All Sciences Proceedings

1st International Conference on Pioneer and Innovative Studies

June 5-7, 2023 : Konya, Turkey

ICPIS

All Sciences Proceedings
http://as-proceeding.com/

https://as-proceeding.com/index.php/icpis

Top-Down Approaches in Human Pose Estimation: A State-of-the Art Review

Yusuf Enes Bölükbaşı*, Rayan Abri*

[1]*Computer Engineering, Ostim Technical University, Turkey*

yusufenesboukbasi@gmail.com

Abstract **–** This paper offers a comprehensive exploration of top-down approaches in human pose estimation, a key facet of computer vision. These approaches primarily focus on identifying the human subject in an image or video, followed by determining the spatial configuration of their body joints. Such techniques are instrumental in an array of sectors, from healthcare and sports analytics to entertainment and security systems.

The document delves into the foundations of top-down pose estimation, presenting a review of established and emerging models. It explicates the role of key performance metrics, including Average Precision (AP), AP at specific Intersection over Union (IoU) thresholds (AP50, AP75), Average Recall (AR), and AR at an IoU of 0.50, in appraising the efficiency and reliability of these models.

The paper underscores the substantial strides made in top-down pose estimation and discusses their efficacy in managing diverse real-world scenarios. It draws attention to the various challenges associated with these techniques, such as handling occlusions, processing images or videos with multiple individuals, and addressing computational constraints.

In conclusion, while top-down approaches in pose estimation have shown notable progress and promise, there exist avenues for further research and development. This paper intends to provide a foundational understanding of these techniques and a platform for future advancements in the field.

Keywords – Pose Estimation, Top-Down Approach, Computer Vision, Deep Learning

I. INTRODUCTION

In the exciting world of computer vision, figuring out a person's position and orientation in images or videos is a hot topic. This is known as human pose estimation, and it's super important for a range of cool applications like video games, animation, health-related tech, and analyzing sports.

There are two main ways we go about estimating a human's pose: the 'top-down' approach and the 'bottom-up' approach. With the top-down approach, we first spot the person in the photo and then try to estimate their pose. The bottom-up methodology, however, presents a distinct perspective. Here, we first detect parts of the body and then link them together to guess the pose.

This paper will zoom in on the top-down approach. We've made some big strides in this area, and our systems have gotten way better at estimating poses accurately and efficiently. But, like most things, it's not perfect. We still have to figure out how to handle situations where parts of the body can't be seen, or when we're dealing with bodies of different shapes and sizes.

We'll take a deep dive into how the top-down approach has changed over time, exploring the big breakthroughs and how they're being used today. We'll also take a good hard look at the problems that

we're still trying to solve and think about what the future might hold for this field.

The aim of this review is to give a full picture of where we're at with top-down human pose estimation and to highlight the areas where we can hopefully make some improvements.

II. Materials And Methods

A. Overview Of Top-Down Methods

Top-down pose estimation methodologies, which are trained on the comprehensive COCO dataset, have significantly progressed due to the development of numerous cutting-edge models [1]. These methods first utilize a human detection system to recognize individuals in an image, followed by the estimation of the pose for each detected person. This two-step mechanism effectively differentiates between individuals in crowded environments.

HRNet is among the high-performing models because it maintains high-resolution representations throughout the entire process [3]. This unique approach deviates from the traditional models, making it particularly beneficial for tasks like pose estimation where intricate accuracy across all scales is indispensable.

Models such as ResNeXt and SE-ResNet have evolved innovatively from well-established networks [2]. ResNeXt introduces a novel dimension, "cardinality" – the set of transformations – and has been fundamental in enhancing the performance of pose estimation. Conversely, SE-ResNet incorporates a mechanism that recalibrates feature maps channel-wise, which boosts their representational quality [5]. By concentrating on the most crucial features and suppressing the less useful ones, these models deliver improved results on pose estimation tasks.

HrFormer, which merges novel technologies, integrates the transformer architecture into a high-resolution structure [3]. This blend of local feature extraction via convolutional neural networks (CNNs) and transformers' ability to capture long-range dependencies contributes to substantial improvement in pose estimation results. Similarly, VitPose, another exceptional model, replaces the traditional convolutional layers with transformer layers. These transformer layers manage long-term dependencies more efficiently, enhancing the precision of pose estimation [4].

