Robust multiple people tracking and re-identification in cluttered environment

Zhou, Zhi

2016

Zhou, Z. (2016). Robust multiple people tracking and re-identification in cluttered environment. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/68933>

<https://doi.org/10.32657/10356/68933>

Robust Multiple People Tracking and Re-identification in Cluttered Environment

Zhou Zhi

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirements for the degree of
Doctor of Philosophy

2015

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Associate Professor Teoh Eam Khwang for his invaluable guidance and continuous support throughout my four years of PhD study. I have learnt a lot from his technical insights, literary and organizational expertise which would be great treasure in my future work.

I also wish to express my grateful to my co-supervisor Dr Wang Yue, currently a scientist at the Institute for Infocomm Research (I2R), A*STAR. He has supported me greatly with brilliant ideas, wise advice and enthusiastic discussion. This research work would not be possible without his guidance.

Many Thanks to the staffs at Internet of Things Lab (formerly known as Computer Vision Lab), Mr Sow Peck Heng and Mr Yuen Sien Huan, for their technical support during the four years. I thank all my friends in the Lab, Huang Guangpu, Lian Zhichao, Shao Zhifei, Mohammad Ghahramani, Sai Phyo Kyaw, Seyed Mohammad Hassan Anvar and Seyedehsamaneh Shojaeilangari, for their kind help and sharing of research experiences. Many thanks to all my friends in Internet of Things Lab, Xu Jinming, Cheng Zhenghua, Zhao Rui, Manoj Ramanathan, Zhu Qingchang, Zhang Yong, Wu Xiaoping, Zhou Hongming, Bai Zuo, Wang Dongze, Ding Jie, Li Xuhao, Ren Ye and Sumit Bam Shrestha, with them I spend many great times.

I would also like express my thanks to my roommates, Wu Kai, Weng Renliang, Qi Yongbin, Zhou Dexiang, and all my friends in NTU and Singapore. Thank them all for sharing the wonderful days in the past four years with me.

I'd like to express my special thanks to my parents for their constant support, encouragement and love in my life. They are always the source of my determination and

perseverance in pursuing all my endeavors. I am forever indebted to them.

Finally, I would like to acknowledge the School of Electrical and Electronic Engineering, Nanyang Technological University, for providing the award of a scholarship and the research facilities for my work in NTU.

Summary

As the computational ability develops in computers, there has been an increasing interest to detect, label and track people by computer vision algorithms in surveillance systems. To liberate human operators from the tedious task watching over a set of monitors, as well as improving the efficiency of surveillance system, an automated system that can process tasks automatically is desired. Computer vision techniques can be used in many tasks in surveillance such as abnormal event detection, searching and tracking of specific target (for example, a criminal suspect), activity analysis and so on. Though developed for decades, there are still some challenges remain unsolved. Occlusion is one of the essential challenges, including partial occlusion, full occlusion and mutual occlusion between multiple targets. Besides, building a model to identify a target is challenging as well, especially under various conditions such as illumination change, various view angles and pose change. The objective of this research is to develop methods that can help dealing with occlusion conditions in people tracking and extracting appropriate descriptor from the target.

Basically there are three major contributions presented in this thesis. In the first contribution, a key-points based single target tracking method is proposed. The proposed method aims to solve the partial occlusion problem during tracking in a cluttered environment. An object model and a surrounding background model are used simultaneously to store key-points extracted from the target and surrounding background, respectively. SURF key-points are located, they are further evaluated and on-line learned by Random Ferns. These labeled key-points can be used to improve the tracking performance as well as the learning in Random Ferns. Tracking of targets under partial occlusion depends on the locating of key-points belonging to the target. Long-term tracking is achieved by combining detection and tracking. Experiments have been conducted on videos where targets are under different occlusion conditions. The proposed method shows superior performance while under occlusion.

In the second contribution of this thesis, a salient parts based framework follows tracking-by-detection method is proposed to enhance the performance on multi-people tracking. As the saliency of an object or area is the quality to stand out from its neighborhood, it is an important component of the human vision system. In order to solve occlusion caused by background object and mutual occlusion between targets, salient parts inside the target along with salient parts around the target are extracted to assist tracking. Salient parts inside the target are considered as representative parts, while salient parts around target can be used as context information. Short-term tracking of salient parts is applied when targets fail to associate with detections. Tracking of target can be improved with a supporting model indicating the spatial relationship between salient parts and the target. The supporting model is on-line updated. To evaluate the performance, extensive experiments have been conducted on several public datasets like Town Center, PETS09 and TUD, which contain complicate conditions including mutual occlusion and trajectory intersection. A comparative study with other multi-people tracking methods is provided and the result shows the promising performance of the proposed framework.

The third contribution of the thesis is to study descriptor of targets for people re-identification in real-time tasks. As biometric information is not available in applications where the camera resolution is low, this research exploits semantic features in constructing the people descriptor. A framework is proposed to extract descriptors in real scene. Semantic feature based description is more invariant to changing factors than directly using low level features such as color and texture. The ensemble of a set of semantic features can achieve good performance in people re-identification. Selected semantic features are extracted from certain parts of human body. All features are then combined to form the final descriptor. This framework is further used in experiment for people re-identification on two static images based datasets and one video based dataset. The result shows the superior performance of the proposed method on varying conditions like view of angles. Besides, Open ID recognition is also included in the experiment.

In summary, two frameworks and one object tracking method are proposed in this research work. The key-points based object tracking method can track the target when it is under partial occlusion. By combining detection, it can achieve long-term tracking. In the first framework, a salient part based multi-people tracking method is proposed. By tracking salient parts inside and around the target patch,

problems such as partial occlusion and mutual occlusion can be dealt with. In the second framework, people descriptor with a set of semantic features is extracted from surveillance video. It can be used to assist the construction of human model and target re-identification in surveillance system. Experimental results show promising performance of proposed frameworks.

Contents

Acknowledgements	i
Summary	iii
List of Figures	xi
List of Tables	xix
List of Abbreviations	xxii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Challenges	2
1.3 Objective	3
1.3.1 Developing a Key-points Based Method for Long-term Object Detection and Tracking	3
1.3.2 Developing an Framework for Multi-people Tracking	4
1.3.3 Exploit People Descriptor Extraction and Its Application	4
1.4 Our Contributions	4
1.5 Organization of the Thesis	6

2	A Literature Review	8
2.1	Introduction	8
2.2	Single Target Tracking	8
2.2.1	Frame to Frame Tracking	9
2.2.2	Tracking by Matching with Template	10
2.2.3	Tracking by On-line Updating Model	11
2.2.4	Object Tracking Combined with Object Detection	14
2.2.5	Salient Based Tracking	16
2.3	Human Detection	17
2.3.1	Histogram of Oriented Gradient	17
2.3.2	Deformable Part Model	19
2.4	Multi-people Tracking	20
2.4.1	Tracking by Detection	20
2.4.2	Optimization of Trajectory Connection	22
2.4.3	Tracking by Body Parts	25
2.5	People Re-identification	27
2.5.1	Model with Low Level Features	27
2.5.2	Re-identification with Supervised Learning	29
2.5.3	Semantic Feature Based Re-identification	30
2.6	Concluding Remarks	32
3	Key-points Based Single Target Tracking	34
3.1	Introduction	34
3.2	Key-points Based Object Tracking	35
3.2.1	Overview of the Method	35
3.2.2	Tracking by Detection	36
3.2.3	Evaluation and On-line Learning of Key-points	37

3.2.4	Object Model and Surrounding Background Model	40
3.2.5	Position Adjustment After MF	42
3.2.6	Adaptive Bounding Box to Scale Change	43
3.3	Experimental Result and Discussion	43
3.3.1	Occlusion Evaluation	43
3.3.2	Comparison with Other Methods	44
3.3.3	Discussion	45
3.4	Concluding Remarks	46
4	Salient Parts Based Multi-people Tracking	47
4.1	Introduction	47
4.2	Multi-people Tracking with Salient Parts	48
4.2.1	Overview	48
4.2.2	Tracking Model Description	50
4.2.3	Tracking Process	50
4.2.4	Salient Part Extraction	55
4.2.5	Single Layer Salient Parts	56
4.2.6	Double Layer Salient Parts	57
4.2.7	Occlusion Reasoning	59
4.2.8	Supporting Model	60
4.2.9	ID Re-identification	61
4.3	Datasets and Evaluation Criterion	62
4.3.1	Datasets	63
4.3.2	Evaluation Criterion	64
4.4	Experimental Results and Discussion	65
4.4.1	Parameter Settings	65
4.4.2	Evaluation on ID Re-identification	65
4.4.3	Qualitative Analysis	67
4.4.4	Comparison on Tracking Performance Evaluation with Other Methods	68
4.5	Concluding Remarks	73

