Computer Vision for Assisted Indoor Rock Climbing

Computer Science Expository Work - December 2023

Asad Khan
University of Toronto
Toronto, ON M5S 1A1
asadk.khan@mail.utoronto.ca

Laura Madrid
University of Toronto
Toronto, ON M5S 1A1
laura.maldonado@utoronto.ca

Lucas Noritomi-Hartwig
University of Toronto
Toronto, ON M5S 1A1
lucas.noritomi@utoronto.ca

Supervisor: Professor Lisa Zhang
University of Toronto
Toronto, ON M5S 1A1
lczhang@cs.toronto.edu

Repository

https://github.com/Laura05010/Computer-Science-Expository-Work.git

Abstract

In this report, we introduce the Indoor Rock Climbing Assistance Tool (IRCAT) to aid visually impaired individuals to perform indoor top-rope rock climbs. The system implements pose estimation and object detection for capturing each instance of the climbing route, and conveys distance information to the climber via audio feedback. The system also integrates audio input from the climber, enabling the system to dynamically switch between conveying the distance to the next hold for different limbs.IRCAT proves useful in a controlled setting, and with further development can become a helpful tool in various rock climbing scenarios. Our goal with this project is to help explore existing methods and implement a useful tool to give suggestions to visually impaired rock climbers.

1 Introduction

Indoor rock climbing originated in the 1970's. Since then, it has gained massive popularity by providing a physical and mental stimulus for individuals of diverse age groups and abilities. The world of indoor rock climbing is constantly developing and is particularly of interest to computer scientists due to its various aspects, inviting exploration into areas such as path generation, pose estimation, hold detection, and human-object interaction for research and study. In regards to finding rock holds, and inferring a climbing path, climbers heavily rely upon their sight. However, those who are visually impaired must rely on their other senses to perform the same tasks. Our project consisted of building IRCAT to effectively assist visually impaired indoor rock climbers by making next step suggestions during their top rope or auto-belay ascent.

2 Related Works

This literature review aims to summarize the current technologies available and approaches explored that were found to be useful in implementing the pose estimation, object detection, path generation, and feedback system components involved in IRCAT.

2.1 Pose Estimation Overview

As climbers ascend a wall, their body movements and positions, "poses", play a crucial role in determining their success and safety. In the context of rock climbing, a pose specifically refers to the position of a human subject's body, joints, and limbs at any moment in time relative to the climbing wall that they are ascending. Pose estimation uses video data and detection algorithms to create an accurate digital representation of a climber's body position in space. To develop a system for assisting visually impaired climbers, being able to accurately estimate these poses, even in instances where certain limbs and joints are occluded behind other body parts, is crucial.

To assess the current state-of-the-art in pose estimation technology, we look into the main components of what comprises an accurate and efficient pose estimation system. First, we look into the type of video capture methods that are being used. Traditional capture methods mainly comprise of RGB Video capture which provides a 2D representation of the environment. However LiDAR-based video capture provides a 3D representation of the environment, by estimating the depth of each point in space to reconstruct the climber's point cloud and their environment's point cloud in 3D-space. For higher-budget setups, IMU sensors can be used to provide real-time motion capture data of the climber's position and orientation.

As mentioned above, RGB video is a staple of video capture and is the most common amongst commercial recording devices which provides a 2D representation of the environment. Recently the emergence of depth-sensing technology has become prevalent amongst common recording devices Beltrán et al. (2022). While the upside of these video capture devices are evident due to their accessibility and ease of use, their limitations become evident as a greater distance between the camera and the subject are observed, as is the case with Beltrán et al. (2022) which saw poor results with the iPad Pro at 6 meters. The iOS documentation itself recommends a 5.5m maximum from the object, which is a cause for potential issues as climbers naturally move further away from the camera as they ascend up the wall, with shorter top rope climbing walls being between 25-30 feet Crux (2020). Implementing IMU sensors is not feasible due to cost constraints and the short comings with LiDAR-based video capture led to the conclusion that the best option in the case of rock climbing was to stick with the reliable technology of RGB video capture with better performance at greater distances.

