ENABLING HIGH-RESOLUTION POSE ESTIMATION IN REAL TIME USING ACTIVE PERCEPTION

Theodoros Manousis, Nikolaos Passalis and Anastasios Tefas

Computational Intelligence and Deep Learning Group, AIIA Lab,
Department of Informatics, Aristotle University of Thessaloniki
Thessaloniki, Greece
E-mails: {tmanousis, passalis, tefas}@csd.auth.gr

ABSTRACT

Deep Learning (DL) models have enabled very accurate pose estimation. However, most of the existing approaches require images of relatively high resolution, since locating body parts and joints accurately is challenging, which increases the computational cost of these approaches. To overcome this limitation in this paper we propose an active perception method for high-resolution pose estimation that enables efficiently selecting the most appropriate image region for analysis and then employing a bottom-up pose estimator on the corresponding region. This allows for significantly improving the efficiency of pose estimation by selectively analyzing in high resolution only the parts of the image that contain humans. To ensure the computational efficiency of the proposed method we propose using low-resolution heat maps extracted using the same pose estimation model in order to guide the active perception process. The proposed method is model agnostic since it can be combined with any bottom-up pose estimation model in order to enable high-resolution analysis. We have experimentally evaluated the proposed method using a well-known pose estimation model, Lightweight OpenPose, demonstrating its effectiveness on three high-resolution variants of the COCO2017 dataset.

***Index Terms*—** pose estimation, high resolution, active perception, deep learning

1. INTRODUCTION

Human pose estimation is a challenging problem that concerns locating the human body parts, as well as the overall pose of humans. The advent of Deep Learning (DL) models has enabled very accurate pose estimation [1, 2, 3]. This led to a wide range of applications of human pose estimation spanning over many different fields, such as sports [4], healthcare, and others [5, 6]. Most of the existing approaches require images of relatively high resolution, since locating body parts and joints accurately is a challenging task. Indeed, the wide availability of high-resolution (HR) sensors enables accurate pose estimation, even on mobile devices that are nowadays equipped with HR cameras.

However, this also comes with a significant computational cost, since processing and analyzing HR images require powerful hardware. The deployment of such algorithms on mobile devices becomes even harder when real-time constraints exist [7]. In such cases, most algorithms typically downscale the input image in order to meet real-time requirements and keep the computational power within the envelope of the device at hand. As we also experimentally demonstrate in this paper, this also comes with a significant reduction in pose estimation accuracy. As a result, we end up having to choose between accurate, yet slow algorithms for pose estimation (if high-resolution images are fed to the model) or faster, yet less accurate pose estimation (if the original images are downscaled before being fed to the model).

At this point, it is worth noting that even though the original image is high resolution, usually humans only cover a small part of the input image. However, existing methods either analyze the whole frame at once (bottom-up) [8, 9], regardless of the number of people that appear, or require the use of human detectors (top-down) [10] to first detect and crop the regions where persons appear. The first category spends a lot of time analyzing the whole image even though persons typically do not appear in most of the frame. On the other hand, the second category has the potential to speed up inference by only analyzing the parts of the image where persons appear. However, it requires the use of person detectors that are also computationally heavy and typically unable to process high-resolution images in real-time on embedded devices [11]. Furthermore, these approaches do not scale well, since the running time increases proportionally to the number of detected persons. Therefore, a question that naturally arises is if it is possible to only analyze the parts of the images that are of interest for human pose estimation while maintain-

This work was supported by the European Union's Horizon 2020 Research and Innovation Program (OpenDR) under Grant 871449. This publication reflects the authors' views only. The European Commission is not responsible for any use that may be made of the information it contains.

