Enhancing Mask R-CNN for Pose Estimation in Dynamic Environments

Shreyas Chigurupati[1]
schigurupati@wpi.edu

Dev Soni[1]
djsoni@wpi.edu

Ankit Talele[1]
amtalele@wpi.edu

Jeel Chatrola[1]
jchatrola@wpi.edu

***Abstract*—The Real-time Pose Correction Feedback System is an innovative application designed to assist users in performing physical exercises with proper form, leveraging the capabilities of the Mask R-CNN framework with ResNet50 for human pose estimation. Utilizing real-time video analysis, the system detects key points of the user's posture and compares them against ideal exercise forms. Discrepancies are highlighted through visual feedback, guiding users to adjust their posture accordingly. This project aims to create an accessible tool that promotes health and fitness, while also serving as an educational platform for demonstrating the practical application of deep learning and computer vision. We use the sport of Yoga to perform analysis of subject poses for proper posture and make expert suggestions and imporvement enhancements.**

I. Introduction

Physical exercise, when executed with incorrect form, can lead to suboptimal results or even injury. Addressing this, our project introduces the Real-time Pose Correction Feedback System, a computer vision-based application that bridges the gap between technology and physical well-being. Building upon the foundations laid by the Mask R-CNN framework, which revolutionized object segmentation and detection in static images, we extend its application to dynamic analysis of image frames for the purpose of human pose estimation. By analyzing a sequence of image frames in real-time, our system identifies key body joint positions, compares them against a dataset of correct postures, and provides immediate visual feedback for posture correction. This project aims not only to enhance individual fitness routines but also to serve as a practical exploration of deep learning and computer vision techniques, providing a valuable learning experience for our team and a stepping stone into the broader field of health informatics.

Classical human pose estimation papers have utilized the essential joint framework primarily for providing accurate pose estimations[1][10] but have not leveraged this information to provide feedback to individuals[2][5]. In contrast, our novel architecture learns human body poses, analyzes them, and compares them with ideal poses, which are also learned through expert sportspersons' stances. We use the ideal pose and the real-time angles of joints in captured image frames to suggest corrections, enabling the individual to better perform the sport, such as Yoga in our case. This approach can be highly beneficial in many sports where funding constraints make professional coaching unattainable. Our model can analyze any professional player's pose from image frames, learn the ideal pose for the sport, and then provide feedback based on the specific sport requirements in the given images.