5	People Descriptor for Re-identification	74
5.1	Introduction	74
5.2	A Framework for People Descriptor Extraction	75
5.2.1	Overview	75
5.2.2	Pre-processing	77
5.2.3	Body Segmentation	79
5.2.4	Semantic Feature Extraction	81
5.2.5	Combination of Semantic Features and Re-identification	84
5.3	Datasets and Evaluation Criterion	86
5.3.1	Datasets	86
5.3.2	Evaluation Criterion	87
5.4	Experimental Results	87
5.4.1	Performance Evaluation on Static Image Datasets	88
5.4.2	Single Frame vs Multi-frames for Human Appearance Model Construction	91
5.4.3	Evaluation of the Performance on Different View Angles	93
5.4.4	Comparison with Other Methods on Multi-views	103
5.4.5	Evaluation on Open ID Recognition	104
5.5	Concluding Remarks	111
6	Conclusion and Recommendations for Future Work	112
6.1	Conclusion	112
6.2	Recommendations for Future Works	113
6.2.1	Exploit Advanced Method for Salient Part Extraction	113
6.2.2	Salient Part Extraction with Adaptive Shape	114
6.2.3	Extension of Semantic Feature Extraction	114
6.2.4	Multi-layer People Descriptor	114
6.2.5	Application of People Descriptor	115

Appendix A	116
Appendix B	122
Author's Publications	126
Bibliography	140

List of Figures

2.1	On-line boosting for tracking (extracted from [9]).	12
2.2	Multiple instance learning (extracted from [10]).	13
2.3	Using Random Ferns to classify patches (extracted from [11]).	14
2.4	Use supporters to estimate the location of target (extracted from [12]).	15
2.5	Context Tracker with assistance of supporters (extracted from [13]). .	16
2.6	Extraction of HOG descriptor (extracted from [3]).	18
2.7	Example of HOG descriptor.	18
2.8	Pyramid of Deformable Part Model (extracted from [44]).	19
2.9	Part filters with constrained locations in Deformable Part Model. . .	20
2.10	Pictorial Structure (extracted from [48]).	21
2.11	Movements performed by MCMCDA (extracted from [51]).	22
2.12	Optimization across all frames to form trajectories (extracted from [16]).	23
2.13	Components of energy function (extracted from [19]).	24
2.14	Movements in achieving minimum energy function (extracted from [19]).	24
2.15	Online learning appearance model (extracted from [55]).	24
2.16	Discriminative part-based appearance models (extracted from [21]). .	26
2.17	Multi-part Multi-people tracking (extracted from [20]).	26
2.18	Models built in Multi-part Multi-people tracking (extracted from [20]).	26

2.19	Symmetry-driven accumulation of local features (extracted from [67]).	28
2.20	Extraction of Mean Riemannian Covariance (MRC) patches (extracted from [76]).	29
2.21	Patch matching (extracted from [76]).	29
2.22	Soft-biometric traits (extracted from [90]).	31
2.23	Segmentation of 7 horizontal regions for attributes extraction (extracted from [23]).	32
3.1	The overview of the key-points based tracking method.	35
3.2	Flow chart of Median Flow (extracted from [29]).	36
3.3	Detection after tracking failure. The tracker gradually drifts away from frame 157. It loses the target at frame 169. But the target is found by the detector at frame 176.	38
3.4	Patch comparison between target patch and patch with occlusion. . .	38
3.5	Transform from decision tree to fern.	38
3.6	Fern extracted from key-point.	39
3.7	Key-points stored in object model and background model. Red points are points with object label, while blue points are points with background label. In (a), one point is out of the bounding box, but is still recognized as object point. In (b), some background points in the bounding box are detected.	42
3.8	Position Adjustment. The first image shows the location after MF tracking, the second images shows candidates detected by key-points models, and the third image shows the target location determined by candidates and the tracking result of MF.	43
3.9	Results in different occlusion conditions. First row is under minor occlusion, second row is under normal occlusion, the last row is under severe occlusion.	44
3.10	Results of comparison with other methods. (a) and (e) are the results of TLD, (b) and (f) are results of Context Tracker, (c) and (g) are results of MIL, (d) and (h) are results of the proposed method. . . .	45

4.1	The overview of the salient parts based multi-people tracking method.	49
4.2	False Positive detections from DPM.	51
4.3	Salient parts determination during the transition from Candidates to Tracker. The first four columns are from frames of candidates, while the last column is the first frame of the tracker. Only those salient parts that constantly appear in candidate frames are kept in the tracker model, while occasional appeared salient parts are discarded.	52
4.4	Comparison of tracking parts and tracking the whole body. The yellow bounding box in the right frame is the tracking result of Median Flow on the whole body area, while the red bounding box is the target location indicated by the tracking result of green part.	53
4.5	Examples of salient parts tracking.	54
4.6	Flow chart of the saliency detection.	56
4.7	Examples of inner salient parts detection. (a) is the original image. (b) and (c) are saliency maps formed by color and orientation respectively. (d) is the result of detected salient parts, where yellow boxes denote color salient part and green boxes denote orientation salient part.	57
4.8	Examples of two layers salient parts detection. (a) is the original image. (b) and (c) are saliency maps formed by color and orientation respectively. (d) is the result of detected salient parts. Red box shows the inner layer and black box shows the outer layer. Yellow boxes denote color salient parts and green boxes denote orientation salient parts.	58
4.9	Examples of mutual occlusion.	59
4.10	Occlusion reasoning. For each target model, appearance histogram in head, torso and leg areas are recorded. The overlapped area is divided into different parts, occlusion reasoning is performed based on posterior probability.	59
4.11	Supporting model. Target location is indicated by tracking results of salient parts when detection association fails.	60
4.12	Example of Structural Local Sparse Appearance Model. The human patch is divided into 3 by 3 blocks with overlap, sparse code is generated in each block and concatenated to form the sparse code of the target.	61

4.13	Example frames of sequences.	63
4.14	Examples of ID re-identification of the proposed method. Left: results from method without ID re-identification (2L_SPT). Right: results of method with ID re-identification (2L_SPT+IDRI). First columns show the tracking before track failure, IDs are shown along with the target. The second columns show the track failure and third columns show the detection of the candidate after track failure. The last columns show the result when the target is tracked again. In the method with re-identification, it can be observed that the same IDs (in the first column) can be resumed in the last column.	66
4.15	Comparison between DPM detection, single layer salient parts tracking (1L_SPT) and double layer salient parts tracking (2L_SPT). The first row are the results from DPM detection, second row are the results from 1L_SPT, while the last row are the results from the proposed method.	68
5.1	Overview of the proposed framework	76
5.2	Examples of the preprocessing work. Original images are shown in row 1. Row 2 are results of background subtraction, the target is shown in red bounding box. Row 3 are masks of cropped bodies. Rows 4 and 5 are masks and images of body after rotation correction	78
5.3	Examples of body segmentation. First row are human body masks, while second row are upper and lower body segmentation results.	80
5.4	Examples of body segmentation with half body visible.	80
5.5	Examples of hair segmentation in head area.	81
5.6	Examples of soft-biometric traits in head area.	82
5.7	HSV color space.	83
5.8	Examples of skin detection.	83
5.9	Examples of traits extraction in torso and leg areas. For the torso area, first row of texts denotes the clothes color, second row denotes sleeve length, third row denotes clothes pattern, forth row denotes whether carrying bag. For the leg area, first row of texts denotes pant color, second row denotes pant length.	84

5.10	Scenarios of surveillance cameras in SAIVT-SoftBio database.	86
5.11	Comparison on CMC curves between the proposed method and other methods on VIPeR dataset.	88
5.12	Comparison on SRR curves between the proposed method and other methods on VIPeR dataset.	89
5.13	Comparison on CMC curves between the proposed method and other methods on iLIDS dataset.	90
5.14	Comparison on SRR curves between the proposed method and other methods on iLIDS dataset.	90
5.15	CMC curves of single frame based descriptor and multi-frame based descriptor.	92
5.16	SRR curves of single frame based descriptor and multi-frame based descriptor.	92
5.17	Severe illumination conditions. The first two images are from one target and the last two are from another target. As can be seen, target in camera 5 suffers bright lighting condition, while target in camera 4 suffers darkness.	94
5.18	CMC curves of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views (Camera 3 vs Camera 8).	94
5.19	SRR curves of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views (Camera 3 vs Camera 8).	95
5.20	CMC curves of proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).	96
5.21	SRR curves of proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).	96
5.22	CMC curves of multiple soft-biometric traits based descriptor and single feature based descriptors with opposite views (Camera 1 vs Camera 3,8).	97
5.23	SRR curves of multiple soft-biometric traits based descriptor and single feature based descriptors with opposite views (Camera 1 vs Camera 3,8).	98

