ConvPose: Unveiling Human Pose Estimation with CNN for Joint Location Regression

Kumar Saikat Halder
khalder@ualberta.ca
Department of Computing Science
University of Alberta, AB, CA

Puja Saha
psaha03@uoguelph.ca
School of Engineering
University of Guelph, ON, CA

***Abstract*— Human pose estimation, which is critical for understanding body movements from images and videos, has received a lot of attention in a variety of fields. This project delves into cutting-edge deep learning techniques, using a CNN for precise regression to predict 14 critical joint locations in human body. The model achieves an impressive Mean Squared Error (MSE) of 0.0878, demonstrating robust accuracy even with unknown data. The study examines these approaches in depth, focusing on data handling, model development and inference strategies. This study adds to the discussion of human pose estimation by meticulously dissecting methodologies and their performance under various conditions. It provides valuable insights into current deep learning solutions as well as direction for future advancements in this critical domain.**

Keywords—Pose Estimation, CNN, Régression, Joints

I. INTRODUCTION

Human pose estimation, which is tasked with pinpointing human joint positions, is at the forefront of computer vision research. Figure 1 depicts the difficulties inherent in this task: complex articulations, the complexities of minute and barely visible joints, occlusions, and the critical need for contextual comprehension. Historically, the field has been focused on navigating a wide range of articulated poses. Part-based models [15] [8] have been instrumental in naturally addressing articulation complexities. Recent advances have ushered in models with simplified inference techniques, indicating a significant trend in the field's evolution.

Fig. 01. Illustrating Human Poses

However, the efficiency attained by these methods is limited by their expressiveness. These approaches, which are frequently reliant on local detectors that primarily focus on individual body parts, tend to model only a fraction of the intricate interactions between these parts. Despite efforts to address these constraints, as shown in Figure 1, efforts to propose holistic pose reasoning methods have run into difficulties, yielding limited success in practical real-world applications.

II. RELATED WORK

The idea of representing articulated objects, particularly human poses, as a graph of parts is not new in computer vision. Initially introduced by Fishler and Elschlager as Pictorial Structures (PSs), Felzenszwalb and Huttenlocher later made these structures practical and manageable using the distance transform technique. This breakthrough paved the way for the development of a wide range of PS-based models that proved useful in practical applications.

This tractability, however, comes with inherent limitations, most notably tree-based pose models with simple binary potential that do not account for image data. As a result, research efforts have focused on improving the representational capacity of these models while maintaining tractability. Earlier attempts at this involved the use of more sophisticated part detectors.

Several models have recently emerged that express intricate joint relationships in the context of human pose estimation. Yang and Ramanan proposed a mixture model of parts, whereas Johnson and Everingham investigated mixture models on a larger scale by incorporating Pictorial Structures (PSs) mixtures. Tian et al. advanced the field by using a hierarchical model to capture richer higher-order spatial relationships. Another method for capturing these higherorder relationships is to use image-dependent PS models estimated by a global classifier. These accomplishments reflect the ongoing evolution and diversification of approaches for modeling complex joint relationships for human pose estimation.

In a similar nearest neighbor setup, Shakhnarovich et al. used locality-sensitive hashing. Gkioxari et al.'s semiglobal classifier demonstrated promise, but with less expressive representation than our method, which was tested primarily on arms. Ionescu et al. concentrated on 3D pose regression, whereas similar works employed CNNs with Neighborhood Component Analysis. None, however, used cascades, which distinguished our approach.

III. Experimental setup

This study adopts a holistic perspective on human pose estimation, capitalizing on recent advancements in deep learning. We propose a method based on a Convolutional Neural Network (CNN). CNNs have demonstrated exceptional performance in visual classification tasks [14] and, more recently, in object localization [22, 9]. However, the application of Deep Neural Networks (DNNs) for precise localization of articulated objects has remained largely unexplored. In this project, we aim to address this question and present a straightforward yet powerful formulation of holistic human pose estimation using a CNN.

We define the pose estimation problem as a joint regression task and demonstrate how to effectively frame it in the context of CNNs. The location of each body joint is regressed using the full image as input and a 48-layered network. This formulation offers two key advantages. Firstly, the CNN can capture the complete context of each body joint, as each joint regressor utilizes the entire image as a signal. Secondly, the approach is considerably simpler to formulate compared to methods based on graphical models. There is no need for explicit design of feature representations and detectors for parts, nor is there a need to explicitly design a model topology and interactions between joints. Instead, we show that a generic convolutional CNN can be learned for this problem.

