

Robust Online Visual Tracking



Zhibin Hong

Faculty of Engineering and Information Technology

University of Technology, Sydney

A thesis submitted for the degree of

Doctor of Philosophy

2015

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Student: Zhibin Hong

Date: 22/09/2015

I would like to dedicate this thesis to my loving parents
Liduan Lin and Cheng Hong

Acknowledgements

I would like to take this good opportunity to appreciate my advisors, several professors, my colleagues, my friends and my family for their significant help during my doctoral study in University of Technology (UTS), Sydney, Australia.

First of all, I would like to express my sincere appreciation and deep gratitude to my advisor supervisor **Prof. Dacheng Tao**. He is the one who led me to the field of academic research. He always gives me plenty of freedom to explore and timely constructive suggestions to help me out of difficulties. I can always benefit and learn a lot from various detailed discussions with him. It is hard to imagine I could have finished this thesis without his high scientific standards, unlimited patience, generous support, constant encouragement and guidance. I also wish to express my sincere appreciation to **Prof. Lianwen Jin** who was my advisor when I was in South China University of Technology for master study. I am so grateful for his generosity, guidance and support. He is the one so important to my career and life track. I would not have come to the field of computer vision and would not have come to UTS and met Prof. Dacheng Tao without him. I am also deeply indebted to **Dr. Xue Mei** who led me to the field of visual tracking. His expertise, guidance and encouragement significantly helped me for the completion of this thesis and kept me away from many detours during the journey of exploration. I sincerely appreciate him for his valuable time for beneficial discussions, constructive suggestions and timely support. I would also like to express my appreciation to **Dr. Chaohui Wang** as a close mentor, collaborator and elder brother, for his kindly support and timely help. In addition, I appreciate Prof. Dacheng Tao, Prof. Lianwen Jin,

Dr. Xue Mei and Dr. Chaohui Wang for their guidance, suggestions and help to my future career in the final stage of PhD study.

I have been fortunate to work in UTS and Centre for Quantum Computation and Intelligent Systems (QCIS). I wish to express my appreciation to Prof. Chengqi Zhang, for his full support for me to attend those top conferences including ECCV, ICCV and CVPR, and for his amazing work of gathering so many brilliant researchers in QCIS. I am grateful to Prof. Xingquan Zhu, Prof. Massimo Piccardi, Prof. Maolin Huang, Dr. Danil Prokhorov, Dr. Lin Chen, Dr. Jun Li for their help during my doctoral study. Moreover, I also want to give special thanks to my excellent collaborators: Jia Wu, Zhe Xu, Zhe Chen, for their brilliant work and timely support, and to my dear colleagues and friends I met in QCIS: Prof. Weifeng Liu, Dr. Wei Bian, Dr. Tianyi Zhou, Dr. Naiyang Guan, Dr. Nannan Wang, Dr. Meng Fang, Dr. Mingsong Mao, Dr. Hongshu Chen, Dr. Shirui Pan, Mingming Gong, Tongliang Liu, Changxing Ding, Maoying Qiao, Dr. Guodong Long, Dr. Jing Jiang, Barbara Munday, Chunyang Liu, Dr. Shengzheng Wang, Dr. Yong Luo, Lianyang Ma, Xiaoyan Li, Fei Gao, Jie Gui, Bozhong Liu, Prof. Bo Du, Dianshuang Wu, Dr. Ting Guo, Dr. Lianhua Chi, Kailing Guo, Ruxin Wang, Qiang Li, Weilong Hou, Shaoli Huang, Chang Xu, Chen Gong, Sujuan Hou, Haishuang Wang, Qin Zhang, Yali Du, Zhongwen Xu, Dr. Yi Yang, Hao Xiong, Prof. Shigang Liu, Prof. Xianhua Ben, Prof. Wankou Yang, Xiyu Yu, Jiang Bian, Guoliang Kang, Liu Liu, for the inspiring discussions, kind support and companionship.

I am also grateful to all the other friends: Zhida Guo, Qihao Chen, Xueshen Xian, Shixin Lin, Hepeng Ling, Yingying Li, Ming Chen, Sheng Dai, YingLi Li, Ruifang Chen, Shi Chen, Jiawei Xuan, Yanhua Zhang, Shanlin Ye, Xueying Chen, Tao Wang, Weiqi Tang, Chunyu Wang, Ang Li, Qihan Zhao, Danyan Lei, for their support and company during both joyful and stressful times.

Finally, I would like to express my deeply felt gratitude to my fam-

ily: my parents, my grandparents, my lovely brother, my uncles and aunties, my cousins, for their endless love, encouragement and full support throughout my study and life.

Abstract

Visual tracking plays a key role in many computer vision systems. In this thesis, we study online visual object tracking and try to tackle challenges that present in practical tracking scenarios. Motivated by different challenges, several robust online visual trackers have been developed by taking advantage of advanced techniques from machine learning and computer vision.

