WeMoove Final Report

Anna Guo Jenna Jacob Katie Wang Natalie Leal Blanco
UT Austin: Computer Vision Course
annayuab@icloud.com jenna.jacob@utexas.edu katiewang@utexas.edu natalie.leal@utexas.edu

Abstract

For many individuals facing health hurdles, from common ailments to severe injuries, physical therapy represents a beneficial routine that provides a tangible path to recovery. Because adherence to prescribed recovery plans is critical, we present a novel application of computer vision techniques to supplement the traditional physical therapy experience. Whereas many recent attempts to integrate technology have involved passive data-logging or costly equipment, we aim to provide a cheap but interactive experience revolving around two key features: 1) offer quantitative feedback by scoring patients' poses and movements against their exercise videos and 2) empower users to curate personalized therapy routines by filtering a database of physical therapy videos with specific poses. In providing a measure of progress and facilitating selection of progressive exercises, we hope to encourage adherence to therapy plans and foster accelerated healing.

1. Introduction

Over 50 million Americans seek physical therapy every year, whether that be from musculoskeletal conditions, car accidents, or sports injuries. Evidently, there is a large need for physical therapy[9]. However, only 35% of PT patients actually adhere to their assigned plan[11]. This is a significant problem since the majority of patients are not following through with their exercises and thus decrease the effectiveness of their therapy plan.

We aim to use Computer Vision to partition each exercise movement into more easily digestible chunks by creating a model that can learn (vectorize) each exercise movement from two given videos, an instructional one and a patient one, and compare the two. By developing a tool that provides feedback as a result of analyzing the two videos, which ideally should match perfectly, adhering to an exercise plan becomes much more feasible, enjoyable, and even motivating.

2. Literature Review

Because patient adherence to physical therapy is such a large problem in the healthcare industry, we aim to improve this by assisting physical therapy patients through computer vision techniques. This will positively impact their wellbeing and treatment success rates. Our review explores strategies to enhance adherence, covering existing attempts, key techniques, challenges, and the technical gap. From VR to advanced computer vision, our examination aims to make physical therapy more accessible and effective.

2.1. Existing Attempts

Several existing attempts have been made to tackle the challenge of improving patient adherence, each with its own approach and limitations. One such approach involves utilizing monitoring technology, such as pedometers, which have proved to be user-friendly. Though such monitoring technologies are able to monitor, or track, the patients' physical activity decently well, there was not much of an effect in terms of improving their adherence to their exercise plans[1]. As a result, patients using such technologies still struggle with adhering to their exercise plans.

Another attempt that has been made is virtual and augmented reality treatment. Through this method, patients are able to put on their VR headsets and transport into any environment they would like to complete their exercises in. The possibilities are endless: they could surround themselves in peaceful, serene nature, or they could opt for a more competitive gaming atmosphere. Physical therapists have found that this actually makes their treatment plans more appealing and thus improves their patient adherence rates[8]. However, it is important to note that this is significantly more expensive. The cost associated with this method is typically too high for the majority of patients and thus makes this approach less practical.

Unlike the aforementioned methods, our proposed method is low-cost, accessible, and effective. Implementing the aspect of gamification allows for physical therapy to be fun yet productive. Patients become competitors, working

hard to beat their previous scores. Additionally, our method could be expanded to offer further personalized feedback based on the analysis of the learner's own video and the given instructional video, ideally without the requirement for perfect lighting or camera angle conditions.

2.2. Useful Techniques

There are many existing computer vision techniques that can be applied to creating an aid for learning exercises. One such technique involves exploring 3D models instead of traditional 2D skeletonization methods to create more comprehensive representations of the exercise movements. Unlike 2D skeletonization, which captures only the spatial coordinates of key body joints in a single plane, 3D modeling techniques can capture the full spatial extent of exercise movements, including depth information[15]. This allows for a more accurate and detailed representation of the individual's body and movements in three-dimensional space, which is particularly valuable for capturing complex and multi-dimensional exercise movements[16].

In addition to 3D modeling, other computer vision techniques such as action recognition, pose estimation, trajectory analysis, and video segmentation and tracking can provide valuable insights into the dynamics of exercise movements.

1. Action recognition algorithms can automatically identify and classify different gestures and even exercise sequences, allowing for automated analysis of exercise patterns and styles.
2. Pose estimation techniques can accurately estimate the skeletal poses of people from video frames, providing a structural representation of their movements that can be used for further analysis and interpretation.
3. Trajectory analysis techniques can track the trajectories of specific body parts or exercise movements over time, allowing for the analysis of exercise movement patterns, velocities, and accelerations[2]. This provides valuable insights into the dynamic aspects of these exercises, helping learners to better understand the timing and pacing of the movements.
4. Video segmentation and tracking methods also serve to segment the videos into individual sequences or movements, allowing for more focused analysis and comparison of specific elements.

