Final project report

Pose-Based Human Localization and Height Estimation from FIFA Player Images

1. Introduction

Height reference markers are commonly attached to doors or walls in public spaces to assist surveillance systems in estimating a suspect's height from CCTV footage. However, in many real scenarios such explicit references are unavailable, motivating interest in estimating human height directly from images.

A key challenge in height estimation from images is the reliance on accurate human localization. It becomes highly unstable when images contain only partial views of a person. As a result, the focus of this project shifted from direct height estimation to pose-based human localization. This shift is motivated by the assumption that body part proportions correlate with height when the pose is correctly identified and sufficiently visible.

In this project, human body keypoints were first extracted using a pretrained pose estimation model. Based on these keypoints, images were classified into pose categories such as full-body, face-only, and partially visible poses. Height estimation was then performed on images using a simple regression model on keypoint coordinates.

Although the project did not achieve highly accurate height estimation, the pose classification model achieved stable performance in distinguishing different body visibility conditions. Furthermore, despite the relatively small size of the filtered training data, the height regression model showed that pose-based preprocessing can improve the stability of height estimation from images.

2. Background

2.1 COCO Keypoint Representation

The human pose information is extracted using a COCO-pretrained keypoint detection model based on a 17-keypoint representation. Its format is specifically designed for human pose estimation tasks, where the goal is not only to detect a person in an image, but also to localize major body joints such as the head, shoulders, hips, knees, and ankles.

Each detected person is represented by 17 keypoints, and each keypoint consists of an (x, y) coordinate in the image space. This representation provides a structured description of human body geometry, which is more informative than simple bounding boxes when analyzing body posture or proportions.

2.2 Limitations of Pretrained Keypoint Models

Although pretrained COCO keypoint models provide a strong baseline, they are not optimized for height estimation tasks. During experiments, several limitations were observed:

- inaccurate keypoint placement when the lower body is not visible
- visibility confidence scores of predicted coordinates assigned to 0 or 1, especially to 1 even for incorrect detections

These issues indicate that pretrained detectors alone are insufficient for reliable height estimation. This motivates the two-stage approach adopted in this project: (1) training a pose classification model to assess pose quality and body visibility, and (2) performing height estimation conditioned on the predicted pose category.

2.3 Pose Diversity in Web-Collected Image Data

Large-scale image datasets contain a heterogeneous mixture of poses and crops, as they are typically scraped from web sources such as Google or Wikipedia. As a result, these datasets include full-body images, face-centered images, and partially occluded bodies within the same collection.

This variability makes it difficult to treat all samples as equally valid inputs for height estimation. By categorizing images based on body visibility and pose quality, the problem can be formulated as a pose classification task rather than an implicit data filtering problem. The goal of this labeling process was not to directly improve height prediction, but to enable a model to learn meaningful distinctions between different pose qualities using keypoint geometry alone.

3. Data Resources

3.1 Data Sources

Two publicly available datasets from Kaggle were used, related to FIFA football players. The first dataset contains player metadata, including name, height, age, and other attributes, while the second dataset provides image folders of professional football players collected from various sources.

The height information in centimeters serves as the ground truth for the height estimation task. The image dataset consists of thousands of player images with large variations in camera angle, resolution, pose, and cropping. Although these datasets were not originally designed for height estimation, they were chosen due to the availability of both visual data and reliable height labels for the same individuals.

After matching player names between the two datasets, a total of over 40,000 labeled images were obtained as the initial image-level dataset before pose-based filtering.

3.2 Data Preprocessing and Name Matching

A major challenge in constructing the dataset was inconsistencies in player name formatting across data sources. Folder names in the image dataset often contained additional tokens such as country names, group labels, or encoding artifacts, while player names in the metadata dataset followed a standardized format.

To address this issue, player names were normalized by fixing encoding errors caused by mixed character sets, removing directory-specific tokens and parentheses, and converting text to lowercase and standardizing whitespace.

After normalization, player names were matched using exact string matching, when possible, followed by approximate string matching as a fallback. This process resulted in a clean image-level dataset where each image is associated with a single player and a corresponding height label.

3.3 Extracting COCO Keypoints and Filtering Usable Images

Human pose information was extracted from each image using a pretrained COCO keypoint detection model based on a Keypoint R-CNN architecture. For each detected person, the model outputs 17 keypoints corresponding to major body joints, including the nose, shoulders, hips, knees, and ankles.

Each keypoint is represented by an (x, y) coordinate in image space. All images were resized to a fixed resolution of (256 × 256) pixels prior to pose extraction to standardize coordinate scales across samples.