Initiating with an initial pose estimation based on the full image, we employ CNN-based regressors to refine joint location predictions using higher resolution sub-images. Our approach demonstrates state-of-the-art or superior performance on four widely used benchmarks, surpassing all reported results. We showcase the effectiveness of our approach on images of people with significant variation in appearance and articulations. Additionally, we demonstrate generalization performance through cross-dataset evaluation.

IV. Dataset

The Leeds Sports Pose (LSP) dataset is an important benchmark in human pose estimation containing 10,000 labeled images for comprehensive analysis and algorithm development. Each image is meticulously annotated, with 14 joint locations detailed. Notably, these annotations keep left and right joints labeled consistently, ensuring a personcentric perspective. An example of images of this dataset has shown in Fig.02.

Fig.02. An Example of LSP Dataset

V. Methodology

The methodology outlines a comprehensive approach for human pose estimation using a Convolutional Neural Network (CNN). Beginning with data preparation, a specialized dataset ('PoseLandmarksDataset') is structured, presumably consisting of images and associated landmarks of the joint locations defining human poses. Since the images were of different sizes, transformations were used for preprocessing, which includes tasks like resizing, cropping, and converting to tensors, which are required for subsequent neural network training. Examples of image after preprocessing is given in Fig.03.

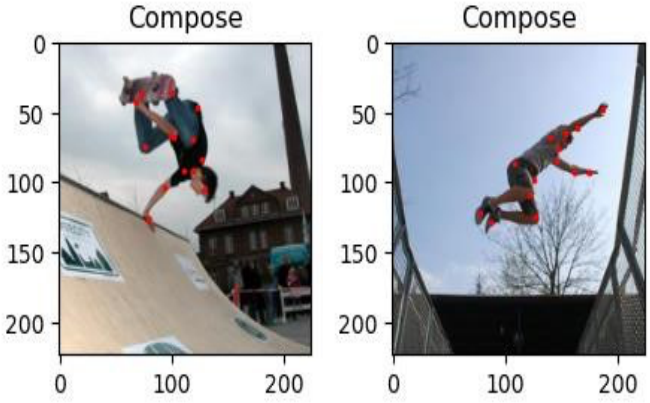

Fig.03. Processed Images Along with Their Key points Indicating the Joints.

The CNN architecture, which employs a layered structure (Table.01) that integrates multiple convolutional layers, serves as the foundation of this methodology. These layers are augmented with batch normalization, which ensures stable learning, and ReLU activations, which introduce critical non-linearity to the network. They are sequentially integrated and allow for feature downsampling, eventually converging into fully connected layers capable of deciphering intricate data patterns. The network's input shape is '(3, 224, 224)' when it operates on 224x224 pixel color images.

The dataset was methodically segmented into distinct portions to ensure a robust evaluation: 70% for training, 15% for validation, and an equal 15% for testing.

During the training phase, the Stochastic Gradient Descent ('SGD') optimizer was used in conjunction with the Mean Squared Error ('MSE') loss. This combination guided the network's learning process, allowing it to reach optimal solutions more quickly. The training-validation cycle was orchestrated by a meticulously orchestrated training loop that spanned 30 epochs. This iterative approach allowed for continuous refinement of the model, which improved its ability to predict the critical 14 joint locations.

In human pose estimation, regression is an oftenused approach to directly predict the coordinates of key points on the human body. Instead of segmenting the image into different body parts, the model directly outputs a set of continuous values corresponding to the coordinates of key points such as joints. Here's a breakdown of how regression was typically applied in our human pose estimation project:

Table.01. Architecture of the Network

Layers	Kernel, Stride, Padding	Output Size	Number of Param
C2D	K= 3, S=1, P=1	[-1, 64, 224, 224]	1,792
BN2D	-	[-1, 64, 224, 224]	128
ReLU	-	[-1, 64, 224, 224]	0
C2D	K= 3, S=1, P=1	[-1, 64, 224, 224]	36,928
BN2D	-	[-1, 64, 224, 224]	128
ReLU	-	[-1, 64, 224, 224]	0
MP2D	K= 2, S=2	[-1, 64, 112, 112]	0
C2D	K= 3, S=1, P=1	[-1, 128, 112, 112]	73,856
BN2D	-	[-1, 128, 112, 112]	256
ReLU	-	[-1, 128, 112, 112]	0
C2D	K= 3, S=1, P=1	[-1, 128, 112, 112]	147,584
BN2D	-	[-1, 128, 112, 112]	256
ReLU	-	[-1, 128, 112, 112]	0
MP2D	K= 2, S=2	[-1, 128, 56, 56]	0
C2D	K= 3, S=1, P=1	[-1, 256, 56, 56]	295,168
BN2D	-	[-1, 256, 56, 56]	512
ReLU	-	[-1, 256, 56, 56]	0
C2D	K= 3, S=1, P=1	[-1, 256, 56, 56]	590,080
BN2D	-	[-1, 256, 56, 56]	512
ReLU	-	[-1, 256, 56, 56]	0
C2D	K= 3, S=1, P=1	[-1, 256, 56, 56]	590,080
BN2D	-	[-1, 256, 56, 56]	512
ReLU	-	[-1, 256, 56, 56]	0
MP2D	K= 2, S=2	[-1, 256, 28, 28]	0
C2D	K= 3, S=1, P=1	[-1, 512, 28, 28]	1,180,160
BN2D	-	[-1, 512, 28, 28]	1,024
ReLU	-	[-1, 512, 28, 28]	0
C2D	K= 3, S=1, P=1	[-1, 512, 28, 28]	2,359,808
BN2D	-	[-1, 512, 28, 28]	1,024
ReLU	-	[-1, 512, 28, 28]	0
C2D	K= 3, S=1, P=1	[-1, 512, 28, 28]	2,359,808
BN2D	-	[-1, 512, 28, 28]	1,024
ReLU	-	[-1, 512, 28, 28]	0
MP2D	K= 2, S=2	[-1, 512, 14, 14]	0
C2D	K= 3, S=1, P=1	[-1, 512, 14, 14]	2,359,808
BN2D	-	[-1, 512, 14, 14]	1,024
ReLU	-	[-1, 512, 14, 14]	0
C2D	K= 3, S=1, P=1	[-1, 512, 14, 14]	2,359,808
BN2D	-	[-1, 512, 14, 14]	1,024
ReLU	-	[-1, 512, 14, 14]	0
C2D	K= 3, S=1, P=1	[-1, 512, 14, 14]	2,359,808
BN2D	-	[-1, 512, 14, 14]	1,024
ReLU	-	[-1, 512, 14, 14]	0
MP2D	K= 2, S=2	[-1, 512, 7, 7]	0
Dropout	-	[-1, 25088]	0
Linear	-	[-1, 512]	12,845,568
ReLU	-	[-1, 512]	0
Linear	-	[-1, 28]	14,364

Total params: 27,583,068
Trainable params: 27,583,068
Non-trainable params: 0
C2D- 2D Convolution, BN2D- 2D Batch Normalization, MP2D- 2D Max Pooling

Visualized the model's predictions alongside the ground truth key points to understand its strengths and weaknesses. Regression-based approaches are computationally efficient and may perform well, especially when you have a large dataset with accurately annotated key points. However, they may be more sensitive to outliers and might require careful handling of noisy or ambiguous annotations. The choice between segmentation and regression depends on the characteristics of your data and the specific goals of your pose estimation project.

VI. Performance Analysis

We encountered performance fluctuations in our initial model iterations, prompting us to integrate TensorBoard for detailed visualization. The use of TensorBoard greatly simplified monitoring of critical performance metrics, particularly training and validation losses. This visualization provided invaluable insights into the model's evolution over time. We initially chose a training duration of 200 epochs. However, TensorBoard analysis revealed an interesting trend: the gap between training and validation losses began to widen significantly after about 25 epochs. Fig.04. depicts the dynamic trends of training and validation losses during the model's execution of 200 epochs.

Fig.04. Visualization of Loss During Training

This visualization proved to be a valuable tool in finetuning our training strategy and ensuring model robustness. This observation prompted us to rethink our strategy. We conducted focused training sessions in subsequent iterations, limiting the model's training to 25 epochs. This strategy enabled us to stop training early, avoiding overfitting and increasing efficiency.

Our performance was evaluated using mean squared error (MSE), which was calculated using the Euclidean distance between annotated and predicted points, effectively denoting the precise joint locations in the human body. While testing with unseen dataset we have found a Mean Square Error (MSE) of 0.0878 which was quite accurate

prediction proving the robustness of our model. Fig.05. shows the predicted points in a unseen image from test dataset.

