods surpassing probabilistic and statistical learning models in the accuracy of estimation. However, both traditional and deep-learning methods that aim to solve the problems outlined by HPE generally follow a pipeline of image processing before attempting to tackle the main problem. These iterative architectures are commonplace, with critical steps such as feature extraction at the forefront of most implementations.

Traditional methods, as stated above, largely fall into a line of statistical and probabilistic models, often using graphical models to represent the relations between joints. These methods are not very generalizable. They also typically struggle with performance due to specific variable requirements and the need for a large pipeline of processing tasks to properly extract features and ultimately estimate the pose. Felzenszwalb & Huttenlocher [7] introduced a widely-used technique in 2-dimensional HPE known as the Pictorial Structures (PS) model; this technique uses an undirected graph in which the vertices correspond to the body parts, and the graph edges between the body parts represent joints. Later research in traditional HPE methods tended to utilize the PS strategy following processing steps, such as background subtraction and edge detection.

Modern deep-learning approaches are more varied; researchers have employed top-down frameworks involving regression-, heatmap-, and video-based methods, as well as bottom-up frameworks, of which information is more limited [2]. In general, deep-learning methods utilize a neural network of some sort to estimate the position and/or area of the body parts and joints. In terms of top-down frameworks, heatmap-based methods are more popular when compared with their regression-based counterparts due to their higher accuracy; the regression-based methods also often neglect to output anything but a set of 2D image coordinates for each joint, missing detail related to the center and size of the body parts themselves [2]. Deep-learning methods are known for their superior accuracy and ability to generalize, as opposed to traditional methods.

With the application and advancement of deep-learning techniques in human pose estimation, the problem has become easier to solve without excessive specialization for certain problems, and with a reduction of necessary pipeline steps and thus, parameters. However, traditional techniques surpass their deep-learning counterparts in some regards; the deep-learning methods utilize black-box models which eliminates the possibility of understanding the under-the-hood process, while traditional methods have structured and well-known steps in their pipelines. In addition to their limited understandability, while deep-learning models can be more performant, their complexity in terms of space (memory) used cannot be understated. Some approaches, such as model compression-based methods [2], aim to alleviate some of this issue by reducing model complexity, with some decline in accuracy.

In this paper, we focus on traditional methods of performing the human pose estimation task in computer vision, exploring potential methods to optimize and/or improve their outputs for the rock-climbing image domain. We attempt to introduce additional pre-processing methods to augment the existing pictorial structures model (PSM) technique. This includes the addition of a deep-learning super-resolution technique to improve the detected feature space and the implementation of a SIFT-based [11] pre-processing step to identify clusters of features, ideally to determine the approximate location and bounding box of a human in an image.

II. Literature Review

A. Deep Learning Methods

Deep Learning methods started gaining popularity in the 2010s after the release of ImageNet, a large database consisting of millions of images for visual recognition research, and AlexNet, the convolutional neural network (CNN) that achieved significantly higher accuracy than its predecessors and established CNNs as a powerful tool for image processing. In this section, we give a quick outline of the deep-learning techniques used for human pose estimation.

a) Pose as Compositional Tokens (PCT)

The authors of [3] proposed the Pose as Compositional Tokens (PCT) representation that considers joint dependency. The PCT method is the current state-of-the-art on the MPII dataset. In this method, the human body is first decomposed into smaller parts called token features by a compositional encoder. Each token feature encodes a sub-structure of the pose, which is then quantized using a shared codebook. This converts them into discrete indices. The encoder, codebook, and decoder are trained together to minimize reconstruction error and ensure accurate pose representation. The encoder, codebook, and decoder are trained together by minimizing reconstruction error, ensuring highly accurate pose representation. For the classification task, the model predicts the category of each token. Finally, the decoder network uses those predictions to reconstruct the full pose.

Strengths: This method is robust against occlusions and works for both 2D and 3D pose estimation. Moreover, it has a faster inference speed than the state-of-the-art ViTPose (Huge).

Weaknesses: The method is not well-adopted, and so demonstrations of its efficacy are only known based on the original work.

b) Regression-based Methods

Regression-based deep-learning human pose estimation methods typically employ an interactive architecture that extracts image features [1]. Convolutional neural networks (CNNs) are often employed in regression-based methods to perform highly efficient pose estimation. In respective reviews, Kulkarni et al. [6] and Chen et al. [1] discuss several implementations of CNNs to perform human pose estimation, including, among others, BlazePose, DeepPose, OpenPose, and Graph convolutional network (GCN), etc. These methods will often use a training set of annotated images to perform supervised learning for human-pose estimation. In their testing of techniques against several evaluation metrics, data sets, and applications, Kulkarni et al. found that BlazePose has notable efficiency in mobile applications while preferring OpenPose for multi-person tracking. Chen et al. note that regression-based methods are often limited by their inability to consider the area of body

parts, only outputting a set of 2D picture coordinates for each joint.

Strengths: Regression-based methods are efficient at performing simple pose estimation tasks and are smaller than most other models used in the discipline.