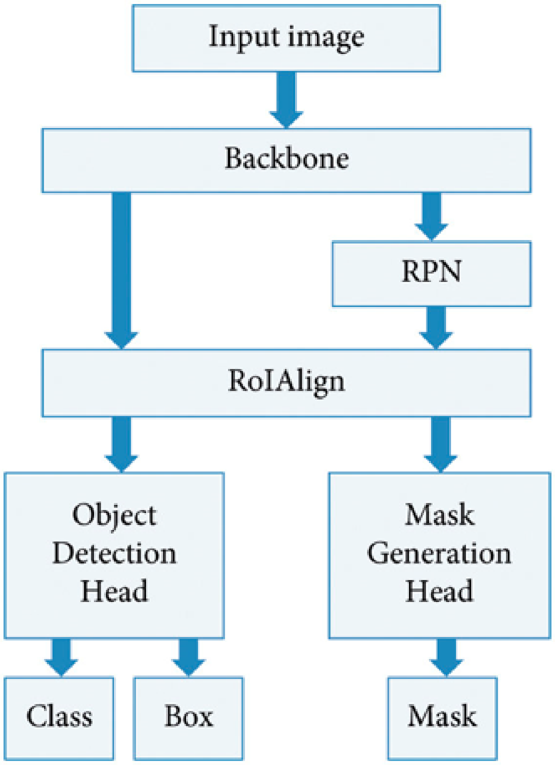
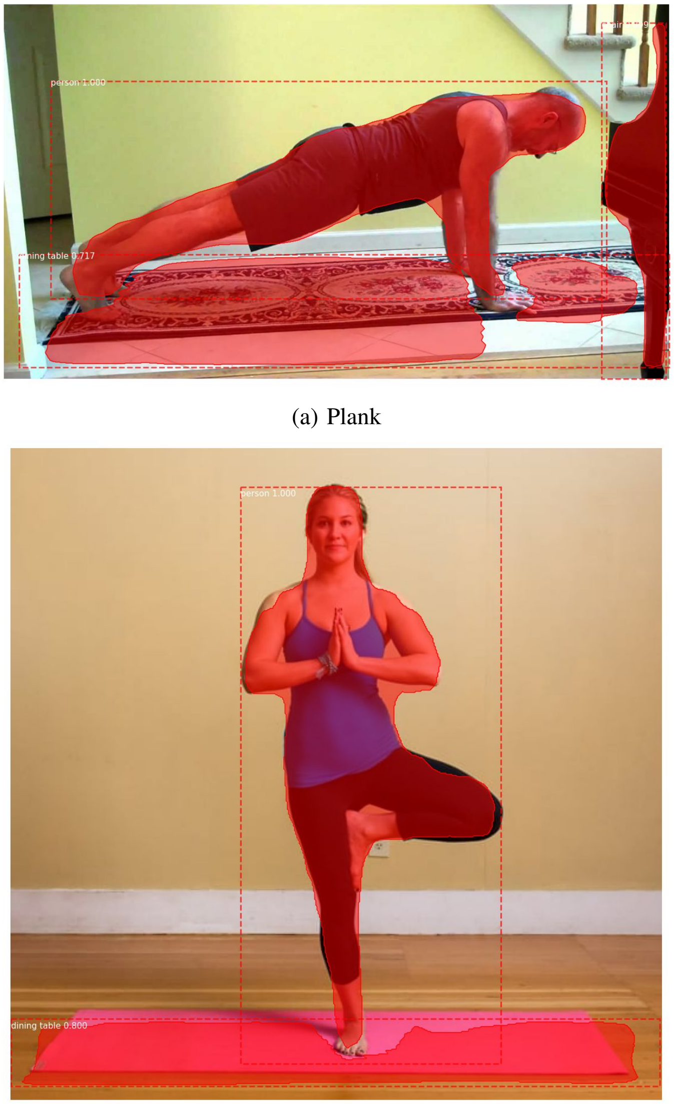


Fig. 1: Mask R-CNN Model

Our system sets itself apart by not only recognizing accurate human poses but also by integrating a comparative analysis with ideal postures derived from expert practitioners in various sports, starting with yoga. This analytical capability is crucial in scenarios where access to professional coaching is limited by financial or geographical barriers. Moreover, our model's adaptability to different sports represents a significant advancement. It can learn the intricacies of various athletic postures by analyzing image frames of professional athletes, thereby creating a versatile database of ideal poses. This feature enables the system to offer customized feedback across a spectrum of physical activities, significantly broadening the potential applications of our technology in sports training and rehabilitation. By harnessing deep learning algorithms and cutting-edge computer vision techniques, our project not only fosters technological progress in health informatics but also democratizes access to high-quality athletic training.

A. Research Contribution

We propose a novel architecture wherein we train the model by inputting the professional persons poses and then through feedback module we analyze the pose and comment on improvement on the pose performed for YOGA. This way the sportsperson gets the feedback on the pose real time and can improve on their form instantly rather than waiting for some expert advice. The images in Fig.2 shows the mask applied to our dataset by using the pre-trained Mask RCNN model.



(a) Plank

(b) Tree

Fig. 2: Mask R-CNN Segmentation

B. Pose Estimation - ResNet50

- **GlobalAveragePooling2D:** Reduces each feature map to a single value by averaging, which helps to minimize overfitting by reducing the model's parameters.
- **Dense Layer:** A fully connected layer with 1024 units and 'relu' activation that transforms the features for the learning of keypoint relationships.
- **Dropout:** A regularization technique with a rate of 0.3 to prevent overfitting by randomly dropping units during training.
- **Final Output Dense Layer:** Produces the predicted keypoint coordinates with a size of `num_keypoints` $\times$ `2` and linear activation.