The Hourglass network, renowned for its symmetric design around the network's mid-depth, grasps complex relationships within the image data [1]. By processing the image through multiple resolution down sampling and subsequently reconstructing the image back to its original resolution, this model adeptly handles the intricacies of pose estimation.

In contrast, models like RSN and SCNet follow a different tactic. RSN employs a stepwise training process where residual steps are gradually added during the training, thereby enhancing the model's learning capability [6]. SCNet amalgamates spatial and channel-wise attention in a two-branch network, emphasizing pertinent spatial locations and channel-wise features, resulting in precise pose estimations, particularly in complex scenes involving multiple people [7].

Lastly, ViPnas showcases the potential of Neural Architecture Search (NAS) by exploring the automatic search for efficient pose estimation models [8]. By integrating pose-specific operations into the search space, it often generates models that perform better on pose estimation tasks.

The continuous development of these top-down approaches demonstrates the immense potential in the field of pose estimation. With the ongoing advancements, we can anticipate increasingly precise and efficient solutions capable of managing more complex scenarios in the future.

B. Applications and Use Cases of Top-Down Approaches

The versatility, resilience, and precision of top-down pose estimation techniques have led to their wide application across numerous domains.

Augmented Reality (AR) and Virtual Reality (VR): Top-down pose estimation's capability to accurately determine human poses is indispensable in AR and VR use cases. It facilitates real-time interactions between users and their virtual surroundings, thereby elevating the immersive

experience. Instances include virtual clothes trial, object manipulation, or engaging in interactive gaming.

Healthcare and Rehabilitation: The utility of top-down pose estimation is significant within healthcare. Applications include patient supervision, gait analysis, and monitoring rehabilitation progress. For instance, a model could observe a patient's posture evolution over time, providing feedback for movement or posture improvement. In surgical procedures, these methods can assist in preoperative planning and intraoperative guidance.

Sports Coaching and Performance Review: Accurate human pose estimation offers substantial advantages for athletes and trainers. Through movement analysis of athletes, trainers can offer insightful feedback to enhance performance and avert potential injuries. The technique finds applicability across various sports, like football, basketball, and gymnastics.

Security and Safety: Within security systems, top-down techniques can identify individuals and estimate their poses to infer their activities and behaviors, crucial for maintaining safety. For example, these methods can detect abnormal or suspect activities, falls in elderly individuals in homes, or pinpoint potentially dangerous situations in workplaces.

Human-Computer Interaction (HCI): The ability to comprehend and interpret human gestures and poses can render HCI more natural and engaging. This can be particularly advantageous in creating more effective and accessible interfaces for individuals with physical impairments.

Animation and Gaming: Pose estimation is a potent instrument for character animation in video games and films. Top-down methods can capture actor’s movements and convert them into realistic animations. This approach has transformed character animation, leading to more realistic and dynamic movements.

Autonomous Vehicles and Robotics: Autonomous vehicles and robots can leverage pose estimation to better comprehend their surroundings. For instance, an autonomous vehicle can predict pedestrian behavior using these techniques, and robots can employ it for safe human interaction.

Through these myriad applications, top-down pose estimation techniques hold the potential to transform various sectors, enhancing productivity and user experiences while also unlocking new opportunities. The ongoing evolution of these models foretells even more exciting applications in the future.

III. Results

A. Performance Evaluation of Top-Down Approaches

Evaluating the performance of a pose estimation system is a crucial step in its development and refinement. It's not just about whether the system can detect poses but also how accurately and reliably it can do so under varying conditions. There are several metrics commonly used for evaluating pose estimation systems, including Average Precision (AP), AP50, AP75, Average Recall (AR), and AR50.

Average Precision (AP): AP is calculated as the area under the precision-recall curve. The precision-recall curve is drawn by plotting precision (P) against recall (R) at various threshold levels.

Mathematically, Precision and Recall are calculated as:

Precision (P) = True Positives / (True Positives + False Positives)

Recall (R) = True Positives / (True Positives + False Negatives)

AP50 and AP75: AP at a specific Intersection over Union (IoU) threshold, say 0.5 or 0.75, is the Average Precision calculated for detections with IoU above that threshold.

Average Recall (AR): AR is calculated similarly to AP, but instead of plotting precision against recall, you plot recall against the number of detections per image (DPI).

