

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

Certificate	ii
Abstract	iii
Acknowledgments	v
List of Publications	vi
List of Figures	xi
List of Tables	xiv
1 Introduction	1
1.1 Aims and Motivations	1
1.2 Literature Review	4
1.2.1 Overview	4
1.2.2 2D Human Pose Estimation	5
1.2.3 3D Human Pose Estimation	9
1.3 Thesis Organization	14
2 Exploring Temporal Consistency for 2D Human Pose Estimation in Videos	15
2.1 Introduction	15
2.2 Related Work	18
2.2.1 Image-based Pose Estimation	18
2.2.2 Video-based Pose Estimation	19

2.3	Problem Formulation	20
2.4	Base Pose Estimation Network	22
2.5	Temporal Consistency Exploration	24
2.6	Multi-Scale Temporal Consistency Exploration	27
2.7	Video-based Pose Estimation Network	28
2.8	Results and Discussion	30
2.8.1	Datasets	30
2.8.2	Implementation Details	31
2.8.3	Performance on Video-based Pose Estimation	33
2.8.4	Ablation Study	34
2.8.5	Runtime Analysis	42
2.9	Conclusion	43
3	Geometry-driven Self-supervised Method for 3D Human Pose Estimation	44
3.1	Introduction	44
3.2	Related Work	47
3.2.1	3D Human Pose Estimation	47
3.2.2	Weakly/Self-supervised Approaches	48
3.3	Overview of the Method	49
3.4	Self-supervised Approach	50
3.4.1	Two-branch Training Architecture	51
3.4.2	Loss Function	51
3.5	Training	54
3.6	Experiments	54

3.6.1	Datasets	54
3.6.2	Implementation Details	56
3.6.3	Ablation Study	56
3.6.4	Analysis of Data Augmentation	57
3.6.5	Comparisons with State-of-the-art Methods	64
3.7	Conclusion	65
4	Self-supervised Method for 3D Human Pose Estimation with Consistent Factorization Network	66
4.1	Introduction	66
4.2	Related Work	68
4.2.1	3D Human Pose Estimation	68
4.2.2	Weakly/Self-supervised Approaches	69
4.2.3	Dictionary-based Approaches	70
4.3	Problem Formulation	71
4.4	Consistent Factorization Network	73
4.4.1	Consistent Factorization Constraint	74
4.5	Hierarchical Dictionary Learning	75
4.6	Training	76
4.7	Experiments	77
4.7.1	Dataset	77
4.7.2	Implementation Details	78
4.7.3	Ablation Study	79
4.8	Comparisons with State-of-the-art Methods	84
4.9	Conclusion	85

5 Conclusions and Future Work	86
5.1 Conclusions	86
5.2 Future Work	87
Bibliography	89

List of Figures

2.1	The base pose estimation network is based on the encoder-decoder architecture, where the encoder network extracts high-level convolutional features and the decoder network recovers high-resolution heatmaps. In addition, we apply the spatial pyramid module and the random erasing technique to improve the robustness of the network.	22
2.2	Architecture of the dilated spatial pyramid module.	23
2.3	Illustration of the TCE module. The TCE module follows a recurrent structure. It predicts the offset field to capture the geometric transformations between adjacent frames and produces the enhanced feature map through the deform and aggregation operations.	24
2.4	Overall architecture of the proposed video-based pose estimation network. The proposed network is based on the encoder-decoder network architecture and extended with the multi-scale TCE module. The multi-scale TCE model fully explores the bidirectional temporal consistency information at multi-scale spatial levels. In this way, our model can generate temporally enhanced feature maps and obtain more precise human pose results.	28
2.5	Precision-recall curves of our method on the Sub-JHMDB and Penn datasets under different PCK thresholds.	34

2.6	Examples of pose estimation results on the Sub-JHMDB and Penn datasets. (row 1,2,3,4,5) Results of challenging samples (i.e., occlusion background and motion blur); (row 6,7,8,9) Results of persons with significant scale variations. Zoom-in for details.	35
2.7	Quantitative analysis of the smoothness of our results.	37
2.8	Visualization of the predicted offset fields.	38
2.9	Plots of the results with different numbers of neighboring frames and temporal stride values.	41
3.1	(a) The network trained with re-projection loss. (b) The proposed network trained with transform re-projection loss. (c) The comparisons of results estimated by the above two networks. The results are shown in two different views.	45
3.2	The overall architecture of our method that follows a two-stage pipeline.	49
3.3	The architecture of the proposed self-supervised training approach.	50
3.4	Illustration of the model pre-training.	53
3.5	Results of different variants in some hard examples.	59
3.6	(a) and (b) are the loss and MPJPE curves of the network trained without and with pre-training respectively.	60
3.7	Quantitative results of our method (trained on the H36M dataset) on the 3DHP dataset.	61
4.1	Architecture of the consistent factorization network.	73
4.2	The encoder-decoder network for hierarchical dictionary learning.	75
4.3	Visualization comparisons of our method versus baseline.	80

4.4	Visualization results of our method (trained on the Human3.6M dataset) on the MPI-INF-3DHP dataset.	82
-----	---	----

List of Tables

2.1	Comparisons with the state-of-the-art methods on Sub-JHMDB dataset using PCK@0.2.	32
2.2	Comparisons with the state-of-the-art methods on Penn dataset using PCK@0.2.	32
2.3	PCK@0.2 of different variants on the split 1 of Sub-JHMDB dataset.	36
2.4	PCK@0.2 of variants with different loss functions on the split1 of the Sub-JHMDB dataset.	39
2.5	Analysis of the random erasing and spatial pyramid on the split1 of the Sub-JHMDB dataset using PCK@0.2.	40
3.1	Detailed results on H36M dataset under Protocol #1 and Protocol #2.	58
3.2	Comparisons of different backbones on the H36M datasets.	60
3.3	Comparisons with recent weakly/self-supervised methods on the H36M dataset under evaluation Protocol #1 and Protocol #2.	63
3.4	Comparisons with recent weakly/self-supervised methods on the 3DHP dataset.	64
4.1	Per-action P-MPJPE of different variants on the Human3.6M dataset.	79
4.2	Comparisons with recent dictionary-based methods on the Human3.6M dataset.	81
4.3	Comparisons with recent weakly/self-supervised methods on the Human3.6M dataset.	83

4.4 Comparisons with recent weakly/self-supervised methods on the
MPI-INF-3DHP dataset. 84