Weaknesses: The regression-based methods are limited in the complexity of their output and do not often take into account the surface area of body parts, simply outputting a set of coordinates for each body part.

c) Heatmap-based Methods

As with regression-based deep-learning methods for human-pose estimation, an iterative architecture is also often utilized in heatmap-based approaches [1]. These methods are also divided by their use of symmetric architectures, where downsampling and upsampling frameworks are used to denote the resolution of features to capture the spatial relationships between joints - asymmetric architectures, - where the upsampling process is lightweight while the downsampling process is heavy - and high-resolution architectures, where high-resolution features are maintained through the entire process. Heatmap-based methods have become more popular than regression-based methods due to improved accuracy but suffer from expensive algorithmic complexity and errors involving feature representations from differing resolutions.

Strengths: The heatmap-based methods are generally more accurate than earlier deep learning methods and also, in general, provide more information than regression-based approaches.

Weaknesses: The heatmap-based approaches are more complex computationally than regression-based methods. They also struggle with inevitable quantization errors [1] due to the nature of taking features from different resolutions.

B. Traditional Methods

Research on human pose estimation and its closely related discipline, human motion analysis, started being published in the 1980s. It was motivated by a wide spectrum of applications, such as surveillance, human-computer interfaces, and athletic performance analysis. There are three popular methods of modeling the human body: The kinematic, planar, and volumetric models [2]. In the kinematic model, the body is represented like a stick figure, with limbs interconnected with joints. The planar model depicts the body by its contours, forming a 2D plane. Finally, the volumetric model represents various body parts as 3D shapes such as cones and cylinders [3].

Many advanced techniques have established themselves as being highly effective. In this section, we provide a brief overview of the traditional methods employed for human pose estimation.

a) Flexible Mixtures of Parts

Yang and Ramanan [4] introduce model articulation using a mixture of small, non-oriented parts. This method enhances conventional pictorial structure models that only record spatial interactions by simultaneously capturing spatial relations between part locations and co-occurrence relations between part mixes. In order to provide realistic posture configurations, the model takes into account co-occurrence dependencies and spatial interactions between pieces.

Strengths: By composing local mixtures, FMP models can model an exponentially large set of global mixtures efficiently. They are also fast enough to search over all locations and scales, both detecting and estimating human poses without any preprocessing.

Weaknesses: The computational complexity of performing the location of parent and child objects in the image, as well as performing the transforms, is very high for the proposed algorithm.

b) Pictorial Structures

Felzenszwalb & Huttenlocher [7] presented a framework using the Pictorial Structure (PS) models for object recognition. The PS model represents objects as a probabilistic graphical model of joints and parts and was widely used before the revolution of deep learning to accomplish the human pose estimation task [8, 9]. Their implementation of the PS model aimed to efficiently solve the energy minimization problem and other problems related to over-parameterization and finding more than one minimal match in an image [7]. When applied to human pose estimation, the PS model identifies people in images by treating the extremities, head, and torso as vertices, with joints being identified as undirected edges.

Strengths: The PS model is capable of robustly estimating the location of body parts and joints between them through energy minimization (an optimization process) and representation of objects based on a probabilistic, undirected, graphical model.

Weaknesses: Although the PS model was significant at the time, it is much weaker when compared with modern methods. It is not particularly tolerant to images with clutter (noise), and many pre-processing steps are often required to make full utilization of the technique.

c) 3D Pictorial Structures

Belagiannis et al. [5] proposed the 3D Pictorial Structures (3DPS) model for multiple human 3D pose estimation from multi-view images. This method first represents the body as an undirected graphical model. It then develops a discrete state space of each body part, which holds the part's position and orientation information. This is formed by the triangulation of corresponding 2D body joints detected in multi-views. The Conditional Random Field (CRF)-based 3DPS model the authors introduce enforces collision, visibility, and kinematic constraints to resolve ambiguities (e.g., incorrect body part associations).

Strengths: It can handle occlusions and identity ambiguities across multiple individuals in the scene. It's also faster, less computationally expensive, and works well on single and multiple humans.

Weaknesses: Since the approach relies on 2D body joint detections, its performance depends on the accuracy of the initial detections. Moreover, modeling both the 3D position and orientation of the body part may be inefficient.

d) Tree-Structured Based Model

Zhang et al. [10] proposed a framework in which the human body is modeled as a three-tier tree structure, with nodes for each of the extremities - two for each limb - and nodes for each of the torso and head. In this model, each body part is modeled by a rectangle with a 2D position, orientation, width, and length. The human pose estimation pipeline involves background subtraction and edge mapping through image segmentation prior to performing any specific analysis to form the tree structure. The estimation step that Zhang et al. proposed involves rules - referred to as grammars - that dictate the tree parsing algorithm. The first grammar involves optimizing the body parts that have the least likely matches, the second states that the initial 'importance' of each body part decreases based on distance from the root node - the torso, in the model - and nodes at each level have the same initial importance as their neighboring nodes on the same level. The third grammar determines whether the 'jump' or 'diffuse' Markov chain dynamic is used in optimizing the selected body part. The fourth and final grammar states that the importance of the body part is propagated to its adjoining body parts - parent and child nodes.