By leveraging these advanced computer vision techniques, learners can gain a deeper understanding of the dynamics and nuances of exercise movements, facilitating more effective learning experiences. These techniques can help learners to break down complex movements into smaller, more manageable components, identify areas for improvement, and refine their technique with greater precision and accuracy. Overall, the integration of computer vision into physical therapy has the potential to revolutionize the way patients approach their recovery plan, making it more accessible, engaging, and effective for patients of all levels.

2.3. Challenges

Several challenges arise from attempting this project. One is the quality of input data, which encompasses various factors such as lighting, angles, occlusions, frame rates, and blurring. These factors can significantly impact the accuracy of movement analysis, as they can obscure or distort key visual information needed for understanding exercise movements. Addressing these challenges may require the development of robust computer vision algorithms capable of handling noisy or low-quality video data, as well as techniques for preprocessing and enhancing the visual information to improve analysis accuracy.

Moreover, handling interactions between multiple persons in an input video presents additional complexities in accurately capturing and representing exercise movements. In some videos, there may be two individuals present that interact with each other in complex ways. It is not uncommon for one to assist the other in certain movements, for example, by helping lift up or stretch their legs. Capturing these interactions and accurately representing them in the analysis poses challenges for computer vision algorithms, as they must be able to differentiate between individuals and track their movements independently while also accounting for their interactions with others. Addressing this challenge may require the development of advanced tracking and segmentation techniques capable of accurately identifying and tracking multiple persons in a dynamic environment.

Overall, addressing these challenges requires a multidisciplinary approach that combines expertise in computer vision, exercise analysis, and human-computer interaction. By developing robust algorithms, incorporating domain-specific knowledge, and leveraging advanced techniques for handling complex visual data, it is possible to create a tool that effectively supports learning exercises from videos using computer vision. However, continued research and development efforts are needed to overcome these challenges and realize the full potential of computer vision technology in the field of physical therapy.

2.4. Technical Gap

The primary aim of this project is to reduce the technical gap between physical therapy videos and effective recovery plans. By enabling the correspondence between individuals with different body proportions captured from various angles, this tool seeks to enhance the accessibility and inclusivity of these exercises plans. This interdisciplinary endeavor involves collaboration between experts in physical therapy, biomechanics, and computer vision, with computer vision providing quantitative analysis capabilities to enhance the understanding and teaching of exercise move-

ments.

In conclusion, developing a tool for learning exercises from physical therapy videos using computer vision holds great potential for revolutionizing physical therapy plans and making them more accessible and doable to a wider audience. By leveraging advanced computer vision techniques and interdisciplinary collaboration, this project aims to bridge the gap between physical therapy videos and effective recovery plans, ultimately empowering individuals to learn and appreciate physical therapy in new and innovative ways.

3. Methods

We needed to accurately detect keypoints to extract poses and gestures from, and to do this we needed a model that could help us achieve that.

3.1. Keypoint Detection with OpenPose

We had three main criteria when selecting an existing library with which to detect body parts.

1. The library had to store model outputs in an easily readable, non-proprietary format, such as JSON, so that we could easily work with it.
2. The model must perform reasonably well on a CPU. When we began our project, we were unsure of how to secure GPU access, and we needed a model that could provide meaningful outputs even if we failed to set up a system for using GPUs.
3. Ideally, the model possessed capability to detect multiple people in a given frame. Although we did not prioritize our product's ability to process videos with multiple people, we wanted the flexibility to add that functionality later without a potential overhead from switching our library of choice.

With the above considerations in mind, we evaluated MediaPipe, MoveNet, and OpenPose.

We first explored MediaPipe, a versatile framework by Google that supported vision, natural language, and audio tasks. With its lightweight design for on-device ML and target towards time-series data, we found its performance promising. However, it appeared to only support single-person pose detection, so we set it aside early on to spare our future selves the pain of transitioning to a different library.

We also investigated Tensorflow's MoveNet, incidentally also part of Google. Built upon MobileNet, it had a satisfyingly low resource consumption. However, with the initial configurations that we had tested, it seemed to have poor accuracy, with keypoints often hovering in the general region of the desired body parts, but not truly aligning with the body joints. Because physical therapy patients may have injuries that cause lower tolerances for certain movements, we decided against using MoveNet.

