Human Pose Estimation And Tracking Via Keypoints Detection

Bagus Hardiansyah
Universitas 17 Agustus 1945 Surabaya, bagushardiansyah@untag-sby.ac.id
Fajar Astuti Hermawati
Universitas 17 Agustus 1945 Surabaya, fajarastuti@untag-sby.ac.id

Abstract

effectiveness of the pose estimation model in accurately predicting various human poses. The consistently high mAP values across all pose classes validate the robustness and precision of the model, making it suitable for real-world applications such as motion analysis, surveillance, and sports performance tracking. Minor discrepancies in scores highlight areas for potential optimization, particularly for poses with slightly lower mAP values. overall mAP of 0.99 demonstrates that the model is highly accurate in estimating poses. Individual class scores show slight variations. For example, "L-bot-pose" has the lowest mAP (0.963). High scores for other classes (e.g., "R-bot-pose" with 0.995) indicate that the model can detect these poses with exceptional accuracy

Keywords: pose estimation, human poses, Pose Tracking

Introduction

Kinematic tracking of individuals is a highly significant task, as people are often interested in monitoring the actions of others. Large-scale, precise, and automated kinematic tracking can enable valuable applications such as surveillance video analysis, studies of human behaviour, and large-scale motion capture. However, existing systems lack the ability to perform kinematic tracking on a large scale, typically demonstrating results on only a few hundred frames. We have developed an accurate and fully automated algorithm capable of processing and evaluating results across oversized frames to address this limitation. This advancement has been made possible through the implementation of deep learning-based architectures [1, 2] and the availability of extensive benchmark datasets, such as the "MPII Human Pose" dataset [3] and the "MS COCO" dataset [4]. These resources have significantly improved the accuracy and scalability of kinematic tracking systems. The body of literature on human tracking is vast and cannot be exhaustively reviewed here. Tracking individuals presents significant challenges due to the rapid movements people can make and the wide variety of poses they can adopt. Crucially, benchmark datasets like "MPII Human Pose" and "MS COCO" have not only supplied the large training sets necessary for training deep learning-based methods but have also introduced detailed metrics. These metrics enable direct and fair performance comparisons among a wide range of competing approaches, driving progress in the field.

The current approach utilizes the configuration in the current frame along with a dynamic model to predict the configuration in the subsequent frame, with these predictions being further refined using image data [5]. Although significant advancements have been made in single-frame multiperson pose estimation, the challenge of articulated multiperson body joint tracking remains largely unaddressed in existing datasets. In this work, we aim to bridge this gap by introducing a new large-scale, high-quality dataset designed for human body keypoint detection, enabling improved pose estimation and articulated tracking.

Related tasks primarily address single-frame person pose estimation, pose estimation in static images, and multi-person articulated tracking. Although the dataset primarily focuses on person-articulated tracking, advancements in single-frame pose estimation are expected to significantly enhance the overall quality of tracking and pose estimation. These interconnected tasks improve the robustness and accuracy of articulated tracking systems.

Methodology

Human pose estimation and tracking detecting and localizing key body parts, such as joints (e.g., elbows, knees, wrists), and connecting them to form a structured pose. Deep learning methods are commonly used for keypoint detection, estimating the positions of these joints in images or videos. This process is fundamental in applications such as activity recognition, augmented reality, and sports analytics. Single-frame pose estimation focuses on analyzing one image at a time, estimating the pose independently for each frame in a video sequence [6].

Deep learning-based pose estimation typically employs convolutional neural networks (CNNs) or transformer architectures. These models process an image and output a set of keypoints representing body joints. A popular method is using heatmaps, where each joint corresponds to a probability distribution over the image space, indicating the likely location of the joint. Models such as OpenPose, HRNet, and PoseNet have driven significant advancements in single-frame pose estimation by leveraging techniques like multi-scale feature extraction, multi-stage refinement, and supervised learning on large annotated datasets [4].

