Enhancing Human Pose Estimation: The Role of DeepPose as a Holistic Method and the Potential of End-to-End Learning Approaches

***Abstract*— This project explores two prominent human pose estimation models: DeepPose and Learning Human Pose Estimation Features with Convolutional Networks. The goal is to evaluate the effectiveness of these models in estimating human body keypoints from images and to investigate the feasibility of applying an end-to-end learning approach to the DeepPose model. We examine how the end-to-end approach can streamline pose estimation by eliminating the need for intermediate steps, such as manual feature extraction or multi-stage processing. Through comparative analysis, we demonstrate that end-to-end learning can improve model performance and efficiency, offering a promising direction for future research in human pose estimation.**

I. Introduction

Human Pose Estimation (HPE) is a key task in the field of computer vision aimed at detecting and localizing human joints from images. The problem can be approached using deep learning methods that leverage convolutional neural networks (CNNs) and Deep Neural Networks (DNNs).

we can group deep learning methods in human pose estimation into holistic and part-based methods [1]. DeepPose is a holistic model, and the other one is a part-based method [1].

II. DeepPose

A. Introduction

The DeepPose model represents a holistic approach to human pose estimation, addressing the challenge of estimating human body joint locations, even when some joints are occluded or not visible in the image [2]. In scenarios where certain body parts are hidden from view, as one can see in Fig.1, holistic reasoning becomes crucial for accurately predicting the positions of these occluded joints [2]. The pose estimation task is framed as a DNN-based regression problem, where the network learns to directly predict the locations of body joints [2]. This formulation offers two key advantages:

1) The DNN model has shown outstanding performance on object localization [3]. It is able to capture the global context surrounding each joint, improving the accuracy of joint localization [2].
2) The DNN-based approach provides a more straightforward and less computationally complex alternative compared to methods that rely on graphical models or other structured prediction techniques [2].

These benefits make the DeepPose model an appealing solution for human pose estimation tasks, especially in environments where occlusion is common.

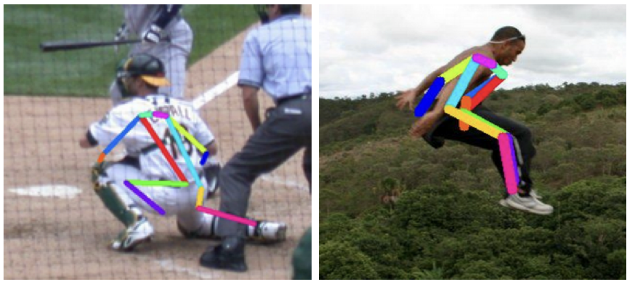

Fig. 1. Many joints are barely visible in certain images, requiring holistic reasoning to predict their locations based on the visible parts of the body and the person's motion or activity [2].

B. Metrix

To facilitate comparison with results, we will use two widely recognized evaluation metrics. The Percentage of Correct Parts (PCP) measures the detection rate of limbs, where a limb is considered detected if the distance between the predicted joint locations and the true joint locations is no more than half the length of the limb [4]. PCP was initially the preferred metric, but it has the drawback of penalizing shorter limbs, such as the lower arms, which are typically more challenging to detect [2].

To address this limitation, a newer metric has emerged, which evaluates joint detection rates using an alternative criterion [2]. In this approach, a joint is considered detected if the distance between the predicted and true joint locations is within a specific fraction of the torso diameter [2]. By adjusting this fraction, detection rates can be reported for different levels of localization precision [2]. This metric, known as Percent of Detected Joints (PDJ), mitigates the issues associated with PCP by applying a consistent detection threshold across all joints [2].

C. Functionality

Basically DeepPose is an approach for human pose estimation using deep neural networks (DNNs), specifically focusing on how pose estimation is framed as a regression problem and how the model is trained and refined using a cascade of pose regressors, and each cascade stage progressively improves the joint locations [2].

Stage 1 starts by using the full image or a person detector to crop out a bounding box around the human subject [2]. The first DNN regressor predicts the joint locations [2].

Subsequent Stages (Stage 2 and beyond) focus on refining the joints by cropping out regions around the predicted joint locations from the previous stage [2]. This allows the

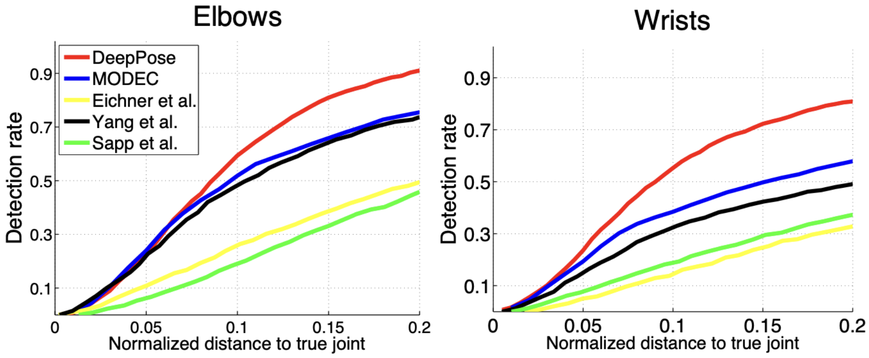

Fig. 2. Percentage of Detected Joints (PDJ) on the FLIC dataset for the elbow and wrist joints. We compare the performance of DeepPose, after two cascade stages, with four other methods [2].

