Proceedings of CONF-CDS 2025 Symposium: Application of Machine Learning in Engineering
DOI: 10.54254/2755-2721/2025.AST26531

Research on Human Pose Estimation Model Based on Convolutional Neural Network

Zhiyun Zhang

College of Zidingxiang, Harbin Institute of Technology, Weihai, China
zzyzhanghit@163.com

Abstract. Human pose estimation is a key area in computer vision with applications in human-computer interaction, intelligent monitoring, and motion analysis. Pose estimation aims to automatically identify human body key points in images, which has significant practical value. Recent advances in convolutional neural networks (CNNs) have made deep learning the dominant approach for pose estimation. This paper proposes a simplified CNN-based model for human pose estimation. Unlike traditional complex architectures, our model improves computational efficiency while retaining strong learning ability. To evaluate its performance, we created a random pose dataset with 100 samples, each consisting of a zero image and 17 randomly generated key point coordinates, normalized in the [0, 1] range. The model is trained using mean square error (MSE) as the loss function, and the Adam optimizer for weight updates. Through backpropagation, the model learns to minimize the gap between predicted and real key points. Experimental results show that, despite the dataset's simplicity, the model effectively locates key points and demonstrates strong learning capabilities. In the future, we plan to use more complex real-world datasets and optimize the network structure to improve performance in more challenging scenarios. In conclusion, the proposed CNN-based model demonstrates its potential for posture estimation tasks.

Keywords: human pose estimation, convolutional neural network, key point prediction, mean square error, data visualization

1. Introduction

With the rapid development of computer vision [1] technology, human posture estimation [2] has become an important research direction in intelligent systems. Human posture estimation refers to the spatial structure of the human body by analyzing the position of key points of the human body in images or videos, and then achieving an accurate description of human movement. It is widely used in human-computer interaction [3], intelligent monitoring [4], motion analysis [5], virtual reality [6] and other fields, and has become an important research topic in the field of modern artificial intelligence. Posture estimation can not only help computers "understand" the behavior of the human body, but also provide a more intelligent and natural interaction method for various intelligent systems.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.AST26531

In the field of human-computer interaction, the application of pose estimation is of great significance. Through posture estimation technology, computers can recognize and understand human movements in real time, and then respond accordingly, making the interaction between humans and computers more intuitive and natural. For example, gesture recognition based on posture estimation [7] allows users to control smart devices through body language, achieving contactless operation and improving the efficiency and experience of human-computer interaction. In the field of intelligent monitoring, posture estimation technology can help monitoring systems identify human activity trajectories and analyze abnormal behaviors, thereby improving public safety protection capabilities. Especially in applications such as security monitoring, smart home, and health monitoring, real-time estimation [8] of human posture is crucial for the early warning and handling of events.

Although human pose estimation technology has made significant progress, it still faces many challenges. Firstly, the accuracy and robustness of pose estimation remain a challenge in the case of complex backgrounds, multiplayer scenes, and partial occlusion. For example, in the case of multi-person interaction or group movement, how to accurately distinguish and estimate the posture of each person is still a difficult point in current research. Secondly, due to the inaccurate annotation of existing datasets and insufficient pose diversity, many pose estimation models have large errors in practical application. The lack and diversity of datasets make the generalization ability of the model poor, making it difficult to cope with the pose estimation problem in complex environments. In order to understand these challenges, how to build more realistic and diverse attitude datasets and how to design more efficient and accurate deep learning models have become hot topics in current attitude estimation research. In this paper, a simple pose estimation model based on convolutional neural networks (CNNs) [9] is proposed, and a simplified random dataset is used for training and testing. Our model effectively predicts the position of key points in the human body by constructing a simplified network architecture, and visualizes the predicted posture skeleton diagram [10], which intuitively presents the prediction effect of the model. The main contributions of this paper include: designing a simplified random dataset to verify the validity of the model architecture; A simple attitude estimation method based on CNN is proposed, which aims to improve the computational efficiency and prediction accuracy of the model. The feasibility of the proposed method under simplified conditions is verified by experiments, and the experimental results are analyzed in detail.Be advised that papers in a technically unsuitable form will be returned for retyping. After returned the manuscript must be appropriately modified.

2. Related work

2.1. Traditional attitude estimation methods

Early studies of human pose estimation mainly relied on traditional image processing techniques and statistical modeling methods. The original pose estimation method mostly used template matching, which directly compared with the predefined pose template to infer the position of human joints in the image. However, this method performs poorly when dealing with complex scenes, dynamic poses, and occlusion, and its accuracy is greatly reduced. Pose estimation methods based on probabilistic graph models, such as conditional random fields (CRF) and Markov random fields (MRF), model human pose by establishing spatial relationships between joints. To a certain extent, this type of method improves the robustness of occlusion problems and posture changes [11], but its computational complexity is high and the accuracy of data annotation is highly dependent, so it is difficult to cope with complex changes in real scenes. Although traditional methods can achieve

DOI: 10.54254/2755-2721/2025.AST26531

certain results in some specific scenarios, they cannot solve the challenges caused by background complexity and diverse gestures due to the lack of sufficient expressive ability.