Strengths: The constraints provided by the tree-structure model proposed by Zhang et al. are powerful in detecting poses, performing well even when body parts are occluded by other objects in the scene. It also demonstrated good performance in noisy images. In terms of speed, it outperformed contemporary pose estimation techniques by several minutes, but the number of processing iterations was much higher.

Weaknesses: The tree-structure model proposed by this paper was not utilized in significant proportions by works in the following years and has not been tested by enough independent researchers to prove its efficacy. While it is a promising model, the implementation of this technique is likely to be complex and may not prove to be effective. Many advanced techniques have been developed now that perform better.

III. System Design

A. System Design & Architecture

Fig. 1 shows our proposed pipeline. The proposed system is composed of a deep-learning super-resolution technique and SIFT feature extraction/clustering to determine positional information about people in images, whose outputs are used to perform pose estimation using a Pictorial Structures model. Inputs are single, still images, and outputs include images with the identified body pose in the form of a stick figure with different colors composing different body parts. Other internal inputs include the appearance model used for pose estimation, parameters related to foreground highlighting, SIFT feature recognition, extraction, and clustering, and the bounding box within which to start the determination of the human pose.

The system pipeline performs the following tasks in order: super-resolution upscaling of the image to improve feature extraction, SIFT feature extraction and clustering to determine the bounding box surrounding the person, and finally, human pose estimation via the pictorial structures implementation.

Fig. 1 The proposed human pose estimation pipeline.

B. Algorithms

a) Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN)

The ESRGAN is a generative adversarial network (GAN) designed to upscale low-resolution images into high-resolution ones. While the high-level architecture was kept the same as its predecessor, SRGAN, the ESRGAN improved on it by employing the use of Residual-in-Residual Dense Blocks (RRDB) as the basic network building unit. These blocks integrate multi-level residual networks with dense connections. It also does not have any batch normalization layers, unlike the SRGAN. This makes the network more stable and capable of capturing complex features while also reducing memory usage and computational complexity in tasks like super-resolution. Moreover, the proposed RRDB has a more complex structure than the original residual block in the SRGAN, using more layers and connections to boost performance.

Besides this, the SRGAN's discriminator was also improved by basing it on the Relativistic GAN, which tries to predict the probability that a real image is relatively more realistic than a fake one. This way, during adversarial training, the generator can benefit from the gradients from both generated and real data, whereas in SRGAN, only the generated part takes effect. This discriminator modification aids in learning more intricate textures and sharper edges.

The pre-trained ESRGAN model we implemented was from [17]. This model is trained on the Flickr2K dataset [18], a common dataset used for image super-resolution tasks. In our implementation, the scaling factor used was n=2; in other words, the resolution of the image would be doubled as compared to the reference image.

b) Scale-invariant feature transform (SIFT) Clustering

The Scale-Invariant feature transform (SIFT) [11] is a method developed to encapsulate and transform features into identifiable information vectors. The SIFT method can be used to perform object recognition and detection between images following feature extraction. The SIFT is

additionally photometrically and geometrically invariant. Other than its uses in object recognition and detection, the key points (features) identified by the SIFT algorithm are useful for identifying areas of dramatic change (a high-magnitude intensity gradient) in the image.

The SIFT algorithm implemented in our pipeline detects features and is followed by clustering the points. Clusters are created based on the properties of each key point's location within the image, the radius of the key point, and the sigma of the Gaussian function used to detect the sift features.

We determine the largest cluster to identify the humanoid within the image and use the cluster's left, top, width, and height to determine the bounding box in which to look for both the human and their pose. Fig. 1 contains an example result of the SIFT feature detection and the subsequent clustering. Fig. 2 shows the derived bounding box from the clusters.

Fig 1. Detected SIFT features and clusters.

Fig 2. Derived bounding box from clusters.

c) Calvin pose estimation (Pictorial Structures Model)

The Pictorial Structures model (PSM) is the core human-pose estimation portion of our system. The PSM model treats human pose estimation as an energy minimization problem, typically representing the human body as an undirected graph. Body parts (vertices) and joints (edges) are oriented in such a way as to minimize a predefined energy function.

The PSM used was implemented by Eichner & Ferrari [13], in the form of the "Calvin upper-body detector." (Calvin detector). In addition to the PSM, the Calvin detector utilizes pre-trained appearance models [15] and other pre-processing techniques, such as part-specific color estimation, to improve the estimate of the human (and their associated pose) in the image. In its final stage, it uses Ramanan's [16] articulated human pose image parsing algorithm to search for the human pose. Fig. 3 shows the presented pipeline in Eichner et al.'s [14] implementation of the PSM.