5.24	CMC curves of proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).	99
5.25	SRR curves of proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).	99
5.26	CMC curves of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views (Camera 3,8 vs Camera 5,6,7).	100
5.27	SRR curves of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views (Camera 3,8 vs Camera 5,6,7).	101
5.28	CMC curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 3,8 vs Camera 5,6,7).	102
5.29	SRR curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 3,8 vs Camera 5,6,7).	102
5.30	CMC curves of proposed method, SDALF, RDC and SDC with multiple views.	103
5.31	SRR curves of proposed method, SDALF, RDC and SDC with multiple views.	104
5.32	CMC curves of proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).	106
5.33	RDFP curves of proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).	106
5.34	RDFAR curves of proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).	107
5.35	CMC curves of proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).	107
5.36	RDFP curves of proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).	108
5.37	RDFAR curves of proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).	108
5.38	CMC curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).	109

5.39	RDFP curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).	109
5.40	RDFAR curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).	110
A.1	CMC curves of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views (Camera 5 vs Camera 7).	116
A.2	SRR curves of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views (Camera 5 vs Camera 7).	117
A.3	CMC curves of proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).	118
A.4	SRR curves of proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).	118
A.5	CMC curves of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views (Camera 4 vs Camera 5,6,7).	119
A.6	SRR curves of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views (Camera 4 vs Camera 5,6,7).	119
A.7	CMC curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).	120
A.8	SRR curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).	121
B.1	CMC curves of proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).	122
B.2	RDFP curves of proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).	123
B.3	RDFAR curves of proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).	123
B.4	CMC curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 1 vs Camera 5,6,7).	124

B.5	RDFP curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 1 vs Camera 5,6,7).	124
B.6	RDFAR curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 1 vs Camera 5,6,7).	125

List of Tables

3.1	Numbers of correct tracking frames in test videos.	45
4.1	Comparison on tracker numbers. The closest numbers to the ground truth are shown in bold.	67
4.2	Comparison on Town Center dataset. The best is shown in bold, the second best is underlined and the third best is in italic	69
4.3	Comparison on PETS09 S2L1. The best is shown in bold, the second best is underlined and the third best is in italic	70
4.4	Comparison on PETS09 S2L2. The best is shown in bold, the second best is underlined and the third best is in italic	71
4.5	Comparison on PETS09 S2L3. The best is shown in bold, the second best is underlined and the third best is in italic	71
4.6	Comparison on TUD Stadtmitte. The best is shown in bold, the second best is underlined and the third best is in italic	72
5.1	Comparison of AUCs (in percentage) on VIPeR dataset, the best is shown in bold.	89
5.2	Comparison of recognition rate (in percentage) in lower rank scores on VIPeR dataset, the best is shown in bold.	89
5.3	Comparison of AUCs (in percentage) on iLIDS dataset, the best is shown in bold.	91
5.4	Comparison of recognition rate (in percentage) in lower rank scores on iLIDS dataset, the best is shown in bold.	91
5.5	AUC of single frame based descriptor and multi-frame based descriptor.	92

5.6	View differences between each pair of cameras in SAIVT-SoftBio database, where s represents similar views, o represents opposite views and d represents dissimilar views.	93
5.7	AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views (Camera 3 vs Camera 8).	95
5.8	AUCs of proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).	97
5.9	AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with opposite views (Camera 1 vs Camera 3,8).	98
5.10	AUCs of proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).	99
5.11	AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views (Camera 3,8 vs Camera 5,6,7).	101
5.12	AUCs of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 3,8 vs Camera 5,6,7).	101
5.13	AUCs of proposed method, SDALF, RDC and SDC with multiple views.	104
5.14	Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).	105
5.15	Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).	105
5.16	Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).	106
5.17	Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).	107
5.18	Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).	108
5.19	Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).	109

A.1	AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views (Camera 5 vs Camera 7).	117
A.2	AUCs of proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).	117
A.3	AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views (Camera 4 vs Camera 5,6,7).	120
A.4	AUCs of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).	120
B.1	Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).	122
B.2	Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).	123
B.3	Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with dissimilar views (Camera 1 vs Camera 5,6,7).	125
B.4	Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with dissimilar views (Camera 1 vs Camera 5,6,7).	125

List of Abbreviations

AdaBoost	Adaptive Boosting
AUC	Area under the Curve
DPM	Deformable Part Model
CMC	Cumulative Matching Characteristic
ELF	Ensemble of Localized Features
HOG	Histogram of Oriented Gradients
HSV	Hue-Saturation-Value
ID	Identity
MF	Median Flow
MIL	Multiple Instance Learning
MOTA	Multiple Object Tracking Accuracy
MOTF	Multiple Object Tracking F-measure
MOTP	Multiple Object Tracking Precision
NN	Nearest Neighbor
OF	Optical Flow
PCA	Principle Component Analysis
RDC	Relative Distance Comparison
RDFAR	Ratio of Detection rate to False Acceptance rate
RDFP	Ratio of Detection rate to False Positive rate
RF	Random Ferns
SDALF	Symmetry Driven Accumulation of Local Features
SDC	Saliency Distance Comparison

SLSAM	Structural Local Sparse Appearance Model
SRR	Synthetic Recognition Rate
SURF	Speeded Up Robust Features
SVM	Support Vector Machines
TLD	Tracking-Learning-Detection

Chapter 1

Introduction

1.1 Background and Motivation

After decades of development, computer based systems have been broadly applied in our daily life. Nowadays, computer vision techniques are able to conduct tasks which were done by human, such as monitoring in industry manufacturing or in surveillance of public areas. There has been an increasing interest to introduce computer vision algorithms to the surveillance system. In traditional surveillance system, human resources were employed to watch over dozens of monitors, that would be costing as well as inefficient. Therefore, intelligent system which is able to automatically conduct surveillance tasks is highly desired. The detection and tracking of people is crucial in surveillance tasks. For example, if the system is monitoring a forbidden region, any people detected in the camera must be tracked and identified in case stranger is intruding. Another example is the detection of abnormality or abnormal actions of individuals in public areas. To detect abnormality, the movement of crowd should be tracked. It will be notified if any abnormal movement such as riot of the crowd happens. Or in the public area, people are detected and their actions are analyzed. If actions are recognized as dangerous, the system can keep tracking of these people and informs security forces.

In order to make surveillance system more intelligent, in this thesis, we aim to develop methods that can provide robust tracking of multi-people in cluttered environment, as well as extracting robust and discriminative people descriptors for target identification.

1.2 Challenges

Though studied for decades, human detection and tracking still face many challenges. One major problem in human detection and tracking are the selection of features and the construction of target model. Better performance as well as more efficient computation could be provided by selecting appropriate features. Many features have been used for object detection and tracking, these features are based on points [1, 2] or regions [3, 4, 5]. Most of existing features exploit raw information like color, orientation and gradient. They are the low level descriptions of a particular point or area. Since these features focus on local information in a small area, they are not invariant to external changes and inevitably face challenges from varying illumination, object appearance changes, distortion and occlusion.

Another challenge is to build an appropriate model for the target. Fixed models [6, 7] were used at early stages of study in human detection and tracking. Since the model is taken from the first frame, the tracking will soon fail due to factors like varying object appearance and environment. In some applications, fixed models are trained with supervised learning methods. But these methods demand a lot of samples and training needs to be done beforehand. Even though, models still perform poorly facing these challenges. Adaptive models [8, 9, 10] were proposed to solve the changing appearance of target model. However, “drift” problem happens. The target model gradually drifts away from the true target during the learning progress. Because on-line learning is based on the samples extracted during the tracking period, samples extracted are not guaranteed to be correct. False positive samples are likely to be included when training samples are collected during the tracking. Therefore, as tracking continues, more and more false positive samples participated in the on-line learning of object model, the model gradually drifts away from the ground truth. Recently, detection is combined with tracking methods [11, 12, 13]. In this way, both fixed and adaptive models are provided. Detection and tracking are performed simultaneously. Tracking by detection methods outperforms those with only fixed or adaptive models and achieves long-term tracking. However, they still face challenges under occlusion conditions, where detection and tracking may both fail.

For the tracking of multiple targets, more problems raise besides those in single target tracking. These extra challenges include mutual occlusion between different targets, ID switch caused by trajectory intersections and so on. Some of existing works depend on short-term tracking methods like Kalman filtering and data association with detections [14, 15]. These methods could be affected by occlusions, where

the appearance changes a lot and makes tracking and detection less reliable. Some methods consider multi-people tracking as an optimization problem [16, 17, 18, 19]. Detections need to be provided across all frames beforehand. Then by considering costs like spatial locations, scale change, appearance change and so on, an overall energy function is minimized to obtain the optimal trajectories. While receiving good performance on conditions like intersected trajectories and mutual occlusion, these methods need the information across a batch of frames. Therefore, future information is demanded in these methods to predict target location in current frame. There were also methods tracking people based on body parts [20, 21] proposed, especially under partial occlusion. But they define body parts as fixed areas of the target patch, appearance of body parts varies drastically during tracking.

1.3 Objective

The objective of this research is to develop framework for robust multi-people tracking, aiming to deal with problems like occlusion. Besides, method to extract people descriptor from real scene is exploited. This has been applied to solve target re-identification problem.