Single-frame pose estimation models operate independently on each frame of a video, disregarding temporal relationships between consecutive frames. While this simplifies computation, it may lead to inconsistencies or jitter when tracking poses across frames, especially in videos with rapid movements. To address this, post-processing techniques like smoothing or filtering are applied to estimated keypoints, improving temporal consistency in video analysis [7].

To handle variations in human appearance, poses, and environmental conditions, deep learning models depend on robust training datasets featuring diverse poses and scenarios. Annotated datasets like COCO, MPII, and Human3.6M provide rich resources for training, helping models generalize effectively. Data augmentation techniques, such as rotation, scaling, and flipping, further enhance model robustness against changes in scale, orientation, and lighting. Pretrained models are often fine-tuned for specific applications to achieve better domain-specific performance [8].

Despite significant progress, single-frame pose estimation faces challenges in real-time applications where continuity and accuracy are essential. Integrating multi-frame analysis or temporal information through pose tracking can address these limitations. Nonetheless, single-frame methods remain valuable for scenarios requiring individual pose snapshots, offering efficient and accurate solutions for applications such as static posture analysis, gesture recognition, and image-based fitness apps [9].

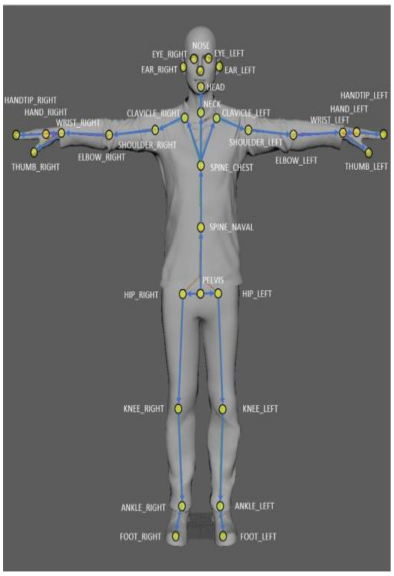

Figure 1. Illustration Joints tracked by the body tracking

This image represents a keypoint-based human pose estimation framework, where specific anatomical landmarks (keypoints) of the human body are identified and localized in a visual representation. These keypoints correspond to critical joints and body parts, such as the head, shoulders, elbows, wrists, pelvis, knees, and ankles, among others. Each keypoint is marked by a yellow dot, and the connections between these points form a skeletal structure, visualizing the overall pose of the person.

Key Features of the Image:

1. Keypoints:

The yellow dots highlight key anatomical landmarks, such as the nose, eyes, ears, shoulders, elbows, wrists, hips, knees, ankles, and feet. These keypoints are essential for understanding the posture and alignment of the body.

2. Connections:

The blue lines connecting the keypoints represent the skeletal structure of the human body. These connections illustrate the spatial relationship between keypoints, forming a virtual "stick figure" that outlines the pose.

3. Pose Estimation:

The framework likely uses a deep learning model trained on annotated datasets (e.g., COCO or MPII) to detect and estimate the location of these keypoints in an image. The model outputs a 2D or 3D representation of the human pose.

4. Applications:

This type of keypoint detection can be applied to activities like action recognition, motion analysis, sports performance tracking, healthcare monitoring, and gesture-based interaction systems.

5. Keypoint Labels:

Each keypoint is labeled with its corresponding body part (e.g., NOSE, SHOULDER_LEFT, ANKLE_RIGHT). This ensures clarity in identifying specific joints for further analysis.

In essence, the image showcases the results of a pose estimation model, which provides a detailed skeletal representation of a human subject by localizing and connecting body keypoints. This visualization makes it easier to analyze human movement, gestures, or postures in various fields like robotics, computer vision, and biomechanical studies.