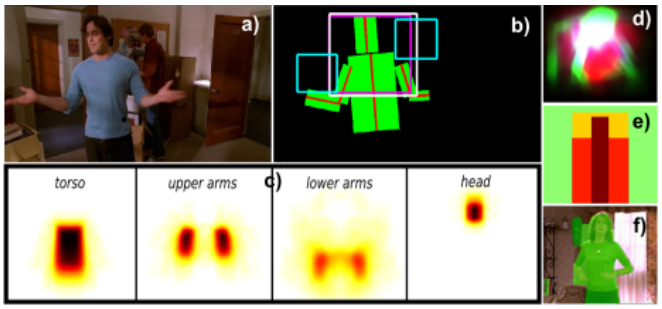

Figure 3. Eicher et al. [14] Pictorial Structures Model pipeline

IV. Results

Fig 4 shows a crop of the output of a rock-climber image after being passed to the ESRGAN to upscale it. As can be seen, the ESRGAN does improve the resolution of the original ground truth image, however minor it may be. The upscaling is more clearly visible in areas such as the climber's right hand, where the hand is more distinguishable from the blue hand-hold as compared to the reference image. In addition, the climber's back also seems less blurry than in the reference image. This may help the pose estimator later on in the pipeline to classify that region as the torso and more easily separate it from the head and neck region.

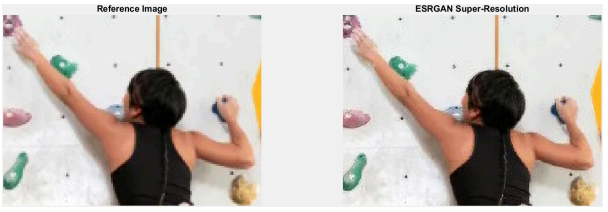

Fig. 4. The output of the ESRGAN along with the ground truth image.

Fig. 5 shows the result of the SIFT feature detection, clustering, and the resulting bounding box. Fig 6. shows the overall result of our pipeline. The PSM pose estimator is unable to accurately estimate the pose of humans within the image. The identified position, orientation, and size of the body parts are not at all accurate to the human seen in the image. Indeed, the positions of the body parts in relation to each other are unreasonable, with each body part arranged haphazardly, and the scale of the depicted body by the stick figure is much larger than it should be. Given these details, it's clear that the pose estimator, given the bounding box

provided by the SIFT clustering, did not obtain a good enough hint of the overall human shape present in the image.

Fig. 5. The detected SIFT features, clusters, and bounding box.

Fig. 6. The pose estimation output of the system's pipeline as a whole. (Purple - Head, Red - Torso, Yellow – Lower Arms, Green - Upper Arms - Blue - Upper Legs, Cyan - Lower Legs)

Other images of rock climbing show similar results. Fig. 7 shows an instance where the bounding box generated from the SIFT features fails to encapsulate the human. Fig. 8 displays an instance where the image is far too filled with feature information, leading to a bounding box that encapsulates the entire image dimensions. Intuitively, these bounding boxes lead to poor pose estimation results.

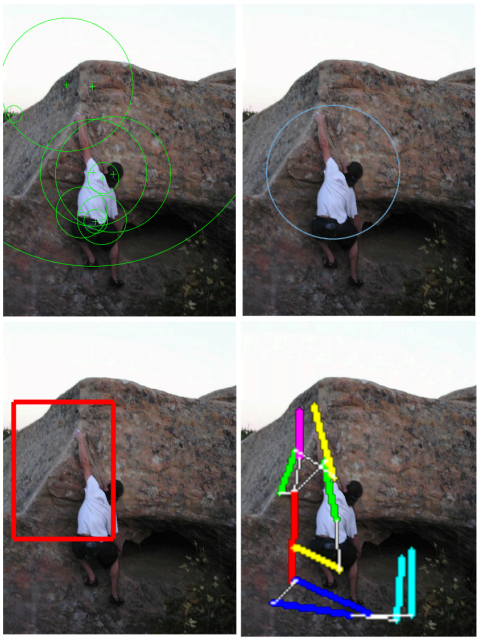

Fig. 7. Pipeline results for a sample image that fails to generate a bounding box that encapsulates the person.

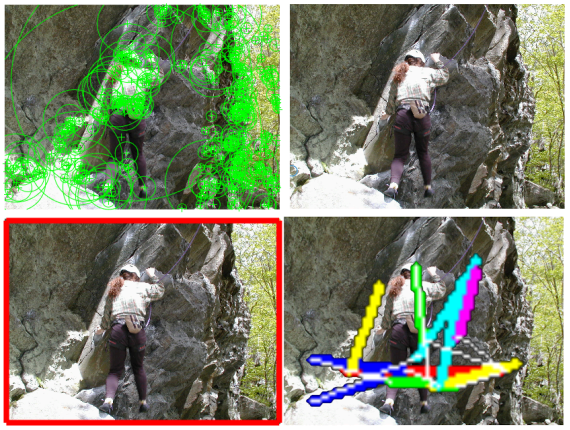

Fig. 8. Pipeline results for another sample image which encapsulates the entire image.

V. DISCUSSIONS

A. *Analysis*

Our results are unsatisfactory and do not meet the expectations of the human pose estimation task due to a variety of factors. In particular, our approach to person detection using SIFT features to create a bounding box did not result in successful detection. For our selected domain, utilizing the upper-body pose estimation features of the Calvin detector (which estimates the bounding box of the upper-body, subsequently fed into the pose estimation algorithm) proved marginally effective in the pose estimation task. See Fig. 9 for a comparison between the results of our method and the utilization of the Calvin detector, based on an image provided by the authors.