1.3.1 Developing a Key-points Based Method for Long-term Object Detection and Tracking

In existing long-term tracking methods, detection and tracking are both exploited. However, they face occlusion problem because of the feature selected. Most existing methods select features in patches to build the target model. Patch samples were on-line learned during the tracking period. But the appearances of patches change easily when targets are under occlusion, which leads to the failure of detection and tracking.

This research intends to develop a method which can solve the occlusion problem and improve the performance based on existing long-term tracking methods. Key-points are extracted from both the target and surrounding background. Random Ferns are employed to learn these key-points and are further used for classification of object key-points and background key-points. By excluding background key-points from updating the target model, model drift could be prevented.

1.3.2 Developing an Framework for Multi-people Tracking

Existing methods on multi-people tracking can be divided into two categories. One is focus on the frame to frame tracking along with detection. Locations of people in current frame are decided by models constructed from past frames, no future information is used. Most of these methods face the problem of partial occlusion. The other methods consider multi-people tracking as optimization problem over a period of time. With the assistant of human detection in each frame, Dynamic Programming methods or Linear Programming methods are used to associate detections in different frames and form consecutive tracklets, aiming to reach the maximum probability or minimum cost. However, these methods rely on a batch of frames which may cause time delay in application.

In this research work, we propose a salient parts based multi-people tracking framework. With salient parts extracted inside and around the target, supporting model is used to locate the target when partial occlusion happens or detection association fails. Besides, ID identification is performed to reduce break trajectories.

1.3.3 Exploit People Descriptor Extraction and Its Application

In most of the existing people re-identification methods, descriptors are based on low level features which describe color, orientation or gradient of certain area. Low level features are good at describing local information and thus good at discriminating different targets. However, they are not invariant to changes of factors like illumination, view angle and people pose. Recently, attributes [22, 23, 24] based descriptors receive more and more attention, and they show competitive performance.

In this research work, we exploited the extraction of human descriptors in a real time system. A descriptor can be obtained by combining a set of semantic features, which is extracted by simple methods designed specifically. The application of extracted descriptors in people re-identification problem is also exploited.

1.4 Our Contributions

The contributions of this research work are listed in this section.

(1) For the first contribution, we aim to solve the partial occlusion problem during long-term tracking. Considering the poor performance of patch represented model in long-term tracking, we use key-points to represent the target to be tracked. The model is on-line learned by Random Ferns. Besides, object points and surrounding background points are used to improve the tracking accuracy and the learning in Random Ferns. Novelties in this framework are:

- Use key-points to represent the object for long-term tracking. Use Random Ferns to on-line learn these key-points.
- Both object and surrounding background models are proposed to store object points and temporary surrounding background points. They are further used to improve the tracking accuracy and the learning of Random Ferns.

Experiments have been carried out on videos with different occlusion conditions. The proposed method shows satisfying performance dealing with occlusions.

(2) In the second contribution, we extend the idea in the first part to solve partial occlusion and mutual occlusion problems in multi-people tracking. Instead of using key-points, salient parts are employed as few key-points can be extracted for low resolution targets. These salient parts are used as representative parts and context information to indicate the location of targets. Novelties in this framework are:

- Introduce salient parts into people tracking. Salient parts are discriminative parts that can be representative for the target.
- Two layers (inner layer and outer layer) of salient parts are extracted. The outer layer salient parts are extracted around the target and used as context information.
- Supporting model is employed along with short-term tracking of salient parts to locate the target. The supporting model is on-line updated.

Extensive experiments have been conducted on public datasets like Town Center, PETS09 and TUD. The experimental results show that it is effective using salient parts in multi-people tracking. The proposed framework achieves superior performance in comparison with other multi-people tracking methods.

(3) In the third contribution of the research work, we exploited semantic features for the extraction of people descriptor. This framework aims to describe a human appearance for the people identification in a disjoint camera surveillance system. And its application in people re-identification is exploited.

- Semantic feature based descriptor is extracted from the video with a set of efficient methods.
- Semantic feature based people descriptor is used to solve people re-identification problem in cluttered environment.

Experiments have been conducted on static image based datasets as well as video based database. Various conditions on view of angles are evaluated. The proposed framework shows promising performance. Open ID recognition is also evaluated in the experiment.

1.5 Organization of the Thesis

The rest of the thesis is organized as follows.

Chapter 2 gives a literature review to related works. First, existing tracking methods on single target are introduced. The tracking is based on general object, but it can be applied for people tracking. It is followed by a review on methods in multi-people tracking. Besides, human detection methods are reviewed as it can provide assistance to people tracking. In addition to tracking methods, people description methods are introduced in Section 2.5. People descriptors are commonly used for people re-identification, but they can also be used in other applications. Finally, existing methods on people re-identification are addressed.

Chapter 3 describes the proposed method for single target tracking. The frame to frame tracking is performed along with target detected by Random Ferns. Object model and surrounding background model are employed to store key-points from the target as well as surrounding background. These key-points are on-line learned by Random Ferns to gain the confidence score. The Random Ferns is used for further target detections. The proposed framework aims to deal with partial occlusion happened in tracking. Experimental results are presented and discussion is provided.

Chapter 4 introduces the proposed framework for multi-people tracking. First, the overview of the framework is addressed. Then salient part detection is described.

Salient parts detected will be further used to assist multi-target tracking. Following are two models under the same framework. One is constructing the model based on salient parts inside the targets only. The other one exploits salient parts inside as well as in close neighborhood to the target. Then the supporting model is introduced, which provides spatial relationships between salient parts and targets by using salient parts as context information. A sparse appearance model is employed to perform ID re-identification to link different tracklets of the same target. Experimental results and discussion are provided as well.

Chapter 5 presents the research conducted on extraction of people descriptors in real scene. First, a framework proposed to extract semantic features based people descriptor is introduced. Experimental results in people re-identification along with discussion are provided in the following part.

Chapter 6 concludes the thesis by summarizing our contributions and provides some discussion on recommendations for future works.

Chapter 2

A Literature Review

2.1 Introduction

After decades of development, extensive researches have been carried out on object detection and tracking, especially for human target. In this chapter, a briefly review is conducted on existing works in single target tracking, human detection and multi-people tracking. Besides tracking, we also give a review on methods describing people in images and cameras, including attributes based method and other methods used in people re-identification.

2.2 Single Target Tracking

Though object tracking has been studied for decades by a lot of researchers [25], it still faces many problems due to conditions like cluttered background and moving camera. In this section, existing methods tracking single target are reviewed. Some methods focused on frame to frame tracking [26, 27]. These methods predict the location of the target in current frame based only on the information from the previous frame. They are efficient in short-term tracking, but the tracking will drift away from the target eventually without the assistance of appearance model. Some methods used template matching for tracking based on fixed appearance model. The tracking could be robust as long as the appearance of target remains unchanged. However, for non-rigid targets like human, the appearance varies during the tracking process. Thus to deal with varying appearance, some methods developed on-line learning for the target appearance model [8, 9, 10]. The problem for these methods

is that the target model on-line learned may drift away from the ground truth due to the involvement of unwanted negative samples in learning. Recent trend in object tracking is to combine detection and tracking [11, 12, 13]. An on-line learned model is used for tracking, but a fixed appearance model is remained for detection in case the tracking fails or tracking model drift away.

2.2.1 Frame to Frame Tracking

Most early methods exploit the temporal relationship between two consecutive frames to predict the motion. The tracking is performed by indicating the location in current frame using only the information relation between previous frame and current frame.

One of the classic frame to frame tracking methods called Optical Flow (OF) was proposed by Lucas and Kanade [26]. The assumption used in OF is that the flow remains constant in a local neighborhood.

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \quad (2.1)$$

Therefore, the least squares criterion can be used to solve the basic Optical Flow equations for all the pixels in a certain area. The moving vector along spatial coordinate V_x , V_y are obtained from the solution.

$$\frac{\partial I}{\partial x} V_x + \frac{\partial I}{\partial y} V_y + \frac{\partial I}{\partial t} = 0 \quad (2.2)$$

OF is simply used for a fixed size, so it can not deal with scale changes of target. Besides, it is not working well on uniform regions or in illumination changing scenario. Some variations of OF such as [28, 29] were developed to improve the short-term tracking performance.

