FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

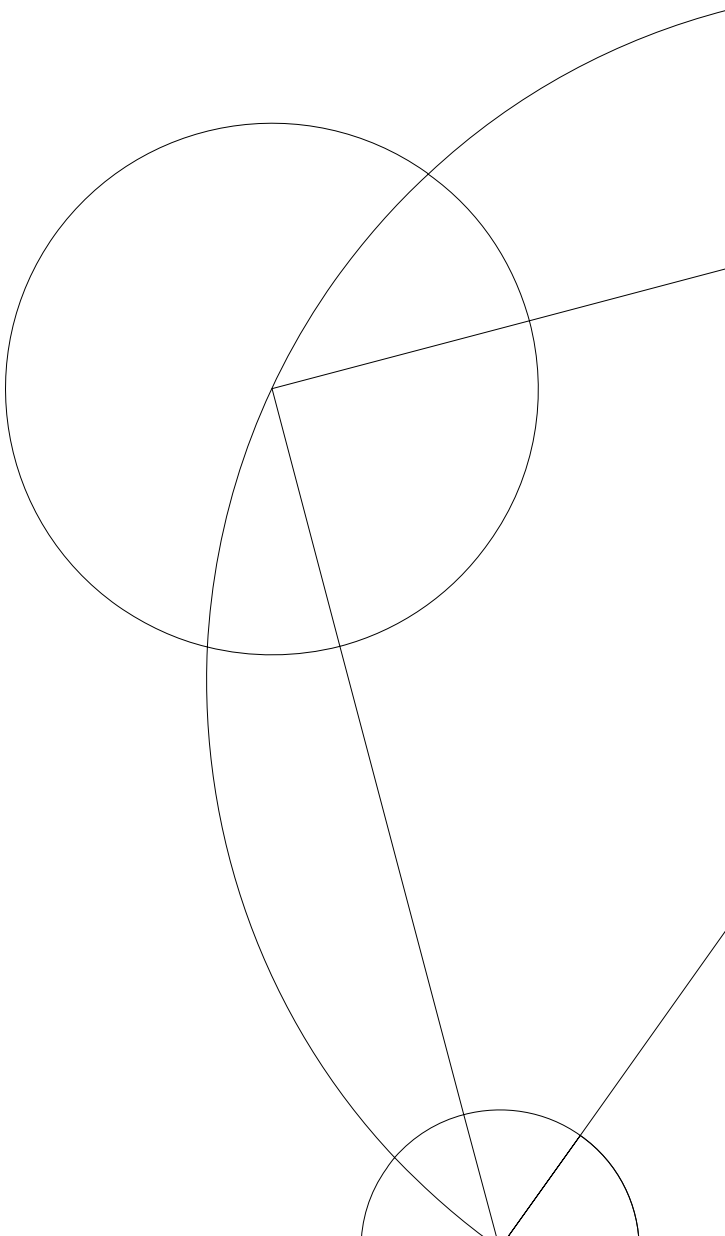
Master Thesis

2D Tracking in Climbing

Using Temporal Smoothing

André Oskar Andersen (wpr684)
wpr684@alumni.ku.dk

2023



Supervisor
Kim Steenstrup Pedersen kimstp@di.ku.dk

Abstract

Preface

Acknowledgement

Contents

Notation

1 Introduction

1.1 Related Work

2-dimensional pose estimation can be divided into either being image-based or video-based, where the methods in the latter case use the tempoeral information of the video to perform the pose estimation.

Image-based methods were initially based on the geometry between the joints of the taget image [10, 11, 16]. Following this, were the convolutional-based methods, that used convolutional neural networks [6] to perform the pose estimation [13, 8, 2, 4]. More recent methods use transformers [12] to deliver state-of-the-art results [14, 15].

Early video-based methods used 3-dimensional convolutions to capture the temporal information between neighboring frames [9, 3]. Other methods use LSTM's [5] to capture this temporal information [7, 1]. Like in the case of image-based methods, transformers [12] have recently been introduced to the video-based methods to capture the temporal information and deliver state-of-the-art-results [17].

1.2 Problem Definition

1.3 Reading Guide

2 Deep Learning Theory

The following section covers the most important background theory for the experiments in Section 5. This includes an introduction to various types of neural networks, as well as an introduction to the optimization of such networks.