II. Brief Explanation of ResNet

ResNet, short for Residual Network, revolutionized the field of deep learning by enabling the training of very deep neural networks. Here's a brief overview:

- ResNet50 is a variant of the ResNet family with 50 layers deep, widely used for various computer vision tasks.
- It introduces **residual blocks** with skip connections that facilitate the training of deeper networks by addressing the vanishing gradients problem.
- These skip connections allow the addition of more layers without a degradation in performance, ensuring only improvements or at least neutrality in network capability.
- The network is composed of convolutional layers, batch normalization, and ReLU activations within its residual blocks.
- Originally trained on ImageNet, ResNet50 serves as an effective feature extractor for tasks such as object detection and pose estimation.

III. Related Work

Our approach significantly diverges from the Pose as Compositional Tokens (PCT)[1] method by focusing not just on accurate pose estimation[1][2][3] but also on real-time feedback and correction. While PCT innovatively uses discrete tokens to represent interdependent joint structures, enhancing pose estimation under occlusion, our system extends beyond identification[7], comparing detected poses with ideal models and providing immediate corrective suggestions. This added layer of interactive guidance is tailored specifically for athletic training and rehabilitation, leveraging pose data for practical application rather than just classification.

IV. Dataset

For our project we focused on a carefully selected subset of the larger yoga dataset, comprising 500 images. This subset was carefully chosen to represent a diverse range of yoga poses, ensuring a comprehensive coverage of the dataset's potential applications. To facilitate accurate and efficient machine learning and computer vision tasks, we utilized the VGG Image Annotator, an online tool renowned for its precision and user-friendly interface. The annotation process involved detailed labeling of each image, capturing essential features and characteristics relevant to our study's objectives and then exporting it in the COCO format into a json file. This process not only enhanced the dataset's utility for our specific project requirements but also made it a valuable resource for future studies in the domain of yoga pose recognition and analysis. By leveraging the VGG annotator's robust framework, we ensured that each image in our subset was annotated with high accuracy and consistency, laying a solid foundation for the subsequent stages of our project.

V. Proposed Method

Our project adopts the Mask R-CNN framework, a state-of-the-art technique for object detection and instance segmentation, as the cornerstone for human pose estimation. Mask R-CNN extends the capabilities of Faster R-CNN by adding a branch for predicting segmentation masks on each Region of Interest (RoI), which is crucial for distinguishing individual instances of objects—particularly useful in human pose estimation.

A. Model Architecture

The core of our approach is a Mask R-CNN model complemented by the ResNet50 architecture. This combination is selected for its proven efficiency in feature extraction and adaptability to varied object detection tasks. The ResNet50 model, pre-trained on the ImageNet dataset, serves as the backbone for feature extraction. On top of this, we implement a fully convolutional network (FCN) that predicts segmentation masks for each detected human figure. The Mask R-CNN model is further fine-tuned to distinguish human figures and delineate their poses accurately. Our complete architecture is shown in Fig. 3.

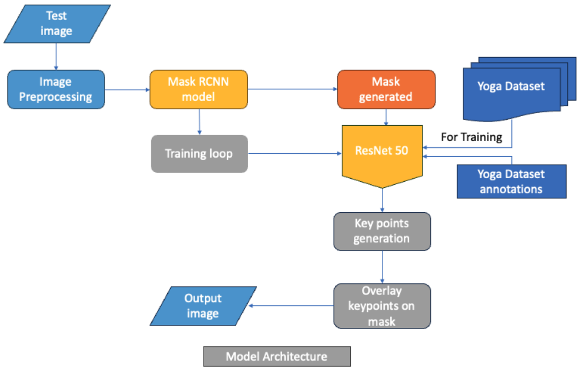


Fig. 3: Architecture

B. Keypoint Detection

For the task of pose estimation, we introduce a keypoint detection mechanism within the Mask R-CNN framework. The model is trained to identify specific keypoints that represent human joints, such as elbows, knees, wrists, etc. We employ a specialized dense layer in the network that outputs the coordinates of these keypoints. This layer is trained to predict the x and y coordinates of each joint, forming the basis of our pose estimation.

C. Feed back system

- Unique combination of computer vision and machine learning in the Real-time Pose Correction Feedback System.
- Focus on yoga to compare joint angles between a novice and an expert using image frames.
- Utilization of a skeleton structure with key joints such as elbows and knees.
- Employment of computer vision to determine joint positions from static images.
- Calculation of angles at these joints and quantification of the differences for precise feedback on posture discrepancies.
- The process applies computer vision for insightful posture analysis, though it is not a learning mechanism.
- Innovation lies in merging machine learning for pose estimation with computer vision for real-time comparative feedback.
- Creates a powerful tool for sports training and health informatics.

