How Good Is Your Pose?
Pose Estimation for Weight Lifting Form Correction

Alexander Martin *†‡ Michael Kingsley *† Vashisth Tiwari *†

Abstract

Pose estimation is a well-researched field in computer vision with many published architectures for the task. However, these methods are not widely used in day-to-day applications due to model computing cost or the general public's lack of understanding of the methods. Currently, the largest and most widely advertised deployment of pose estimation in a non-technical field is the use of automatic offside detection in the FIFA 2022 World Cup [2].

Inspired by the use of these methods in a widely spectated sport, we introduce another use of pose estimation in athletics that anyone can use: pose estimation for weightlifting form. In this paper, we explore this for the three widely performed exercises: bench presses, deadlifts, and squats. The framework uses a CNN to classify the lifts and ViTPose to identify key joints and to make suggestions. We find that pose estimation can be successfully used for online coaching without human supervision and can even be used in competition athletics to flag lifters with illegal forms.

1. Introduction

Form is crucial in athletic performance. Minor changes in form can increase power, speed, and reduce the risk of injury for athletes. Form is easy to critique in weight lifting due to the standards from Olympic and competition lifting. There is less variability in form for weight lifting than in other sports where more advanced bio-mechanics are at play. In order to improve form, athletes must practice and focus on proper technique. A qualified coach or trainer can help athletes understand the proper form and provide feedback and guidance on how to improve.

In this paper, we find that pose estimation can be successfully used for online coaching without human supervision and can even be used in competition athletics to flag lifters with illegal form. We propose the use of Deep CNN to classify exercises and then apply a 2D human pose estimation model to provide ways to score the form of lifters in lieu of a personal trainer.

First, the CNN will help identify which exercises the lifter is doing and can be used as a starting point for the form analysis. The CNN will be trained on a dataset of images of people performing either barbell bench press, barbell squat, or barbell deadlift. Based on this training, we can identify which of these three labels best predicts the exercise the lifter is doing.

Once the pose has been identified using the CNN, a 2D human pose estimation model will be used to estimate the lifter's pose and analyze their form. The pose estimation model will output 17 key points that represent the lifter's body position, discussed in section 4.2. These key points can then be used to compare the lifter's form to the ideal form for the exercise and score the lifter's form. The scoring will be based on the lifter's deviation from the ideal form and can be used to provide feedback on how to improve.

Overall, this paper proposes the use of a Deep CNN and 2D human pose estimation to analyze and score the form of lifters. The proposed system has the potential to provide feedback and guidance on how to improve form in a more efficient and cost-effective way than traditional methods.

2. Background

2.1. Image Classification

Convolutional Neural Networks (CNNs) are a type of deep learning neural network that are mainly used for image classification. They are composed of multiple layers of neurons, each of which is responsible for a certain task. The neurons in the first layer of a CNN are responsible for detecting the lowest level features in an image, such as edges, corners and basic shapes. As the image is passed through each layer, more complex features such as parts of objects are detected. The last layer of the network is used to classify the image based on the features it has detected.

ResNet [3] is a type of CNN that uses residual learning to ease the training of networks that are substantially deeper than those used previously. In a traditional CNN, each layer

*University of Rochester, Rochester, NY 14627, USA
†equal contribution
‡corresponding author: amart50@u.rochester.edu

The code for this project can be found here: https://github.com/alexmartin1722/liftingpose is responsible for learning a certain set of features. This can become difficult when the network is too deep, as the layers may not be able to learn the desired features. In a ResNet, instead of learning a completely new set of features in each layer, each layer is responsible for learning a set of residual functions. These residual functions represent the difference between the desired output and the predicted output from the layer before it. By learning these residual functions, the network can better optimize the weights in each layer.

2.2. Pose Estimation

Pose estimation is an important focus in the computer vision community due to its large range of real-world applications. Pose estimation aims to automatically predict and track human posture by localizing joints and defining limb orientation. Although numerous methods have been developed to address this challenge, most of them rely on complex and hand-crafted models, which are expensive to train and require a large amount of data. Recent works have shown that plain vision transformers [6] can be used to achieve excellent performance in visual recognition tasks, however, their potential for pose estimation has not been fully explored.