2.2 Hold Detection

To support visually impaired rock climbers, the tool needs to identify the intended climbing route and track the positions of holds. Indoor climbing gyms typically mark routes with rock holds of matching color, texture, or size. While blind climbers rely on touch to discern routes through hold shapes and textures, our camera relies solely on pixel-based information.

In their study, Ayesha Arif and Kim (2021) showcased superior performance in human-object interaction (HOI) detection, surpassing prevailing methods by integrating techniques from pose estimation and object detection. This amalgamation involved combining K-Means clustering and YOLO alongside innovative methodologies like Fuzzy C-Means for super-pixels and Random Forests for object segmentation. Moreover, Sarah Ekaireb and Manjunath-Murkal (2020) demonstrated remarkable achievements through the training of a neural network using a Roboflow model, achieving an impressive Mean Average Precision (mAP) of 96.3% in holds detection and segmentation. Further advancements involved training a neural network for multi-class classification, distinguishing different hold colors to delineate separate climbing paths, yielding an accuracy of 99.3%. Notably, challenges arose when similar colors—such as green and turquoise—were occasionally misidentified as part of the same path, as detailed in the paper.

From the findings in Ayesha Arif and Kim (2021), we opted to integrate YOLOv8 for robust object detection. Additionally, drawing from the insights in Sarah Ekaireb and Manjunath-Murkal (2020), we decided to use the CV2 HSV library for precise color identification, bypassing the need for an extra color detection model.

2.3 Audio Feedback Systems

The essence of real-time feedback in climbing motion analysis lies in its ability to offer climbers instantaneous insights into their movements, techniques, and interactions with the environment.

Such feedback, when optimized, can significantly enhance the climbing experience, ensuring safety, improving performance, and aiding in technique refinement.

One of the first approaches established was by *MetaHolds: A Rock Climbing Interface for the Visually Impaired* Ilich (2008) paper which focused on auditory aids for those with varying degrees of vision loss. Speakers were placed on each hold to give the climber information about the type of hold, the positioning of the hold and the route it is located on. The climbing holds were categorized by 5 different types (pinch, crimp, sloper, pocket, jug) and each type had a specific sound effect. The climber would wear a sensor on top of their helmet that would check the positioning of the head and once the head was placed in front of the hold, the speaker for that hold would be activated. After the experiment, various participants suggested that the sound of the tone should vary based on distance and it was noted that sometimes participants forgot which holds the jugs meant. Ilich (2008) also found that verbal affirmation should only be given when the climber is on the correct hold and has correctly handled the hold based on its type. Overall, the participants did not feel distracted by the audio and understood its clear correlation to the 5 categories of holds.

Drawing from Ilich (2008), IRCAT aims to convey precise hold indications along the climbing route, implementing auditory feedback that adjusts pitch according to the distance between the climber's limb position and the target rock hold.

3 Methods

Building upon the insights and advancements detailed in the related work section, we now delve into the methodology employed in our project. IRCAT had three main components: human body pose estimation to keep track of the climber's pose along the climbing wall, rock hold object detection and colour detection for identifying the rock holds pertaining to the climber's route, and audio feedback as the method of interaction used between the climber and the system.

3.1 Human Body Pose Estimation

We decided to use the MediaPipe Framework to get the climber's pose. MediaPipe's pose object is capable of detecting 32 landmarks in the body, however our tool only uses the landmarks pertaining to the hands and feet.