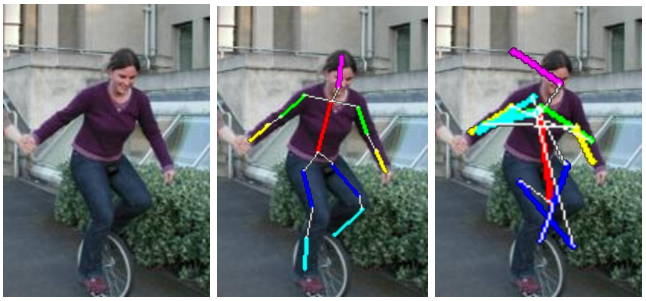

Fig. 9. a) Base image, b) Calvin detector & pose estimator results c) SIFT-features bounding-box and pose estimator results

Additionally, the bounding box derived from the SIFT feature clusters was not suited for the implementation of the PSM pose estimator, as it would contain, usually, the entire body of the person. The PSM pose estimator operates properly only when the bounding box contains the head and upper torso of the person. Fig. 10 is an example of the bounding box created using manual ROI selection and the result of the pose estimation resulting from that. Manual selection outperforms the SIFT clustering, as seen in Fig. 11. It is interesting to note that while the SIFT clustering implementation does not create an adequate bounding box for the pose estimator, the Calvin detector is unable to find a person within the sample image seen in Figs. 10 and 11.

Fig. 10. Manually defined bounding box

Fig. 11. Detected pose from the manually defined bounding box

The ESRGAN was believed to be able to fix this issue by restoring resolution to the head and shoulders and forcing the sift features to cluster and concentrate on the head and shoulder area. However, this does not work for complex images where rock formations are also prominent and create areas of interest.

Ultimately, our method fails to achieve proper pose recognition due to the fact that the pose estimation algorithm utilized relies on the usage of domain-specific models, that is, a pre-defined model of how a human body should look, which does not fit our domain of rock climbing images. The available models for full-body pose estimation of the algorithm are highly restrictive, only producing satisfactory results in images that contain forward-facing and upright humans.

Models for the algorithm that were better suited to our task were unavailable to us during the development of our pipeline. However, even with these models, the algorithm would likely have failed to achieve significantly better results due to the irregularity and complexity of the human poses present in rock-climbing images. Specific image characteristics that contribute to the failure are the orientation of the body itself, which varies significantly in rock climbing and is frequently non-forward-facing and non-upright, as well as the occlusion of the person's limbs in most scenarios.

However, our system's underperformance is not the fault of the methods that were utilized as part of the human pose estimation task. It is likely that we underestimated the complexity under the hood of these traditional methods and additionally overestimated their generalizability, leading to unsatisfactory results by default. Since these methods have largely gone dark over the past decade in terms of both new research and usage, it is difficult to know whether or not attempting to optimize or otherwise improve them in some way is a relevant task for researchers.

B. Unsuccessful Approaches

a) Haar cascade detector

The Haar cascade detector is a model available in MATLAB to detect bodies in photographs. It is a machine learning-based approach that is typically used for detecting faces. The cascade function is trained by using several positive and negative images. It can also be used to detect human bodies. The class used is vision.cascadeObjectDetector. It uses Haar cascades with the Viola-Jones algorithm. The model accepts the value of a classification model, which determines what feature you are looking for in the image. The options for this parameter include UpperBody and specific facial features. We tried to use it for detecting the lower body of the rock climbers even though it is not a pre-defined parameter. This was attempted by detecting the upper body by creating a boundary box around it. We then chose the boundary box with the largest area, which indicates that it is indeed the upper body. Based on the boundary box, we determine the lower half of it as the lower body, draw a boundary box around it, and visualize it. This method did not work as the main issue was that in most images, a human was not even detected in the first place, even though there was indeed a human body in the image. The second issue was that even when the human was detected, the box was more than often not drawn properly around the human.

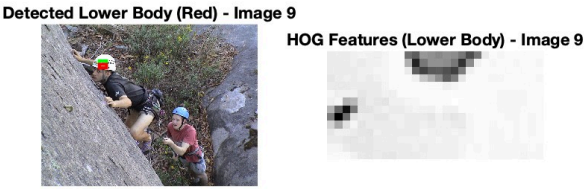

Fig. 12. Boundary box from Haar and HOG features of the box

b) Histogram of Gradients (HOG)

We also tried using Histogram of Gradients (HOG) for detecting the lower body in the images. There is a vision.peopleDetectorACF model in MATLAB, which uses Aggregated Channel Features to detect the human. Once we detected the lower body of the human in the image, we used HOG's feature descriptor to extract the HOG features. We used the extractHOGFeatures function from MATLAB to extract features from the images. However, just like Haar, it was not able to detect the humans at all, even though there was clearly a human rock climbing. In the off chance that it

did detect there was a human, it did not draw a bounding box anywhere in the image.