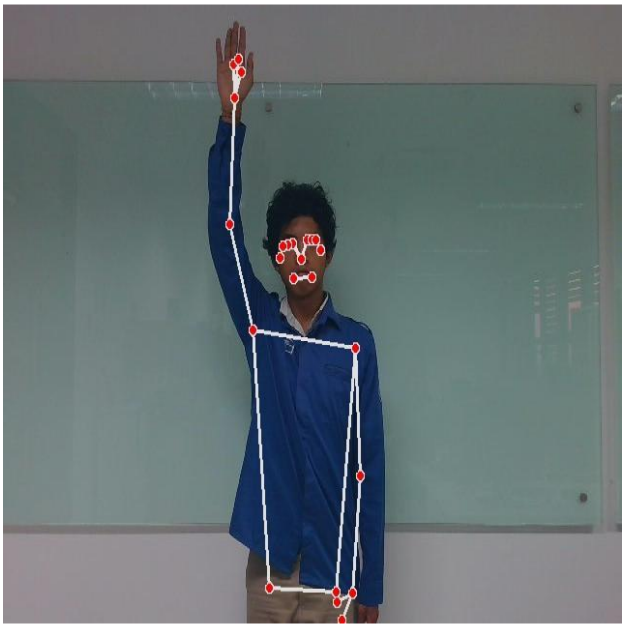

Figure 2. Pose Estimation the Body Tracking

This image of human pose estimation through keypoints detection, where red dots represent key body joints (nose, shoulders, elbows, wrists, hips, knees) and white lines connect these points to form a skeletal structure. Using deep learning models the system identifies the position of each joint in a single frame, accurately capturing the pose of the person, such as their raised arm. This technique is widely used in applications like activity recognition, motion analysis, gesture control, and healthcare, providing a detailed understanding of human posture and movement through visual analysis.

Results and Discussion

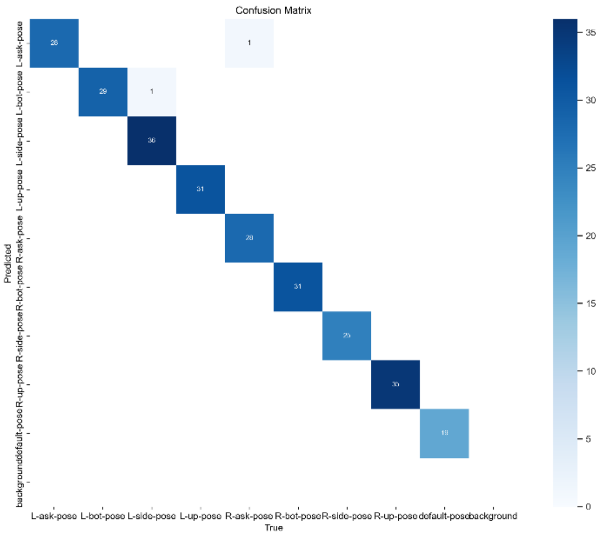

Figure 3. Confusion Matrix

This represents a confusion matrix for a validation dataset, typically used to evaluate the performance of a classification model. The confusion matrix summarizes the prediction results of the model, showing the relationship between actual (true) labels and predicted labels for different classes. Each row corresponds to the actual class, while each column corresponds to the predicted class, with the diagonal elements representing correct predictions (true positives) for each category.

From the confusion matrix, we can observe that the majority of predictions lie on the diagonal, indicating that the model performs well for most classes. For instance, classes like "L-ask-pose," "L-side-pose," and "R-up-pose" show high true positive counts, demonstrating that the model accurately classifies these poses in most cases. Off-diagonal elements represent misclassifications, where the model incorrectly predicts a pose as belonging to another class. For example, there are a few misclassifications where "L-bot-pose" is confused with "L-side-pose"

The confusion matrix also highlights areas where the model may require improvement. Misclassifications can occur due to overlapping features between poses, insufficient data for certain categories, or complex variations in pose appearances. To address these issues, techniques such as augmenting the training data, fine-tuning the model, or using advanced architectures can help improve classification performance. The matrix is a valuable tool for pinpointing specific areas for optimization and guiding improvements in model design, as shown in Figure 3.