2.2. Advances based on deep learning

This With the rapid development of deep learning technology, human pose estimation methods based on convolutional neural networks (CNNs) have gradually become mainstream. The deep learning method automatically learns the features of images through a multi-layer network, avoiding the limitations of manual feature design in traditional methods, and greatly improving the accuracy and robustness of pose estimation. OpenPose is one of the most well-known methods, which enables both single and multi-person pose estimation through an end-to-end convolutional neural network architecture. The Part Affinity Fields (PAFs) [13] technology introduced by OpenPose [12] enables the effective modeling of spatial relationships between joints, thereby improving the accuracy of pose estimation in complex scenarios. Another commonly used method is the Hourglass network [14], which gradually refines the joint position by extracting multi-scale features layer by layer, especially in multi-person pose estimation. Compared with traditional methods, deep learning methods can handle more complex backgrounds, multi-person occlusion, and pose changes, and have become the mainstream technology in the field of pose estimation. However, most of these methods rely on large-scale labeling datasets [15], such as COCO, MPII, etc., which makes the quality and diversity of the dataset critical to the accuracy and generalization ability of the model.

2.3. Datasets and simplified experiments

Although deep learning-based methods have achieved remarkable results in pose estimation, they often rely on large-scale, accurately labeled training datasets, which places high demands on the cost and quality of data labeling. Although the existing datasets (such as COCO and MPII) cover more posture and scene changes, there are still problems such as inaccurate annotation and insufficient data diversity, which poses a challenge to training high-precision models with good generalization ability. To solve this problem, some researchers have begun to explore the use of simplified and synthetic datasets to verify the effectiveness of the model. By generating random key point coordinates and simple background images, researchers can avoid the interference of background complexity and occlusion problems in real-world data, and focus on the optimization of model architecture and training strategies. These simplified datasets can help quickly evaluate the advantages and disadvantages of different pose estimation models, and provide ideas for subsequent research based on real datasets. This study adopts this simplified dataset design method to verify the performance of the model under simplified conditions through randomly generated pose samples, and lays the foundation for further model improvement and practical application.

3. Method

3.1. Dataset design

In this study, to simplify the experiment and focus on the design of the model architecture, we designed a simplified randomly generated pose dataset. The dataset contains 100 samples, each consisting of an all-zero image and a set of randomly generated keypoint coordinates. The images are 256x256 pixels in size, and each image contains 3 channels (RGB), simulating the basic characteristics of a real image. The key point coordinates correspond to the positions of 17 human

DOI: 10.54254/2755-2721/2025.AST26531

joints, such as head, shoulder, elbow, knee, etc., and these coordinates are normalized to the range of [0, 1] in each sample, representing the position of the relative image size.

This randomly generated dataset does not contain actual image content and, therefore, does not represent real-world human poses. Its main purpose is to verify the model's learning ability and architecture design under simplified conditions. By removing complex background and environmental distractions, we were able to focus on how the network architecture can efficiently handle human pose estimation tasks. In addition, the simplification of data also avoids the impact of annotation errors and noise on the training process. Although the data set is not directly related to real-world data, it provides us with a concise experimental framework for testing the accuracy and learning ability of the model under basic conditions.

Although this dataset simplifies many complex factors, it can still help us evaluate the performance of convolutional neural networks under different environmental conditions, especially when the dataset is simple and noisy. In this way, we can observe how the model responds to the generation of different postures and verify its ability to predict key points. This lays the foundation for subsequent research based on real data, helping us understand the performance and potential of network architectures under simplified conditions.

3.2. Model structure

The pose estimation model proposed in this study is based on the classical convolutional neural network (CNN) architecture, which aims to provide an efficient and accurate pose estimation solution. The model is mainly composed of three convolutional layers, a pooled layer and two fully connected layers. The function of the convolutional layer is to extract the spatial features of the input image, and the pooling layer helps to reduce the spatial size of the image, thereby reducing the amount of computation and preventing overfitting. Each convolutional layer is connected with a ReLU activation function, which is used to introduce nonlinear factors and improve the expressiveness of the network. The design of the convolutional layer enables the network to automatically extract the local features of the image through the local receptive field, and abstract richer high-level features layer by layer.

