Human Pose Estimation from a Single Image

Yikai Wang
Student id: 2020233280

Chuanhao Hao
Student id:

Abstract

1. Introduction

Human pose estimation, which has been extensively studied in computer vision literature, involves estimating the configuration of human body parts from input data captured by sensors, in particular images and videos. Human pose estimation provides geometric and motion information of the human body which has been applied to a wide range of applications such as human-computer interaction, motion analysis, augmented reality(AR), virtual reality(VR), healthcare,etc. With the rapid development of deep learning solutions in recent years, such solutions have been shown to outperform classical computer vision methods in various tasks including image classification, semantic segmentation and object detection. Significant progress and remarkable performance have already been made by employ deep learning techniques in human pose estimation tasks. However, challenges such as occlusion, insufficient training data, and depth ambiguity still pose difficulties to be overcome. 2D human pose estimation from images and videos with 2D pose annotations is easily achievable and high performance has been reached for the human pose estimation of a single person using deep learning techniques. 2D single-person pose estimation is used to localize human body joint positions when the input is a single-person image. A good human pose estimation system must be robust to occlusion and severe deformation, successful on rare and novel poses, and invariant to challenge in appearance due to factors like clothing and lighting. Early work tackles such difficulties using robust features and sophisticated structured prediction: the former is used to produce local interpretations, whereas the latter is used to infer a globally consistent pose.

However, this conventional pipeline has been greatly reshaped by convolutional neural networks[4], a main driver behind an explosive rise in performance across many computer vision tasks. Recently pose estimation[9] systems have universally adopted convolutional neural networks as their main building blocks, largely replacing hand-crafted features and graphical models; this strategy has yielded drastic improvements on standard benchmarks. In our project, we

2. Related Work

Traditionally 2D human pose estimation method adopt different hand-crafted feature extraction techniques for body parts, and these early works describe human body as a stick figure to obtain global pose structures. Recently, deep learning-based approaches have achieved a major breakthrough in human pose estimation by improving the performance significantly.

Using AlexNet[3] as the backbone, Toshev et al[10] proposed a cascaded deep neural network regressor named DeepPose, with the introduction of "DeepPose" , research on human pose estimation began shifting from classic approaches to deep learning networks. Toshev et al use their network to directly regress the x,y coordinate of joints. The work by Tompson et al[9] instead generates heatmaps by running an image through multiple resolution banks in parallel to simultaneously capture features at a variety of scales. A critical feature of the method proposed by Tompson et al[9] is the joint use of ConvNet and a graphical model. Their graphical model learns typical spatial relationships between joints. Others have recently tackled this in similar ways[6] with variation on how to approach unary score generation and pairwise comparison of adjacent joints. Based on GoogLeNet[8], Carreira et al.[1] proposed an Iterative Error Feedback(IEF) network, which is a self-correcting model to progressively change an initial solution by injecting the prediction error back to the input space. Sun et al.[7] introduced a structure-aware regression method called "compositional pose regression" based on ResNet-50[2]. This method adopts a re-parameterized and bone-based representation that contains human body information and pose structure, instead of the traditional joint-based representation. Luvizon et al.[5] proposed an end-to-end regression approach for human pose estimation using soft-argmax function to convert feature maps into joint coordinates in a fully differentiable framework.

3. Data

Max Planck Institute for Informatics(MPII) Human Pose Dataset: This is a popular dataset for evaluation of articulated human pose estimation. The dataset includes around 25000 images containing over 40000 individuals with annotated body joints. The images were systematically collected by a two-level hierarchical method to capture everyday human activities. The entire dataset covers 410 huamn activities and all the images are labeled. Each image was extracted from a YouTube video and provided with preceding and following un-annotated frames. Moreover, rich annotations including body part occlusions, 3D torso and head orientations are also labeled.

4. Methods

5. Experiments

6. Conclusions

References

[1] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks mark. *Commun. ACM*, 60(6):84–90, 2017. 1

[4] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[5] Diogo C Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, 2019. 1

[6] 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*, pages 4929–4937, 2016. 1

[7] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017. 1

[8] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[9] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27:1799–1807, 2014. 1

[10] 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*, pages 1653–1660, 2014. 1