In particular, we propose a robust distracter-resistant tracking approach by learning a discriminative metric to handle distracter problem. The proposed metric is elaborately designed for the tracking problem by forming a margin objective function which systematically includes distance margin maximization, reconstruction error constraint, and similarity propagation techniques. The distance metric obtained helps to preserve the most discriminative information to separate the target from distracters while ensuring the stability of the optimal metric.

To handle background clutter problem and achieve better tracking performance, we develop a tracker using an approximate Least Absolute Deviation (LAD)-based multi-task multi-view sparse learning method to enjoy robustness of LAD and take advantage of multiple types of visual features. The proposed method is integrated in a particle filter framework where learning the sparse representation for each view of a single particle is regarded as an individual task. The underlying relationship between tasks across different views and different particles is jointly exploited in a unified robust multi-task formulation based on LAD. In addition, to capture the frequently emerging outlier tasks, we decompose the representation matrix to two collaborative components which enable a more robust and accurate approximation.

In addition, a hierarchical appearance representation model is proposed for non-rigid object tracking, based on a graphical model that exploits shared information across multiple quantization levels. The tracker aims to find the most possible position of the target by jointly classifying the pixels and superpixels and obtaining the best configuration across all levels. The motion of the bounding box is taken into consideration, while Online Random Forests are used to provide pixel- and superpixel-level quantizations and progressively updated on-the-fly.

Finally, inspired by the well-known Atkinson-Shiffrin Memory Model, we propose MUlti-Store Tracker, a dual-component approach consisting of short- and long-term memory stores to process target appearance memories. A powerful and efficient Integrated Correlation Filter is employed in the short-term store for short-term tracking. The integrated long-term component, which is based on keypoint matching-tracking and RANSAC estimation, can interact with the long-term memory and provide additional information for output control.

Contents

Contents	viii
List of Figures	xi
List of Tables	xvii
1 Introduction	1
1.1 Background	1
1.2 Literature survey of Online Visual Tracking	4
1.2.1 Generative Methods	5
1.2.2 Discriminative Methods	11
1.3 Summary of Contributions	14
1.4 Publications Related to this Thesis	17
2 Distracter-Resistant Tracking via Dual-Force Metric Learning	18
2.1 Introduction	19
2.2 Related Work	22
2.3 Particle Filter	25
2.4 Dual-Force Metric Learning	26
2.4.1 Dual-Force Formulation	27
2.4.2 Reconstruction Error Constraint	31
2.5 Distracter-Resistant Tracker	33
2.5.1 L1 Minimization Tracking	33
2.5.2 Distracter-Resistant Tracker	34
2.6 Experiments	35

2.6.1	Qualitative Comparison	36
2.6.2	Quantitative Comparison	38
2.7	Conclusion	40
3	Robust Tracking via Multi-Task Multi-View Joint Representation	41
3.1	Introduction	41
3.2	Related Work	45
3.3	Multi-task Multi-view Sparse Tracker	47
3.3.1	Sparse Representation-based Tracker	48
3.3.2	Robust Multi-task Multi-view Sparse Learning with Least Absolute Deviation	48
3.3.3	The General Form and Special Cases	52
3.3.4	Optimization with Approximated Least Absolute Deviation	53
3.3.5	Outlier Rejection	57
3.3.6	Tracking using Robust Multi-task Multi-view Sparse Representation	58
3.3.7	Template update	58
3.4	Experiments	59
3.4.1	Implementation Details	61
3.4.2	Evaluation on Publicly Available Sequences	62
3.4.3	Evaluation on Noisy Sequences	66
3.4.4	Evaluation on CVPR2013 OOTB	69
3.4.5	Evaluation on ALOV++ Dataset	74
3.4.6	Discussion	75
3.5	Conclusion	76
4	Non-rigid Object Tracking using Multilevel Quantizations	79
4.1	Introduction	80
4.2	Related Work	82
4.3	Tracking with Multilevel Quantizations	84
4.3.1	Multilevel Quantizations Model	84
4.3.2	Online Color-Texture Forests	87

4.3.3	ORF Training and Occlusion Handling	89
4.4	Experiments	91
4.4.1	Implementation Details	91
4.4.2	Tracking non-rigid Objects	93
4.4.3	Evaluation on CVPR2013 OOTB	94
4.5	Conclusions and Future Work	96
5	Robust Tracking using Multi-store Memory Model	98
5.1	Introduction	99
5.2	Related Work	101
5.3	The Proposed Multi-store Tracker	103
5.3.1	Short-term Integrated Correlation Filters	104
5.3.2	Short-term Processing of Keypoints	106
5.3.3	Long-term Memory Updates	110
5.3.4	Output and Short-term Memory Refreshing	112
5.4	Experiments	112
5.4.1	Evaluation on CVPR2013 OOTB	113
5.4.2	Evaluation on ALOV++ Dataset	115
5.5	Conclusion	116
6	Conclusions	118
	References	121