After the convolutional layer is the pooled layer, which can reduce the spatial size by downsampling the feature map, and at the same time retain important feature information in the image. In this way, the model is able to extract important information related to human posture from a more compact feature map. The fully connected layer receives the feature maps output by the convolution and pooling layers, further integrates these features, and finally outputs the (x, y) coordinates of the 17 joints through the Sigmoid activation function. Thanks to the Sigmoid function, the network is able to normalize the coordinates of the output to the range of [0, 1], ensuring that the output of each joint coordinate is within a reasonable range, which is suitable for the single-person pose estimation task. The simplicity and efficiency of the model's structure make it excellent in terms of training speed and computational resource consumption, while ensuring strong learning ability.

3.3. Training and testing

In the training process, the mean square error (MSE) is used as the loss function, and the Adam optimizer is selected for model optimization. The MSE loss function calculates the square difference between the predicted key points and the true key points, with the aim of minimizing the prediction error and making the output of the model as close to the true value as possible. Adam Optimizer

DOI: 10.54254/2755-2721/2025.AST26531

automatically adjusts the learning rate during the training process to adapt to different training stages, improve the convergence speed, and effectively avoid the problem of too large or too small learning rate.

Each round of training uses backpropagation algorithms for gradient calculation and weight updates. The backpropagation algorithm calculates the influence of each weight on the final loss function, gradually adjusts the parameters of the network, so that the loss value gradually decreases, and finally reaches the network convergence. During the testing phase, we evaluated the model using the same random dataset and output the predictions for each sample. Through the results of the test phase, we can evaluate the prediction accuracy of the model in practical applications, especially in the case of data simplification, whether the model can effectively learn the law of joint positioning. During training, we monitor the loss changes in each round and evaluate the generalization ability of the model through the validation set. This training strategy allows us to quickly obtain an effective attitude estimation model under simplified conditions, which can provide a reference for subsequent experiments on more complex datasets.

4. Experimental results and analysis

4.1. Training process

In this experiment, we trained a human pose estimation model based on convolutional neural network (CNN). During the training process, the loss value of the model gradually decreases with the increase of training rounds, which indicates that the network is gradually optimizing its prediction ability. In the early stage of training, the loss value is high because the weight initialization of the network is relatively random. However, as the training progresses, the network gradually adjusts its weight through backpropagation, and the loss value gradually decreases and stabilizes. Figure 1 shows the change of loss value during training, which gradually decreases and eventually stabilizes as the training progresses.

In the early stages of training, the model begins to learn from the data, but the loss is high due to the random network parameters. After several rounds of training, the model's loss gradually decreases and gradually converges. This suggests that even if the dataset is simple (randomly generated data), the network can learn efficiently and eventually converge to a reasonable predictive value. In this way, we verify the effectiveness of the network architecture, proving that convolutional neural networks can effectively learn laws and optimize predictions even under the condition of simplifying data sets.

Table 1. Changes in loss values during training

Training rounds	Loss value	Convergence
1	1.823	High
5	1.200	Steadily declining
10	0.720	began to stabilize
15	0.385	Convergence to lower levels
20	0.015	Completely convergent

As can be seen from the data in Table 1, the loss value decreased significantly in the previous rounds of training, demonstrating that the model can quickly adapt and learn from randomly generated data. In the later stage of training, the loss value tends to be stable, indicating that the

DOI: 10.54254/2755-2721/2025.AST26531

network parameters have been adjusted to the appropriate value, and the model converges and can make stable prediction of key points. Even under the condition of simplifying the data set, the model can still effectively extract useful features from the data and complete the task of pose estimation.

4.2. Test results

After the model was fully trained, we tested it using the same randomly generated dataset, and the results showed that the model was able to accurately predict the location of key points despite the simplification of the dataset. The key points in the test dataset are randomly generated and therefore not the same as the actual image data. Nevertheless, the model is able to learn certain patterns from these simplified data, and the prediction results on the test set are more accurate. Table 2 shows the error comparison between the model's predictions and the real data during the testing phase.

Table 2. Model prediction vs. real key point coordinates (sample 1 to sample 4)

Sample number	True coordinates (x, y)	Predicted coordinates (x, y)	Error (Euclidean distance)
1	(0.1, 0.2)	(0.095, 0.215)	0.0312
2	(0.4, 0.5)	(0.405, 0.48)	0.0294
3	(0.6, 0.3)	(0.61, 0.28)	0.0410
4	(0.7, 0.7)	(0.69, 0.69)	0.0141

As can be seen from Table 2, the model has a small prediction error on the test set, especially in the case of sample 1 and sample 2, where the predicted joint coordinates are almost exactly the same as the real coordinates. Although the test dataset is generated entirely randomly and does not contain complex background information, the model is still able to extract patterns from this data and predict the location of key points more accurately. The samples with small errors in the table show that the network architecture has strong generalization ability, can adapt to changes in simplified datasets, and has certain learning ability.