2.3. Pose Datasets

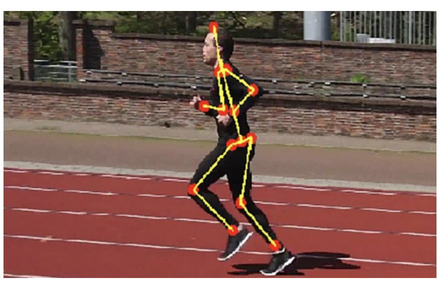


Figure 1. An example annotation from MPII

Most datasets annotate 17 key points, while others expand their annotations up to 133 (COCO Whole Body [4]). The most relevant dataset for our use case is the OCHuman dataset [7]. This dataset includes annotations of bodies that are occluded by objects or the frame, which is perfect for weightlifting in a commercial setting because you cannot guarantee an object-free filming environment. Most times, lifters are slightly covered by squat racks, safety bars, benches, and large weights. It is important to be able to predict and still provide feedback for joints that are not perfectly pictured in the frame, so the OCHuman is a great benchmark to test the pose estimation model on. The model used for pose estimation in this paper (ViTPose) is trained on this dataset for the very reasons discussed above.

2.4. Models

2.4.1 ViTPose

ViTPose [5] is a pose estimation model consisting of plain and non-hierarchical vision transformers, where the backbones are pre-trained with masked image modeling pretext tasks.

It adopts a simple architecture, with de-convolution and prediction layers, and decoders without skip-connections and cross-attentions. Given a person instance image, ViTPose embeds the image into tokens via path embedding. After this, the embedded tokens are processed by several layers consisting of self-attention layers and feed-forward networks. [5]

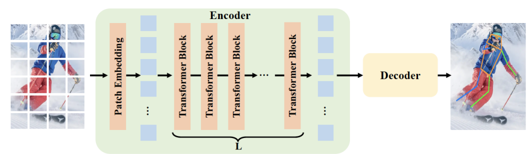


Figure 2. ViTPose Framework

ViTPose outperforms representative methods on the MS COCO keypoint detection benchmark. The largest model sets a new state-of-the-art, 80.9 AP on the MS COCO test set as well as high performance on OCHuman with 91.6 AP.

3. Dataset

For this paper, we collected a new dataset for classifying and estimating pose on lifts. Using Python's Beautiful Soup API we web-scraped 100 photos each for barbell bench press, barbell squat, and barbell deadlift. After gathering the data and checking that they are all suitable for classification and pose estimation, we used OpenCV to transform our data. Keeping a copy of the original data, first, we flipped the data horizontally, then we proceeded to randomly rotate our data between 30 degrees and 180 degrees. Consequently, we had 900 photos as our final data set, 300 of each lift.

To score a person's lift against a lifter with good form, we annotate gold data that we've found of lifting coaches and athletes performing the exercise. We then annotate each lift in the same style as COCO body annotations. We've provided three images (one at each angle [front, left, right]) for each lift that we classify in our CNN in section 4.1.

4. Framework

To be able to score a lifter's form, we perform three tasks. First, we classify the lift into three categories, squat, bench, and deadlift (section 4.1). Second, we perform pose estimation on the lift to get the joint locations of the lifter 4.2. After we have the lifter's keypoint data, we then score the lifter's form against the three gold annotations for that lift (section 4.3).

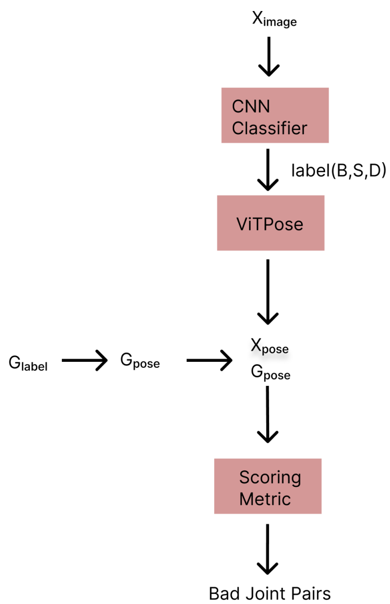


Figure 3. Framework
(the input image is fed to the CNN to decide the type of the lift, then annotated images are obtained using ViTPose which are compared against gold standards to provide feedback)

4.1. CNN

To classify the lift types, we attempt to implement two architectures for a CNN. One is a CNN based off the architecture from project one, and the other is based on the ResNet [3] architecture. The CNN used in project one consists of three convolutional layers with RELU activations. The loss criteria were changed from BCE (binary cross-entropy loss) to cross-entropy loss, given we have three lifts for the classification.

4.2. ViTPose

For pose estimation, we rely on the newest state-of-the-art model for human pose estimation, ViTPose [5]. ViTPose is the perfect model for this because of its high throughput and AP on the OCHuman dataset.

To use this model, we configure it to annotate 17 key points on a person: nose, eye, ear, shoulder, elbow, wrist, hip, knee, ankle[1]. For the full pose estimation, we employ a top-down pose estimation method, utilizing the ViT backbone from the ViTAE transformer [6]. A top-down process is a two-step method, first object detection is used to predict where the target is and a bounding box is annotated around the target. Once the target is predicted, the pose estimation is performed on the target.