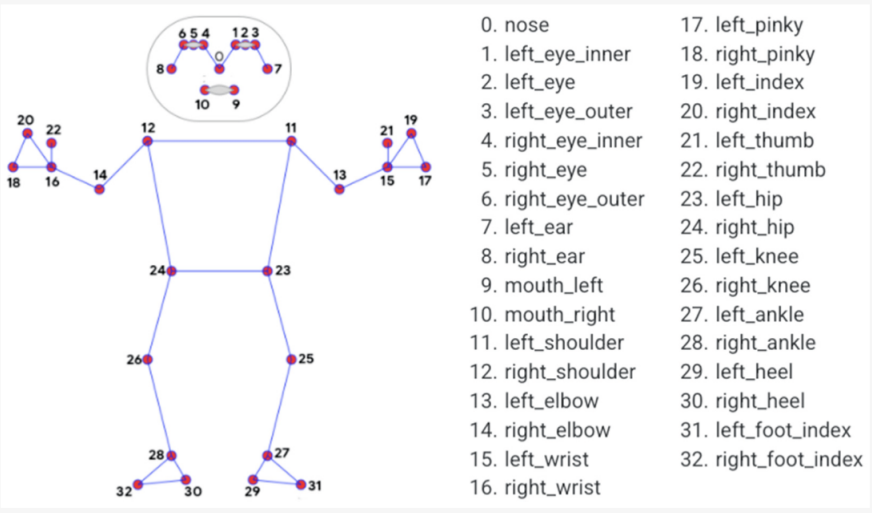


Figure 1: MediaPipe pose landmarks

For the following limbs we took the following landmarks:

Limb	Right foot	Left foot	Right hand	Left hand
Landmarks	"right_ankle", "right_heel", "right_foot_index"	"left_ankle", "left_heel", "left_foot_index"	"right_pinky", "right_index", "right_thumb", "right_wrist"	"left_pinky", "left_index", "left_thumb", "left_wrist"

We aimed for a singular representative point for each limb, selecting the centroid of the polygon formed by the chosen landmarks to fulfill this purpose.

3.2 Rock Hold Object Detection

For YOLOv8 model training, we utilized the Roboflow dataset from the Climbing Hold Detection Computer Vision Project, comprising 404 annotated images depicting various hold types (Crimp, Jug, Pinch, Pocket, Sloper). While our focus didn't involve the hold types, the model demonstrated proficiency in detecting diverse hold variations, though it encountered challenges with very small foot holds.

To provide real-time guidance, we utilized the iPhone 13 for camera capture. During live feed processing, computations were performed on each frame, based on the previously mentioned aspects of the project. Initially, our simultaneous implementation of hold detection and pose estimation caused route instability. To resolve this, we introduced a calibration step, involving a second climber (often the belayer), to update and approve the route before the climb commenced. Throughout the 30-second calibration period, the YOLOv8 model continuously generates predictions through the `calibrate_holds` function. We selectively utilize detections with a confidence threshold exceeding 75%, employing cv2's box annotator to mark the identified holds. Concurrently, we integrate route detection using nine distinct color masks. These masks are individually crafted by defining Hue, Saturation, and Value ranges for designated colors:

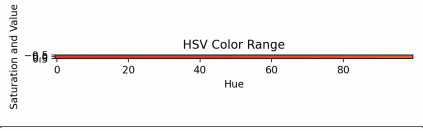
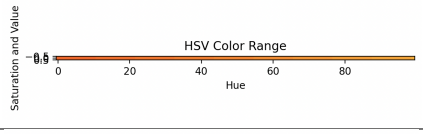
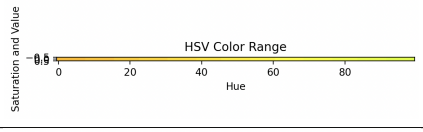
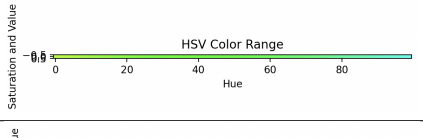
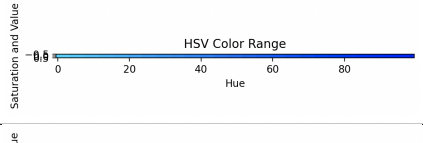
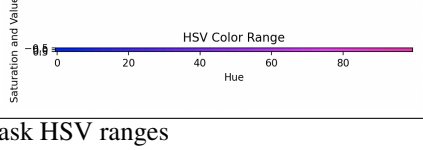
Colour	**Lower to Upper ranges**	**Visual representation of ranges**
Red	[0,100,100] to [10,255,255]	
Orange	[11,100,100] to [20,255,255]	
Yellow	[22,100,100] to [35,255,255]	
Green	[40,50,0] to [88,255,255]	
Blue	[95,50,50] to [120,255,255]	
Purple	[121,50,50] to [160,250,250]	