Finally, we decided to try OpenPose. We quickly learned of its ability to output key points from each frame of a video as a JSON file. Additionally, not only does OpenPose enable multi-person pose detection, but it also appeared to support 3D keypoint detection when two camera perspectives were provided. While we did not plan for the possibility of 3D keypoint detection, we were intrigued by the potential, and believed its progress to be evidence that OpenPose was still being maintained. Hence, we chose OpenPose for keypoint detection.

While OpenPose does have a higher performance overhead than the other two models, we consider it a worthy trade for the improved accuracy and the increased options for input videos (many of which may have more than one person). In fact, while testing OpenPose's hyperparameters, we learned that lowering the video resolution caused a significant speedup in processing time, but we were ultimately against reducing the resolution in our input videos because our project performance doesn't depend on low latency, but it does benefit from higher accuracy.

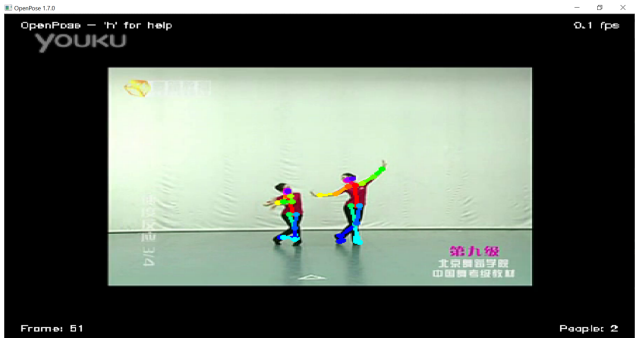

Figure 1. OpenPose in action: pose detection on a frame from an input video of two people. OpenPose accurately detects the key points on both people despite their slightly unorthodox body positions.

3.2. Feature Transform

For a given video, OpenPose outputs a directory of json files, with one for each frame. Before passing these outputs to a second machine learning model, we wanted an algorithm that could map similar poses and movements to the same values, regardless of camera angle or relative location of the person to the camera. (Due to the nature of OpenPose outputting x, y values for keypoints, we have already circumvented the issue of changes in illumination.) Our keypoint-to-angle algorithm aims to:

- Greatly reduce the role of absolute pixel locations in identifying features
- Make the model invariant to differences in the user's body structure
- Detect straight-line, elliptical, and hybrid motion

We choose to transition away from absolute positions due to inspiration from SIFT's usage of illumination gradients over absolute color values. Absolute positions could

cause our model to unintentionally train on an individual's location in the camera view., y-positions. Similarly, we hoped to make our model invariant to the user's height and distance from the camera (people in the foreground tend to appear taller than people in the background).

Thus, we decided to add some math equations to transform our features from absolute x, y positions to angles between detected key points. Additionally, this will likely aid the model in recognizing elliptical motions, which are prevalent in physical therapy exercises, as a continuous motion, as opposed to a given, singular elliptical motion being calculated as multiple straight-line movements. Additionally, this encodes information about body parts relative to each other. For instance, moving one's right hand downwards on the left side of one's body could be differentiated from moving one's right hand downwards on the right side of one's body through the angle of the right elbow relative to the right shoulder.

3.3. Pose and Gesture Extraction

By the temporal nature of videos and humans' inability to teleport, many frames have body angles that are close in value to those of the neighboring frames, similar to how in images, many pixels may have color values close to their neighbors. It would be inefficient to always work with all of the frames in a given video. In fact, in images, features are often only extracted for corners. We suggest the idea of extracting poses and gestures from videos.

While the algorithm we've designed is inspired by the Harris Corner Detector, poses and gestures differ from corners in that we're not looking for a gradient across all dimensions. Intuitively, this makes sense: even if a person were to stand-in place (not change the angles in their lower body), we would still consider a hand-wave to be a gesture, and raising different hands would be considered different poses.

To reiterate, unlike a corner, which only has two dimensions x and y, a pose or gesture would have as many dimensions as there are angles. Additionally, while corners are detected based on the ratio of the eigenvalues, we're more concerned with having one strong edge or many weak edges in identifying a gesture: requiring all angles to change would omit basic gestures like a simple hand wave because the majority of angles in the body wouldn't change. Hence, we use Euclidean Distance as a "change in body position" score.

Another point of interest is that poses would have a low "change in body position" score. Going back to the corner detection analogy, poses are analogous to a flat area within an image, yet they arguably define a person's physical actions more than the brief moments of highest "change in body position". Thus, to extract poses from angles, we actually apply non-minimum suppression (instead of the Harris Corner Detector's non-maximum suppression) after calculating the Euclidean Distance between angles in adjacent frames.

