

UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

**Exploring Region-based Deep Learning to
Understand Objects in Real-world Scenarios**

by

Ruiheng Zhang
Supervisor: A/Prof. Min Xu

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Sydney, Australia

2021

Certificate of Authorship/Originality

I, Ruiheng Zhang declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering, Faculty of Engineering 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. This document has not been submitted for qualifications at any other academic institution.

Also, 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, China.

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

Signature: Production Note:
 Signature removed
 prior to publication.

© Copyright 2021 Ruiheng Zhang

To my loving parents, my dear grandparents, my beloved wife and daughter.

ABSTRACT

Exploring Region-based Deep Learning to Understand Objects in Real-world Scenarios

by

Ruiheng Zhang

Supervisor: A/Prof. Min Xu

One way to infer about the real scenes is by understanding the object that presents in it, involving object localization, object recognition, object tracking, etc. Despite many advances in computer vision techniques, object understanding in real-world scenarios still remains many challenging tasks. There is no universal algorithm that can solve all of the scenarios with their own practical difficulties. This dissertation focuses on exploring region-based deep learning to understand objects in three typical real-world scenarios.

The first part of the dissertation studies facial landmark detection in the condition of lack of finely labeled training data. We generate weakly labeled training data to replace finely labeled data using generative adversarial networks. Then, we propose a region-based convolutional neural network to detect facial components and landmarks simultaneously. Notably, our approach can handle the situation when large occlusion areas occur, as we localize visible facial components before predicting corresponding landmarks. Extensive evaluations on several datasets indicate the effectiveness of the proposed approach.

In the second part, multi-player identification and tracking tasks in sports video are discussed. We build a robust multi-camera multi-player tracking with identification framework, from player detection, to identification, to tracking. To handle the identity switches, we design a distinguishable deep representation for player identity, considering pose-guided partial features, team class, and jersey number. For

data association, a robust multi-player tracker incorporating with player identity is further developed to produce identity-coherent trajectories. Experiment results illustrate that our framework handles the identity switches effectively, and outperforms state-of-the-art trackers on the sports video benchmarks.

Finally, we study vehicle detection in infrared images with poor texture information, low resolution and high noise levels. To deal with these difficulties, we propose a backbone network to exploit discriminative features, composing of a frequency feature extractor, a spatial feature extractor and a dual-domain feature resource allocation model. Hypercomplex Infrared Fourier Transform is developed to calculate the infrared intensity saliency, while a convolutional neural network is used to extract feature maps in the spatial domain. To efficiently integrate and recalibrate the frequency and spatial features, we propose a Resource Allocation model for Features based on the well-designed attention blocks. The experiments substantiate the merits of the proposed method through comparisons with state-of-the-art methods.

Dissertation directed by Associate Professor Min Xu
School of Electrical and Data Engineering

Acknowledgements

First and foremost, I want to extend my heartfelt gratitude to my supervisor, A/Prof. Min Xu. Her kind supervision, valuable suggestions and warm encouragement helped me to overcome difficulties and to successfully complete this thesis. She had empowered and inspired me to work on multiple research projects which led to publications in top journals and conferences. She will always be my most beloved supervisor. I have learned so much from her, not only how to be a successful scholar but also how to be a kind person in life. I also would like to appreciate my co-supervisor Dr. Xiaoying Kong for providing me with continuous support throughout my study and research.

Second, I wish to express my sincere appreciation to Prof. Lixin Xu and A/Prof. Chengpo Mu, my Dual-PhD supervisors at Beijing Institute of Technology. Although we are far away from each other, they gave me many suggestions and much help.

Third, I would like to pay tribute to Prof. Dacheng Tao, Prof. Dayong Jin, Prof. Linlin Ge, Prof. Sean He, Prof. Qiang Wu, Prof. Shenghong Hu, Dr. Yu Peng, Dr. Wenguan Wang, Dr. Yang Yu, and Dr. Zhengyu Yu for their inspiring words, continued support, and elaborate instructions about my presentations and academic reports. They had a significantly positive influence on me during my Ph.D. career.

Forth, I thank my fellow lab mates and my excellent friends: Tianrong Rao, Haimin Zhang, Lingxiang Wu, Wanneng Wu, Yukun Yang, Lei Sang, Zhongqin Wang, Xiaoxu Li, Qiyu Liao, Shuo Yang, Jiahao Xia, Caoyuan Li, Yang He, Yaxin Shi, Shilei Zhou, Huan Li, Wei Wei, Xiaofeng Xu, Qi Zhang, Zhibin Li, Yan Huang, Huidong Xu, Deyang Liu, Yuexi Zhang, Jiayan Qiu, Ye Shi, Yulong Sun and others that I cannot list them all for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the funs we have had.