Another classic method is Mean-shift tracker [27]. The mean-shift tracker is a density estimation based method that maximizes the similarity in a local area iteratively. The histogram of target model in the feature is computed as

$$\hat{q}_u = C \sum_{i=1}^n k(\|x_i^*\|^2) \delta[b(x_i^*) - u] \quad (2.3)$$

where C is a normalization factor and is calculated as

$$C = \frac{1}{\sum_{i=1}^n k(\|x_i^*\|^2)} \quad (2.4)$$

In each iteration, histograms of the target in last frame and the window around the object location in current frame are compared. Using the same kernel profile, but with bandwidth h , the histogram of target candidate is given by

$$\hat{p}_u(y) = C_h \sum_{i=1}^{n_h} k\left(\left\|\frac{y - x_i}{h}\right\|^2\right) \delta[b(x_i) - u] \quad (2.5)$$

$$C_h = \frac{1}{\sum_{i=1}^{n_h} k\left(\left\|\frac{y - x_i}{h}\right\|^2\right)} \quad (2.6)$$

Then for each pixel in the patch, a weight is assigned by

$$w_i = \sum_{u=1}^m \sqrt{\frac{\hat{q}_u}{\hat{p}_u(y_0)}} \delta[b(x_i) - u] \quad (2.7)$$

Pixels with feature corresponding to high frequency shown in previous frame will get a higher weight. Besides, a spatial kernel is used for weighting, so that pixels closer to the target centroid can get higher weights. Then the location after shift will be calculated as

$$\hat{y}_1 = \frac{\sum_{i=1}^{n_h} x_i w_i g\left(\left\|\frac{\hat{y}_0 - x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n_h} w_i g\left(\left\|\frac{\hat{y}_0 - x_i}{h}\right\|^2\right)} \quad (2.8)$$

The mean-shift vector is computed and points to the direction where the histogram similarity is increasing. This process is repeated until convergence is achieved.

2.2.2 Tracking by Matching with Template

Different with methods in section 2.2.1, some methods performed tracking by detecting the target with a fixed appearance model in each frame. Candidates are matched with the template in each frame and the one with the closest distance in feature space is picked as tracking result. The template can be set as the object in the first frame, or it can be learned by off-line training methods before tracking [30, 31].

However, off-line training needs the template model to be learned beforehand, which is not available for target without prior information.

In a sequence of images extracted from a video $I_n(x)$ where $x = (x, y)^T$ represents the pixel coordinates and $n = 0, 1, 2, \dots$ is the frame number. In tracking with template matching, model $T(x)$ is extracted from the first frame $I_0(x)$. The shape of the target can be any shape, for example, it could be a rectangular or ellipse region, it could also be a face shaped region. The goal is to find the best match to the template in each frame. Many kinds of distance can be used to measure the similarity between the model and candidate region. Take the sum of squared error as a measure example, the target location in a new frame is calculated as

$$p_n = \arg \min_p \sum_{x \in T} |I_n(W(x; p)) - T(x)| \quad (2.9)$$

The assumption behind template matching is that the appearance of the object keeps unchanged during the tracking process. This assumption may be reasonable for a short period of time. However, the tracking in the video faces changes from background such as illumination conditions, interaction between object and background, even the appearance of the object itself changes. Eventually the template can no-longer accurately represent the target.

2.2.3 Tracking by On-line Updating Model

It is not accurate to represent the target with a fixed model if the appearance of the target keep changing, especially for non-rigid targets. Therefore, some tracking methods focus on building an on-line updating model, so that the model can learn the change of object appearance gradually. On-line learning tracking methods start with an initial classifier, which is trained beforehand or extracted from the first frame. The classifier is used to evaluate unlabeled samples extracted in following frames. Then estimated labels are assigned to these samples and they are used as new training set to update the classifier.

Grabner [9] extended the famous learning method Adaptive Boosting (AdaBoost) [32] to on-line learning (Figure 2.1). Different with the off-line case, which uses all samples to train one weak classifier in each iteration, the on-line case uses one sample in each new frame to update all weak classifiers and the corresponding voting weight. The on-line method selects one weak classifier from the feature pool to update at

Figure 2.1: On-line boosting for tracking (extracted from [9]).

each frame. A strong classifier is formed by combining all weak classifiers selected from all frames before.

Start with a fixed number of selectors, each with its own feature pool. One sample is picked in a new frame. Selectors are updated with the importance weight λ of the current sample. λ is calculated by considering error rate of previous frames and the sign of the sample in this frame. The estimation of the probability distributions of positive and negative samples is processed by standard Expectation Maximization [33]. The weak classifier with the smallest error is selected by the selector

$$\arg \min_m (e_{n,m}), e_{n,m} = \frac{\lambda_{n,m}^{wrong}}{\lambda_{n,m}^{corr} + \lambda_{n,m}^{wrong}} \quad (2.10)$$

And the linear combination of all selectors with its importance weight forms the

strong classifier

$$h^{strong}(x) = \text{sign}\left(\sum_{n=1}^N \alpha_n \cdot h_n^{sel}(x)\right) \quad (2.11)$$

A drawback for the on-line AdaBoost is that the limited discriminative ability due to the limited number of weak classifiers, as well as the fixed number of selectors.

Ensemble Tracking [8] is also an extension work of AdaBoost to the on-line learning. It constantly updates a set of weak classifiers to separate the foreground object from the background. Each weak classifier is trained on positive and negative examples. Positive samples are extracted from target location, while negative samples are extracted from areas around the target. The model of target is represented by a strong classifier contains T weak classifiers. During the updating process, only the “best” K weak classifiers are kept, according to their weight. $T - K$ new weak classifiers will be added to the model from the feature pool.

Figure 2.2: Multiple instance learning (extracted from [10]).

In Multiple Instance Learning (MIL) [10], the training examples are delivered by a bag of spatially related units, rather than independent training examples (Figure 2.2). This is done by taking the patch at the current tracker location and patches very close to the location as a bag of positive examples, and sampling the neighborhood around the tracker location for bags of negative samples. When traditional

on-line learning methods are used, it may not obtain the optimal positive sample if the tracker location is not precise (Figure 2.2 (A)). So the drift problem happens over time. Besides, if multiple positive examples extracted separately (Figure 2.2 (B)) from neighborhood around the tracker location are provided to update the model, it will undermine the discriminative ability. To some extent, MIL tracking can solve this problem by using spatially related samples and therefore improve the tracking performance.

Nebehay [34] introduced a key-point based long-term tracking method. All key-points belonging to the target vote for the center in each frame. The geometric constellation of the key-points is considered in current frame and key-points are used as context information to indicate the target location. Besides, outlier key-points are identified by clustering and excluded from voting to ensure stable long-term tracking.

2.2.4 Object Tracking Combined with Object Detection

Recently, methods combining detection and tracking are popular. By exploiting the help of specific target detector, long-term object tracking can be achieved since tracker can be resumed after tracking failure.

Kalal [11] developed a framework, which combines object detection and tracking to deal with long term object tracking. It uses Median Flow (MF) [29] for short-term object tracking. At the mean-time, it builds an adaptive object model by learning the positive and negative samples extracted on-line.

Figure 2.3: Using Random Ferns to classify patches (extracted from [11]).

During the tracking, the bounding box of the object obtains a confidence score from the Random Ferns (RF) [35], which is a variation of Random Forests [36] and it is an

efficient way to learn and classify object (Figure 2.3). Random patches are extracted from a target bounding box to obtain features and form the Random Ferns. When the tracker fails, the detector will decide whether and where the target is. When another object is detected with a higher confidence score other than the result of tracker, the target being tracked will be changed to the object detected. In this case, tracking can be resumed if the target reappears from occlusion or tracker drifts away from the true model.

Instead of detecting the target directly, Grabner [12] performed tracking by detection where context information is detected and used to predict the target location indirectly (Figure 2.4). Grabner used the local features inside and around the object to build a motion model, which indicates the location of the object. When tracking is failed, this model is used to find the location of the object, by detecting the local features and using the spatial relationship between them and the object to compute the object location.

Figure 2.4: Use supporters to estimate the location of target (extracted from [12]).

This method introduced the definition of “supporters” which are useful features to predict the target location. These supporters might come from target object points, shadows, other objects interacting with the target, or any regions that provide information about the object positions. The spatial relation between these supporters and the target is not affected by motions and change of backgrounds. They are also learned on-line in order to make better estimation of the target location. However, it is expensive to detect and match all of local features in each frame, and estimation of target under total occlusion or out of the field of view is still difficult.

Dinh [13] extended Kalals work, by detecting the supporters around the tracking object and tracking the distracters which has similar appearance of the target being tracked, the context tracker can prevent drift and distraction from other objects (Figure 2.5). Local key-points around the target are used as supporters if they are consistently appeared in video for a short time span. By recording these supporters, distraction can be excluded since distracter has different spatial relations with supporters. Regions with similar appearance of target and consistently co-occur with

high confidence score are considered as distracters. The tracker must keep tracking these distracters to avoid drifting by learning them as negative samples.

Figure 2.5: Context Tracker with assistance of supporters (extracted from [13]).

There are also some key-points based tracking by detection method. Hare [37] employed key-points for tracking. On-line updating of model is achieved by using linear structured SVM. And binary descriptors such as BRIEF and BRISK are used for real-time performance. Yang [38] proposed a key-points based tracking method. Tracking is done by matching key-points, while templates and knowledge-based methods are used to construct detector. An error collector is employed to record false positive key-points from tracker and detector, which reinforce the on-line training of key-point based tracker and detector.