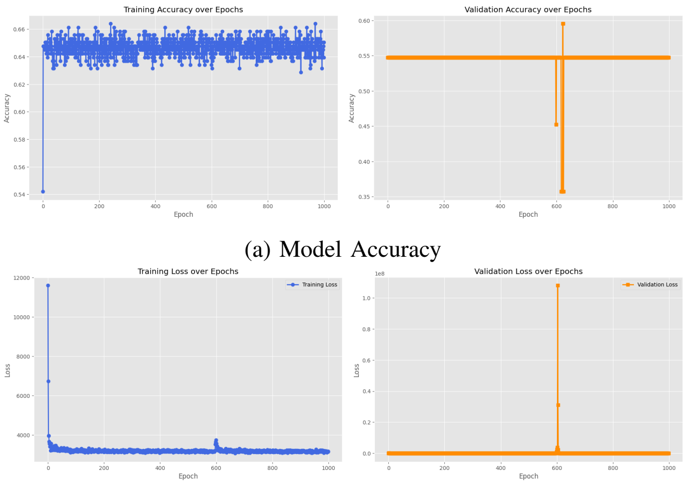
D. Training and Optimization

The model is trained on a dataset comprising diverse images annotated with human figures and their corresponding keypoints. During training, we utilize various data augmentation techniques to ensure the robustness of the model against different orientations, scales, and environmental conditions. We also implement strategies like dropout and fine-tuning the last layers of the network to prevent overfitting and enhance the learning of pose-specific features. The training process is executed with a focus on achieving a balance between detection accuracy and computational efficiency, making the model suitable for real-time applications.

Through this methodology, our project aims to harness the power of Mask R-CNN not just for object detection but also to extend its utility to the nuanced task of human pose estimation. The proposed approach is designed to provide a comprehensive solution that is both accurate in pose detection and efficient in processing, catering to the growing demands of real-time applications in various fields.

VI. Results

Our Mask R-CNN model on human pose estimation presents a comprehensive analysis of the model's performance over training epochs. The accuracy graphs as shown in Fig.4(a) revealed the model's consistent performance on training data and a moderate yet robust accuracy on validation data. The training accuracy plot indicates a stable convergence with minor fluctuations, showcasing the model's ability to learn and adapt to the training data. In contrast, the validation accuracy plot reveals a wider range of variability, suggesting the model's sensitivity to the validation data's complexity and diversity. This behavior is mirrored in the loss plots (Fig.4(b)), where the training loss shows a consistent decrease, indicating effective learning, whereas the validation loss exhibits spikes which could be attributed to the model encountering previously unseen poses or variations in the data.



(a) Model Accuracy

(b) Model Loss

Fig. 4: Performance plots

Qualitatively, the model's pose predictions aligned well with the ground truth, successfully identifying keypoints across various poses.

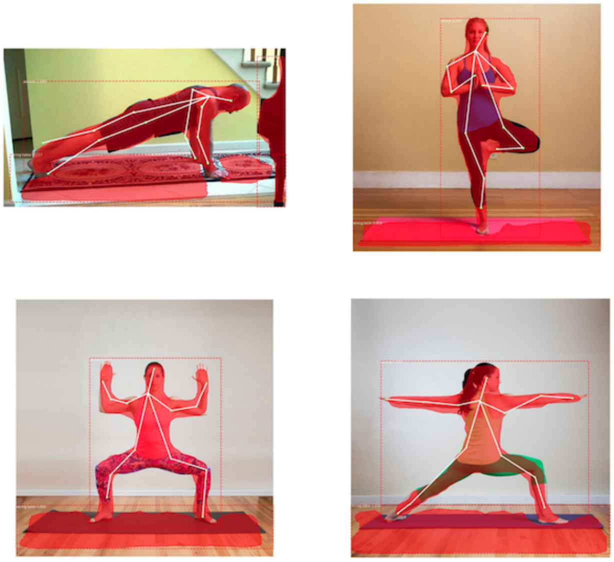


Fig. 5: Output of different poses with masks and keypoints (a) Plank, (b) Godess, (c) Tree, (d) Warrior

Furthermore, the visual results provided by the pose estimation system demonstrate high precision in detecting human poses. The illustrated examples of various yoga poses as shown in Fig. 5 highlight the model's capability to accurately map the keypoints of the human body, even in complex postures. These visual results not only validate the quantitative findings but also demonstrate the practical utility of the model in real-world applications, from fitness and health monitoring to advanced human-computer interaction scenarios. The model's adeptness at both learning from the data and generalizing to new instances holds promise for future enhancements and broader applications in the field of computer vision.