AR50: This is the Average Recall calculated for detections with an IoU above 0.50.

Model	Input Size	AP	AP50	AP75	AR	AR50
HrNet width 32	256x192	0.749	0.906	0.821	0.804	0.945
HrNet width 32	384x288	0.761	0.908	0.826	0.811	0.944
HrNet width 48	256x192	0.756	0.908	0.826	0.809	0.945
HrNet width 48	384x288	0.767	0.911	0.832	0.817	0.947
ResNext with 50 layers	256x192	0.715	0.897	0.791	0.771	0.935
ResNext with 50 layers	384x288	0.724	0.899	0.794	0.777	0.936
ResNext with 101 layers	256x192	0.726	0.900	0.801	0.781	0.939
ResNext with 101 layers	384x288	0.744	0.903	0.815	0.794	0.939
ResNext with 152 layers	256x192	0.730	0.903	0.808	0.785	0.940
ResNext with 152 layers	384x288	0.742	0.904	0.810	0.794	0.940
HrFormer Small	256x192	0.738	0.904	0.812	0.793	0.941
HrFormer Small	384x288	0.757	0.905	0.824	0.807	0.941
HrFormer Base	256x192	0.754	0.906	0.827	0.807	0.943
HrFormer Base	384x288	0.774	0.909	0.842	0.823	0.945
ViTPose-S	256x192	0.739	0.903	0.816	0.792	0.942
ViTPose-B	256x192	0.757	0.905	0.829	0.810	0.946
ViTPose-L	256x192	0.782	0.914	0.850	0.834	0.952
ViTPose-H	256x192	0.788	0.917	0.855	0.839	0.954
ViTPose-H*	256x192	0.790	0.916	0.857	0.840	0.953
HourGlass 52-depth	256x256	0.726	0.896	0.799	0.780	0.934
HourGlass 52-depth	384x384	0.746	0.900	0.812	0.797	0.939
SeresNet with 50 layers	256x192	0.729	0.903	0.807	0.784	0.941
SeresNet with 50 layers	384x288	0.748	0.904	0.819	0.799	0.941
SeresNet with 101 layers	256x192	0.734	0.905	0.814	0.790	0.941
SeresNet with 101 layers	384x288	0.754	0.907	0.823	0.805	0.943
SeresNet with 152 layers	256x192	0.730	0.899	0.810	0.787	0.939
SeresNet with 152 layers	384x288	0.753	0.906	0.824	0.806	0.945
rsn with 18 layers	256x192	0.704	0.887	0.781	0.773	0.927
rsn with 50 layers	256x192	0.724	0.894	0.799	0.790	0.935
2xrsn with 50 layers	256x192	0.748	0.900	0.821	0.810	0.939
3xrsn with 50 layers	256x192	0.750	0.900	0.824	0.814	0.941
ScNet with 50 layers	256x192	0.728	0.899	0.807	0.784	0.938
ScNet with 50 layers	384x288	0.751	0.906	0.818	0.802	0.942
ScNet with 101 layers	256x192	0.733	0.902	0.811	0.789	0.940
ScNet with 101 layers	384x288	0.752	0.906	0.823	0.804	0.943
ViPNAS-MobileNetV3	256x192	0.700	0.887	0.783	0.758	0.929
ViPNAS-Res50	256x192	0.711	0.894	0.787	0.769	0.934

IV. DISCUSSION

The discussion of top-down pose estimation models, their applications, and future prospects opens up a range of crucial aspects and considerations. As we review the examples of these models, there's a clear sense of the significant strides made in this field.

One salient point is the robustness and versatility of the top-down approach in dealing with different scenarios and use-cases. Despite the inherent challenge of differentiating between individuals in crowded scenes, these models have proved their efficacy by separating the task into human detection and pose estimation.

The innovation of integrating transformer architectures into pose estimation models is another point of discussion. Transformer-based models like HrFormer and VitPose have shown that long-range dependencies can be managed more efficiently, leading to improved precision in pose estimation.

The incorporation of stepwise training processes, as seen in RSN, and dual attention mechanisms like in SCNet, are notable advancements. However, questions may arise as to how these advanced techniques can be optimized further for better performance and accuracy.

The application of Neural Architecture Search (NAS) in models like ViPnas is a promising development. However, one might ponder over the challenges in its implementation, the computational resources required, and how this technique can be made more accessible for broader use.

