

Visual Object Tracking



Zijing Chen

Faculty of Engineering and Information Technology
University of Technology Sydney

A thesis submitted for the degree of

Doctor of Philosophy

2018

Certificate of Original Authorship

I, Zijing Chen, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

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

This thesis is the result of a research candidature conducted with another University as part of a collaborative Doctoral degree.

This document has not been submitted for qualifications at any other academic institution.

Production Note:

Signature: Signature removed prior to publication.

Date: 29/10/2018

Acknowledgements

I would like to express my gratitude to all those who helped me finish my doctoral studies.

First and foremost, I want to extend my heartfelt gratitude to my supervisor, Professor Dacheng Tao. His kind supervision, valuable suggestions and warm encouragement helped me to overcome difficulties and to successfully complete this thesis. His foresight and sagacity inspired me to find inspiration in cutting-edge research, which made the process of research a great joy. He deserves to be admired, not only for his great achievements in artificial intelligence, computer vision and machine learning, but for his diligence and courage which inspires others to expend a greater effort on their research.

Second, I would like to thank my co-supervisor, Dr Jun Li, who helped me pursue my passion for research. With incredible talent, he is full of inspiring and effective ideas which helped me survive the tedious and burdensome research tasks. He is extremely patient and is generous with his time, helping me to make gradual progress.

Third, I wish to express my sincere appreciation to Prof. Xinge You. He is a Dual-PhD supervisor at Huazhong University of Science and Technology. Although we are far away from each other, Prof. You gave me many suggestions and ideas via emails and numerous phone calls.

Fourth, I would like to pay tribute to Prof. Sean He, Dr Ling Chen, Dr Wei Liu, and Dr Guoqiang Zhang 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.

I would like to thank my excellent collaborators. Dr Ruxin Wang and Dr Qiang Li helped me both in my research work and life in Sydney. Dr Xiubao Jiang, Dr Weihua Ou, Dr Hasan Abdulwahid gave me many tips to help me find solutions, especially in solving complicated mathematical problems related to my research. Mr Huan Fu and Mr Baosheng Yu guided me to the right path on deep learning. I also worked closely with Mr Boxuan Zhong who is always optimistic and cheerful.

Special thanks to my husband Zhe Chen, who helped me to overcome depression, insomnia and pessimism. He was the one who encouraged me to grow as an individual and helped me to replace insecurity with confidence.

Finally yet importantly, I would like to express my special thanks to my parents. I really appreciate their support and the confidence they have placed in me.

Abstract

Visual object tracking is a critical task in many computer-vision-related applications, such as surveillance and robotics. If the tracking target is provided in the first frame of a video, the tracker will predict the location and the shape of the target in the following frames. Despite the significant research effort that has been dedicated to this area for several years, this field remains challenging due to a number of issues, such as occlusion, shape variation and drifting, all of which adversely affect the performance of a tracking algorithm.

This research focuses on incorporating the spatial and temporal context to tackle the challenging issues related to developing robust trackers. The spatial context is what surrounds a given object and the temporal context is what has been observed in the recent past at the same location. In particular, by considering the relationship between the target and its surroundings, the spatial context information helps the tracker to better distinguish the target from the background, especially when it suffers from scale change, shape variation, occlusion, and background clutter. Meanwhile, the temporal contextual cues are beneficial for building a stable appearance representation for the target, which enables the tracker to be robust against occlusion and drifting.

In this regard, we attempt to develop effective methods that take advantage of the spatial and temporal context to improve the tracking algorithms. Our proposed methods can benefit three kinds of mainstream tracking frameworks, namely the template-based generative tracking framework, the pixel-wise tracking framework and the tracking-by-detection framework. For the template-based generative

tracking framework, a novel template based tracker is proposed that enhances the existing appearance model of the target by introducing mask templates. In particular, mask templates store the temporal context represented by the frame difference in various time scales, and other templates encode the spatial context. Then, using pixel-wise analytic tools which provide richer details, which naturally accommodates tracking tasks, a finer and more accurate tracker is proposed. It makes use of two convolutional neural networks to capture both the spatial and temporal context. Lastly, for a visual tracker with a tracking-by-detection strategy, we propose an effective and efficient module that can improve the quality of the candidate windows sampled to identify the target. By utilizing the context around the object, our proposed module is able to refine the location and dimension of each candidate window, thus helping the tracker better focus on the target object.