Table 1: Colour mask HSV ranges

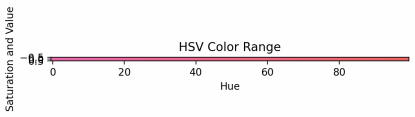
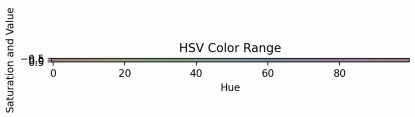
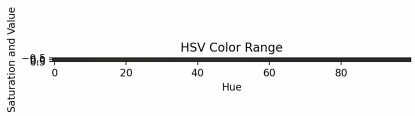
Colour	Lower to Upper ranges	Visual representation of ranges
Pink	[165,50,70] to [180,160,250]	
White	[0,0,100] to [180,40,155]	
Black	[0,0,0] to [180,40,50]	

Table 2: Colour mask HSV ranges continued

The detected holds are sorted based on these color masks, or given the uncoloured category if no mask corresponds to the detection in the `identify_routes` function. Subsequently, they are organized according to their respective colored routes, culminating in the maximum number of identified routes by the end of the calibration process.

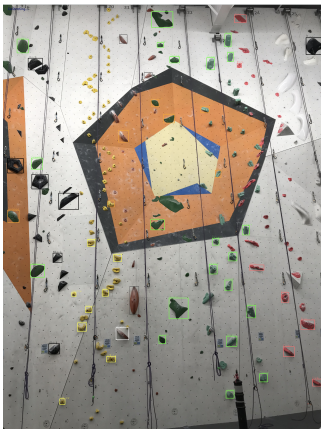
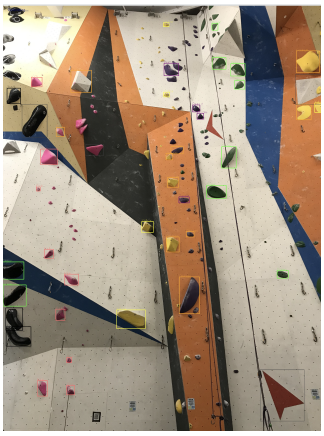
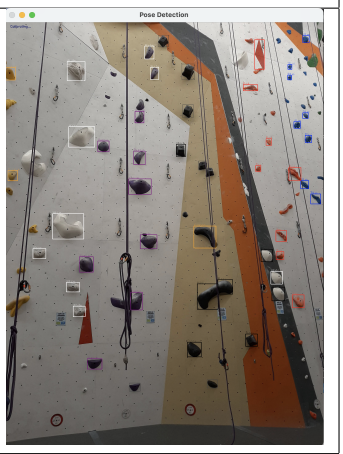
Wall 1	Wall 2	Wall 3
		

Table 3: Examples of detected and annotated walls after calibration

As previously stated, the model exhibited limitations in accurately detecting footholds, prompting the need for user input to both update and select the climbing route. After the calibration stage, the user is prompted to select one of the possible routes on the wall based on the colours detected through the `add_detections` function. This can be seen in Figure 2.

Subsequently, the user can augment the climbing route by selecting holds that haven't been annotated. Upon clicking on a particular hold, a turquoise box representing the average size of holds along that route will appear.This can be seen in Figure 3.

Upon completion of this task, the user is directed by the `remove_detections`function to remove holds if needed, by clicking within the vicinity of an annotated hold displayed on the screen. These holds are marked in magenta, signifying their intended deletion. This can be seen in Figure 4.