Images were retained only if exactly one person was detected. This step removed images containing multiple people, ensuring that each sample corresponds to a single individual, and therefore, a single height label.

3.4 Pose-Based Labeling

Vertical distances between key body parts, such as nose, shoulders, hips, knees, and ankles, were computed using keypoint coordinates. These distances were used to infer pose quality. Based on simple geometric heuristics, each image was assigned to one of the following categories:

- full_body: all major body segments are visible with a plausible vertical order
- upper_body: lower body segments are truncated
- face: facial region dominates the detected pose
- lower_cut: lower body is cut off near the bottom of the image
- full_body_bent: full body is present but in a bent posture
- malformed: physically implausible keypoint configurations

Images labeled as malformed were excluded from further analysis.

In early experiments, keypoints corresponding to missing body parts were masked as NaN. However, this masking strategy was later removed, as it made pose classification degraded the performance of downstream height regression.

During later experiments, pose labels were further simplified into three categories—full_body, partial_body, and face—due to classification ambiguity between upper-body and lower-cut poses.

3.5 Final Dataset Composition

After extraction and filtering, the dataset was significantly reduced in size. This reduction highlights a trade-off between pose quality and dataset size, which significantly impacts the reliability of height regression.

The final labeled dataset contains several hundred images, with a strong imbalance across pose categories. Full-body images form most usable samples for height estimation.

This dataset was used in two stages:

1. pose-based image classification to assess pose quality
2. height estimation on images classified as valid full-body poses

In addition, baseline experiments were conducted without pose filtering to provide a comparative reference. Although the resulting dataset is small compared to the original image pool, it provides higher-quality supervision for analyzing the relationship between human pose and height estimation.

Pose category	full-body	partial-body	face
1165	336	815	14

- Table 1 : Final Dataset Composition

4. Methods

4.1 Pose Keypoint Extraction

Human pose information was extracted using a pretrained COCO Keypoint R-CNN model with a ResNet-50 backbone and Feature Pyramid Network. This model was selected due to its widespread use as a standard baseline for human pose estimation tasks. Pretrained weights from the COCO dataset were used without additional fine-tuning.

Input images were resized to a fixed resolution prior to inference to standardize coordinate scales, normalized to the [0, 1] range, and provided to the model as a list of tensors.

For each detected person, the model outputs 17 keypoints corresponding to major human body joints. Each keypoint is represented by a triplet (x, y, confidence), where (x, y) denotes image space coordinates and the confidence value reflects the model's internal visibility estimate. Only the (x, y) coordinates were used in subsequent stages. Confidence scores were excluded from both pose classification and height regression, as they were observed to be assigned to extreme values and provided limited discriminative information in practice.

4.2 Pose-Based Image Classification

Pose classification was performed using only keypoint geometry as input features. No pixel-level image information was used. Pose labels were generated automatically using heuristic, rule-based criteria defined on keypoint coordinates during dataset definition. These rules were designed to reflect body visibility and pose completeness, such as whether lower body joints were present in a plausible vertical order.

The classification model was implemented as a multi-layer perceptron consisting of fully connected layers with ReLU activations and dropout regularization. This lightweight architecture was sufficient to capture geometric differences between pose categories while minimizing overfitting, given the limited size of the labeled dataset.

4.3 Height Estimation Model

Height estimation was formulated as a regression problem conditioned on pose quality. In the pose-based setting, height prediction was performed only on images classified as valid full-body poses. This design explicitly separates pose-related uncertainty from the height regression task.

For comparison, a baseline regression model was also trained without pose-based filtering, using identical input features but without conditioning on pose classification.

4.3.1 Model Architecture and Training Procedure

Height regression was implemented using a feedforward neural network composed of fully connected layers with ReLU activations and dropout. Mean squared error was used as the optimization loss, while mean absolute error was used for evaluation to provide interpretable error values in centimeters. Model selection was performed based on validation performance.

The same regression architecture was used for both pose-based and baseline experiments to ensure a fair comparison. To further control for dataset size effects, the baseline regression model was trained on a randomly sampled subset of the full dataset, matched in size to the pose-filtered full-body dataset.

5. Results and Evaluation

5.1 Pose Classification Performance

The pose classification model was trained to distinguish three pose categories: full_body, partial_body, and face. Classification performance was evaluated using accuracy on a validation set.