Looking at the diverse applications of these models, it's evident that top-down pose estimation is becoming integral in many fields. As these models continue to improve, one could discuss the potential challenges in their real-world implementation. For instance, privacy and ethical considerations in surveillance applications, or the accuracy required for healthcare applications.

Another point for discussion could be the need for more comprehensive and diverse datasets to train these models, which could help them generalize better and handle more complex and varied scenarios.

Finally, as top-down pose estimation models continue to evolve, it would be interesting to discuss future prospects and the potential for these models to revolutionize more industries. There might also be scope for these models to be combined with other AI technologies, such as natural language processing, for even more nuanced and holistic solutions.

In conclusion, the advancements, applications, and potential of top-down pose estimation models open up a broad array of discussion points, posing intriguing questions about the future of this promising field.

V. Conclusion

In summary, top-down pose estimation models have made a significant impact across various fields, ranging from healthcare to gaming, autonomous vehicles to sports performance analysis. These models, such as HRNet, ResNeXt, HrFormer, VitPose, Hourglass, SE-ResNet, RSN, SCNet, and ViPnas, have demonstrated remarkable advancements in handling the complexities of pose estimation tasks, offering increased precision, robustness, and efficiency.

The unique advantage of these models lies in their approach of partitioning the process into two principal stages: first identifying humans, and then estimating their poses. This characteristic effectively eliminates the challenge of distinguishing between individuals in crowded scenes, which has traditionally been a stumbling block in pose estimation tasks.

The continuous evolution and improvement of these models, marked by the incorporation of transformer architectures, recalibration of feature maps, stepwise training processes, dual attention mechanisms, and even Neural Architecture Search (NAS), showcase the immense potential in the field of pose estimation.

In the realm of applications and use-cases, top-down approaches have proven to be transformative. From enhancing user experiences in AR and VR, monitoring patients in healthcare, providing performance analysis in sports, ensuring safety in surveillance systems, enriching human-computer interactions, to delivering lifelike animations in the gaming and film industry - these models have opened up new possibilities and solutions.

While the progress made so far has been significant, it's clear that this is a field in constant development. With the ongoing research and advancements, we can look forward to even more accurate, efficient, and versatile top-down pose estimation models capable of handling increasingly complex scenarios in the future. Their potential to reshape various industries is vast, making them an exciting and promising area of study and application.

VI. Acknowledgment

I would like to express sincere gratitude to Ordulu Teknoloji and Rayan Abri for their invaluable support and contributions throughout this study. The insights, expertise, and unwavering dedication they brought to this research have been fundamental in its success. This acknowledgment not only represents my personal appreciation but also emphasizes their significant roles in propelling this study forward. Without their assistance, the completion of this work would not have been possible.

VII. References

[1] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep High-Resolution Representation Learning for Visual Recognition. Retrieved from https://arxiv.org/abs/1908.07919

[2] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2016). Aggregated Residual Transformations for Deep Neural Networks. Retrieved from https://arxiv.org/abs/1611.05431

[3] Ma, X., Chen, J., Li, X., Sun, P., Wang, T., Wang, J., & Wang, Z. (2021). HRFormer: High-Resolution Transformer for Dense Prediction. Retrieved from https://arxiv.org/abs/2110.09408

[4] Chen, Y., Zhang, H., Wang, C., Xu, C., Zhang, Y., & Zhu, J. (2022). ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. Retrieved from https://arxiv.org/abs/2204.12484

[5] Hu, J., Shen, L., & Sun, G. (2017). Squeeze-and-Excitation Networks. Retrieved from https://arxiv.org/abs/1709.01507v4

[6] Tang, W., Yu, P., & Wu, Y. (2020). Learning Delicate Local Representations for Multi-Person Pose Estimation. Retrieved from https://arxiv.org/abs/2003.04030

[7] Tian, Y., Li, H., & Li, C. (2020). SPCNet:Spatial Preserve and Content-aware Network for Human Pose Estimation. Retrieved from https://arxiv.org/abs/2004.05834

[8] Fu, H., Zhang, H., Lin, J., Li, H., Ma, X., Cao, X., & Zhang, X. (2021). ViPNAS: Efficient Video Pose Estimation via Neural Architecture Search. Retrieved from https://arxiv.org/abs/2105.10154