network to focus on high-resolution details of the body parts, improving localization precision [2]. At each stage, a new network is trained to predict the displacement of the joint from its previous stage estimate. These refined locations are then used to generate sub-images for the next stage, improving localization at finer scales [2].

Please refer to Table I for an example. It presents the results from the previous three stages, where a clear trend is observed: an increasing number of stages leads to a higher PCP percentage.

D. Evaluation

The results we obtained on the FLIC dataset are shown in Fig.2, where we compare our method with four other approaches. Our improvements are particularly noticeable in the low-precision domain, where the focus is on detecting approximate poses rather than precisely localizing the joints [2]. At a normalized distance of 0.2 on the FLIC dataset, we observe a 0.15 increase in detection rate for the elbow and a 0.2 increase for the wrists compared to the next best-performing method [2]. This demonstrates that, as a holistic approach, DeepPose outperforms others in general human pose estimation.

Method	Arm		Leg		Ave.
	Upper	Lower	Upper	Lower	
DeepPose-st1	0.5	0.27	0.74	0.65	0.54
DeepPose-st2	0.56	0.36	0.78	0.70	0.60
DeepPose-st2	0.56	0.38	0.77	0.71	0.61

TABLE I
PERCENTAGE OF CORRECT PARTS (PCP) AT 0.5 ON LSP(ANOTHER DATASET) FOR DEEPPOSE [2].

III. END TO END INSPIRATION

Now, let's explore the part-based method. This approach involves training several smaller CNNs for independent binary classification of body parts, followed by a higher-level weak spatial model that helps eliminate significant outliers and ensures global pose consistency [1].

One of the key limitations of earlier methods for tackling full pose estimation using end-to-end learning approaches, particularly deep networks, was the scarcity of labeled data [5]. It is worth noting that the authors, who proposed this part-based method, address the problem of limited labeled data for pose estimation by using larger, more detailed dataset and by making better use of the rich structure of pose annotations to train better models [5]. The dataset used here is FLIC.

Given that DeepPose is also capable of working with the FLIC dataset, this presents an interesting opportunity to explore how end-to-end learning could be implemented within DeepPose as well. The rich and structured annotations in FLIC could potentially enable a more holistic approach to pose estimation, bypassing the need for manual feature engineering. This leads us to believe that, with the right adjustments, DeepPose could benefit from a similar strategy of leveraging larger datasets and sophisticated annotations to better train a model using end-to-end learning. By integrating these techniques into DeepPose, we might not only improve pose detection accuracy but also enhance the model's ability to generalize across different poses, potentially closing the gap between part-based methods and fully end-to-end systems.

IV. CONCLUSION

DeepPose, as a holistic method, holds a unique and irreplaceable role in human pose estimation, particularly in its ability to effectively estimate joints that are not visible in the image. The results demonstrate a distinct advantage in detecting approximate poses, which, in certain practical applications, may prove to be more important than precisely localizing individual joints. This advantage lies in DeepPose's ability to infer the global human pose from partial or occluded body parts, making it robust to variations in pose visibility.

However, one limitation of DeepPose is its lack of research into the application of end-to-end learning methods. In contrast, another study on a part-based method that incorporates end-to-end learning shows promising results, suggesting that this approach could enhance pose estimation performance by directly learning the mappings from input images to joint locations without relying on manual steps.

REFERENCES

[1] A. D. E. P. Athanasios Voulodimos, Nikolaos Doulamis. (2018) Deep learning for computer vision: A brief review. Accessed: 2024-11-16. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1155/2018/7068349

[2] A. Toshev and C. Szegedy. (2014) Deeppose: Human pose estimation via deep neural networks. USA. 2-s2.0-84911381180. [Online]. Available: https://doi.org/10.1109/CVPR.2014.214

[3] C. Szegedy, A. Toshev, and D. Erhan. (2013) Object detection via deep neural networks. [Online]. Available: https://papers.nips.cc/paper/2013

[4] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "Articulated human pose estimation and search in (almost) unconstrained still images," ETH Zurich, D-ITET, BIWI, Tech. Rep. Technical Report No. 272, 2010. [Online]. Available: https://www.biwi.ethz.ch/

[5] A. Jain, J. Tompson, and M. Andriluka. (2014) Learning human pose estimation features with convolutional networks. [Online]. Available: https://openreview.net/forum?id=ryxriwHeZ