Contents

Contents	viii
List of Figures	x
List of Tables	xvi
1 Introduction	1
1.1 Background	1
1.2 Motivation of This Study	5
1.3 Summary of Contributions	7
1.4 Publications Related to the Thesis	8
2 Literature Review	10
2.1 Generative Tracking Methods	11
2.2 Discriminative Tracking Methods	15
2.2.1 Traditional Methods	15
2.2.2 Convolutional Neural Network-based Methods	19
2.3 Other Related Research Areas	25
3 Template-Based Tracking with Sparse Representation	29
3.1 Introduction	29
3.2 Sparse Tracking with Mask templates	30
3.3 Performance boosting	34
3.4 Dynamically Modulated MST with Efficient Solver	36
3.5 Experimental Results	39
3.6 Conclusion	51

CONTENTS

4	A Robust Tracker Based on a Bi-channel Fully Convolutional Neural Network	54
4.1	Introduction	54
4.2	Generic Pixel Level Tracker	55
4.3	Experiment	60
4.4	Conclusion	62
5	Learn to Focus on Objects for Tracking-by-Detection	66
5.1	Introduction	66
5.2	The Transformation Model	70
5.2.1	TRM with hand-crafted features	72
5.2.2	TRM with CNN features	77
5.3	Experiments	80
5.3.1	Compared With Traditional Proposal Methods	81
5.3.2	Compared With Regional Proposal Network	90
5.3.3	Improvement On Tracking	93
5.4	Conclusion	97
6	Conclusions	102
	References	105

List of Figures

1.1	An example of object states in online object tracking. The first row presents a few frames starting from time $t = 0$, and the second row lists the ground truth states of the target correspondingly. The states in the form of a bounding box are illustrated with rectangles, while the states in the form of binary masks are illustrated in white areas. Either the red rectangle or the white mask inside represents the initial state of the target. Correspondingly, the yellow rectangles or the related masks are the desired output of a tracker in each new frame.	2
1.2	Example of challenging issues in object tracking. The target in the first row suffers from scale change and shape variation. The target in the second row suffers heavy occlusions, illumination changes and cluttered background surroundings.	4
1.3	Spatial and temporal context. The patches (denoted by orange circles) give spatial context for surrounding spatial positions. For temporal context, the context patches of the neighbouring time frames of the sequence in the same spatial location are collected. .	6
3.1	Frame differences with a various number of interval frames (e.g. m_2 and m_5) describe multiple scales of the evolution of the motion of corruptions over a period of time.	32
3.2	Establishing mask templates based on temporal context. The mask template can capture the on-going changes in the target area. . .	32
3.3	Tracking results of different methods on parts of the selected sequences.	42

LIST OF FIGURES

3.4	Center location error for each test sequence. The result of MMST, which is marked by red lines, has the lower error rate on average for the test sequences.	45
3.5	Overlap rate for each test sequence. The result of MMST is marked by red lines and has higher overlap rate on average.	46
3.6	Statistic results of center location error of all trackers.	52
3.7	Statistic results of overlap rate of all trackers.	53
4.1	The processing flow of the bi-channel fully convolution neural network. Based on the input information, low-level and high-level temporal information are extracted and analysed in corresponding branches. By fusing the results of two branches, the foreground area of the target can be identified.	56
4.2	The working flow of the low-level branch: the optical flow data is extracted by a fully convolutional neural network with a clustering operation afterwards, so that foreground and background areas can be separated.	57
4.3	The working flow of the high-level branch. It adopts the fully convolutional neural network to predict the decrease and increase (red and blue) of the foreground mask of the target. By adding the predictions to the previous foreground mask, an initial estimation of the target can be obtained.	57
4.4	Architecture of CNN of the semantic branch. We add batch normalization to the five convolutional layers adapted from FCN. Five up-sampling operations are applied to make the final output the same shape as the input image.	60
4.5	Qualitative comparison among trackers. Our output is marked in red shadow. The result of the other trackers are shown by bounding boxes.	63
4.6	ROC, from the beginning to the 60% of a video sequence. Our output is shown by black lines marked with stars. The rest of the other trackers are shown by curves in color.	64

LIST OF FIGURES

4.7	ROC, from the 60% to the end of a video sequence. Our output is shown by black lines marked with stars. The rest of the other trackers are shown by curves in color.	65
5.1	Focus on objects with transformation models. The transformation model transforms partially aligned original proposals and focus them on objects.	71
5.2	The working flow of improving proposals based on hand-crafted features. Original proposals have been generated on the input image. With the processing of the translation model, the locations of these proposals are adjusted to better align the target. Then with the deformation model, the scale and shape of the proposal are amended to better focus on the object.	73
5.3	Area arrangement of feature extraction for one proposal. Above figures show how the characters of the context around and within an original proposal (represented by the shadowed area marked as ‘O’ in (a)) are extracted. As in (a), surrounding areas are organized in a grid. These grid cells can be combined to generate surrounding features. For example, if we use braces to represent a combination of cell areas, then features can be extracted from {A, B, C, D, O, E} and {F, G, H} separately. In addition, the proposal area, O, is partitioned in two passes, one horizontally and the other vertically, generating internal horizontal (b) and vertical (c) features.	74
5.4	Explore the context in multiple spatial ranges. Three kinds of kernels are utilized here. The bar-shaped convolution kernels in (a) and (b) are designed to extract vertical patterns and horizontal patterns separately. The area kernel in (c) is used to explore the pattern of surrounding areas. Besides, to analyze the context information in multiple spatial ranges, the Atrous convolution is applied on these kernels.	79