We believe that the Haar cascade classifier model and the ACF followed by the HOG pipeline are not working well with our images because the humans are in rock climbing positions, and the classifier works better with humans in normal standing positions. In rock climbing, humans tend to be in more dynamic and varied positions, which the estimator fails to detect.

c) Control point based rotation transform

We considered using control points to have the user select two points corresponding to the shoulders in order to rotate the image and place the body in a vertical position which the calvin pose estimator prefers. But besides not providing better or satisfactory results, it is not an improvement to the existing pipeline which is entirely autonomous.

d) Manual Region of Interest Selection

While not technically unsuccessful, selecting a manual region of interest provided better results than the Calvin upper body detector, it made the SIFT clustering technique, one of the few completed technical implementations, redundant/ unnecessary. We were unable to find a traditional computer vision technique that could be implemented within the period of the study that performed the task of segmenting the head and shoulders.

VI. FUTURE WORK

Based on our findings, we strongly believe that traditional computer vision techniques are fundamentally inadequate for the task of human pose estimation. Our experiments demonstrated that classic feature extraction methods, segmentation, and pictorial structure-based approaches fail to provide the robustness and adaptability required for accurate pose estimation across diverse environments and human subjects.

Through our literature review, we observed that even the traditional techniques in human pose estimation still relied on supervised learning models to refine their predictions. The Calvin detector, for example, incorporated pre-calculated parameters derived from an extensive dataset of labeled human pose images. This reliance on data-driven pipelines shows the need for machine learning, particularly deep learning, for this problem.

Based on our previous findings, one approach that may have aided the pose estimator in properly delineating the body parts could have been fine-tuning the ESRGAN model on images of humans rock climbing in various environments, positions, attire, and styles. Fine-tuning the pre-trained ESRGAN model on images that it would expect to see would help it upscale the resolution of those images in areas that are more relevant to pose estimation, i.e., the parts of the human body. The current ESRGAN model we used was pre-trained on a general dataset of images that included pictures of a wide variety, such as people, sculptures, animals, buildings, vehicles, fauna, and many others. These images are not specific to any application or field. Therefore, the model may have performed inadequately in upscaling images of humans rock climbing. Fine-tuning the model on these types of images may be beneficial in making the images sharper, allowing the SIFT feature extractor to detect more relevant human features and, hence, improving the performance of the overall pipeline and pose estimator.

Given that human pose estimation inherently depends on understanding complex spatial relationships and variations in human movement, approaches that rely solely on traditional computer vision are attempting to revive methodologies that were abandoned for a reason. These older approaches struggle with occlusions, variations in lighting, body shapes, and perspectives, making them ineffective for real-world applications.

While we acknowledge the value of traditional computer vision techniques in preprocessing, the core task of pose estimation requires models that can learn directly from data rather than relying on manually designed heuristics. Future work in this area should focus on utilizing deep learning architectures, such as convolutional neural networks (CNNs) and transformer-based models, which have demonstrated state-of-the-art performance in human pose estimation tasks.

Furthermore, a promising avenue for future research lies in hybrid approaches that combine the efficiency of traditional computer vision for preprocessing with the power of deep learning models for pose estimation. Exploring self-supervised learning or transfer learning techniques using pre-trained models on large-scale datasets could also enhance pose estimation accuracy while reducing data annotation costs.

Ultimately, we conclude that deep learning is a more viable path forward for accurate and reliable human pose estimation. Future research should prioritize exploring more efficient architectures, optimizing computational performance, and developing models that generalize well across different human subjects and environments.

REFERENCES

[1] H. Chen, R. Feng, S. Wu, H. Xu, F. Zhou, and Z. Liu, "2D Human pose estimation: a survey," *Multimed. Syst.*, vol. 29, no. 5, pp. 3115–3138, Oct. 2023, doi: 10.1007/s00530-022-01019-0.

[2] S. Dubey and M. Dixit, "A comprehensive survey on human pose estimation approaches," *Multimed. Syst.*, vol. 29, no. 1, pp. 167–195, Feb. 2023, doi: 10.1007/s00530-022-00980-0.

[3] J. K. Aggarwal and Q. Cai, "Human Motion Analysis: A Review," *Comput. Vis. Image Underst.*, vol. 73, no. 3, pp. 428–440, Mar. 1999, doi: 10.1006/cviu.1998.0744.

[4] Y. Yang and D. Ramanan, "Articulated Human Detection with Flexible Mixtures of Parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013, doi: 10.1109/TPAMI.2012.261.

[5] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3D Pictorial Structures for Multiple Human Pose Estimation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1669–1676. doi: 10.1109/CVPR.2014.216.