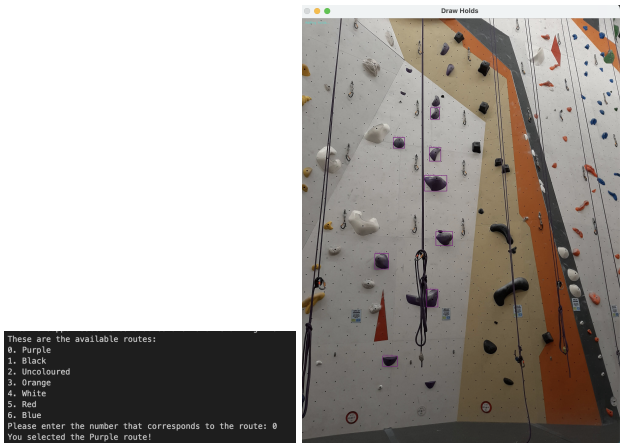


Figure 2: Available routes from Wall 3 & Selected route

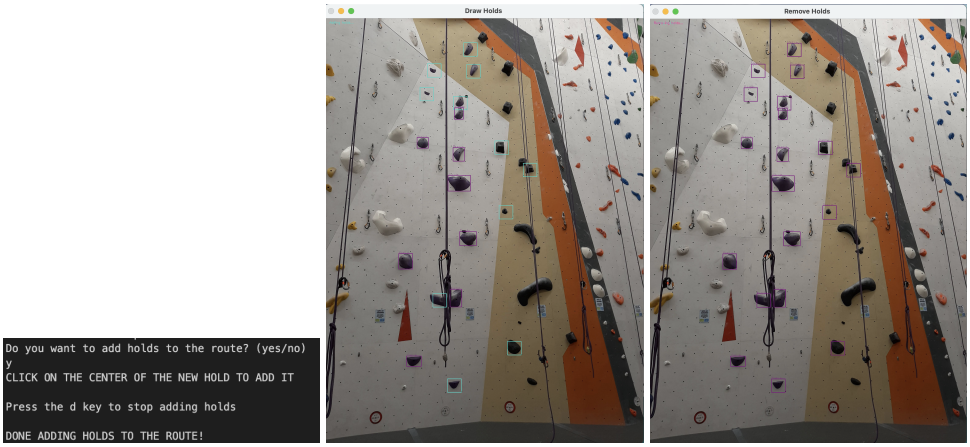


Figure 3: Add prompt & User selected holds & Updated route

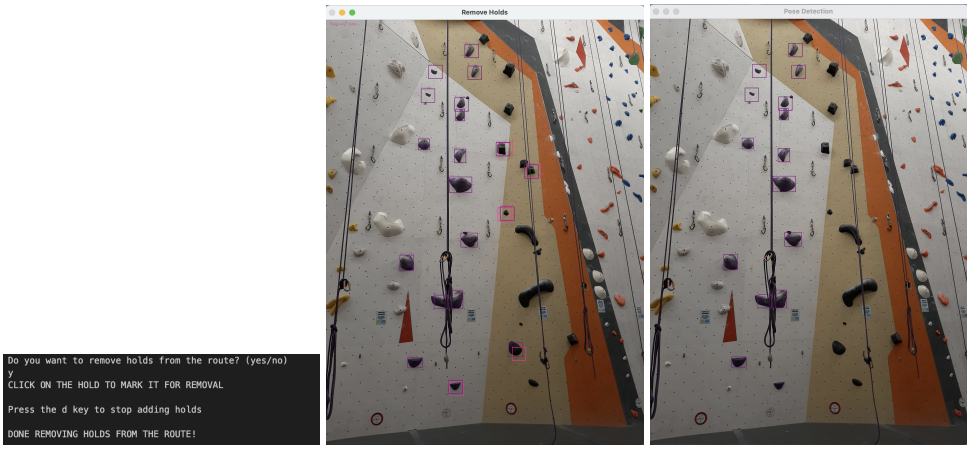


Figure 4: Remove prompt & User selected holds & Final route

3.3 Path Generation

The path generation component of IRCAT is a critical feature, facilitating a sequential identification of holds for the climber as they make progress up the rock climbing wall. Upon successfully grab-

bing a hold, the system employs a series of functions to identify the next target hold, prioritizing proximity to the current position of the limb.

When a climber grabs a hold, the IRCAT system initiates a sequence to determine the next target hold. This sequence involves the following steps:

1. **Detection of Hold Grabbing:** The function `check_grab_hold` calculates the distance between the climber's limb and the selected target hold over a period of 3 seconds. If the distance remains within a distance threshold of roughly 50 pixels (We use pixel measurements to gauge distances in our system), the hold is considered close enough to have been grabbed.
2. **Finding the Closest Next Hold:** Once a hold is grabbed, `find_closest_hold` is triggered. This function iterates through all detected holds, excluding those already grabbed, to find the nearest hold relative to the current position of the limb. It calculates the distance between the limb and each potential hold, selecting the one with the minimum distance as the next target.
3. **Updating Target Hold:** The identified closest hold then becomes the new `TARGET_HOLD`, updating the system's focus for the climber's next move.