[1] All are annotated left/right except for nose

4.3. Scoring Metric

The scoring metric used in this paper utilizes a technique from natural language processing called inter-annotator agreement (IAA). One of the limitations of the CNN classification is that if we were to filter each lift by its angle (or range of angles) human annotators would have to spend lots of time classifying the lift in a range of angles. This also would increase the complexity of the CNN and reduce its overall accuracy in classifying the lift. To solve this problem we use a scoring metric on each of the gold annotations and utilize annotator agreement to decide if the feedback should be passed along to the user as seen in Algorithm 1. One should note that if a joint is not annotated by ViTPose, those joints are ignored in the scoring metric.

Algorithm 1 Scoring Metric

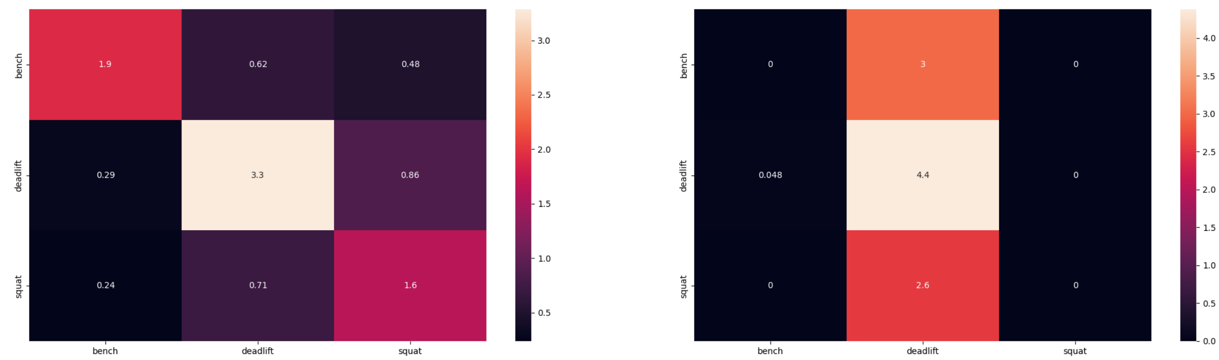
for each critical pair in keypoints **do**
  Get the Euclid distance between two joints
  $distance = \sqrt{(\frac{x_2 - x_1}{width})^2 + (\frac{y_2 - y_1}{height})^2}$
  for each gold annotation **do**
    Compare each ratio with gold ratios
    if outside of size threshold **then**
      lift is bad form
    end if
  end for
  if Lift agrees with front **then**
    consider lift a front lift and good form
  end if
  if left angle and right angle agree **then**
    consider lift a side lift good form
  end if
  if $\frac{2}{3}$ lifts call it bad form **then**
    report those joints as bad form
  end if
end for

The critical pairs of joints vary for each lift. In bench press, the critical joint relations are (wrist, wrist), (wrist, elbow), (elbow, elbow), and (elbow, shoulder). For deadlifts, the critical joints are (wrist, wrist), (hip, hip), (ankle, ankle), and (knee, knee). For squat, the critical joints are (wrist, wrist), (knee, knee), (ankle, ankle), and (hip, knee).

5. Results

5.1. CNN

For the CNN pose classification, we find that our CNN framework performs well, while our ResNet implementa-



(a) CNN Confusion Matrix

(b) ResNet Confusion Matrix

Figure 4. Confusion Matrix for CNN Implementations on the validation set (true labels on the x-axis, predicted labels on the y-axis)

Model	Bench	Squat	Deadlift
CNN	**63.5%**	63.0%	**74.2%**
ResNet	0%	**100%**	0%

Table 1. CNN Classification vs ResNet Classification Accuracy (The table shows the percentage of lifts classified correctly for each of the three lifts)

tion struggles. In Figure 4a and Table 1 we see that our classification method is successful in differentiating the types of lifts and provides good results. Our ResNet implementation is obviously incorrect as it can only classify the deadlift pose as seen in Figure 4b.

In our implementation, it can be seen in Figure 4 that the classification between deadlift and squat encounters the most confusion due to the visual similarity and setup of the two lifts.

It should also be noted that we tried many implementations of CNN architectures, as well as prebuilt networks and networks with pre-trained weights that we fine-tuned on our task. Even with these methods, we were not able achieve high classification scores for the squat lift and most models tend to classify it as a deadlift.