2.1 Feedforward Neural Networks

-

2.2 Convolutional Neural Networks

2.3 Recurrent Neural Networks

2.3.1 Long Short-Term Memory Unit

2.3.2 Gated Recurrent Unit

2.4 Transformer

2.5 Training a Neural Network

3 Models

The following section covers the theory behind the various models that will be introduced in Section 5.

3.1 Mask R-CNN

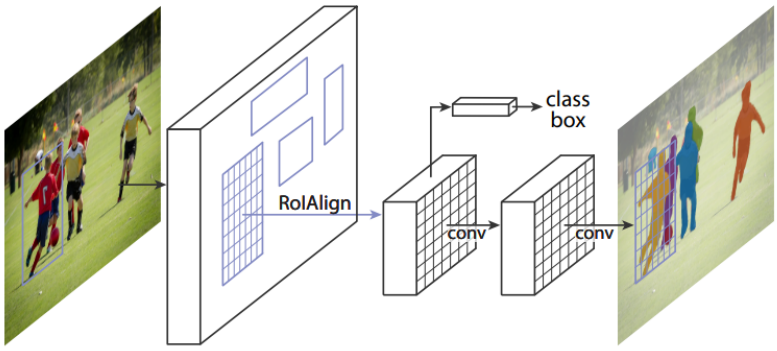


Figure 1: The Mask R-CNN framework for instance segmentation [4].

When we will be performing the pose estimation in Section 5, our developed methods will be a variation of the *Mask R-CNN*, introduced by He *et al.* in 2018 [4]. The following subsection explains the architecture of the Mask R-CNN and is based on an interpretation of He *et al.* [4].

3.2 UniPose-LSTM

3.3 DeciWatch

4 Dataset

5 Experiments

6 Discussion

7 Conclusion

8 References

[1] Bruno Artacho and Andreas Savakis. *UniPose: Unified Human Pose Estimation in Single Images and Videos*. 2020. DOI: 10.48550/ARXIV.2001.08095. URL: https://arxiv.org/abs/2001.08095.

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields*. 2018. DOI: 10.48550/ARXIV.1812.08008. URL: https://arxiv.org/abs/1812.08008.

[3] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. *Detect-and-Track: Efficient Pose Estimation in Videos*. 2017. DOI: 10.48550/ARXIV.1712.09184. URL: https://arxiv.org/abs/1712.09184.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. *Mask R-CNN*. 2017. DOI: 10.48550/ARXIV.1703.06870. URL: https://arxiv.org/abs/1703.06870.

[5] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[6] Yann LeCun, Yoshua Bengio, et al. "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.

[7] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. *LSTM Pose Machines*. 2017. DOI: 10.48550/ARXIV.1712.06316. URL: https://arxiv.org/abs/1712.06316.

[8] Alejandro Newell, Kaiyu Yang, and Jia Deng. *Stacked Hourglass Networks for Human Pose Estimation*. 2016. DOI: 10.48550/ARXIV.1603.06937. URL: https://arxiv.org/abs/1603.06937.

[9] Tomas Pfister, James Charles, and Andrew Zisserman. *Flowing ConvNets for Human Pose Estimation in Videos*. 2015. DOI: 10.48550/ARXIV.1506.02897. URL: https://arxiv.org/abs/1506.02897.

[10] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. "Poselet Conditioned Pictorial Structures". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 588–595. DOI: 10.1109/CVPR.2013.82.

[11] Yuandong Tian, C. Lawrence Zitnick, and Srinivasa G. Narasimhan. "Exploring the Spatial Hierarchy of Mixture Models for Human Pose Estimation". In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 256–269. ISBN: 978-3-642-33715-4.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. 2017. DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762.

[13] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. *Convolutional Pose Machines*. 2016. DOI: 10.48550/ARXIV.1602.00134. URL: https://arxiv.org/abs/1602.00134.

[14] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. *ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation*. 2022. DOI: 10.48550/ARXIV.2204.12484. URL: https://arxiv.org/abs/2204.12484.

[15] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. *TransPose: Keypoint Localization via Transformer*. 2020. DOI: 10.48550/ARXIV.2012.14214. URL: https://arxiv.org/abs/2012.14214.

[16] Yi Yang and Deva Ramanan. "Articulated Human Detection with Flexible Mixtures of Parts". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.12 (2013), pp. 2878–2890. DOI: 10.1109/TPAMI.2012.261.

[17] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. *DeciWatch: A Simple Baseline for 10x Efficient 2D and 3D Pose Estimation*. 2022. DOI: 10.48550/ARXIV.2203.08713. URL: https://arxiv.org/abs/2203.08713.