The path generation process works in tandem with the pose estimation and audio feedback systems. Pose estimation tracks the climber's movements, providing real-time data to the path generation component. Subsequently, the audio feedback system communicates the next target hold to the climber, using changes in pitch to indicate the proximity to the target.

3.4 Audio Feedback and Interaction

After the climber's pose and the rock holds have been detected, the system must convey how to perform the following move to the climber in a non-invasive, non-distracting, and intuitive manner, as well as take in live audio input from the climber to allow for adjustments to be made to the system during the climb.

To ensure that the feedback system is non-invasive to the climber, we implement it solely using audio, which requires minimum equipment (only earbuds), and whose volume can be adjusted to the comfort of the climber. The feedback system consists of only simple sine waves so that if others are speaking to the climber, the climber must only discern the words being spoken through the sine waves, instead of more complex signals.

To indicate to the climber where to reach for the next rock hold, we implement a sequence of sine waves whose period (or pitch) depends on the L^2 distance of the limb from the target hold given the x and y coordinate distances in the function `play_distance`. The pitch of the sound, otherwise the period of the sine wave, is computed using a variation of the 2-dimensional exponential function, and using the frequency of the note C-natural, $c = 261.63$ (the baseline), and tuned parameters:

$$f(x, y) = c \cdot \left(b \cdot e^{-a \cdot \left(x^2 + y^2\right)} + 1 \right)$$

where $a = 0.083666$, $b = 3$, and c is the frequency of C-natural. We choose the note of C-natural as it is a relaxing tone that is less likely to agitate the ears when played often. Below we include a plot of the curve, where the center of the rock hold is imagined to be at the origin.

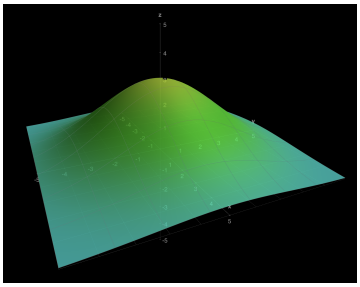


Figure 5: Plot of frequency function

Similar to a metal detector, the pitch represents the proximity of the limb to the target hold. This simple design is intended to be as intuitive for users, requiring the least amount of training before usage.

This implementation uses the sounddevice python library, which offers an easy method for constructing a sine wave given a computed frequency. A separate thread waits for the coordinate distances to be sent every 5 frames, and then calculates the sine wave and plays it for a short interval.

During the climb, the climber may want to switch the limb that the system is using to target the next rock hold. This may be to gain a better position before performing the next larger step towards completing the climb. In any case, the system must easily take input with little effort from the climber, and be specific enough not to easily falsely assume an input. An initial thought was to detect a gesture on whichever limb is chosen. However, this would require that the chosen limb was not occupied with maintaining a hold. It would also run the risk of falsely classifying a regular movement as a gesture, which would cause confusion to the climber.

It was found that, similar to system feedback, input from the climber would best be taken in the form of audio using the microphone built into the earbuds, instructing the system on which limb to focus. This system used the audio_feedback and speech_recognition python libraries, and depended on Google's Web Speech API to convert audio to text for keyword searching.

4 Results

This system was tested with the help of Professor Daniel Zingaro. During testing, we found that IRCAT was able to detect the pose of the climber with high accuracy, and was even able to maintain the structure and direction of the pose when the body was turned at various angles, and when certain parts were occluded by either other body parts or other objects.

IRCAT was able to accurately detect rock holds along the climbing wall, and classify them by colour accurately. There were instances where IRCAT would face some difficulty with detecting the smaller footholds when the camera was placed at a reasonable distance for capturing the entire climbing wall. However, in these instances, the manual selection and deselection feature worked very well to refine the climbing route detection. The path generation component of IRCAT performed well during testing. It was able to easily tell the system which rock hold along the route was the next target after waiting for the 3 seconds where the extremity of focus is close enough to the current target hold.

The audio feedback of IRCAT performed well, and was able to intuitively convey the distance to the next hold to the climber. An issue faced was that the sound emitted was discrete and, when played quickly enough to match the speed of the frames, would end up sounding like clicks. This was mediated by only processing the sound every 5 frames. Even still, the discrete sine waves would have been more useful if made continuous, making the change in pitch more clear to the climber. Another issue found was the audio playback would be sporadic when connected to earbuds, likely due to the many sine waves being sent at a rapid pace over bluetooth.