[6] S. Kulkarni, S. Deshmukh, F. Fernandes, A. Patil, and V. Jabade, "PoseAnalyser: A Survey on Human Pose Estimation," SN Comput. Sci., vol. 4, no. 136, 2023, doi: 10.1007/s42979-022-01567-2.
[7] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial Structures for Object Recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, Jan. 2005, doi: 10.1023/b:visi.0000042934.15159.49.
[8] V. Ferrari, M. Marín-Jiménez, and A. Zisserman, "2D human pose estimation in TV shows," in *Statistical and Geometrical Approaches to Visual Motion Analysis*, D. Cremers, B. Rosenhahn, A. L. Yuille, and F. R. Schmidt, Eds. Berlin, Heidelberg: Springer, 2009, vol. 5604, *Lecture Notes in Computer Science*. doi: 10.1007/978-3-642-03061-1_7.
[9] M. Eichner, M. Marín-Jiménez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images," *International Journal of Computer Vision*, vol. 99, pp. 190–214, 2012. doi: 10.1007/s11263-012-0524-9.
[10] X. Zhang, C. Li, X. Tong, W. Hu, S. Maybank, and Y. Zhang, "Efficient human pose estimation via parsing a tree structure based human model," *2009 IEEE 12th International Conference on Computer Vision*, Kyoto, 2009, pp. 1349-1356, doi: 10.1109/ICCV.2009.5459306.
[11] Lowe, David. "Distinctive Image Features from Scale-Invariant Keypoints". *International Journal of Computer Vision*. 2004. vol. 60. pp. 91-110. doi: 10.1023/B%3AVISI.0000029664.99615.94.
[12] M. Andriluka, L. Pishchulin, P. Gehler and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3686-3693, doi: 10.1109/CVPR.2014.471.
[13] M. Eichner, V. Ferrari, "Human Pose Estimation in Still Images V 1.05," *Calvin Research Group*, Accessed Feb. 2025, https://calvin-vision.net/bigstuff/calvin_upperbody_detector/.
[14] M. Eichner, M.J. Marín-Jiménez, A. Zisserman, and V. Ferrari, "2D Articulated Human Pose Estimation Software v1.22", *Visual Geometry Group*, Zurich, Accessed Feb. 2025, https://calvin-vision.net/bigstuff/articulated_human_pose_estimation_code.
[15] M. Eichner and V. Ferrari, "Better Appearance Models for Pictorial Structures" *British Machine Vision Conference* (BMVC), 2009, doi: 10.5244/C.23.3
[16] D. Ramanan, "Learning to parse images of articulated bodies," in Advances in Neural Information Processing Systems, vol. 19, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2006. doi: 10.5555/2976456.29765
[17] manoreken (2025). ESRGAN Single Image Super Resolution Matlab port (https://www.mathworks.com/matlabcentral/fileexchange/111175-esrgan-single-image-super-resolution-matlab-port), MATLAB Central File Exchange. Retrieved April 1, 2025.
[18] Flickr2K Dataset. Retrieved March 31, 2025. https://www.kaggle.com/datasets/daehoyang/flickr2k

