Thesis Title

Francesca Marsicano

Master Thesis
March 2022

Prof. Dr. My Prof

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

cgl
computer graphics laboratory

Abstract

Multi-person pose estimation in images and videos is an important yet challenging task with many applications. Despite the large improvements in human pose estimation enabled by the development of convolutional neural networks, there still exist a lot of difficult cases where even the state-of-the-art models fail to correctly localize all body joints. This motivates the need for an additional refinement step that addresses these challenging cases and can be easily applied on top of any existing method. In this work we analyze the different training protocols to extend the capabilities of he current architecture without drawing on additional ground truth labels or network stacked architecrure

Contents

List of Figures

List of Tables

Introduction

The goal of human pose estimation is to localize semantic keypoints of a human body. Recently, many methods utilize deep convolutional neural networks (CNNs) a Conventionally, the pose refinement has been mainly performed by multi-stage architectures

Existing methods for 3D motion capture of athletes can be classified into those that use motion sensors and those that use visual information only. The methods that use motion sensors can detect the accurate movement of the human skeleton by attaching markers to the human body. However, dedicated equipment such as motion sensors are expensive and the environment for measurement is limited. We refine OpenPose with our newly introduce sports dataset and a augmentation technique to improve the quality of 2D pose estimation in extreme poses. In addition, markers must be attached on the human body, which requires expert knowledge and may not be comfortable for the athlete. As a consequence, non-invasive techniques such as 3D pose estimation from multi-view RGB images are preferred ([MCL19], [SVB+19] , [XYN+20], [XZY+20], [LKP+20]

[29],[13]). In general, existing methods first estimate the 2D pose of the person in each 2D image and then triangulate the 2D skeletons to create a 3D skeleton. By doing so, a marker-less, low-cost system that can be used anywhere can be built. In such systems, accurate and robust 2D pose estimation is critical.

If taken into account the ultimate goal of using such systems in video production the attention

s. First, collecting accurate 3D pose annotation for RGB images is expensive and time-consuming

The first stage locates 2D human key-points from appearance information, while the second stage lifts the 2D joints into 3D skeleton employing geometric information. S

Related Work

Sample references are [ZRB+04] and [Alt89].

2.1 Supervised learning

2.1.1 Unsupervised learning

2.1.2 Model refinement

g. Carreria et al. [6] iteratively estimated error feedback from a shared weight model. The output error feedback of the previous iteration is transformed into the input pose of the next iteration, which is repeated several times for progressive pose refinement.

2.1.3 Knowledge distillation

Different from traditional knowledge distillation - a knowledge transformation methodology among networks, which forces student neural networks to approximate the softmax layer outputs of pre-trained teacher neural networks, the proposed self distillation framework distills knowledge within network itself. Within this framework enhancement of performance have been achieved in the field of classification [ZSG+19]

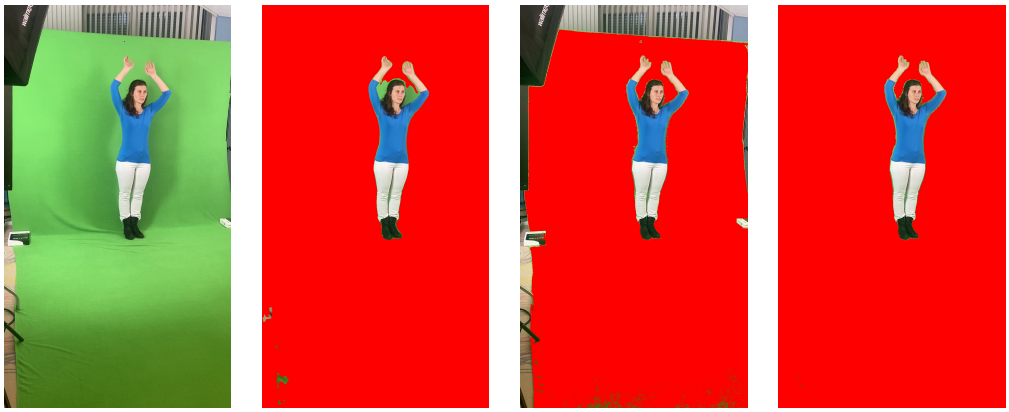
Your Central Work

OpenPose [CHS+19] is one of the most popular methods to obtain a 2D skeleton from a single color image. OpenPose consists of a deep convolutional neural network (CNN) that is trained on a large human database annotated with 2D skeletons. Remarkably, OpenPose shows high accuracy even for images that contain multiple people. It can detect poses in real time with high accuracy but in case a person in in an extreme position if fails to estimate the correct pose. Some of the most common cases of failure are due to ambiguity between left and right side, crossing of arms and legs, upside down position and twisting motions. In this work we propose to refine the network on a single target by retraining on ad hoc created datasets. In the following sections we first go through 3.1 how we annotated the image collection, continue in section 3.2 over the results obtained by training in a supervised fashing and ultimately in the last section **??** side-step the vexation of correct labels by training in an unsupervised manner.