5.2. ViTPose

The pose estimation results we get are very good. Since we do not have a large amount of gold data, we are not able to quantify the accuracy. However, to ensure satisfactory performance, we went through each image by hand. In an ideal experiment, we would supply this to Mechanical Turkers to note the points that are out of position, however, we lack the funding and time for this. Instead, we went through each image and noted the points that were off.

For most lifts, ViTPose does a great job with body coordinates and tends to fail on the ear and nose estimations. This is understandable, as ViTPose does not segment the head for its own pose estimation as other methods do. Other noticeable inaccuracies are some predictions when the limb is not visible. In situations of partial visibility, the results still are good, however when a full limb is not visible, ViTPose annotates with extremely low confidence.

5.3. Scoring Metric

We find the scoring metric successful on the 15 images we took ourselves. It is able to note the regions of poor form with high success. The only time an annotation is scored poorly when it might be good is in the case of a ViTPose annotation with low confidence. This is covered in more depth in Appendex B.

6. Conclusion

In conclusion, our CNN and ViTPose implementations are successful in pose classification and pose estimation, respectively. Our CNN framework is able to differentiate between different types of lifts and our ResNet implementation struggles due to the visual similarity and setup of the two lifts. Our scoring metric is able to accurately note the regions of poor form with high success. The ViTPose annotations are found to have a high success rate. However, there are some inaccuracies in cases of partial limb visibility or when a full limb is not visible. Overall, our implementations are successful and show promise for future projects.

7. Future Work

To further improve the performance of this model, it would be important to classify lifts by the angle of the camera in a more critical way. The current method ignores this constraint and tries to work around it with a unique scoring metric.

It would also be beneficial to annotate a dataset of lifting poses for other use cases and expand the model beyond the bench, squat, and deadlift. When any exercise does not involve advanced bio-mechanics, it would be possible to implement this method on that exercise. Additionally, more work needs to be done on making a more rigorous classifier because our classifier does poorly in differentiating between squats and deadlifts, given their visual similarities.

A. Use Cases

A.1. Qualifying Olympic Lifts

With the new bench press rules being introduced by the International Powerlifting Federation (IPF) in 2023, qualifying bench press has gotten much more difficult [1]. The rule introduced mandates lifters to get their elbows parallel with the top of the shoulder joint to prevent lifters from getting away with smaller ranges of motion.

B. Annotated Pose Estimations

In Figures 5-7, we can see that ViTPose is successful on our own lift examples. It has high success on the limbs in the frame and even limbs occluded by our example lifter. However, It struggles with the face annotation and in Figure 5, you can see a poor annotation of the right wrist and elbow because the full limb is out of frame. It also is less successful on the ankles in Figure 6 that are cut out of the frame.

Figure 5. Deadlift Pose Estimation

Figure 6. Bench Pose Estimation

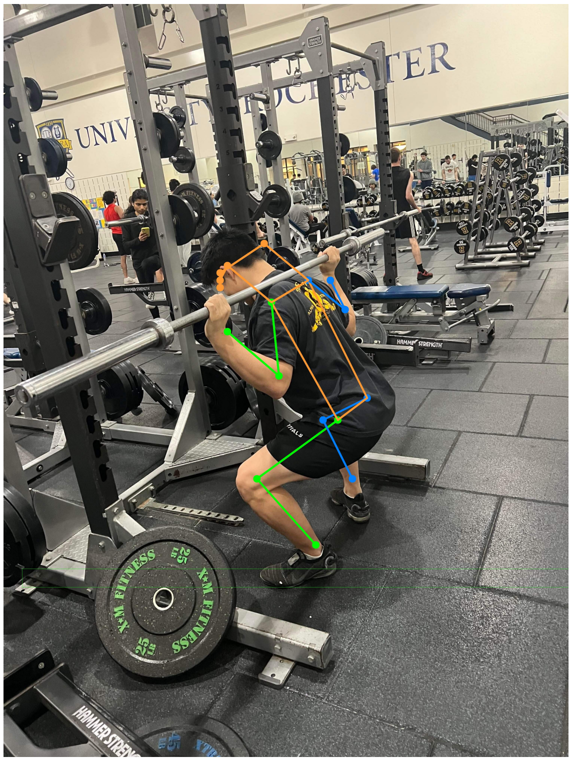


Figure 7. Squat Pose Estimation

References

[1] Jake Dickson. International powerlifting federation unveils bench press rule change for 2023. 5

[2] FIFA. Semi-automated offside technology to be used at fifa world cup 2022. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1, 3

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[5] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*, 2022. 2, 3

[6] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias, 2021. 2, 3

[7] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul L. Rosin, Zixi Cai, Han Xi, Dingcheng Yang, Hao-Zhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. 2018. 2