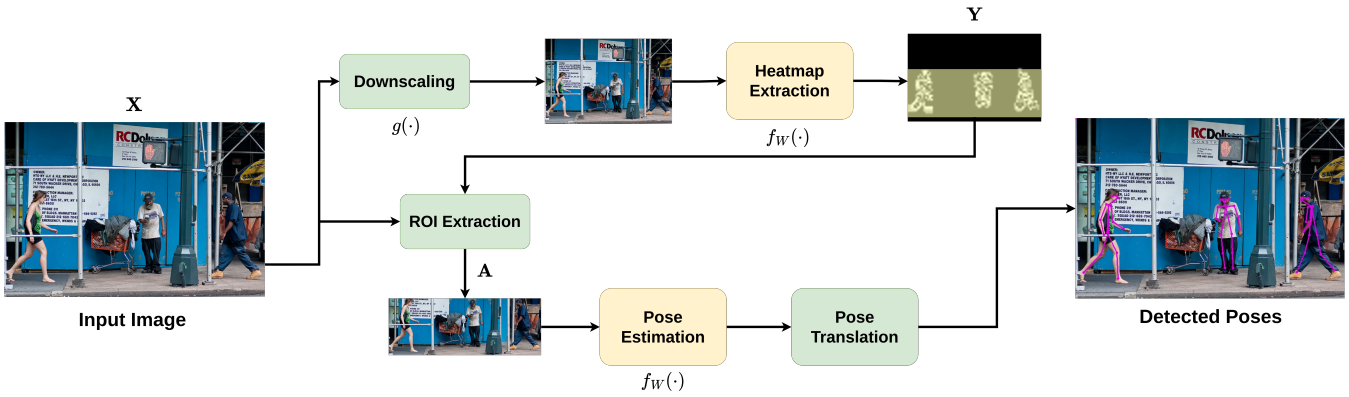


Fig. 1: Pipeline of the proposed method. Note that the same pose estimation model is employed twice, i.e., one time to extract the heatmap that drives the active perception process (selecting the ROI to crop), and one time to extract the final poses.

ing the advantages and speed of bottom approaches. Active perception methodologies [12, 13], which allow DL models to focus on specific parts of the input image, can provide a solution to this problem, by allowing only for analyzing selected regions of interest in the input image. Of course, such models should be fast enough in order to avoid spending more time in selecting such regions compared to the time required for analyzing the input.

The main contribution of this paper is an active perception method for HR pose estimation that enables efficiently selecting the most appropriate image region analysis and then employing a bottom-up pose estimator on the corresponding region. Intuitively, the proposed method works in a similar fashion to the way humans process images, i.e., we first take a brief look into the input image and then further analyze (look) into specific regions. In order to make the proposed method computationally efficient, we avoid having a separate active perception network, as typically happens in active perception [14]. Instead, we take advantage of the existing backbone of a DL-based pose estimation model in order to extract a heatmap that guides the active perception module. We have experimentally found out that this process can be performed using a down-scaled version of the input image, which can significantly speed up the inference of the model. Then, we crop the selected region of interest (ROI) that is fed into the same model in high resolution. This pipeline enables to process high-resolution images significantly faster compared to feeding the original high-resolution image and with significantly better accuracy compared to performing down-scaling. It is worth noting that the proposed method is model agnostic since it can be combined with any bottom-up pose estimation model in order to enable high-resolution analysis. In this paper we have experimentally evaluated the proposed method using a well-known pose estimation model, OpenPose [1], demonstrating its effectiveness on three high-resolution variants of the COCO2017 dataset [15].

The rest of the paper is structured as follows. First, the proposed method is introduced in Section 2. Then, the experimental evaluation is provided in Section 3. Finally, conclusions are drawn in Section 4.

2. PROPOSED METHOD

The main idea of the proposed method is to selectively analyze only the portion of the input images that contain useful information and discard the rest of the input, allowing for both improving inference speed, as well as maintaining accuracy by analyzing the parts of the image where people might appear in high resolution. The outline of the proposed method is presented in Fig. 1.