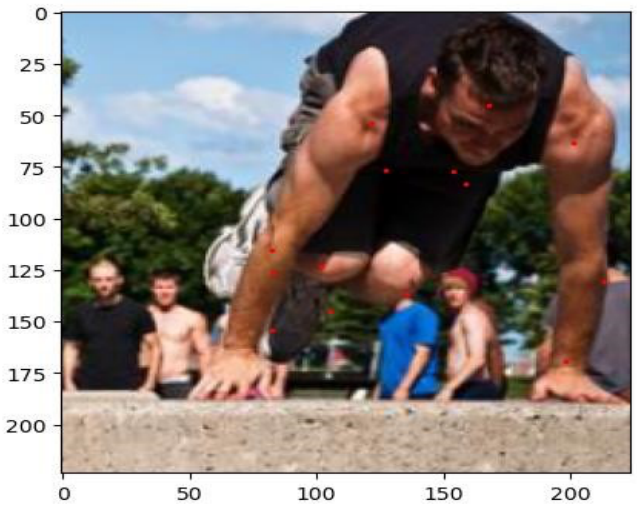

Fig.05. An Unseen Images with Predicted Points Of Joints

VII. CHALLENGES

However, the methodology to find the joints location was hampered by significant challenges and potential limitations. Processing the data and its annotations proved was a significant challenge. Annotating such a large number of images necessitates close attention to detail and precision. The difficulty of ensuring accurate joint labeling and maintaining consistency across a large number of images emphasizes the complexities of this task. Moreover, the model's inherent complexity, as exemplified by its multilayered architecture, raised the possibility of overfitting, necessitating research into regularization techniques such as dropout or weight decay. While the code alluded to data augmentation techniques such as resizing and cropping, a more in-depth understanding of their impact on model generalization was required. Processing of data hampered the data annotations consistency in few cases. Fine-tuning hyperparameters such as learning rate and batch size added a significant impact on model performance.

VIII. CONCLUSION

The Human Pose Estimation project represents a significant advancement in computer vision by elucidating the complexities of determining human body positions in visual media. It sheds light on the potential and challenges of understanding human body language through meticulous data collection, robust model architecture, and deliberate training. Beyond this stage, its impact ranges from improving humancomputer interaction and immersive experiences to assisting healthcare and sports analytics. This advancement not only improves computational efficiency but also strengthens the human-machine bond. This project foreshadows a future in which technology comprehends human body language in great depth, hinting at a time when machines will be empathetic companions. The progress in human posture estimation, with ongoing algorithmic advancements, reveals new perspectives and possibilities beyond this initial exploration.

IX. FUTURE WORK

The future work is planned with optimism and purpose. The process of fine-tuning, which is constantly growing, offers the potential to achieve small but significant improvements. At the same time, the desire to analyze information in real-time presents an opportunity to apply this technology in dynamic and time-sensitive settings. The expansion of multi-person posture estimation broadens the scope of this technology, picturing situations where the interactions between individuals are accurately interpreted without any interruption. The model's adaptability and resilience in the face of various environmental circumstances are driven by its robustness, which is the foundation for its practicality in real-world scenarios.

Fine-tuning: To enhance the model's performance, additional refinement might be pursued. This encompasses the process of fine-tuning hyperparameters, modifying the structure of the model, or employing transfer learning by utilizing a pretrained model on a more extensive dataset.

Real-time Processing: The implementation of real-time human posture estimation is a critical area that needs improvement. This may entail enhancing the model for accelerated inference or investigating algorithms with higher efficiency.

Multi-person Pose Estimation: Expanding the model's capability to process numerous individuals in an image or video is a compelling direction for future research. This would need altering the structure to accommodate overlapping positions and intricate interactions.

Enhancing Robustness to Environmental elements: Improving the model's ability to withstand and perform well in the presence of different environmental elements, such as obstructions, varied attire, and intricate backgrounds, is crucial for practical implementations in real-world scenarios.

REFERENCES

[1] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In BMVC,2010.

[2] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In CVPR, 2011.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS,2012.

[4] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In ECCV, 2002.

[5] R. Nevatia and T. O. Binford. Description and recognition of curved objects. Artificial Intelligence, 8(1):77–98, 1977. [6] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In CVPR, 2013. [7] D. Ramanan. Learning to parse images of articulated bodies In NIPS, 2006.

[8] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In CVPR, 2013.

[9] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In CVPR, 2003.

[10] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 3476–3483. IEEE, 2013.

[11] C. Szegedy, A. Toshev, and D. Erhan. Object detection via deep neural networks. In NIPS 26, 2013. [12] G. W. Taylor, R. Fergus, G. Williams, I. Spiro, and C. Bregler. Pose-sensitive embedding by nonlinear nca regression In NIPS, 2010.

[13] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In ECCV, 2012.

[14] F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. In CVPR, 2013. [15] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In CVPR, 2011.

[16] "LSP Dataset," The dataset is available at https://paperswithcode.com/dataset/lsp.