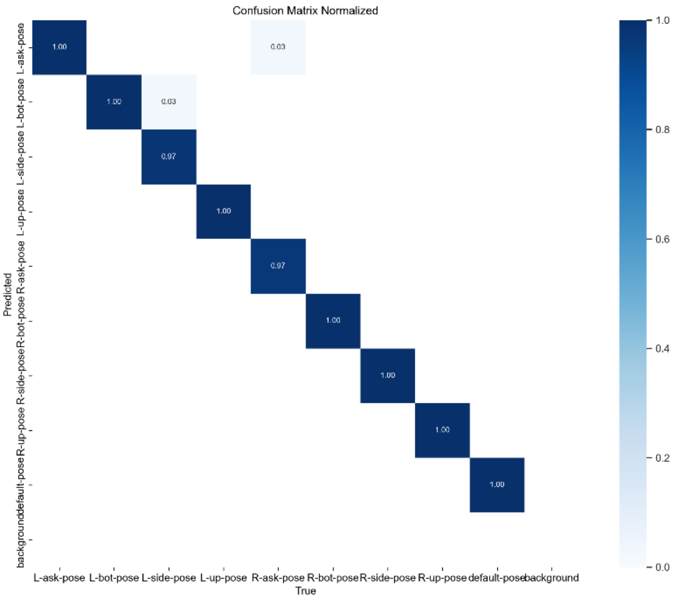

Figure 4. Confusion Matrix Normalized

The normalized confusion matrix shown in the image is a visualization of a classification model’s performance, where each value is normalized to represent the proportion of predictions relative to the total actual samples for each class. This normalization scales the values to range from 0 to 1, making it easier to interpret the model's performance independently of the sample size for each class. For example, diagonal values close to 1 indicate high accuracy for the corresponding class.

In this matrix, the diagonal elements (representing true positives) dominate with values near or at 1.00 for most classes, such as "L-ask-pose," "L-up-pose," and "R-side-pose." This suggests that the model classifies these poses with near-perfect accuracy. Misclassifications are represented by off-diagonal values, which are notably small (e.g., 0.03), indicating that errors are minimal and the majority of the predictions align with the ground truth.

The normalized confusion matrix is particularly useful for identifying classes with imbalanced data, as it allows for a fair comparison of performance across all classes regardless of their sample size. By showing proportions rather than raw counts, it provides a clearer view of the model's relative strengths and weaknesses, making it a valuable tool for performance evaluation and further optimization of classification models.

This table presents the quantitative validated results for pose estimation performance, measured using mean Average Precision (mAP). The mAP metric is a standard evaluation method for assessing the accuracy of pose estimation models, representing the average precision across all keypoints or poses detected. Higher mAP values (closer to 1.0) indicate better performance.

Key Elements of the Table:

1. Class:

This column specifies the type of pose or the category being evaluated. Examples include "L-ask-pose" and "R-bot-pose," representing distinct human body configurations.

2. Image Instances:

This column indicates the number of image samples available for each pose class in the validation dataset. For example, 28 images are used for "L-ask-pose," while 37 are used for "L-side-pose."

3. mAP:

This column reports the mAP value for each pose class. For instance, "L-ask-pose" and "L-side-pose" achieve high mAP scores of 0.994, indicating near-perfect precision in detecting these poses. The overall mAP for all classes is 0.99, signifying that the model performs exceptionally well across all evaluated poses.