The first step of the proposed method is to create a confidence map of humans in the frame (heatmap). Let $\mathbf{X} \in \mathbb{R}^{n \times m \times c}$ be an input image to be analyzed, where n is the height, m is the width of the image and c is the number of channels. To generate the heatmap we employ a pose estimation model $f_W(\cdot)$, where $\mathbf{W}$ are the trainable parameters of the model. Even though the model $f_W(\cdot)$ is trained to detect body parts and joints, it can be directly used to detect regions of potential interest since these appear on humans. Note that during this step, we are not interested in estimating the exact pose of a human. Therefore, we can use a low-resolution version of $\mathbf{X}$ in order to identify potential regions of interest. To this end, we employ a downscaling function $g(\cdot)$ in order to acquire the heatmap as:

$$\mathbf{Y}_H = f_W(g(\mathbf{X})) \in \mathbb{R}^{n' \times m' \times c'}, \quad (1)$$

where n' and m' are the width and height of the extracted heatmap and c' is the number of joints/body parts to be detected (since the detectors typically extract one heatmap of each joint/body part). In this work, we perform simple averaging in order to downscale the image to the desired resolution. In practice, this can be implemented as part of the model

using the appropriate pooling layer. As shown in Fig. 2, we can downscale the image by a factor of 6 and still be able to identify regions where humans appear. To acquire the final heatmap that can be used for active perception we simply sum the confidences for all potential detections as:

$$\mathbf{Y} = \sum_{i=1}^{c'} [\mathbf{Y}_H]_i \in \mathbb{R}^{n' \times m'}, \tag{2}$$

where $[\mathbf{Y}_H]_i$ denotes the i-th slice of heatmap $\mathbf{Y}_H$. This heatmap is further filtered and smoothed by clipping values below a threshold value $c = 0.1$ and using a 2D average filter with kernel $k = 5$. At this point, the heatmap shows an approximation of the human figures and it needs to be found the area of interest to be cropped from the original frame. In order to get the area of interest a simple but effective method is used. We locate the most upper-left and the most down-right pixels of the heatmap with a non-zero value, and we keep those coordinates in order to crop the original image. The area of interest in this heatmap is denoted by $\mathbf{b}_{heatmap}$, defined as:

$$\mathbf{b}_{heatmap} = [x_{min}, y_{min}, x_{max}, y_{max}], \tag{3}$$

where $x_{min}, y_{min}, x_{max}, y_{max}$ are the top-left and bottom-right coordinates.

Note that the coordinates of the bounding box boundaries expressed in the heatmap correspond to the resized image size. Therefore, to crop the area of interest of the high-resolution input image properly, these coordinates must be transformed to match the real input image size. This step is crucial to ensure that the method accurately identifies the area of interest and produces reliable pose estimation results. Therefore, we define the area in the original high-resolution image to be cropped as:

$$\mathbf{b}_{ROI} = [x'_{min}, y'_{min}, x'_{max}, y_{max}], \tag{4}$$

where

$$[x'_{min}, x'_{max}] = [x_{min}, x_{max}] \cdot \frac{m}{m'} s, \tag{5}$$

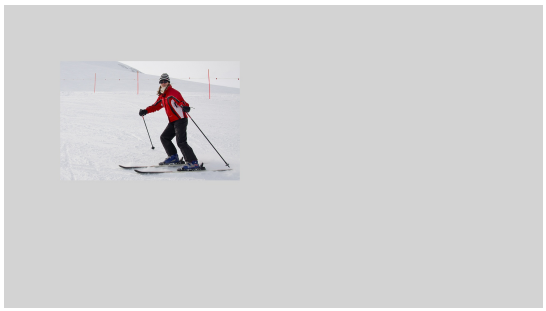
$$[y'_{min}, y'_{max}] = [y_{min}, y_{max}] \cdot \frac{n}{n'} s, \tag{6}$$

s is the scaling factor that is applied to the input image in order to be resized on the desired value. This scaling factor should also take into account possible down-sampling performed by the model as the heatmap is extracted. We also denoted by $\mathbf{A}$ the final region of interest extracted from the original image $\mathbf{X}$ after cropping the area $\mathbf{b}_{ROI}$.