List of Figures

1.1	An example of online visual object tracking discussed in this thesis. Given a bounding box of an object in the first frame of a video, the task of a tracker is to locate the target in subsequent video frames.	2
1.2	Some examples of challenges in online visual object tracking.	3
1.3	An example of target templates and trivial templates used in L1 tracker [131]. Original figure is from (Mei & Ling 2009) [131].	7
2.1	Some examples of distracters. Red windows highlight the tracking objects, while green windows are the areas which have the greatest similarity to the objects (calculated by Euclidean distance) and can easily be locked onto.	20
2.2	Mutual effects of negative and positive samples in different methods: (a) Typical generative trackers, which only focus on the positive space. Some patches from the background that are similar to the target may be selected. (b) Most discriminative methods, which only create mutual effects between positive and negative space. Some negative samples are pushed away from positive samples during optimization, while some hard negative samples may be still near the positive space. (c) Our metric. Distances between positive and negative samples are maximized. Similarity is propagated in the negative space and therefore connections are built between samples to effectively separate positive and negative samples.	24

2.3	Figures to illustrate the effect of similarity propagation. (a) Without negative samples, the tracker locks onto the distracter (highlighted by dashed rectangle window). (b) to (h) Iteration of similarity propagation. The distracter is linked by other negative samples when convergence is reached. The cyan points denote the center locations of negative samples.	28
2.4	Qualitative results of DFMLDR compared with different algorithms. Frame numbers are shown in the top left of each figure. Note that (D) contains the results of two sequences, i.e. the first row for Tiger1 and the second row for Tiger 2.	37
2.5	Position error (in pixel) plot of each tracker on eight tested sequences for quantitative comparison.	39
3.1	Flowchart to illustrate the proposed multi-task multi-view tracking framework.	44
3.2	Illustration for the structure of the learned coefficient matrices \mathbf{P} and \mathbf{Q} , where entries of different color represent different learned values, and the white entries in \mathbf{P} and \mathbf{Q} indicate the zero rows and columns. Note that this figure demonstrates a case that includes four particles and three views, where the second particle is an outlier whose coefficients in \mathbf{Q} comprise nonzero values.	49
3.3	A schematic example of the learned coefficients. We visualize the learned coefficient matrices \mathbf{P} and \mathbf{Q} for all particles across all views, which are color histograms, intensity, HOG and LBP, respectively. Each matrix consists of four column parts corresponding to four different views, where the brighter color represents larger value in the corresponding matrix element. The seventh template in the dictionary is the most representative (which is circled in green in the shown intensity templates $\mathbf{D}^{(2)}$) and results in brighter values in the seventh row of \mathbf{P} across all views (they are associated by the line with two arrows), while some columns in \mathbf{Q} have brighter values which indicate the presence of outliers.	51

3.4	Examples of detected outliers. The green bounding boxes denote the outliers and the red bounding box denotes the tracked target. The outliers are detected out of 400 sampled particles. There are two outliers in the left frame and six outliers in the right frame.	58
3.5	Qualitative results of MTMVTLS and MTMVTLAD compared to different algorithms. Frame indexes are shown in the top left of each figure.	60
3.6	Qualitative results of MTMVTLS and MTMVTLAD compared to different algorithms. Frame indexes are shown in the top left of each figure.	64
3.7	Some examples of the contaminated sequences.	67
3.8	Qualitative results of MTMVTLS and MTMVTLAD compared to different algorithms on <i>EXTsequences</i> . Frame indexes are shown in the top left of each figure.	70
3.9	Precision plots and success plots on the CVPR2013 tracking benchmark. The values appearing in the legend of the precision plot are the precision scores in the threshold of 20, while the ones in Success plots are the AUC scores. Only the top 10 trackers are presented, while the other trackers can be found in [181]. The trackers appearing in the legend are as follows: Struck [68], SCM [198], TLD [86], LRT [197], VTD [98], VTS [99], CXT [45], CSK [74], ASLA [84].	72
3.10	The success plots for BC and DEF subsets of CVPR2013 tracking benchmark. The value appearing in the title is the number of sequences in the specific subset. The values appearing in the legend are the AUC scores. Only the top 10 trackers are presented, while the other trackers can be found in [181]. The trackers appearing in the legend are as follows: DFT [105], LSK [119], CPF [147].	73
3.11	The success plots of MTMVTLAD and its baseline variants on the CVPR2013 tracking benchmark. The values appearing in the legend are the AUC scores.	74