I thank, School of Electrical and Data Engineering, Faculty of Engineering and IT, University of Technology Sydney for providing the infrastructures, computing power, and financial support.

Finally, I am grateful for my dear family, especially my father Bin Zhang, my mother Juan Li, my grandpa Wenhua Li, my grandma Yachun Liu, my father in law Shengchun Shi, my mother in law Conghua Lin, my uncle Donghai Zhang and Yunhai Ma, my brother Zhe Yuan and Zichao Ma, who had been supportive during my PhD candidature stages, researching and writing my thesis. I am sincerely grateful for everything and especially the opportunities given in this life, from being born to the completion of this Doctor of Philosophy thesis.

Special thanks to my wife Dr. Qiaolin Shi, who is always being with me. She was the one who encouraged me to grow as a real man and gave the pretty birth to my beloved daughter Yuxi Zhang.

Ruiheng Zhang
Sydney, Australia, 2020.

List of Publications

Journal Papers

- J-1. **R. Zhang**, L. Wu, Y. Yang, W. Wu, Y. Chen, M. Xu, ‘Multi-camera Multi-player Tracking with Deep Identification in Sports Video’, *Pattern Recognition*, Volume 102, 2020, 107260, ISSN 0031-3203.
- J-2. **R. Zhang**, C. Mu, M. Xu, L. Xu, Q. Shi, J. Wang, ‘Synthetic IR Image Refinement using Bidirectional Mappings with Adversarial Learning’, *IEEE Access*, 2019, 7: 153734-153750.
- J-3. **R. Zhang**, C. Mu, M. Xu, L. Xu, X. Xu, ‘Facial Component-Landmark Detection with Weakly-supervised LR-CNN’, *IEEE Access*, 2019, 7: 10263-10277.
- J-4. **R. Zhang**, L. Xu, Z. Yu, Y. Shi, C. Mu, M. Xu, ‘Deep-IRTarget: An Automatic Target Detector in Infrared Imagery using Dual-domain Feature Extraction and Allocation’, *IEEE Transactions on Multimedia*, 2021.
- J-5. Q. Liang, W. Wu, Y. Yang, **R. Zhang**, Y. Peng, M. Xu, ‘Multi-Player Tracking for Multi-View Sports Videos with Improved K-Shortest Path Algorithm’, *Applied Sciences*, 2020, 10, 864.
- J-6. Q. Zhang, **R. Zhang**, L. Ge, M. Xu, ‘A High Accuracy Burned Area Detection Framework Based on The Joint Processing of Sentinel-1&2 Data’, *Remote Sensing of Environment*. (under review)
- J-7. Y. Yang, **R. Zhang**, W. Wu, M. Xu, ‘3D Localization for Multiple Players in Multiview Sport Videos with Deep Identification Reasoning’, *Pattern Recognition*. (under review)

Conference Papers

- C-1. **R. Zhang**, M. Xu, Y. Shi, J. Fan, C. Mu, L. Xu, ‘Infrared Target Detection using Intensity Saliency and Self-Attention’, *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.
- C-2. Y. Yang, **R. Zhang**, W. Wu, Y. Peng, M. Xu, ‘Multi-camera sports players 3D localization with identification reasoning’, *2020 25 th IEEE International Conference on Pattern Recognition(ICPR)*. IEEE, 2020.
- C-3. Y. Yang, M. Xu, W. Wu, **R. Zhang**, and Y. Peng, ‘3D Multiview Basketball Players Detection and Localization Based on Probabilistic Occupancy’, *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2018.