Then, the area of interest that is extracted from the original high-resolution image is fed to the pose estimation model again:

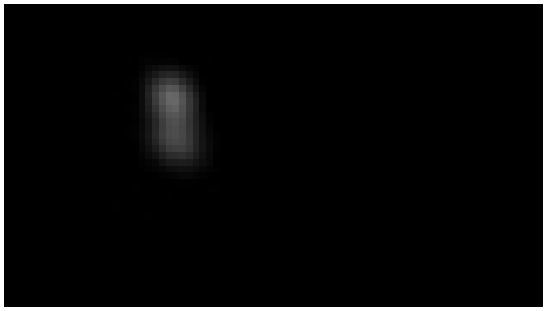
$$\mathbf{Y'}_H = f_W(\mathbf{A}) \in \mathbb{R}^{n'' \times m''}, \tag{7}$$

where $\mathbf{Y}'$ is the new heatmap produced from the estimator and $\mathbf{A}$ is the new (n'', m'') cropped image, where n'', m'' are its width and height dimensions. Then, this heatmap is appropriately processed based on the method used for pose estimation, e.g., by applying non-maximum suppression and incorporating information from Part Affinity Fields for the OpenPose algorithm [1]. Finally, note that the poses obtained from the previous step are calculated on a resized cropped image, where the predicted keypoints are represented in a local system of coordinates. Therefore, it is necessary to transform these keypoints back into the original coordinate system of the high-resolution input image before returning the predicted poses. This can be trivially implemented by performing the appropriate translations based on $\mathbf{b}_{ROI}$.



(a) Initial input image

(b) Heatmap when HR image is fed to the model



(c) Heatmap from input image when resized before fed to the model

Fig. 2: Comparing heatmaps generated when the input image (a) is directly fed to the model (b) or downscaled by a factor of about 4 (c). We can see that even though the downscaled image is not appropriate for localizing the keypoints, it is adequate for detecting regions of interest.

Table 1: Comparing the proposed method to the LW OpenPose approach. The average perception (0.5:0.95) is reported. We also report the speed (FPS) in parentheses. Higher values indicate better performance.

Dataset	LW OpenPose	Proposed
COCO (Original)	0.400 (35 FPS)	**0.415** (35 FPS)
COCO (720p)	0.288 (27 FPS)	**0.359** (31 FPS)
COCO (1080p)	0.172 (23 FPS)	**0.312** (40 FPS)
COCO (1440p	0.111 (20 FPS)	**0.274** (54 FPS)

Table 2: Comparing different resizing options for the second stage of the proposed method, along with the impact of directly using high-resolution images (Raw HR).

Dataset	COCO (1080p)	COCO (1440p)
Proposed (Setup 1)	0.312 (40 FPS)	0.274 (54 FPS)
Proposed (Setup 2)	0.393 (25 FPS)	0.334 (38 FPS)
Proposed (Setup 3)	**0.408** (14 FPS)	**0.342** (24 FPS)
Raw HR	0.424 (2.7 FPS)	OUT OF MEMORY

3. EXPERIMENTAL EVALUATION

The proposed method was evaluated using the COCO2017 dataset, which contains 5000 images, with almost 50% depicting at least one person. Furthermore, to simulate the effect of having high-resolution images we created a synthetic dataset based on the COCO2017 since no large-scale high-resolution datasets for pose estimation exist to the best of our knowledge. Therefore, three additional datasets were created by randomly placing the original images onto a white canvas of larger resolutions of 1280×720 (720p dataset), 1920×1080 (1080p dataset), and 2560×1440 (1440p dataset), respectively. The annotations were appropriately translated in order to match the generated images. For all the conducted experiments using the proposed method, we performed downscaling both for the original image and the extracted region of interest to a height of 360 pixels. Furthermore, for all the conducted experiments we employ a fast and lightweight bottom-up pose estimation, the Lightweight OpenPose (LW OpenPose) [16]. The inference speed was measured using an 8-core workstation with an NVIDIA 2070 Super GPU with 8GB of VRAM.

First, in Table 1 we compare the proposed method to the baseline Lightweight OpenPose approach, which resizes the input images to a height of 368 before processing. In all evaluated cases, including the original dataset the proposed method leads to significant performance improvements while meeting or even exceeding in many cases the real-time requirements ($\geq$ 25 FPS). For example, for the highest resolution dataset (1440p) the average precision increases from 0.111 to 0.274. Please also note that the inference speed increases in the proposed method as the resolution increases. This is a (positive) side effect of the employed active perception procedure, since if there are no regions of interest identified during the first step, then the second step is not activated. At the same time, note that the proposed method can avoid the costly rescalings involved in the original LW OpenPose approach, which can lead to increasing the inference speed (for the LW OpenPose).