3.1 Dataset creation

Each dataset is comprised of a collection of poses, i.e. a series of images of a subject moving in front of camera. These have been extracted from a recorded video taken in two different settings: single camera view and multi camera view.

In the first case seven videos are captured in front of the same background with the same subject with various types of clothing. In the second setting there are 4 subjects, with 3 types of different outfits on 4 types of different backgrounds. Some examples can be seen in Fig. 3.2. Three of them serve only for evaluation (office, white wall, cubard) while the greenscreen is used for training. Since this can be easily masked out it is possible to train with different kind of background augmentations. The segmentation algorithm is based on chroma key compositing: the RGB image is converted to the HSV colorspace where all hue values corresponding background color can be easily identified. The region of the background which is not covered



Original frame (a) (b) (c)

Figure 3.1: *Generated mask and their binary composition overlayed on the original image* (a) *Deeplab output correctly identifies the figure but presents spurious mistake around the arms and at the very bottom.* (b) *Color segmentation presents much sharper boundaries but it's action is limited to the greenscreen.* (c) *Composition of the two binary masks.*

by the greenscreen has to be removed with another approach. Therefore an additional mask is generated with DeepLab [CPK+16] a semantic segmentation network is generated. This has good understanding of the contextual information to extract the foreground correctly but does not output sharp edges. The ultimate result is obtained by binary composition.

To fast track the generation of labels we make use of machine predictions instead of relying on human annotations.

These come in two variants:

- labels are generated with baseline OpenPose network. This relies on a single camera view setting
- labels are obtained throughun the projection of the 3D skeleton back into the camera view.

Such examples may act as noise and pollute the learned model if the model is not rich enough to capture such appearance variability.

One could start by arguing that the reason is not that datasets are bad, but that our object representations and recognition algorithms are terrible and end up over-learning aspects of the visual data that relates to the dataset and not to the ultimate visual task

3.2 Training on 2D data

3.3 Triangulated experiments

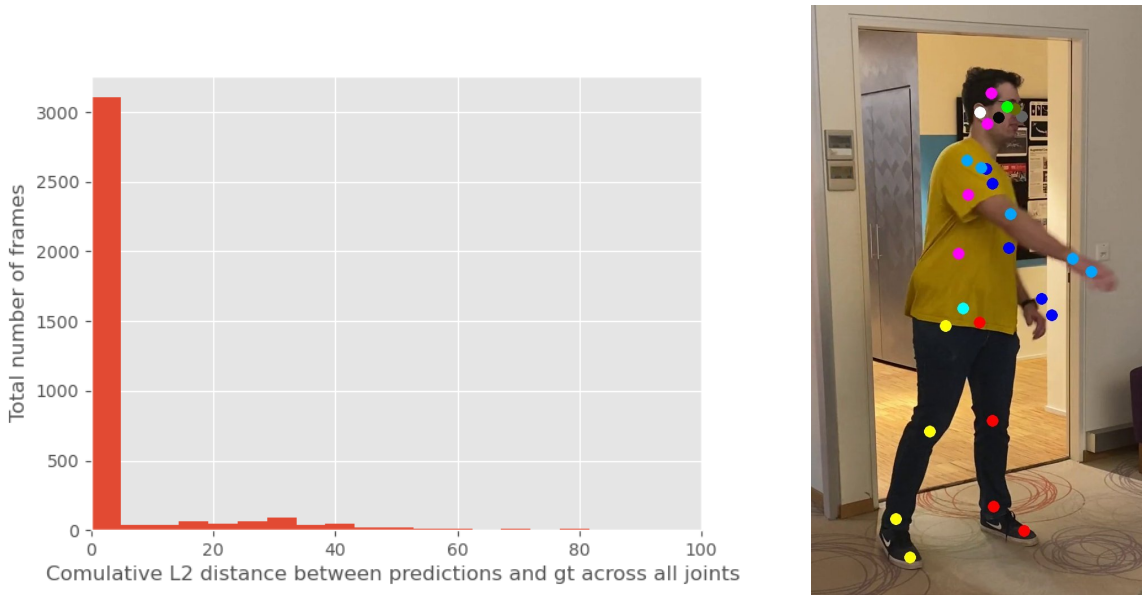
labelsection: training on 3D data

(a)

(b)

(c)

Figure 3.2: *Samples from the created dataset.* (a) *Four different appearance in the single camera view.* (b) *One take from each camera from the dual camera setup with greenscreen as background, two takes with a background "in the wild"* (c) *Other subjects.*



(a) (b)

Figure 3.3: (a) *Distribution of the errors between the triangulated projected point and ground truth. Only the comulative errors across all joints is considered here. Deviations of around 20 pixels from ground truth, (example in figure* (b)*) can be considered relevant : during the generation of ground truth labels small deviations as the well most face keypoints were left unchanged. Around 550 frames, out of the 3685 needed adjustment..*