3.12	The survival curves for top ten trackers in ALOV++ dataset. The average F -scores over all sequences are specified in the legend. The trackers appearing in the legend are as follows: Struck [68], FBT [141], VTS [99], TLD [86], L1O [135], NCC [31], MIL [9], L1T [133], IVT [150]	76
3.13	The respective average F-scores of the proposed MTMVTLAD tracker in 14 ALOV++ challenge subsets.	77
3.14	Failure cases of MTMVTLAD. (a) Failure cases on <i>Skiing</i> and <i>MotorRolling</i> sequences of CVPR2013 benchmark. (b) Failure case in the <i>LongDuration</i> subset of ALOV++ dataset. The numbers appear on the top of each bar is the tracker’s average F -score over 10 sequences of the <i>LongDuration</i> subset.	77
4.1	Illustration of the structure of the proposed hierarchical appearance representation model (left) and a practical example (right). In the proposed framework, a node in the Conditional Random Field (CRF) models each pixel, superpixel, and bounding box. At the pixel level, each pixel receives a measurement from a Random Forest and connects to the corresponding superpixel at the middle level. At the superpixel level, each superpixel also obtains a probability output by another Random Forest and suggests the pixels within the same superpixel to share the same label. At the bounding box level, different candidate bounding boxes (green) are considered, and the best position (red) with the best configuration is found. (a) shows the tracking result (in red bounding box) at Frame #226 in the <i>Basketball</i> sequence. (b) displays the superpixelization of the image. (c) and (d) are the output of the pixel-level RF and final labeling result, respectively, while (e) and (f) are the output of the superpixel-level RF and final labeling result.	80

4.2	Occlusion handling on the <i>Jogging</i> sequence. The index is specified in the top-left of each frame, and the two figures between each frame are the corresponding outputs of the pixel-level RFs and labels x_i , respectively. The occlusion is detected from the Frame #049, from which point the RFs stop updating until the target moves out of occlusion.	91
4.3	Qualitative results of MQT on the non-rigid object tracking dataset. Frame numbers are shown in the top left of each figure. Each column contains results of three sequences: (A) Cliff-dive 1, Cliff-dive 2, Mountain-bike; (B) Diving, High Jump, Gym.	94
4.4	Qualitative results of MQT compared to different trackers on the CVPR2013 benchmark. Only top five trackers on success plots are presented. Frame numbers are shown in the top left of each figure. Each column contains results of two sequences: (A) Basketball, David3; (B) David, Tiger1.	94
4.5	Quantitative comparison on CVPR2013 benchmark. The performance score for each tracker is shown in the legend. For each figure, only the top 10 trackers are presented. The trackers appearing in the legend are as follows: MQT (ours), Struck [68], SCM [198], TLD [88], VTD [98], VTS [99], CXT [45], CSK [74], ASLA [84], LOT [144], LSK [119].	96
4.6	Success plots for some challenge subsets of CVPR2013 tracking benchmark. The performance score for each tracker is shown in the legend. The value appears in the title is the number of sequences in that subset. Only the top 10 trackers are presented. The trackers appearing in the legend are as follows: OAB [62], TM-V [39], DFT [105], CPF [147], MIL [9].	97
5.1	The Atkinson-Shiffrin memory model represented by an illustrative neural network. Nodes and their connections inside the short- and long-term stores represent the possible structure of the neural network inside the human brain.	100

5.2	System flowchart of the proposed tracker based on the Atkinson-Shiffrin Memory Model. The short-term processing in short-term store is conducted by an ICF via two-stage filtering. Another set of short-term procedures including keypoint matching, keypoint tracking and RANSAC estimation is conducted by a conservative long-term component in short-term store, and it is able to interact with the long-term memory located in the long-term store. Both the results of short-term processing and long short-term processing are obtained by a controller, which decides the final output and the ICF update.	102
5.3	Tracking results of selected algorithms in representative frames. Frame indexes are shown in the top left of each figure. The showing examples are from sequences <i>Couple</i> , <i>Lemming</i> , <i>Jumping</i> , <i>Jogging</i> , <i>Singer2</i> , <i>Bolt</i> , respectively.	113
5.4	Quantitative comparison on CVPR2013 OOTB. The performance score for each tracker is shown in the legend. For each figure, only the top 10 trackers are presented.	114
5.5	Survival curves for top ten trackers on AIOV++ dataset. The average F -scores over all sequences are specified in the legend.	116

List of Tables

2.1	Average position error (pixels). Bold indicates best performance.	40
3.1	Average overlap & success rates (percentages)	65
3.2	Parameters of synthetic dataset	68
3.3	Average success rates in the contaminated datasets.	68
3.4	Average overlap & success rates (percentages)	69
4.1	Non-rigid object tracking: percentage of correctly tracked frames.	92