LIST OF FIGURES

5.5	Focus proposal net. With the iterative self-adapting block, the transformation parameters predicted at step k will be combined with parameters obtained at step $k - 1$ which helps the proposal to gradually focus on the target.	79
5.6	The working flow of object detector with FoPN. The detector takes an image as the input and uses a base network to generate feature maps which preserved key information of the target object. Then the proposed FoPN processes with the feature map, and generate refined proposals. These proposals are sent to the detection part (Fast R-CNN), and output the detection result.	81
5.7	Effect of TRM on object coverage by object proposals: visual assessments. Each test image is illustrated with six sub-figures that arranged in three columns and corresponds to 3 stages (from left to right) separately. They are random boxes as initial proposals, adjusted proposals after applying the transformation model, and further refined proposals with the deformation model. The real-life images in the first row show the 5 top-ranked proposals out of all proposals in each stage. The second row shows the degree of coverage by the entire set of proposals in the corresponding stage, where brighter colours are for higher levels of coverage, i.e. when a pixel is included in more windows within the proposal set, its colour will be brighter.	84
5.8	Effect of TRM on object coverage by object proposals with failure cases. The man driving the car is missed because his size is too small compared to the car; two cats are identified as one object because one of them is severely occluded by the other.	85
5.9	Statistical analysis of object coverage by proposals with TRM. The figure shows the statistical analysis of overlap variations of random proposals (blue), translated proposals (green) and deformed proposals (red). By comparing the hists, it is obvious that TRM increase the number of proposals with high overlaps and reduce the number of proposals with low overlaps.	85

LIST OF FIGURES

5.10	Improvements on mean overlap scores (w.r.t. ground truth) when applying TRM on popular object proposal algorithms. The means of K highest overlap scores achieved by original proposals and corresponding transformed proposals are respectively plotted by solid and dashed lines when K varies from 10 to 1000. For each tested proposal generator, it has been shown that the dashed line is always above the solid line, which proves that better coverages on objects can be obtained if proposals are refined by TRM.	87
5.11	Comparison between mean overlaps and a required number of proposals. The transformed proposals (in dashed lines) could achieve the same overlap scores with less number of proposals.	88
5.12	This box plot shows the comparisons in the numbers of proposals to achieve similar performance for using original proposals and transformed ones. The boxes in the same column share a similar range of mean overlap of top 100 proposals. Refined proposals sourced from different proposal generators are coloured differently. For Selective Search, to achieve the similar mean overlap score of top 100 windows when 400 ~ 600 original proposals are generated in total (as the yellow area marks), only 180 ~ 250 transformed proposals (marked by the blue bar) are required for using TRM to improve the proposals. Thus a fewer number of transformed proposals is required to achieve the similar performance.	89
5.13	Performance comparison between the proposed FoPN and RPN. The red bounding boxes are results of the proposed method while the blue dashed bounding boxes are from RPN.	92
5.14	The recall rates of the proposed proposal generation method (TRM) compared to the Region Proposal Network (RPN) [133]. The recall rates of the compared methods are evaluated with top 10, 50 and 100 proposals.	93
5.15	The tracking result on <i>CarScale</i> . The results of the baseline method are marked in black. The results of RPN are in blue and ours TRMT are in red (best viewed in colour).	95

LIST OF FIGURES

5.16	The tracking result on <i>Couple</i> . The results of the baseline method are marked in black. The results of RPN are in blue and ours TRMT are in red (best viewed in colour).	96
5.17	The tracking result on <i>Jogging</i> . The results of the baseline method are marked in black. The results of RPN are in blue and ours TRMT are in red (best viewed in colour).	97
5.18	The tracking result on <i>Human2</i> . The results of the baseline method are marked in black. The results of RPN are in blue and ours TRMT are in red (best viewed in colour).	98
5.19	The overlap rate on <i>CarScale</i>	98
5.20	The overlap rate on <i>Couple</i>	99
5.21	The overlap rate on <i>Human2</i>	99
5.22	The overlap rate on <i>Jogging</i>	100

List of Tables

3.1	Center location error (in pixels). The bold and italic numbers indicate the best and the second-best respectively.	48
3.2	Successful rate (in pixels). The bold and italic numbers indicate the best and the second-best respectively.	49
3.3	Speed (fps). Bold fonts indicate the best performance algorithm.	49
5.1	Detection score for using different numbers of proposals generated by RPN and the proposed method. The detection is performed using the Fast RCNN method. Best scores for each category and final performance are illustrated in bold.	93
5.2	Mean overlap rate. Best scores are illustrated in bold.	96
5.3	Mean frame rate per second.	97