2.2.5 Salient Based Tracking

In this section, a brief review on saliency based tracking methods is presented. The saliency of an object or area is the quality to stand out from its neighborhood, it is an important component when human observe object in real world. After Itti [39] introduced saliency to computer vision, saliency based tracking methods have been developed.

In tracking methods, saliency is usually used to select features or areas which can be most representative for model construction, since the target usually possess some features discriminant from background. Wu [40] proposed a salient block based object tracking method. With saliency detection method similar to [39], salient blocks are extracted using a branch-and-bound search method on the saliency map. Each salient block is tracked by Particle Filter, and all tracking results are combined to determine the target location. Target template and positions of the blocks are adaptively updated during tracking. Ali [41] employed salient feature points for tracking

of partial occluded target. Prominent feature points are extracted in the initialization period and they are tracked by particle filter based method in various attributes like location, velocity, etc. The target location is determined by considering relative locations of all correctly tracked feature points.

Some methods extend the idea in saliency detection, included descriptor saliency and motion saliency to perform tracking of target. Mahadevan [42] proposed a tracking method based on discriminant center surround saliency. Salient features are selected from pool of feature descriptors to represent the target, they are informative for discriminating target and background. Besides, motion saliency is also included to help target tracking or target initialization. In order to address the problem of inaccurate initialization, Yi [43] proposed generalized Hough Transform (GHT) based tracking method by employing selected local features with motion saliency and descriptor saliency. Motion saliency and descriptor saliency are used to select feature points and local feature points vote for the target center through GHT. Descriptor saliency is defined as how much the descriptor coincided with past consensus, while motion saliency emphasizes motions of feature points which are distinctive to all feature points.

2.3 Human Detection

Human is one of the targets that receive most attention from researchers in detection and tracking area. As a complementary part of human tracking, human detection plays an important role. Fortunately, human detection techniques have been rapidly developing in last decade. Shape information has been exploited and proven to be effective in the detection of human. In this section, the most famous human detection methods in recent years are reviewed.

2.3.1 Histogram of Oriented Gradient

Dalal [3] introduced a feature descriptor called Histogram of Oriented Gradients (HOG), describing the global and local shape information for pedestrian detection. HOG descriptors were extracted from positive and negative samples and feed into linear SVM for the training of human classifier.

To extract the descriptor, one patch is divided into connecting small regions called cells in order to capture the local shape information (Figure 2.6). Histogram of

Figure 2.6: Extraction of HOG descriptor (extracted from [3]).

gradient is formed inside each cell by orientation. Then these histograms are concatenated together. To make the descriptor invariant to illumination change, 2 by 2 cells in a local region form a block. Histograms of cells inside one block are normalized together. Blocks with 50 percent overlap are used across the whole patch for the normalization.

One example is shown is Figure 2.7. Figure 2.7(b) shows the HOG descriptor extracted from the original image Figure (2.7(a)). Then after combining with positive weights trained by SVM, the descriptor is shown in Figure 2.7(c). As we can observe, HOG can well describe the local gradient distribution thus can capture the local shape. It is invariant to small scale and illumination changes. With large samples trained by supervised learning methods like SVM, it achieves good performance in upright human detection.

Figure 2.7: Example of HOG descriptor.

2.3.2 Deformable Part Model

Based on HOG, Felzenszwalb [44] developed Deformable Part Model (DPM) as a mixture of multi-scale star-structured models. It is constructed with a root filter along with a set of part filters.

The evaluation of DPM can be divided into two parts. Before HOG features extracted from the image, Gaussian Pyramid is performed on the original image (Figure 2.8).

Figure 2.8: Pyramid of Deformable Part Model (extracted from [44]).

The root filter captures the global shape information, which makes use of HOG over the whole patch. To reduce the computation complexity, the root filter is usually performed on the image with lowest resolution. Thus it is performed on the highest level of image pyramids, which is down sampling result of the original images (top row of Figure 2.8). In this way, the calculation can be reduced.

Part filters with relatively constrained locations are used to capture local shape information (Figure 2.9). A dense set of positions inside on patch are assigned score to be particular part of the object. With learning by Latent SVM [45], scores of positions for part filters are decided, as well as the score from part filters based on the local features. Therefore, for each part filter, it has a geometric configuration indicating the relative position of this part to the root filter.

Figure 2.9: Part filters with constrained locations in Deformable Part Model.

2.4 Multi-people Tracking

In surveillance applications, multi-people tracking is a common task. Different with tracking single person, multi-people tracking problem faces more problems, such as mutual occlusion of targets, object ID switches caused by intersection of trajectories. It is important to track each target robustly and reduce the interference from other targets. Existing multi-people tracking methods can be roughly divided into two categories. One category is to follow short-term tracking and data association with detections. Only information from previous frames are used, locations of people in current frame are decided by short-term tracking or detection by model constructed from past frames. The other category is to consider multi-people tracking as an optimization problem over a period of time. With the assistant of human detection in each frame, Dynamic Programming [46] methods or Linear Programming methods [47] are used to associate detection in different frames and form consecutive tracklets, aiming to reach the maximum probability or minimum cost.

2.4.1 Tracking by Detection

One way to perform multi-people tracking is to follow tracking by detection approach. Tracking is performed from frame to frame for each target, detection is combined with tracking.

Andriluka [14] introduced the tracking by detection approach to multi-people tracking. A part-based pedestrian model similar to pictorial structures model [48] is constructed to detect pedestrians (Figure 2.10).

By using hierarchical Gaussian Process Latent Variable Model (hGPLVM) to model the walking cycle of targets, the pedestrian detection performance is improved. The

Figure 2.10: Pictorial Structure (extracted from [48]).

detection result is combined with frame to frame trackers to provide robust multi-people tracking.

In Breitensteins work [15], Particle Filter is used as tracker to track targets in across frames. The tracking is combined with detection from target specific classifiers. During the tracking process, target specific classifiers are constructed and on-line updated. These classifiers are further used to detect the target as well as distinguish it from other targets. And Hungarian algorithm [49] is used for data association between detections and trackers.

Zhang[50] used an ensemble of templates for the tracking of each target. A set of templates is used to construct an appearance model. The model is on-line updated and used for target detection. Detections are associated with tracking results from mean-shift tracking and Kalman filtering. The birth and death of Trackers are controlled in this method to deal with varying number of targets and eliminate unwanted false detections.

In Segals work [18], Latent Data Association Parametrization and Inference Algorithm are introduced to multi-people tracking. He transferred the tracking problem into a question of linking detection results into longer trajectories. This is done by assigning each track at time t as the continuation of some track at $t - 1$ with considering a set of discrete variables. Associations between observations are implicit, rather than being explicitly sought as in most traditional formulations. A key advantage of the formulation is that the number of tracks which is in fact a model selection problem is determined automatically during inference.

Tracking by detection methods depend largely on the short-term tracking and target model for detection. However, the short-term tracking performance is not stable for

long-term. Therefore, building a robust yet discriminative target model is important for the tracking by detection methods.

2.4.2 Optimization of Trajectory Connection

Another way to process multi-people tracking is to consider it as a global optimization problem. Some methods perform adding, merging and splitting tracklets recursively to obtain a minimal global cost function over a large patch of frames, where global cost function could be defined variously in different methods.

In Ges work [51], Markov Chain Monte Carlo Data Association (MCMCDA) is used to estimate a varying number of trajectories, based on tracklets extracted from the video. First, short tracklets are formed with short-term tracking methods such as mean-shift and particle filters. An appearance based model is constructed based on each tracklet with unsupervised learning of model parameters and object trajectories. Then MCMCDA is performed to process a set of movements between tracklets like Birth/Death, Split/Merge, Extension/Reduction and Switch etc (Figure 2.11).

Figure 2.11: Movements performed by MCMCDA (extracted from [51]).

Benfold [16] also used MCMCDA for data association. He proposed a multi-people tracking method by stably tracking the head of targets. HOG human detection along with KLT [52] tracking is used and a likelihood function is designed to represent the correct data associations and track types. Estimations on head locations are provided for maintaining of target IDs when occlusion or false detection is encountered. And gaps can be filled by merging tracklets with high probability in likelihood function (Figure 2.12).

In Fleuret's work [46], a Probabilistic Occupancy Map (POM) is estimated based on images from multiple cameras with different viewing angles. Locations of people are determined based on POM, then a Dynamic Programming method is applied to

Figure 2.12: Optimization across all frames to form trajectories (extracted from [16]).

form trajectories individually over sequences of frames. A motion model along with an appearance model is used in determining the trajectory of each target.

The multi-people tracking is performed in a 3D grid formed by three simultaneously-recorded cameras in Wu's method [53]. Network-flow data association is used to connect detection results across frames into tracklets. The goal is to minimize the global cost by forming trajectories of objects through the network. Sparsity-driven target models are constructed for each person and they are used in the data association process.

