Sedentary Posture Recognition and Correction Using a Convolutional Neural Network (CNN) and the You Only Look Once Version 8 (YOLOv8) Pose Estimation Model

Justin Rui
University of Toronto Schools
ruiju@utschools.ca

Daniel Ganjali
University of Toronto Schools
ganda@utschools.ca

Henry Tian
University of Toronto Schools
tiahe@utschools.ca

Daniel Cui
University of Toronto Schools
cuida@utschools.ca

Kevin Wen
University of Toronto Schools
wenke@utschools.ca

***Abstract*—Poor posture is a leading contributor to musculoskeletal disorders, significantly affecting quality of life and productivity. This project introduces a deep learning framework to identify anatomical keypoints and offers a system that classifies seated posture as good, fair, or bad while providing user posture-related feedback. Initially, a custom Convolutional Neural Network (CNN) was developed with 47.3% accuracy, but due to practical constraints, the system was integrated with the You Only Look Once Version 8 (YOLOv8) pose mode with 84.9% accuracy. This system operates through a phone camera connected to a main device, achieving a posture detection accuracy of 92.3% at 30 Frames per Second (FPS). With broad applications, such as workplace ergonomics, remote learning, and online physical therapy, this project proposes a non-invasive solution for proactive posture correction.**

I. Introduction

A. Motivation

With the rise of sedentary lifestyles due to digitalization and increased screen time exposure, posture-related health problems have become a concern. Musculoskeletal disorders (MSDs)—a class of disorders including back pain, neck strain, and spinal misalignment—have been directly correlated with poor posture, particularly during extended periods of sitting [1]. Today, MSDs are some of the most harmful and costly conditions—with almost 40% of adults having suffered from back pain in the last 3 months [2]. Furthermore, incorrect posture alone has been shown to result in up to a 29.3% decrease in labour productivity [3]. Therefore, it is clear that there has never been a greater need for accessible real-time posture correction tools as professionals and students spend more time sitting in front of screens.

B. Related Works

Conventional posture assessment techniques, like wearable sensor-based tracking systems or in-person ergonomic assessments, offer important information about body alignment and possible ergonomic hazards. However, those methods have many drawbacks, as they tend to be expensive, invasive, or unsuitable for real-time monitoring. Recent advances in human pose estimation models have enabled automated tracking, with models like OpenPose [4] and AlphaPose [5] producing high-accuracy results in full-body keypoint detection in static images and video frames. However, few existing systems are specifically designed and optimized for seated posture monitoring while providing real time feedback. Computer vision methods, like the aforementioned, have massive potential in providing a solution that is capable of giving feedback for one's posture with a regular webcam or smartphone camera.

C. Problem Definition

Despite recent advancements, there is a lack of real-time, non-invasive solutions specifically tailored for seated posture monitoring. Although there have been models like OpenPose for general purpose tracking, there are few adapted to seated posture correction. This leads to the need for an accessible real-time posture analysis system capable of labeling keypoints and, specifically, using those keypoints to offer posture-related feedback.

II. Methodology

As previously mentioned, our team faced significant complexity and computational challenges during the development of this project, leading us to pivot from our initial custom CNN to the YOLOv8 pose framework. However, we have documented our ongoing progress towards a custom CNN below.

A. Dataset

Both our CNN and the YOLOv8 pose model for keypoint identification were trained on the Common Objects in Context (COCO)-Pose dataset—a subset of the COCO 2017 dataset filtered to human keypoints. This particular set was chosen for the high-quality keypoint annotations and extensive size (59 000 images) [6]. Each training example is annotated with 17 anatomical keypoints, such as shoulders, elbows, hips, and knees, which are later utilized to analyze body posture and identify possible ergonomic improvements [7].

B. Preprocessing

Several steps were taken in the preprocessing, including resizing the image to the input resolution of 256×256, introducing Gaussian noise and augmenting the data—which included random rotation, scaling (0.75 to 1.25x) and vertical reflection. One consideration was that a significant amount of the COCO-Pose dataset has multiple individuals annotated, whereas our system is tailored for a single person. Since this portion of the dataset should ideally be retained, images with multiple annotated individuals were cropped to a bounding box of a single person in the frame.

C. Initial Model Architecture

Initially, our team trained a custom CNN to detect 17 anatomical human keypoints. In this section, we provide a breakdown of the model architecture and development process.

The feature extraction block consists of 5 convolutional layers, each with batch normalization, stride-based down sampling, and a ReLU activation to improve training stability and convergence. As seen below, the dimensions of the image are reduced in each layer while feature depth is increased to continually learn spacial patterns. Each output is fed into the next layer—refining the predictions at each stage.