Meanwhile, we consider a gesture to be the opposite of a pose, a small range of frames when there is a significant "change in body position". Arguably, a gesture would take place for all the frames between two poses, not just for moments of the fastest "change in body position". However, because we applied a threshold that selects only a subset of the most significant poses, the above definition may cause multiple gestures to be recognized as one, so we have currently defined a gesture to be three consecutive frames of significant change in angles. (However, it could be interesting to explore modified algorithms for detecting poses and gestures in future work.)

3.4. Distance Evaluation

Our project uses three versions of Euclidean Distances for various functions: 1) the default formula identifies poses and gestures and suppress the non-maximum/minimum cases, 2) the formula on a subset of columns to calculate a quantitative score to show the user their similarity to the physical therapy video, and 3) a modified combination of floored angle gradients and weighted Euclidean Distance.

3.5. Video Selection

To test our model, we first selected videos from Cornell University's CornellHealth website[18]. These are legitimate physical therapy videos that get prescribed to current patients by their Cornell Health physical therapist, making them an ideal dataset to base our pose estimation model on.

We then completed these exercises ourselves to obtain a sample set of testing videos. We were able to compare the instructional videos with the videos of ourselves following the exercises. The accuracy of our model was determined by comparing our test video with the Cornell video, and determining on our own if the numerical accuracy generated by the model seemed accurate. For example, a patient video following the instructional video perfectly should yield a 100whereas a patient video that does not follow the instructional video well should yield a lower score.

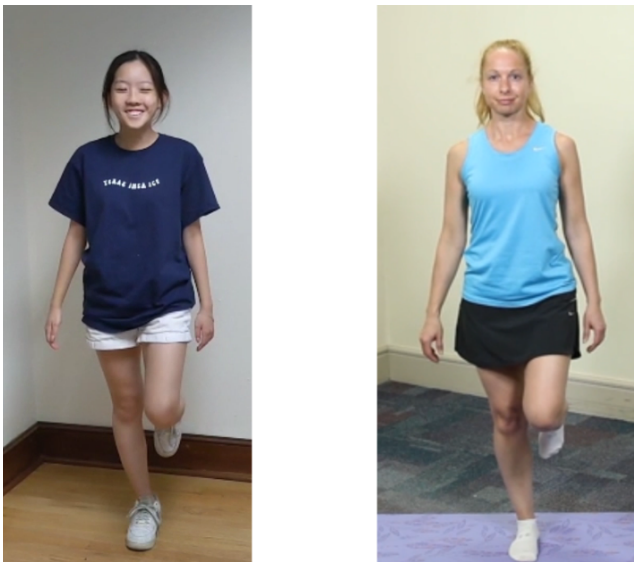

Figure 2. Video Testing: the comparison of one of our team's recreation of the excercise agaisnt the original.

4. Results

We had one of our teammates record physical therapy exercises and used those videos to compare with the Cornell videos. From the human eye, it looked as though our teammates' video was extremely accurate, and our algorithm supported this theory.

Given a pose, we are able to identify similar poses in different videos from a database that we curated. This encourages the patient to be more engaged with the exercises, since the exercises do not have to from the medical Cornell database, but rather also from dance.

4.1. User Customization Options

Our project offers the user some simple customization options so that it can be catered to individuals. These are the difficulty levels and the body part isolation features.

The difficulty is separated into three levels: easy, medium, and hard. Each level has a specific threshold of accuracy that is used to score the patient's video: Easy had a threshold of 60, medium 45, and hard 30. The difference between the angles of the two videos has to be less than or equal to this threshold to be considered "perfect" for each level.

The project can also isolate specific body parts to score. The accuracy of the exercises takes the whole body into consideration, so some body parts that remain idle will raise the accuracy just by being there. This feature allows the patient to decide what they want to be scored, so that if there is a specific body part they want to isolate in their practicing, they are able to get scores of that accordingly. This feature scores the same way as normal but utilizes the body part extraction feature from OpenPose. This can be especially useful if the patient is receiving treatment for a specific body part. For example, if they injured their arm, they may want feedback specifically on their arm to see how they are doing.

Figure 3. Pose Matching: the three images above were from different videos we processed and were all matched as containing similar poses. The first image is from a PT video, the second image is from a Chinese Dance video, and the last image is from an Indian Dance video.

5. Discussion

This project had some unexpected setbacks which we worked towards solving.

To identify poses, we needed to find the distances between the angles of frames, but we originally had some trouble with the euclidean distance formula. The base formula

was too sensitive to noise in the movements of the subject, which meant that the algorithm was overfitted to the data. By flooring some of the angle gradients, we were able to account for noise. This made the algorithm more resistant to slight differences in position, camera angle, orientation of the subject, etc.