Contents

Certificate	ii
Abstract	iv
Acknowledgments	vi
List of Publications	viii
List of Figures	xiv
List of Tables	xviii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Summary of Contributions	4
1.3 Organization	5
2 Literature Review	7
2.1 Deep Learning-based Object Detection	7
2.2 Facial Component and Landmark Detection	10
2.3 Multi-object Identification and Tracking in Sports Video	11
2.4 Infrared Target Detection and Recognition	14
3 Facial Component and Landmark Detection With Weakly-Supervised LR-CNN	16
3.1 Introduction	16
3.2 Data Preprocessing and Augmentation	19

3.2.1	Weakly-supervised Training Data	21
3.2.2	Fully-supervised Training Data	23
3.3	Weakly-supervised LR-CNN Framework	24
3.3.1	Region-based Component Detection	24
3.3.2	Two Branch Landmark Detection	27
3.3.3	Weakly-supervised Learning and Loss	29
3.3.4	Training	30
3.4	Experiments and Discussions	31
3.4.1	Dataset and Evaluation Metrics	31
3.4.2	Comparison with Other State-of-the-art Methods	32
3.4.3	Ablation Study	39
3.5	Conclusion	45
4	Multi-camera Multi-player Tracking with Deep Player Identification	46
4.1	Introduction	46
4.2	DeepPlayer Model for Player Identification	50
4.2.1	Cascade Mask RCNN	51
4.2.2	Pose-guided Partial Feature Embedding	54
4.2.3	Player Identification	56
4.3	Individual POM with ID	57
4.3.1	3D Localization Formulation	58
4.3.2	IPOM with Identified ID	59
4.3.3	IPOM with Ambiguous ID	59
4.4	KSP-ID for Tracking	60

4.5	Experiments and Discussions	64
4.5.1	Dataset	64
4.5.2	Experiments Settings	64
4.5.3	Baselines	65
4.5.4	Evaluation Metrics	66
4.5.5	Results	67
4.5.6	Ablation Study	70
4.6	Conclusion	74
5	Infrared Target Detection using Dual-domain Feature Extraction and Allocation	75
5.1	Introduction	75
5.2	Overall of The Deep-IRTarget Framework	78
5.3	Dual-domain feature extraction	79
5.3.1	Infrared Saliency using Hypercomplex Infrared Fourier Transform in the Frequency Domain	80
5.3.2	CNN-based Feature Extraction in the Spatial Domain	85
5.4	Resource Allocation for Feature	85
5.4.1	CNN-based Feature Integration	86
5.4.2	Channel SE-Attention Block	87
5.4.3	Position SE-Attention Block	91
5.4.4	Allocated Feature	93
5.5	Detection Head	93
5.6	Experiments and Discussions	94
5.6.1	Implementation Details	94

5.6.2	Datasets	95
5.6.3	Baseline	95
5.6.4	Evaluation Metrics	97
5.6.5	Evaluation on MWIR Dataset	98
5.6.6	Evaluation on BITIR Dataset	101
5.6.7	Evaluation on WCIR Dataset	101
5.6.8	Discussions	102
5.6.9	Ablation Study	104
5.7	Conclusion	108
6	Conclusions	109
	Bibliography	111

List of Figures

1.1	In the era of big data, image, video and other media data show ‘explosive’ growth.	2
1.2	Deep learning technology promotes the development of vision application.	3
2.1	Object detection milestones.	7
2.2	The architecture of YOLO and SSD.	8
2.3	The architecture of the RCNN series.	9
3.1	The whole pipeline. Blue boxes and lines represent fully labeled data and fully-supervised learning processing, while red boxes and lines show weakly labeled data and weakly-supervised learning.	19
3.2	Data preprocessing and augmentation.	20
3.3	The proposed LR-CNN architecture. Blue lines represent fully-supervised process, and red lines show weakly-supervised process.	25
3.4	Cumulative error curves. The red line (Ours) is our weakly-supervised method and the green one (Ours(non-weakly)) is our LR-CNN without generated weakly labeled data.	36

3.5	Some detection results on Helen testset (the first row), LFPW testset (the second row) and 300-W testset (the rest row). The different color bounding boxed show facial component detection and the text and number pairs denote the probabilities of bounding boxes belong to the corresponding categories. The predicted landmark coordinates are plotted by different color points corresponding their component.	37
3.6	Some detection results of our method on extreme illuminations and occlusions in first row. The second row shows the result of LDDR. . .	39
3.7	Left image is weakly-supervised result, right image is non-weakly-supervised result.	40
4.1	An example of identity switch in player tracking. The identities of Player 11, Player 14 and Player 15 exchange when players interchange.	47
4.2	The overall framework consists of three modules: (1) the DeepPlayer model for players' 2D localization, instance segmentation and identification, (2) the IPOM model to localize players' 3D coordinates with ID, (3) the KSP-ID model for the ID-enhanced multi-player tracking.	48
4.3	The architecture of the DeepPlayer model. This model consists of two part: (1) the Cascade Mask RCNN for coarse-grained player detection(Cascade Mask-RCNN-P) and fine-grained jersey number recognition(Cascade Mask-RCNN-J); (2) the player mask embedding into the deep representation using PoseID. Finally, the player identity is decided by the jersey number class, the team class and the deep representation.	51
4.4	PoseBox construction. Given a mask, the player pose is estimated by OpenPose. PoseBox1 = torso + arms + legs; PoseBox2 = head + torso + arms + legs; PoseBox3 = head + arms + legs.	56