TABLE I
CONVOLUTIONAL LAYERS IN THE CUSTOM CNN

Layer	Input	Output	Kernel	Stride
Conv1	256×256×3	128×128×64	7×7	2
Conv2	128×128×64	64×64×128	3×3	2
Conv3	64×64×128	32×32×256	3×3	2
Conv4	32×32×256	16×16×512	3×3	2
Conv5	16×16×512	8×8×1024	3×3	2

The heat map prediction up samples the feature map to the desired output of 17 heatmaps—each representing the probability distribution for a given anatomical label. The dimensions of which may be seen below.

TABLE II
TRANSPOSE CONVOLUTIONAL LAYERS IN THE CUSTOM CNN

Layer	Input	Output	Kernel	Stride
Transpose Conv1	8×8×1024	16×16×512	4×4	2
Transpose Conv2	16×16×512	32×32×256	4×4	2
Transpose Conv3	32×32×256	64×64×17	4×4	2

For each heatmap, the coordinate with the highest accuracy is selected. If the accuracy is too low, this value is discarded, and is left undefined. This aim of this process is to select the most probable point, as shown in the function below:

$$(x, y) = \arg\max H(i, j) \tag{1}$$

where (x, y) is the predicted coordinate, and $H(i, j)$ is the intensity at the pixel (i, j).

A Mean Squared Error (MSE) loss function measures the accuracy of the heatmap outputs at the annotated locations, as shown in the formula below:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(H_i - \hat{H}_i)^2 \tag{2}$$

where H_i is the observed heatmap value at the ith pixel, $\hat{H}_i$ is the predicted heatmap value, and n is number of pixels in the given heatmap.

D. Pose Estimation and Posture Classification

Thresholds derived from ergonomic guidelines allow us to categorize postures into good, bad, or fair primarily based on torso and neck deviations [8] [9].

- Good: $S \leq 20\%$
- Fair: $20\% < S \leq 40\%$
- Bad: $S > 40\%$

The neck angle is computed as the angle between the shoulder and ear keypoints relative to the vertical axis, while the torso angle is computed as the angle between the shoulders and the hips. From these two angles, a posture score is calculated as:

$$S = 100 - \frac{(\text{neck deviation} + \text{torso deviation})}{2} \tag{3}$$

where:

- neck deviation: Absolute difference between the neck angle and the ideal angle (0)
- torso deviation: Absolute difference between the torso angle and the ideal angle (0)

Since there is a possibility that a pixel coordinate may be left undefined from the CNN, we use its anatomical counterpart. For example, in a left-side-view image, if the right shoulder is undefined, then the left shoulder is substituted in place of it's counterpart.

E. Model Rationale and Transition to YOLOv8-Pose

Based on our previous iterations and research, we concluded that a CNN was the best model for the project, as overall, it is better at identifying spatial patterns and generalizing across different settings.

Initially, however, a simple binary classifier was developed to categorize a user's posture as good or bad. This classifier worked to a degree, but lacked in returning specific feedback or quantifying the degree of good or bad posture. Therefore, we determined that a key-point-based model followed by angle analysis was optimal, providing specific areas and regions to correct.

Upon testing our custom CNN, we faced significant challenges such as poor accuracy, trouble with generalization, and extremely high computation requirements. Since keypoint detection typically requires deep architectures and extensive large-scale computation, our team transitioned to a pre-trained optimized YOLOv8-pose model [10]. Compared to other models like OpenPose and AlphaPose, YOLOv8 uses much less computational resources without sacrificing accuracy, making it ideal for our application—developed with accessibility in mind. Through using the lightweight YOLOv8 framework, our system was able to run at 30 FPS on standard laptops.

III. Results

A. Performance Metrics

Our custom CNN reached an average keypoint labeling accuracy of 47.3% for the shoulders, ears, hips, and knees (measured by the Percentage of Correct Keypoints (PCK) metric at a threshold of 0.5), while the YOLOv8 model improved this to 84.9%. To evaluate the accuracy of the rule-based posture classification system, 100 curated side-view images for good, fair, and bad postures were labeled, with 300 validation images in total classified with the YOLOv8 model and posture system. An overall accuracy of 92.3% was achieved (277/300).

TABLE III
Posture Classification Accuracy

Posture Category	Accuracy
Good Posture	89%
Fair Posture	92%
Bad Posture	96%

The accuracy for bad posture was notably higher, showing the need for further refinement of posture estimation and a more sophisticated rule-based system. Another major drawback of the current system was the need to curate and solely use images with a well-aligned side-view camera angle.