3.3.1 Validation method

It seems improbable that

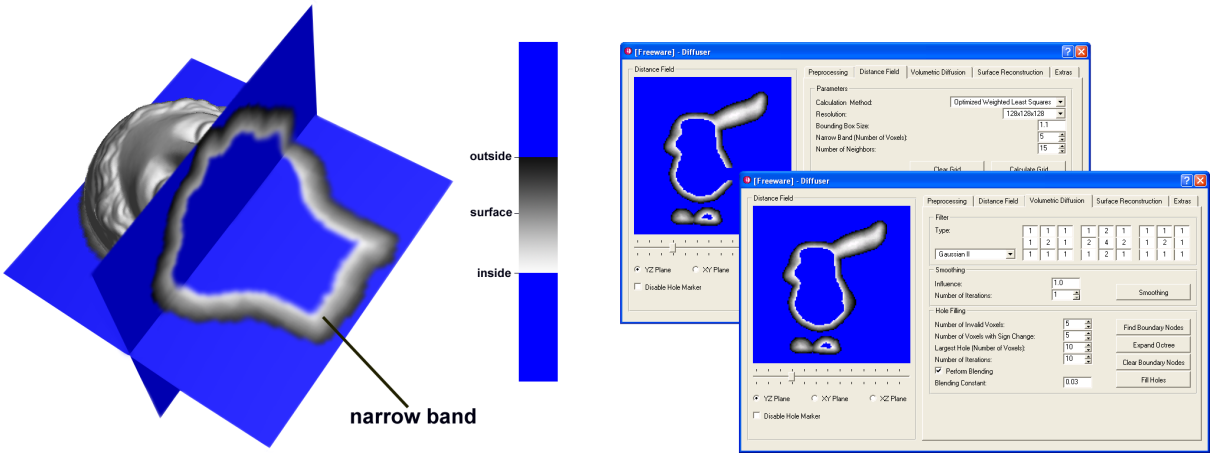
3.3.2

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat, see Table 3.1. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam, see Figure 3.4 (a). Isn't it?

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam

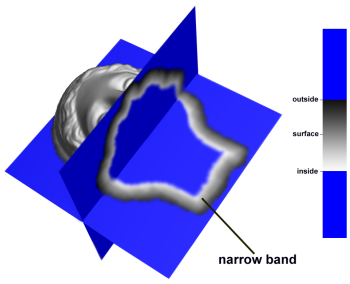
Quant.	*Ingredient*
200g	Weißmehl
1/4	Packung Frischhefe
4EL	lauwarme Milch
4EL	ÃŰl
1TL	Zucker
1TL	Salz
	lauwarmes Wasser

Table 3.1: *Flammkuchenteig. The ingredients have to be carefully chosen.*

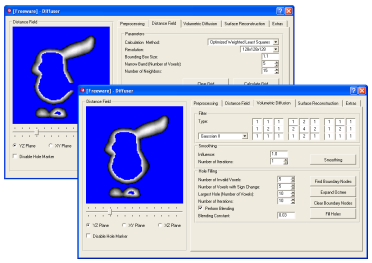


(a) (b)

Figure 3.4: *Volumetric diffusion.* (a) *Slices of the distance volume reveal the narrow band.* (b) *The user interface of the automatic hole filling tool allows to fine-tune the algorithm. The volumetric representation can be previewed before surface reconstruction.*



(a) Caption first.



(b) Caption second.

Figure 3.5: *Caption of both (a), (b).*

3.4 Second Section

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero et accumsan et iusto odio dignissim qui blandit

Bibliography

[Alt89] Simon L. Altman. Hamilton, grassmann, rodrigues, and the quaternion scandal—what went wrong with one of the major mathematical discoveries of the nineteenth century? *A Mathematical Association of America journal*, dec 1989.

[CHS+19] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[CPK+16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.

[LKP+20] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. *CoRR*, abs/2006.07778, 2020.

[MCL19] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single RGB image. *CoRR*, abs/1907.11346, 2019.

[SVB+19] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. *CoRR*, abs/1904.01324, 2019.

[XYN+20] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[XZY+20] Zhenbo Xu, Wei Zhang, Xiaoqing Ye, Xiao Tan, Wei Yang, Shilei Wen, Errui Ding, Ajin Meng, and Liusheng Huang. Zoomnet: Part-aware adaptive zooming neural

network for 3d object detection. *CoRR*, abs/2003.00529, 2020.

[ZRB+04] Matthias Zwicker, Jussi Räsänen, Mario Botsch, Carsten Dachsbacher, and Mark Pauly. Perspective accurate splatting. In *Proceedings of Graphics Interface*, pages 247–254, 2004.

[ZSG+19] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. *CoRR*, abs/1905.08094, 2019.