VII. Discussion

The model's training performance, characterized by a steady decline in loss, suggests that the Mask R-CNN architecture, paired with the ResNet50 backbone, was effective for the task of human pose estimation. However, the oscillations in the validation loss and accuracy imply a potential overfitting to the training data, or conversely, a need for a more diverse and challenging validation set to better assess the model's generalization.

The observed discrepancy between training and validation accuracy indicates that while the model has learned the training data well, its performance on unseen data could be further improved. This could be addressed by implementing more robust data augmentation techniques, exploring regularization strategies, or by collecting more varied and comprehensive training data.

The qualitative results, particularly the model's ability to accurately predict human poses, confirm its practicality. Nevertheless, the quantitative results hint at the possibility of improving the model's robustness to new and diverse datasets.

In essence, the experiments have demonstrated the promise of our approach, while also highlighting areas for improvement. Future work can build on these findings, refining the model to enhance its accuracy and reliability in varied real-world applications.

VIII. Conclusion and Future Work

In conclusion, the application of Mask R-CNN for human pose estimation has proven to be a promising direction, with the model achieving notable accuracy in detecting and estimating human poses in images. The training process has demonstrated the model's capability to learn effectively, while validation has provided insights into the robustness and adaptability of the model to new data.

Despite the successes, there remains room for improvement, particularly in enhancing the model's performance on the validation set. Future work will focus on addressing the overfitting observed during the training phase, potentially through the integration of more complex data augmentation techniques, the introduction of additional regularization methods, and the expansion of the training dataset to encompass a wider array of human activities and environments.

Future enhancements of our Real-time Pose Correction Feedback System aim to integrate user feedback into its learning model, enhancing adaptability for fast-paced sports. With higher fps for real-time analysis, it promises to democratize advanced training, offering affordable, high-quality coaching to athletes across various sports, regardless of their budget.

Further research will also explore the integration of temporal information by extending the model to video data, allowing for dynamic pose estimation over time. This would be a significant step towards applications in motion analysis, augmented reality, and real-time interactive systems.

The ultimate goal is to develop a model that not only excels in accuracy but also in its ability to generalize across various real-world settings, making it a versatile tool for numerous practical applications in the field of computer vision and beyond.

References

[1] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, Han Hu. ”Human Pose as Compositional Tokens” in CVPR 2023

[2] Kaiming He, Georgia Gkioxari, Piotr Dollar and Ross Girshick. ”Mask r-cnn.” Proceedings of the IEEE international conference on computer vision. 2017.

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. ”2D human pose estimation: New benchmark and state of the art analysis”. In CVPR, 2014.

[4] Rıza Alp Güler, Natalia Neverova, Iasonas Kokkinos. ”DensePose: Dense Human Pose Estimation in the wild”. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7297-7306

[5] Wang, Yuejuan, Ji Wu, and Heting Li. ”Human detection based on improved mask R-CNN.” Journal of Physics: Conference Series. Vol. 1575. No. 1. IOP Publishing, 2020.

[6] Cai, Huimin, and Yang Gao. ”Optimization of Human Pose Detection Based on Mask RCNN.” 2021 2nd International Symposium on Computer Engineering and Intelligent Communications (ISCEIC). IEEE, 2021.

[7] Ke Sun, Bin Xiao, Dong Liu, Jingdong Wang. ”Deep High-Resolution Representation Learning for Human Pose Estimation” In CVPR, 2019.

[8] K. He, X. Zhang, S. Ren and J. Sun, ”Deep Residual Learning for Image Recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[9] Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, ”Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields” in CVPR 2017.

[10] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, Erjin Zhou ”TokenPose: Learning Keypoint Tokens for Human Pose Estimation” in CVPR 2017.