Training converged rapidly within a small number of epochs. Validation accuracy exceeded 85% by epoch 2 and stabilized above 95% after epoch 8. The best-performing model achieved a validation accuracy of approximately 97%, showing that keypoint geometry alone provides sufficient information to distinguish pose quality categories.

Pose quality could be reliably inferred from COCO keypoints using a lightweight classifier, providing a stable foundation for downstream height estimation.

5.2 Height Estimation Performance

Two experimental settings were compared:

1. Pose-based regression, where height estimation was performed only on images classified as full_body poses.
2. Baseline regression, where height estimation was performed without pose-based filtering, using an equally sized randomly sampled dataset for fairness.

Model	Dataset Size	Train MSE	Val MAE (cm)
Pose-based	336	~840	15.0
Baseline	336	~838	18.8

- Table 2 : Height Estimation Performance Comparison

5.2.1 Pose-Based Height Regression

In the pose-based setting, training loss decreased steadily across epochs, with validation MAE dropping sharply during early training. After approximately 15 epochs, validation MAE stabilized around 15 cm, with minor fluctuations thereafter. This behavior suggests that the model quickly learns coarse height-related patterns from keypoint geometry but is limited in capturing fine-grained height differences.

5.2.2 Baseline Height Regression

The baseline regression model exhibited a similar learning trajectory but after convergence, the baseline model reached a validation MAE of approximately 19 cm, compared to 15 cm for the pose-based model when trained on datasets of equal size.

5.3 Comparison and Interpretation

When controlling for dataset size, the baseline model achieved lower training error but higher validation MAE, indicating overfitting to pose variability. The pose-based model showed more stable generalization by restricting samples to full-body poses, suggesting improved robustness despite modest numerical gains.

6. Discussion

6.1 Saturation of Height Estimation Performance

The observed validation MAE around 15 cm suggests a fundamental limitation in the available input information rather than insufficient model capacity. While the regression model was able to learn relationships between body proportions and height, fine-grained height differences could not be reliably captured. This limitation is expected given that all inputs are derived from 2D keypoint coordinates without any depth or scale reference.

Even for full-body images, absolute height cannot be uniquely determined from monocular images alone. Variations in camera distance, focal length, and perspective distortion can produce similar keypoint configurations for individuals of different heights. As a result, the model tends to converge toward predicting values near the dataset mean, limiting further improvements in MAE.

6.2 Effect of Dataset Size and Pose-Based Filtering

Pose-based filtering significantly reduced the dataset size by restricting training samples to images classified as full-body poses. While this improved input consistency and reduced pose-related noise, it also limited the diversity and quantity of training data available for regression. The small dataset size increases sensitivity to noise and reduces the model's ability to learn subtle height-related patterns.

To ensure a fair comparison, baseline regression experiments were conducted using randomly sampled datasets matched in size to the pose-based full-body dataset. Interestingly, when baseline model was trained on larger unfiltered datasets, the baseline model achieved lower numerical MAE but exhibited strong regression to the mean average, predicting similar height values for most samples. Although this

improves average error metrics, it undermines the semantic meaning of regression, as individual height differences are no longer meaningfully represented.

6.3 Limitations of Monocular Keypoints and Confidence Scores

Visibility confidence scores produced by the pretrained keypoint model were found to direct toward mostly 1, even for incorrect detections. As a result, confidence values were excluded from both pose classification and regression models. Performance limitations arose from the fundamental ambiguity of the task and the characteristics of the data.

6.4 Implications for Pose-Based Height Estimation

The results indicate that pose-based preprocessing improves robustness by explicitly separating pose quality from height estimation. While numerical improvements over the baseline remain modest, the pose-based approach avoids degenerate solutions dominated by mean predictions and yields more interpretable regression behavior.

Overall, pose-based localization is a necessary but not sufficient condition for accurate height estimation from images. Further improvements would likely require additional sources of scale information, such as multi-view imagery, or camera metadata.

8. Conclusion

Experimental results showed that pose quality can be reliably inferred from COCO keypoints using a lightweight classification model, achieving high validation accuracy despite limited training data. Conditioning height regression on pose quality improved robustness, reducing degenerate mean prediction behavior observed in baseline models trained without pose filtering. However, absolute height estimation from monocular keypoints remained fundamentally limited, with validation error around 15 cm due to scale ambiguity and the absence of depth information.

These findings suggest that pose-based preprocessing is a necessary but not sufficient condition for accurate height estimation from single images. While numerical improvements were modest, separating pose uncertainty from regression led to more meaningful predictions and clearer failure modes. Future work could incorporate additional sources of scale information to further reduce ambiguity in height estimation.