UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Deep Learning based Human Pose Estimation

by

Yang Li

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2020

Certificate of Authorship/Originality

I, Yang Li declare that this thesis, is submitted in fulfilment of the requirements for the award of Ph.D, in the Faculty of Engineer and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree at any other academic institution except as fully acknowledged within the text. This thesis is the result of a Collaborative Doctoral Research Degree program with Beijing Institute of Technology.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 30/09/2020

ABSTRACT

Deep Learning based Human Pose Estimation

by

Yang Li

Human pose estimation is an important research area in vision-based human activity analysis. Human pose estimation aims to estimate the human articulate joint positions in 2D/3D space from images or videos. Due to the complexity of the real environment and the diversity of human poses, vision-based human pose estimation is challenging. Recently, the rapid development of deep learning has much promoted the simulation of the analysis and reasoning capabilities of the human visual system. Therefore, it is of considerable significance to further explore vision-based human pose estimation using deep learning techniques. Specifically, this thesis proposes a series of methods for human pose estimation, summarized as follows:

We propose a video-based 2D pose estimation model, which embeds a multi-scale TCE module into the encoder-decoder network architecture for explicitly exploring temporal consistency in videos. The TCE module applies the learnable offset field to capture the geometric transformation between adjacent frames at the feature level. In addition, we explore the multi-scale geometric transformations at the feature level by integrating the spatial pyramid within the TCE module, which achieves further performance improvements.

We propose a self-supervised approach for 3D human pose estimation, which only relies on geometric prior knowledge and does not require any 3D human pose annotations. To this end, we design the transform re-projection loss, which is an effective technique to exploit multi-view consistency information and constrain the estimated 3D poses during training. Besides, we introduce a root position regression

branch to restore the global 3D poses during training. In this way, the network can reserve the scale information of re-projected 2D poses, which can improve the accuracy of the predicted 3D poses.

We propose a self-supervised 3D human pose estimation method based on the consistent factorization network, which fully disentangles the 3D human shape and camera viewpoint to overcome the projection ambiguity problem. To this end, we design a simple and effective loss function using multi-view information to constrain the canonical 3D human pose. Moreover, in order to reconstruct robust canonical 3D human poses, we represent 3D human pose as a combination of a dictionary of 3D pose basis, and adopt geometric information of 3D human poses to learn a hierarchical dictionary from 2D human poses by solving the NRSfM problem.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor A/Prof. Richard Yi Da Xu and Kan Li (Beijing Institute of Technology) for their continuous support of my Ph.D study and research, for his patience, motivation, and immense knowledge. Their guidance helped me in all the time of my research. I could not have imagined having a better advisor and mentor for my Ph.D study.

My sincere thanks also go to my fellow labmates: Shuai Jiang, Ziyue Zhang, Congzhentao Huang, Chen Deng, Ximeng Zhao, Jason, etc., for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last one year. I have the honour of studying and working with them in the past year which is valuably stamped in my life.

Finally, I must express my very profound gratitude to my parents and to girlfriend for providing me with unfailing support and continuous encouragement throughout my years of study and the process of researching. This accomplishment would not have been possible without them. Thank you.

Yang Li

List of Publications

Journal Papers

J-1. **Li Y.**, Li K., Wang X., Xu, R. Y. D., Exploring temporal consistency for human pose estimation in videos, Pattern Recognition, 2020, 103: 107258.

J-2. **Li Y.**, Li K., Wang X., Recognizing actions in images by fusing multiple body structure cues, Pattern Recognition, 2020: 107341.

Conference Papers

C-1. **Li Y.**, Li K., Jiang S., Zhang Z., Huang C., Xu, R. Y. D., Geometry-Driven Self-Supervised Method for 3D Human Pose Estimation, AAAI, 2020: 11442-11449.

C-2. **Li Y.**, Li K., Wang X., Deeply-Supervised CNN Model for Action Recognition with Trainable Feature Aggregation, IJCAI, 2018: 807-813.

C-3. Huang C., Jiang S., **Li Y.**, Zhang Z., Traish J., Deng C., Ferguson S., Xu, R. Y. D., End-to-end Dynamic Matching Network for Multi-view Multi-person 3d Pose Estimation, accepted by ECCV, 2020.

C-4. Zhang Z., Xu, R. Y. D., Jiang S., **Li Y.**, Huang C., Deng C., Illumination Adaptive Person ReID based on Teacher-Student Model and Adversarial Training, accepted by ICIP, 2020.

Contents

List of Figures

List of Tables