B. Real-Time Performance

The posture classification system was also assessed under several diverse environments for robustness and reliability. Under good lighting and minimal background noise, the system generally performed well. However, there were several instances where this model made errors in classification, particularly in cluttered environments. In environments with multiple individuals, such as at the Canadian Undergraduate Conference on AI (CUCAI), the model occasionally tracked people in the background rather than the target. Furthermore, when important keypoints for the angle calculations, like knees and hips, were fully hidden with no suitable replacement, detection precision drastically decreased. Other common misclassifications include head orientation, where momentary neck angle changes are seen as poor posture; confusing leaning and slouching with one another; and background noise disrupting keypoint identification.

C. Recommendations for Improvement

1) Background Noise
A major consideration before this system can be deployed is reducing and filtering out background noise. Based on our study, we have concluded that isolating the target individual from the background is an essential preprocessing step, which can be done by segmenting the target individual with a CNN or an alternative form of filtering.
2) Rule-Based Angle Analysis
Through testing, we discovered several errors associated with the rule-based posture system in place. Although a strong proof-of-concept was established, it is clear the system is overly simplistic: the thresholds were disrupted by occasional variations and the system was unable to recognize smaller but crucial details such as spine curvature at times. Our team recommends that a larger dataset with a greater number of anatomical keypoints, specific to posture, is utilized to allow for more advanced analysis.
3) Camera Perspective
For angle calculations, this two-dimensional system relies on a well-aligned side-view camera. When incorrectly oriented, the angle measurements are incorrect, leading to classification errors. We suggest a few strategies to mitigate this issue:
    - Rather than calculating angles to the vertical axis, calculate them relative to other keypoints to reduce dependency on camera angle.
    - For moderate amounts of camera warp, transform the image to a “perfect side-view” representation.
    - If the perspective were directly in front of the user, the model would need to recognize three-dimensional keypoint positions, which might be done with the use of depth sensors.

IV. Conclusion

This paper presents a real-time AI-powered posture detection system using a custom CNN and YOLOv8-pose model to classify posture based on neck and torso angles. Although the custom CNN provided valuable insights and research, because of challenges in generalization and computational demands, the YOLOv8 model was critical in deploying a system that offers instant posture feedback at 30 FPS; increasing our keypoint accuracy from 47.3% to 84.9% and allowing us to reach a posture classification accuracy of 92.3%. By developing this project with accessibility at the forefront, we have ensured our system functions on consumer-grade hardware, with instant feedback on a standard phone camera connected to a main device—offering numerous applications in fields such as workplace ergonomics, education environments, rehabilitation, as well as fitness and wellness.

While our results show strong potential, there are several challenges, such as dependencies on well-aligned camera angles, background noise, and obstructed keypoints. To address these issues, our team recommends expanding the dataset to label a greater amount of keypoints, refining the rule-based posture system for greater adaptability, and developing three-dimensional keypoint detection with depth sensors. In the long term, we plan to continue developing the user interface while incorporating more comprehensive user recommendations. By bridging computer vision and health sciences, this paper highlights the growing importance of artificial intelligence in preventative healthcare and ergonomic intervention.

References

[1] N. I. for Occupational Safety and Health, "Step 1: Identify risk factors," https://www.cdc.gov/niosh/ergonomics/ergo-programs/risk-factors.html, n.d., retrieved March 17, 2025.

[2] J. W. Lucas, E. M. Connor, and J. Bose, "Back, lower limb, and upper limb pain among u.s. adults, 2019 (nchs data brief no. 415)," https://www.cdc.gov/nchs/products/databriefs/db415.html, 2021, national Center for Health Statistics.

[3] C. M. Rahman, S. M. Uddin, M. A. Karim, and M. Ahmed, "Evaluation of work postures-the associated risk analysis and the impact on labor productivity," *ARPN Journal of Engineering and Applied Sciences*, vol. 10, no. 6, pp. 2542–2550, 2015.

[4] G. Hidalgo, Z. Cao, T. Simon, S.-E. Wei, H. Joo, Y. Raaj, and Y. Sheikh, "Openpose: Real-time multi-person keypoint detection library for body, face, hands, and foot estimation," https://github.com/CMU-Perceptual-Computing-Lab/openpose, 2018, computer software.

[5] MVIG-SJTU, "Alphapose: Real-time and accurate full-body multi-person pose estimation & tracking system," https://github.com/MVIG-SJTU/AlphaPose, n.d., computer software.

[6] M. Asad, "Coco 2017 keypoints," https://www.kaggle.com/datasets/asad11914/coco-2017-keypoints, 2020, computer software.

[7] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," https://github.com/cocodataset/coco, 2015, computer software.

[8] Ergoweb, "Posture evaluation," https://ergoweb.com/posture-evaluation/, n.d., retrieved March 17, 2025.

[9] J. Huizen, "Sitting positions: Posture and back health," *Medical News Today*, February 9 2023.

[10] G. Jocher, J. Qiu, and Ultralytics, "Ultralytics yolo," https://github.com/ultralytics/ultralytics, n.d., computer software.