Initially, the algorithm returned low accuracy scores for two videos that seemed otherwise quite similar. In order to ensure that the algorithm was working correctly as intended, an example video was run against itself, with an expected and actual accuracy result of 100%. This demonstrated that the algorithm was working but needed to use some sort of threshold or margin of error.

We also conducted manual testing ourselves, making sure that the accuracy score from the algorithm made sense with the two input videos. For example, if the videos are quite similar, we expect a high accuracy score, but if they are vastly different, we expect a low accuracy score.

We tested individuals of different heights to ensure that height did not create a large difference in terms of accuracy score. For example, if the individual in the instruction video is very tall, but the patient is very short and follows the video perfectly, the calculation of the angles should ensure that height is not a factor that would limit the accuracy score.

6. Conclusion

Through this project, we were able to successfully utilize computer vision techniques to create a program tailored to improve patient adherence to recovery plans in physical therapy. We consider this to be the beginning of a greater work with its applicability not just in physical therapy but also in the healthcare industry and beyond.

6.1. Limitations

Although we believe our project proves large progress, it did come with its fair share of limitations. As previously stated the camera angle between the videos must be the same, otherwise the scores will not accurately reflect their performance. This is due to the fact that our Gesture Detection algorithm is overly simple.

Another major limitation we faced in the creation of the project is the time bottleneck from OpenPose. Due to our restrictions with GPU access, we were forced to downscale the data we processed; either through resolution or duration. Downscaling the resolution proved to reduce the most time, however, it also produced questionable keypoints. Thus, for the purpose of testing, we were forced to process shorter videos at full resolution, in order to get the most accurate keypoints from OpenPose; this resulted in a vast time limitation as a 16-second video took 34 minutes to process.

6.2. Implications

For future implementations of this project, we would focus on matching movements, generate difficulty levels for exercises, and expand the scope of the project to also analyze other physical activities.

Currently, our project can match poses, but we would like to further develop this to match movements too. This way, if a patient wanted to practice a specific motion, like raising the arm, they would be able to easily find exercises that would help them practice this.

We would also like to have our project be able to give a difficulty rating for exercises as it analyzes the poses. This would be so that those in the early stages of physical therapy can see which exercises are useful to them.

The final update would be to expand the project's scope to other physical activities. Specifically, we would like to focus on dance. The project would be able to determine a dancer's accuracy, and offer many practice dance movements from various genres that the dancer can practice with. Another implication could be martial arts, which would work in the same way that physical therapy and dance would, but be tailored for any unexpected specifications that martial arts requires.

References

[1] A. Albergoni, F. J. Hettinga, A. La Torre, M. Bonato, and F. Sartor. The role of technology in adherence to physical activity programs in patients with chronic diseases experiencing fatigue: a systematic review. *Sports Medicine - Open*, 5(41), 2019. 1

[2] B.B. Armor, J. Su, and A. Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):1–13, 2016. 2

[3] F. Baradel. *Structured deep learning for video analysis*. PhD thesis, 2004.

[4] S. Bernstein. What is physical therapy? *WebMD*, 2023.

[5] V. Buwaneka, A. Samadder, and K. Sharma. Dancevideo, 2023.

[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[7] C. Chan, S. Ginosar, T. Zhou, and A. Efros. Everybody dance now, 2019.

[8] U. Cumberlands. 7 ways technology is changing physical therapy, 2013. 1

[9] Y. Diena. 74 physical therapy statistics, facts & demographics, 2023. 1

[10] Y. Jafarian and H. Park. Learning high fidelity depths of dressed humans by watching social media dance videos, 2021.

[11] R. Klepps. 7 thought-provoking facts about physical therapy you can't ignore, 2018. 1

[12] K. LeViet. Pose estimation and classification on edge devices with movenet and tensorflow lite. *TensorFlow Blog*, 2021.

[13] Maureentkt. Selecting your real-time pose estimation models. *Medium*, 2021.

[14] MediaPipe. Pose landmark detection guide. *Google for Developers*, 2023.

[15] A. Mumuni and F. Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Elsevier*, 16(1), 2022. 2

[16] OpenCV. Open source computer vision library, 2024. 2

[17] Y. Pang and Y. Niu. Dance video motion recognition based on computer vision and image processing. *Applied Artificial Intelligence*, 37(1), 2023.

[18] Cornell University. Pt exercise videos. *CornellHealth*. 4

[19] Wikipedia. Dbscan.

[20] M. Yang and Z. He. Dance action recognition model using deep learning network in streaming media environment. *PMC*, 2022.