4.5	Overview of the IPOM model. The input includes player’s segmentations and IDs. The 3D localization model processes the players with identified ID and the players with ambiguous ID, followed by a post-process with a threshold. The output is the final 3D localization results.	58
4.6	A simplified flow system for a directed graph. The yellow and blue color represent two different players. Given a depth of one, a player occupying the location i at time t can arrive at one of the three neighbor locations at time $t + 1$. The weight $e_{i,j}^t$ of edge depend on not only the marginal posterior probability ρ_i^t of the presence of the player, but also the proposed player ID correlation coefficient $\varrho_{i,j}^t$ of the node pairs.	61
4.7	Illustrative tracking results on the APIDIS dataset. Different color indicates different team, and the number indicates the identity across frames.	68
4.8	Illustration of the APIDIS dataset in Camera 3 and 6 indicate that the proposed method is able to avoid identity switch among Player 11, Player 14, and Player 15 in the dashed box.	70
4.9	Illustration of our tracking results on the STU dataset in Camera 1, 4 and 6. Red and yellow color boxes indicate black and white teams respectively. The blue dotted boxes show that our method avoids identity switch between Player 6 of the white team and Player 11 of the black team.	71
4.10	Results of the three types of PoseBoxes on the APIDIS and STU dataset. PoseBox1 = torso + arms + legs; PoseBox2 = head + torso + arms + legs; PoseBox3 = head + arms + legs.	73

5.1	The architecture of the Deep-IRTarget framework mainly consists three parts, (1) Dual-domain Feature Extraction(DFE) in frequency and spatial domain; (2)Resource Allocation for Feature(RAF) with channel-wise and position-wise attention; (3) Detection head based on the Region Proposal Network (RPN).	78
5.2	The pipeline of the Hypercomplex Infrared Fourier Transform.	80
5.3	The details of the Channel SE-Attention block and the Position SE-Attention block are illustrated in (a) and (b). Different colored lines represent different streams.	88
5.4	An illustration of the detection results on MWIR, BITIR and WCIR dataset, comparing with other methods. The first group of two rows is related to the MWIR dataset, the second group of two rows is related to the BITIR dataset, the last six to the WCIR dataset. The first column is the ground truth of the samples. Different colored bounding boxes mean different classes.	99
5.5	The performance of the Deep-IRTarget with HFT, PFT, and HIFT on MWIR, BITIR and WCIR.	106
5.6	The performance of the Deep-IRTarget with DANet, SENet, and RAF on MWIR, BITIR, and WCIR.	107

List of Tables

3.1	The landmark indexes and the vector length of each facial component in the 68-landmark annotation.	21
3.2	The component detection result (AP) compared with other methods.	33
3.3	The landmark detection results (average error distance) compared with other methods, on Helen, LFPW and 300-W test set separately.	34
3.4	The component detection result (AP) compared with other methods.	41
3.5	The 51-landmark detection result (average error distance) compared with our methods with variants on different test set.	42
3.6	Detection results of our algorithm on Helen test set using different settings of anchors. The network is ResNet-50.	43
3.7	Detection results with various RoI layers.	43
3.8	Landmark detection results with various architectures.	44
4.1	Training parameters of the DeepPlayer model in three steps.	65
4.2	The characteristics of the datasets and the corresponding parameters.	65
4.3	Quantitative comparison results of the proposed method with state-of-the-art trackers on the APIDIS dataset.	69
4.4	Ablation experiments on the APIDIS dataset.	72
4.5	Ablation experiments on the STU dataset.	72
4.6	Ablation studies on player identification on the STU dataset.	73

5.1	The architecture of CNN-Spatial	86
5.2	The architecture of CNN-Fuse	87
5.3	The performance of our method and the baselines on the MWIR dataset.	100
5.4	The results on the BITIR dataset.	102
5.5	The results on the WCIR dataset.	103
5.6	The ablation study of our method.	105