4.3. Visual analysis

In order to further evaluate the prediction effect of the model, we connect the prediction key points of each sample through visualization to form a skeleton map of the human body. Through this visualization, we can more intuitively see how the model performs in different samples. Table 3 summarizes the skeleton diagram error and completeness scores of the first few samples, helping us to comprehensively evaluate the prediction accuracy of the model and the completeness of the skeleton diagram.

Table 3. Skeletal plot error and completeness score for sample prediction results

Sample number	Predict the number of joints	Total error (pixels)	Skeletal diagram completeness
1	17	2.34	high
2	17	3.12	middle
3	17	4.15	middle
4	17	1.98	high

Table 3 shows the prediction effect of the model in the testing stage, with mostly small prediction errors and high skeletal diagram completeness scores. Samples 1 and 4 have the smallest prediction

DOI: 10.54254/2755-2721/2025.AST26531

error and high completeness of the skeleton diagram, indicating that the model can accurately predict the position of each joint and form a complete skeleton structure. In Samples 2 and 3, despite the slightly larger error, the connection relationship of the overall pose is reasonable, showing the high accuracy of the model in generating the skeleton diagram. Through this visual analysis, we can not only evaluate the prediction results of the model, but also clearly see how the model performs in the simplified dataset.

5. Conclusion and future work

In this paper, a convolutional neural network (CNN)-based human pose estimation method is proposed, and the effectiveness of the method is verified by a simplified stochastic dataset. The experimental results show that the model can accurately predict the coordinates of key points in the human body under simple conditions, and clearly demonstrate the learning effect of the model through visualization. By constructing a simplified dataset, we can effectively evaluate the model's performance under specific conditions, providing a theoretical basis for subsequent deep learning model optimization. Despite the initial success, the method of this paper still has certain limitations. Since a simplified dataset is used, the robustness and accuracy of the model still need to be further improved. With the gradual application of real-world image datasets, complex backgrounds, diverse pose changes, and multi-person interaction will become challenges for pose estimation models. Therefore, future research will focus on the following directions: Training with real datasets: To address the limitations posed by simplifying data sets, future work will focus on considering the use of more diverse and complex real-world image datasets, such as COCO, MPII, etc. These datasets cover a variety of scenarios and human poses, which can help improve the generalization ability of the model, especially in dynamic scenes. Optimize network structure: The current model structure is relatively simplified and has not fully considered some advanced features, such as multi-scale feature extraction and global context information. Future work will explore how to design more complex and refined network structures, such as using advanced architectures such as Hourglass network and ResNet, to improve the prediction accuracy and robustness of models. Integration of other deep learning techniques: We plan to further integrate other deep learning techniques, such as image augmentation and multi-task learning.

References

[1] Voulodimos A, Doulamis N, Doulamis A, et al. Deep learning for computer vision: A brief review [J]. Computational intelligence and neuroscience, 2018, 2018(1): 706834.
[2] Andriluka M, Pishchulin L, Gehler P, et al. 2d human pose estimation: New benchmark and state of the art analysis [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 3686-3693.
[3] Preece J, Rogers Y, Sharp H, et al. Human-computer interaction [M]. Addison-Wesley Longman Ltd., 1994.
[4] Valera M, Velastin S A. Intelligent distributed surveillance systems: a review [J]. IEE Proceedings-Vision, Image and Signal Processing, 2005, 152(2): 192-204.
[5] Aggarwal J K, Cai Q. Human motion analysis: A review [J]. Computer vision and image understanding, 1999, 73(3): 428-440.
[6] Burdea G C, Coiffet P. Virtual reality technology [M]. John Wiley & Sons, 2003.
[7] Mitra S, Acharya T. Gesture recognition: A survey [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2007, 37(3): 311-324.
[8] Paetkau D, Slade R, Burden M, et al. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power [J]. Molecular ecology, 2004, 13(1): 55-65.
[9] Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects [J]. IEEE transactions on neural networks and learning systems, 2021, 33(12): 6999-7019.
[10] Few S. Now you see it: simple visualization techniques for quantitative analysis [M]. Analytics Press, 2009.

DOI: 10.54254/2755-2721/2025.AST26531

[11] Kitano H. Biological robustness [J]. Nature Reviews Genetics, 2004, 5(11): 826-837.
[12] Cao Z, Hidalgo G, Simon T, et al. Openpose: Realtime multi-person 2d pose estimation using part affinity fields [J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(1): 172-186.
[13] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7291-7299.
[14] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation [C]//European conference on computer vision. Cham: Springer International Publishing, 2016: 483-499.
[15] Ferryman J, Shahrokni A. Pets2009: Dataset and challenge [C]//2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance. IEEE, 2009: 1-6.