For system adjustment via audio input from the climber, it was found that the accuracy of IRCAT was high, though its reliability in recognizing voice was low and would require the climber to speak loudly into the microphone to be detected. The system's reliability was significantly reduced further when the audio input/output was connected to the earbuds of the climber rather than the computer speaker.

5 Conclusion and Discussion

We implemented a prototype of IRCAT, which was able to detect rock holds and classify routes, estimate the human body pose, and convey to the climber the distance to the next holds on the selected climbing route.

The pose estimation component of IRCAT performed very well for its intended purpose. However, more work is encouraged to improve its stability when detecting poses. Another aspect that could be improved upon would be to have the "hand" area be calculated from the center of the palm instead of from the center of the entire hand, as rock holds are more typically grabbed by closely matching the center of the hold with that of the palm.

The rock hold object detection system also performed adequately, and the drawbacks mentioned above are able to be addressed by having a more encompassing dataset, including various sizes of rock holds, especially smaller ones, to allow for better detection. We also encourage more work to be done on the calibration process, which can be made to develop a more stable picture of the climbing route by accounting for percentage of frames detecting an area as a rock hold. Another improvement would be to allow for the camera to move during the climb, thus requiring a dynamic detection system to be implemented, while also including a memory of previous rock holds that may be occluded at the current frame.

Future work is encouraged on developing heuristics for path generation, and hopefully taking into account the different levels of reach for different types of limbs (arms vs legs). A more ambitious improvement would be to have a method for IRCAT to decide which limb to switch to depending on the pose of the climber as well as the rock holds near specific limbs. When no obvious "best limb" is detected, the system could revert back to the standard "right hand", "left foot", "left hand", "right foot" pattern.

Further work is needed to improve the audio feedback and interaction system. While the accuracy of the Google Web Search API is helpful, a system such as Hugging Face Speech2Text would be more desirable. The issue faced with implementing Hugging Face was that, by design, it takes in existing audio recordings, and is difficult to implement with live audio. More work on this aspect is encouraged. Feature-wise, other refining such as indicating when a hold is achieved, and when a route is complete would also be helpful for the climber as it may seem like a bug when the pitch suddenly changes, when in reality the limb of focus has switched.

IRCAT has proven to be a useful tool to help those who are visually impaired enjoy the sport of rock climbing. While the sport alone is not a crucial aspect to all in everyday life, it offers a playground where researchers can explore various different types of actions and interactions that may be seen in other areas of life in a single setting. It is our hope that the insight for this project is used to help implement other tools to better assist the visually impaired in everyday life.

References

Mohammed Alarfaj Ahmad Jalal Shaharyar Kamal Ayesha Arif, Yazeed Yasin Ghadi and Dong-Seong Kim. 2021. Human Pose Estimation and Object Interaction for Sports Behaviour. (2021), 1–18. https://digitallibrary.aau.ac.ae/bitstream/handle/123456789/693/Human%20Pose%20Estimation%20and%20Object%20Interaction%20for%20Sports%20Behaviour.pdf?sequence=1&isAllowed=y

Raul Beltrán Beltrán, Julia Richter, and Ulrich Heinkel. 2022. Automated Human Movement Segmentation by Means of Human Pose Estimation in RGB-D Videos for Climbing Motion Analysis. In *VISIGRAPP*. https://api.semanticscholar.org/CorpusID:246859460

Conquer Your Crux. 2020. How high are climbing walls on average? https://www.conqueryourcrux.com/how-high-are-climbing-walls-on-average/

Michael Ilich. 2008. MetaHolds: A rock climbing interface for the visually impaired. *Ilich M* (2008).

Prem Pathuri Priyanka Haresh Bhatia Ripunjay Sharma Sarah Ekaireb, Mohammad Ali Khan and Neha Manjunath-Murkal. 2020. Computer Vision Based Indoor Rock Climbing Analysis. 1 (2020), 1–17. https://kastner.ucsd.edu/ryan/wp-content/uploads/sites/5/2022/06/admin/rock-climbing-coach.pdf