Table 1. Quantitative Validated Results Pose estimation performance (mAP)

No	Class	Images Instances	mAP
	all	264	0.99
1	L-ask-pose	28	0.994
2	L-bot-pose	29	0.963
3	L-side-pose	37	0.994
4	L-up-pose	31	0.995
5	R-ask-pose	29	0.985
6	R-bot-pose	31	0.995
7	R-side-pose	25	0.995
8	R-up-pose	35	0.995
9	default-pose	19	0.995

Interpretation:

1. The overall mAP of 0.99 demonstrates that the model is highly accurate in estimating poses across the dataset.
2. Individual class scores show slight variations. For example, "L-bot-pose" has the lowest mAP (0.963), suggesting that the model may have some difficulty accurately detecting this pose compared to others.
3. High scores for other classes (e.g., "R-bot-pose" with 0.995) indicate that the model can detect these poses with exceptional accuracy.

Conclusion

This article emphasizes the significance of human kinematic tracking using deep learning-based pose estimation techniques. Through advancements in deep learning architectures and the availability of robust datasets like "MPII Human Pose" and "MS COCO," the study addresses critical challenges such as scalability, accuracy, and automated tracking across large frames. By introducing new datasets and focusing on single-frame pose estimation, the work bridges the gap in articulated multiperson pose tracking, enabling progress in activity recognition, motion analysis, and other fields. The methodology, which relies on keypoint detection and subsequent pose estimation, demonstrates the potential of neural networks in accurately capturing human movement across varied poses and conditions.

The results, as illustrated by the confusion matrix, highlight the model's high accuracy in detecting key human poses, with most predictions aligning closely with true labels. Misclassifications are minimal, but they reveal areas for improvement, such as addressing overlapping poses or refining data for specific categories. The normalized confusion matrix further validates the model's performance by showing proportions of correct classifications, ensuring fair evaluations even in the case of imbalanced datasets. These tools underscore the robustness of the system and its ability to generalize effectively across diverse scenarios.

In conclusion, this research demonstrates that deep learning methods, combined with large-scale datasets and robust evaluation metrics, have significantly advanced the field of kinematic tracking and pose estimation. While single-frame pose estimation shows great promise for applications such as gesture recognition and static posture analysis, the integration of multi-frame temporal analysis could further enhance the model's continuity and accuracy in real-time applications. The study provides a strong foundation for future research aimed at refining tracking techniques, improving dataset quality, and exploring new domains of application.

The table reflects the effectiveness of the pose estimation model in accurately predicting various human poses. The consistently high mAP values across all pose classes validate the robustness and precision of the model, making it suitable for real-world applications such as motion analysis, surveillance, and sports performance tracking. Minor discrepancies in scores highlight areas for potential optimization, particularly for poses with slightly lower mAP values.

Acknowledgments

Thank You LPPM Universitas 17 Agustus 1945 Surabaya through the Fundamental Research Grant (Hibah Penelitian Fundamental Sumber Dana Perguruan Tinggi) for funding this research. Thank you dear colleagues and co-writers for the hard work and supports.

References

Reading materials should use the Mendeley Cite Style IEEE, with at least 15 journal articles published in the last 5 (five) years before this manuscript's publication.

[1] K. 01. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.

[2] K. 02. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[3] M. 03. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3686–3693, 2014, doi: 10.1109/CVPR.2014.471.

[4] T. Y. 04. Lin *et al.*, "Microsoft COCO: Common objects in context," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014, doi: 10.1007/978-3-319-10602-1_48.

[5] C. 05. Bregler and J. Malik, "Bregler-Malik98," pp. 1–8, 1998, [Online]. Available: papers://90b07e07-903c-45d4-a899-2629f88b6b69/Paper/p1604

[6] Z. 06. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 1302–1310, 2017, doi: 10.1109/CVPR.2017.143.

[7] K. 07. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 5686–5696, 2019, doi: 10.1109/CVPR.2019.00584.

[8] Y. 08. Liu, C. Qiu, and Z. Zhang, "Deep learning for 3D human pose estimation and mesh recovery: A survey," *Neurocomputing*, vol. 596, 2024, doi: 10.1016/j.neucom.2024.128049.

[9] T. 09. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 4645–4653, 2017, doi: 10.1109/CVPR.2017.494.