Andriyenko [17] proposed a multi-people tracking method based on global optimization method. An energy function is constructed with consideration of detection of people, target dynamic, collision avoidance, persistence of tracklet and regularization (Figure 2.13). With these considered in the energy function, movements on the tracklets like initiating, growing, shrinking, splitting and merging can be performed (Figure 2.14). The objective is to find a set of tracklets achieves minimum energy function. Milan [19] extended this work by including additional appearance component into the energy function, in order to improve the performance against mutual occlusion.

In Bae's work [54], tracklets are obtained first. Then these tracklets are linked based on matching of on-line discriminative appearance model. An incremental linear discriminant analysis based method is proposed to model and on-line update the appearance of target, which will be further used to link tracklets according to the detectability and continuity of tracklet.

Those methods pursuing global optimization with large batch of frames usually require complicated model and heavy calculation. Therefore, some researchers find

Figure 2.13: Components of energy function (extracted from [19]).

Figure 2.14: Movements in achieving minimum energy function (extracted from [19]).

reliable tracklets in a smaller batch of frames first, concatenation is later performed to form longer trajectories. Kuo [55] proposed an online learning appearance model for multi-people tracking. Reliable tracklets linking detection responses are first performed, followed by the linking of tracklets based on the appearance models constructed on each tracklets. For each tracklet, an appearance model is on-line learned by Adaboost. Patches belonging to this tracklet are used as positive samples while patches from other tracklets are used as negative samples. Then the merging of two tracklets is determined by the similarity of two descriptors from the models on-line learned from the two tracklets (Figure 2.15). Yang [56] extended Kuo's work by introducing online learned Conditional Random Field (CRF) model. He transformed the multi-people tracking into an energy minimization problem, and CRF model is used to differentiate pairs of tracklets.

Figure 2.15: Online learning appearance model (extracted from [55]).

Pirsiavash [57] proposed a greedy algorithm based tracking method to track variable number of people from a single camera. He considered multi-people tracking as an integer linear program problem and follows the min-cost flow method. Birth and death of trackers are estimated by evaluating the cost function.

Pursuing a global optimization, Berclaz [58] applies K-shortest Paths algorithm on Direct Acyclic Graphs (DAGs), which is formed based on the POM [46]. Shitrit [59] extended the work in [58] by introducing sparse appearance information of people. This is used to prevent ID switches when the tracklets of two targets have intersects, and therefore improving the performance of trackers .

2.4.3 Tracking by Body Parts

In multi-people tracking, tracking parts of human body can always help improve the accuracy of the tracker.

Wu [60] used the human body part detection in static images to help the tracking of human. Three part detectors detecting head-shoulder, torso and legs are combined to detect human body. For each part detector, two camera views (profile view and frontal/rear view) are considered. By detecting body parts, partial occlusion condition can be dealt with. Part detectors are used when data association with results from combined detector is failed, as a complementary part in the detection process. Otherwise, the mean-shift is used for frame to frame tracking.

In Yang's work [21], Discriminative Part-based Appearance Models (DPAM) is proposed to deal with occlusion problem and help tracking multiple humans in real scenes. The part-based appearance model consists of 15 squares covering the whole body area (Figure 2.16). DPAM explicitly finds not occluded parts by occlusion reasoning and can be online updated, features are extracted from parts instead of from whole human responses.

Izadinia [20] proposed a multi-people tracking method with the assistance of human parts tracking concurrently. K-Shortest Paths optimization is applied for data association of small batch of frames. Besides the association of the human patches, body parts are also associated by considering part appearance, deformation in the star-model of target and temporal consistence (Figure 2.17). For each part, the appearance model is constructed by on-line learning samples with Linear SVM, as shown in Figure 2.18.

Figure 2.16: Discriminative part-based appearance models (extracted from [21]).

Figure 2.17: Multi-part Multi-people tracking (extracted from [20]).

Figure 2.18: Models built in Multi-part Multi-people tracking (extracted from [20]).

2.5 People Re-identification

It is an important topic to build an appropriate model for the target in people tracking. It should be robust to make the tracking consistent. Meanwhile, it should be discriminative so that it will not be confused with other targets. Researchers have been working on identifying people in public places, even under the cover of different cameras in different areas. There are different ways to construct the target model for the identification task.

2.5.1 Model with Low Level Features

Some works focus on developing the features used to construct the human model. Commonly used appearance features include color, texture, shape and so on.

Color is the most commonly used feature in many computer vision applications. One of the earliest works on people re-identification is from Javed [61]. The color histogram is used to represent a person. Color information is then feed into a Bayesian model together with the position and time information to track a person in disjoint cameras. Bird [62] segmented the human body into ten horizontal regions. For each region, the median values in Hue, Saturation and Lightness color planes are calculated. The descriptor is formed by cascading value of median colors from each region. Then Fisher Linear Discriminants was used to perform appearance-based on-line classification. Kao etc. [63] used color information to form a hierarchical color structures for each image. For different people, a set of images is used and makes each node in the hierarchical color structures a Gaussian mixture model. Further calculation of the similarity of two people is based on Bayesian decision. In Madden's work [64], on-line clustering method is employed to classify pixels belonging to a person. Then the color spectrum histogram is used to represent the person. Intensity transformation is applied when matching targets from disjoint cameras. Prosser [65] proposed Cumulative Brightness Transfer Function to make the color representation invariant to illumination changes due to different sites of disjoint cameras.

Texture is another popular feature used in people re-identification. Gray [66] picked an Ensemble of Localized Features (ELF). Features used in this method include texture features like Schmid and Gabor, as well as eight color channels from RGB, YCbCr and HSV color spaces. Farenzena [67] partitioned the person based on the image inside the silhouette (Figure 2.19). A person is segmented into head, torso

and legs. Central axis is also calculated in torso and legs. Then for each part, local features concerning color and texture are used to represent this person. Features used in this method include weighted color histogram, Maximally Stable Color Regions (MSCR) and Recurrent High-Structured Patches (RHSP). Re-identification is done by matching those features in each part of the body.

Figure 2.19: Symmetry-driven accumulation of local features (extracted from [67]).

Berdugo [68] used background subtraction and saliency map to segment the people first, and employed three texture features—oriented gradients, color ratio and color saliency to form a discriminative descriptor. Ma [69] introduced Local Descriptors encoded by Fisher Vectors for people re-identification. A 7-d descriptor is calculated for each pixel including information on coordinate, intensity, first and second order derivatives. These descriptors are extracted on each channel of HSV color space. Then they are fit into Gaussian Mixture Model to obtain Fisher Vectors and concatenated to form the signature. Kviatkovsky [70] exploited a structure of color distributions by using different parts as well as shape context of the object in order to deal with illumination changes.

Besides color and texture, other features like local key-points, shape information and group information are also used in people re-identification. Gheissari [71] proposed a method combining interest point matching and model fitting in corresponding body parts. The body is segmented beforehand into different parts as head, torso and legs. Wang [72] introduced local shape information and appearance context for the representation of people. HOG in the log-RGB color space is used as a local descriptor. Spatial distribution of the appearance inside each local part is considered as

a feature as well. Hamdoun [73] extracted interest points from a person in a short period of time and use descriptors of interest points to model this person. Descriptors of interest points from different persons are used to build a K-D tree. Then for an unknown person, the identification is achieved by matching interest points in the K-D tree. Zheng [74] made use of information from close neighborhood to identify a person in group. Center Rectangular Ring Ratio-Occurrence is used to describe the global relation-ship of the group of people, and Block Based Ratio-Occurrence is used to describe local spatial information. Bak [75] used HOG to detect the human body and performed a color based foreground/background segmentation method. Two kinds of local features a set of haar-like features and Dominant Color Descriptor (DCD) are used. Mean Riemannian Covariance (MRC) patches (Figure 2.20) are extracted from people and the combination of MRC patches are used for people re-identification in [76] (Figure 2.21).

Figure 2.20: Extraction of Mean Riemannian Covariance (MRC) patches (extracted from [76]).

Figure 2.21: Patch matching (extracted from [76]).

2.5.2 Re-identification with Supervised Learning

Some other methods focus on developing novel similarity calculation methods, or exploiting machine learning methods to enhance the people re-identification perfor-

mance.

Gray [77] selected a set of local features and used AdaBoost to train samples. Optimal weights are obtained and assigned to each feature for the combination of final descriptor. Similarly, Prosser [78] used SVM to train samples with the same features used in [77]. In this method, people re-identification is considered as a ranking problem and a novel SVM-based ranking method is developed. In Bak's [75] method, AdaBoost is used to choose the most distinctive features from the feature set to represent the person. Avraham [79] used SVM to train the samples and obtain the transformation function between two cameras. Then for images in one camera, the descriptors can be matched with images in another camera after using the trained transformation function. Brand [80] extended the idea and made use of it on the people re-identification for more than two cameras. Wang [81] trained a distance metric with feature vectors extracted from overlapping areas. It showed good performance on people re-identification under occlusion and scale change. Zheng [82] also trained a metric by learning from samples. The probability is maximized that distances between true matches are smaller than false matches. Zhao [83] proposed a method by learning those salient features and match saliency distribution in datasets. Adjacency constrained patch matching is performed between two images and One-class SVM is used to find out salient score in the training set. Li [84] employed neural network to pick optimal features from a set of low level features for the re-identification task. A large dataset is built for testing of the method. Pedagadi [85] performs unsupervised Principle Component Analysis (PCA) and supervised Local Fisher Discriminant Analysis (LFDA) to extract the descriptor from the high dimensional features of color image. LDFA is employed to learn a distance metric between image pairs in the training set. In Tao's work [86], regularized smoothing KISS metric learning is proposed to extend Kostinger's work [87] to receive more stable result on small size training set.