VII. Appendix

A. Results

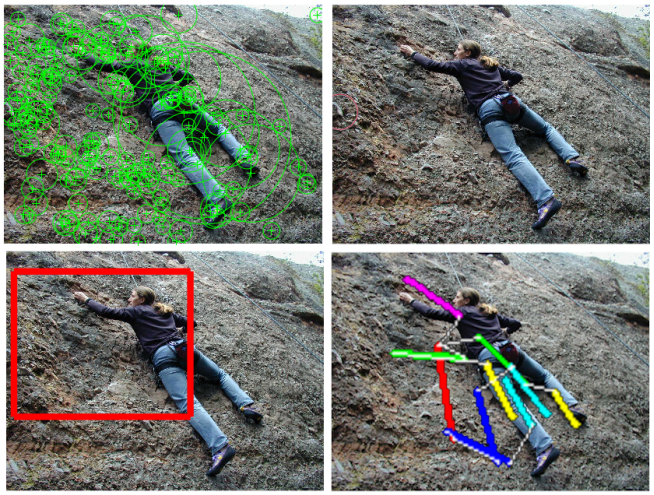

Fig. 13. A pose estimation result with the system's pipeline

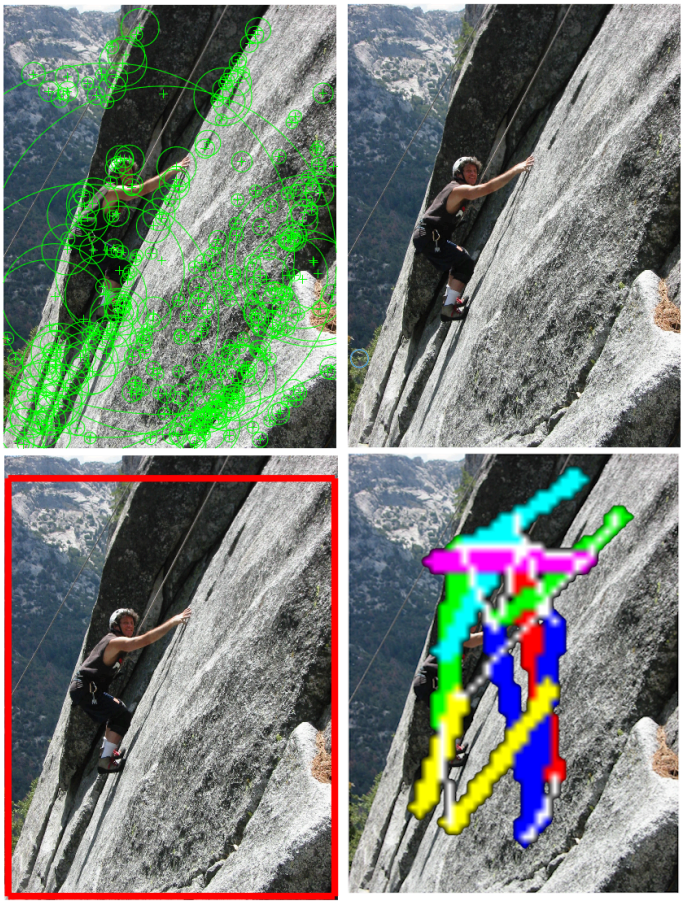

Fig. 14. A pose estimation result with the system's pipeline

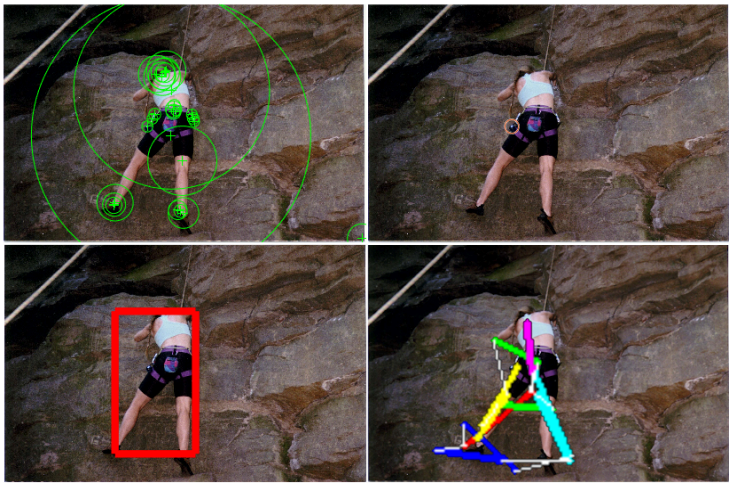

Fig. 15. A pose estimation result with the system's pipeline

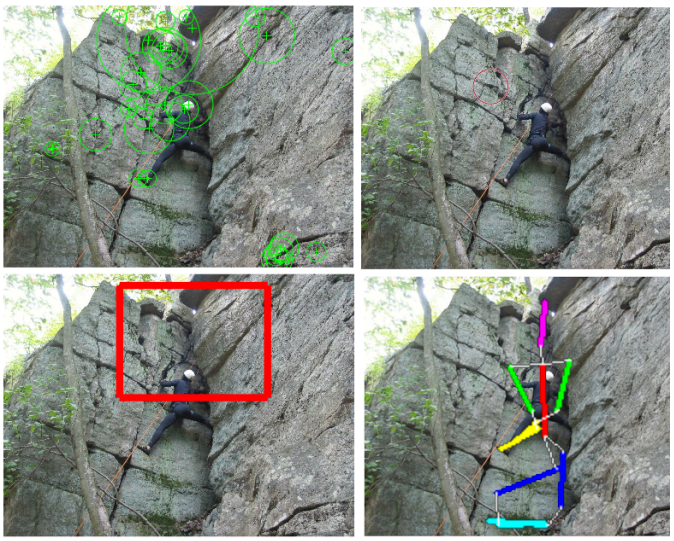

Fig. 16. A pose estimation result with the system’s pipeline

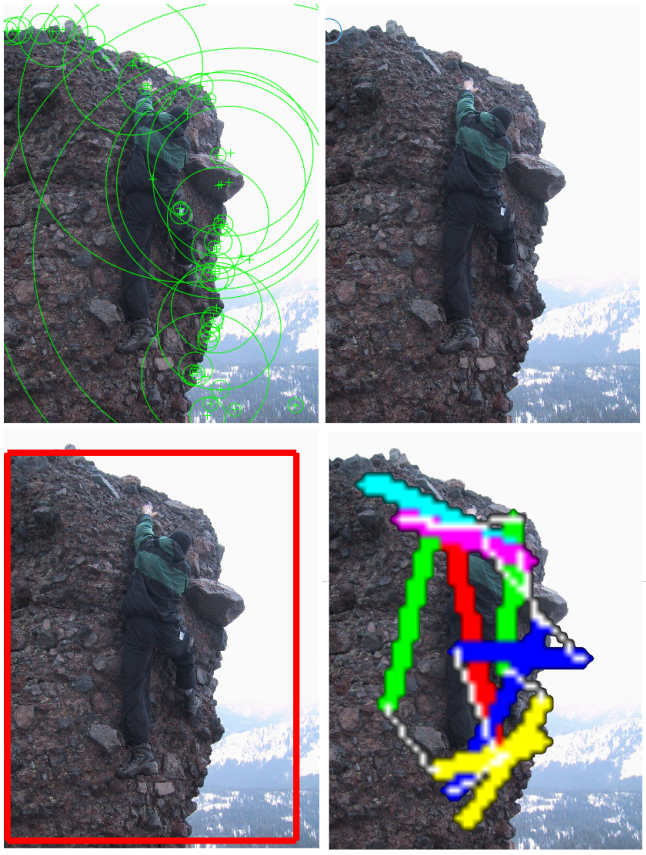

Fig. 16. A pose estimation result with the system’s pipeline