Furthermore, we have noticed that for high-resolution images the heatmap activations are not accurate enough, which could lead to extracting larger areas of interest. To examine if we could overcome this limitation by employing higher resolution crops during the second stage of the proposed method we also performed an additional evaluation, where the cropped area was resized to 540 pixels (Setup 2) and 720 pixels (Setup 3) respectively. The experimental results are reported in Table 2 and compared to the baseline proposed method (Setup 1), as well as to the LW OpenPose without performing any kind of rescaling (abbreviated as "Raw HR"). First, note that increasing the size of the images during the second stage led to increasing the pose estimation accuracy in all stages. However, this also comes with a decrease in the observed speed, which - however - remains comparable with the LW OpenPose while achieving significantly better precision (Table 1). Furthermore, note that even though directly using the raw HR images leads to increased precision, the speed is reduced over one order of magnitude, while using 1440p images was not possible despite using a GPU with 8GB of VRAM.

4. CONCLUSIONS

In this paper, we proposed an active-perception method for HR pose estimation that allows for efficiently selecting the most appropriate image region analysis and then employing a bottom-up pose estimator on the corresponding region. The proposed method was evaluated on three high-resolution variants of the COCO2017 dataset leading to significant improvements, allowing for improving pose estimation precision, while reducing inference time. Furthermore, the proposed method can be tuned to meet the requirements (speed-precision trade-off) by appropriately tuning the resizing dimensions used in the first and second inference steps. The obtained results also highlight the potential of active perception approaches in HR analysis paving way for enabling HR analysis for other tasks, e.g. object detection [17], crowd counting [18], and others. Also, the proposed method used a simple, yet effective process for selecting the region of interest. However, more intelligent approaches could be used to produce more than one region of interest, e.g. by employing neural regions of interest-based proposals [19].

5. REFERENCES

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[2] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 483–499.

[3] Alexander Toshev and Christian Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. jun 2014, IEEE.

[4] Takumi Kitamura, Hitoshi Teshima, Diego Thomas, and Hiroshi Kawasaki, "Refining openpose with a new sports dataset for robust 2d pose estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 672–681.

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 561–578.

[6] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Rui Qian, Tao Wang, Ning Xu, Hongkai Xiong, Guo-Jun Qi, and Nicu Sebe, "Human in events: A large-scale benchmark for human-centric video analysis in complex events," *arXiv preprint arXiv:2005.04490*, 2020.

[7] Haopan Ren, Wenming Wang, Kaixiang Zhang, Dejian Wei, Yanyan Gao, and Yue Sun, "Fast and lightweight human pose estimation," *IEEE Access*, vol. PP, pp. 1–1, 03 2021.

[8] Milan Kresović and Thong Duy Nguyen, "Bottom-up approaches for multi-person pose estimation and it's applications: A brief review," *arXiv preprint arXiv:2112.11834*, 2021.

[9] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4929–4937.

[10] Adrian Bulat and Georgios Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proceedings of the European Conference on Computer Vision*, pp. 717–732. 2016.

[11] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7753–7762.

[12] Ruzena Bajcsy, Yiannis Aloimonos, and John K Tsotsos, "Revisiting active perception," *Autonomous Robots*, vol. 42, pp. 177–196, 2018.

[13] Nikolaos Passalis and Anastasios Tefas, "Pseudo-active vision for improving deep visual perception through neural sensory refinement," in *Proceedings of the IEEE International Conference on Image Processing*, 2021, pp. 2763–2767.

[14] Theodoros Bozinis, Nikolaos Passalis, and Anastasios Tefas, "Improving visual question answering using active perception on static images," in *Proceedings of the International Conference on Pattern Recognition*, 2021, pp. 879–884.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *Proccedings of the European Conference on Computer Vision*, 2014, pp. 740–755.

[16] Daniil Osokin, "Real-time 2d multi-person pose estimation on cpu: Lightweight openpose," *arXiv preprint arXiv:1811.12004*, 2018.

[17] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023.

[18] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 833–841.

[19] Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2017, pp. 924–933.