2.5.3 Semantic Feature Based Re-identification

In the recent years, semantic features like soft-biometric and attributes has attracted more and more attention. Jain et al. [88, 89] defined soft-biometrics as those characteristics from human body that can provide robust but not identical information about the person. Dantcheva et al. [90] redefined soft-biometrics as physical, behavioral or adhered human characteristics. These traits are natural with semantic meaning, and can be used to describe humans. In this way, the range of soft-biometric is extended and not restricted to traits belonging to the human body.

Soft-biometrics are listed in Figure 2.22. These soft-biometric include permanent traits like gender, ethnicity, or traits of human body appearance that do not change easily, such as age, skin color, hair color, eye color, body shapes and so on. Besides soft-biometric features, some appearance attributes remain unchanged in a short-term time, they can also be good complementary to the biometric features. These attributes include wearing of caps and glasses, or information from clothes and accessories.

Soft Biometric trait	Face/Body/Accessory	Nature of value	Permanence	Distinctiveness	Subjective perception
Skin color	Face	Continuous	Medium	Low	Medium
Hair color	Face	Continuous	Medium	Medium	Medium
Eye color	Face	Continuous	High	Medium	Medium
Beard	Face	Binary	Low/Medium	Low	Medium
Moustache	Face	Binary	Low/Medium	Low	Medium
Facial measurements	Face	Continuous	High	Medium	Medium/High
Facial shapes	Face	Discrete	High	High	High
Facial feature measurements	Face	Continuous	High	High	Medium/High
Facial feature shapes	Face	Discrete	High	High	High
Make-up	Face	Discrete	Low	Low	Medium
Ethnicity	Face	Discrete	High	Medium	Medium
Marks	Face/Body	Discrete	High	Medium/High	Low
Gender	Face/Body	Binary	High	Low	Low
Age	Face/Body	Continuous	Low/Medium	Medium	Medium
Height	Body	Continuous	Medium/High	Medium	Medium
Weight	Body	Continuous	Low/Medium	Medium	Medium
Gait	Body	Continuous	Medium	Medium	High
Body measurements	Body	Continuous	Medium/High	Medium/High	Medium
Body shapes	Body	Discrete	Medium	Medium	Medium
Clothes color	Accessory	Discrete	Low	Medium	Medium
Glasses	Accessory	Binary	Low/Medium	Low	Low

Figure 2.22: Soft-biometric traits (extracted from [90]).

Some of semantic features have already been used in people re-identification. In [91], more than twenty of soft-biometric traits like age, gender, race, skin color, hair color and so on were manually extracted by annotators from video. Another biometric feature used is the symmetry gait signatures. It describes the symmetricity of the body in a period. Fourier decomposition is used to remove noise. Then Analysis of Variation (ANOVA) is used to select those most important traits to represent people. Shakhnarovich [92] proposed a framework to do some analysis after the face detected. Gender and ethnicity are two traits chosen. Some learning processes based on AdaBoost or SVM are provided to train binary classifiers. Simple features such as Haar-like features are used. Denman [93] constructed a soft-biometric traits based

model with height and color information in three body parts. They show acceptable performance on people re-identification in multi-camera scenarios. Vaquero [22] introduced a people identification method by using soft-biometric traits in the head area such as facial hair type, type of eyewear and hair type. Adaboost is used to train selected traits. Dantcheva and Dugelay [94] proposed a face re-identification method based on three soft-biometric traits: hair, skin and clothes. For each trait, patches are retrieved from fixed areas. Color and texture are used to describe it. In Layne's works [23, 24], the human is segmented into 7 horizontal regions and 15 binary attributes were chosen by the author (Figure 2.23). All attributes are trained by SVM based on color and texture features in a specific region.

Figure 2.23: Segmentation of 7 horizontal regions for attributes extraction (extracted from [23]).

2.6 Concluding Remarks

In this chapter, some research works related to our study have been reviewed. First we reviewed different approaches in single target tracking, which includes frame to frame tracking method, tracking by fixed model, tracking with on-line updated model and tracking combined with object detection.

Since human is the target of tracking in our study, some powerful human detectors have been reviewed. HOG provides good description in global and local shape information, while DPM puts constraints on the location of body parts and combines them in the human body detection. The accurate detection of these human detectors elevated the performance in people tracking.

We also investigated methods for multi-people tracking. To solve problems like mutual occlusion and ID switch, methods either pursuing optimization across all frames or performing frame to frame tracking were proposed for multi-people tracking. Some methods exploit the tracking of human body parts to assist tracking.

In the last section, research works on people re-identification are reviewed. Different features were extracted and learning metrics were trained to learn the relationships between two sets of images. Besides, semantic features like soft-biometric features and attributes were exploited to construct more robust features for people re-identification. These methods could be used in the construction of target model for applications like people tracking.

Though research works have been carried out on topics mentioned above, some problems still remain. Object tracking under occlusion is not solved by state-of-the-art tracking methods. Most methods focus on building target model based on the entire appearance, while employing local description of target when occlusion happens is usually ignored. Besides, context information is not employed to assist tracking under occlusion. Furthermore, most target re-identification methods are only done based on static datasets. Those conditions required may not be feasible in reality. Thus, these unsolved problems listed above motivated our research work to improve people tracking and re-identification in cluttered environment.

Chapter 3

Key-points Based Single Target Tracking

3.1 Introduction

Object tracking in cluttered environment has been a challenging task in computer vision. Both the models of object and background are changing, sometimes the background is not even static due to the moving camera. This requires to update both models of target and background during the tracking process. In recent years, semi-supervised methods [8, 9, 10] have been proposed to on-line learning the object appearance in order to deal with appearance change. And methods combining tracking with detection [11, 14] receive good performance on long-term tracking of objects. Objects can be adaptively tracked with minor variations. Even if tracking fails due to occlusion or dramatic appearance change, the target can be detected by the detector and resume tracking. However, most of these methods perform patch based on-line learning, it is more suitable for the tracking of rigid object. Since the appearance of non-rigid object keeps changing, it is difficult to build an appearance model that can robustly represent the non-rigid target. Besides, the tracker may fail to track or detect the object when it is partially occluded. Most of the time, detection is only plausible for patch based appearance model when majority of the object is visible. In recent years, some key-points based tracking methods [37, 38] have been developed, key-points are employed to deal with partial occlusion.

In this chapter, a key-point based tracking framework is introduced to perform long-term tracking. It can handle partial occlusion problems as other key-point

based methods, even in cluttered environment. However, different with other key-points based method focusing on object only, the proposed method considers both object and surrounding background during tracking. Object is represented by key-points inside the object. Key-points are on-line learned by Random Ferns (RF) [35] during the tracking, which is an extension of Lepetit and Fua's key-points recognition work [95]. A background model is constructed as well to capture key-points from surrounding background, which can be used to exclude background key-points from false detection as object key-points. With tracking by detection and on-line learning of object and background key-points, the proposed method leads to a better performance than the state-of-art methods when occlusion happens.

3.2 Key-points Based Object Tracking

3.2.1 Overview of the Method

Figure 3.1: The overview of the key-points based tracking method.

Determination of the target location in the proposed method depends on tracking and detection results. In short-term, object is tracked by Median Flow (MF) [29], which is similar to Optical Flow [26]. When tracking fails or tracking result is unconvincing, detector will be used to detect the target. In the first frame, RF, object

model and surrounding background model are initiated. Then in each following frame, if tracking is valid, MF is used to track the object and get a coarse location. Object points and surrounding background points labelled by the two models are used in adjustment for a fine object location. If the tracking result receives a high confidence score from the RF, it is considered as the target location. Positive and negative sample key-points will be extracted and used to update the RF. Points with surrounding background label are excluded from positive samples and object points are excluded from negative samples. The two models are updated as well. If tracking fails or tracking result receives low confidence, RF will be used to detect the object in the next frame. When reliable object is detected, it will be considered as the target and tracking resumes. An overview of the method is given in Figure 3.1.

3.2.2 Tracking by Detection

In the proposed framework, tracking is working along with detection. The short-term tracking of target between frames is performed by Median Flow (MF) [29]. MF is a method developed based on OF (Figure 3.2).

Figure 3.2: Flow chart of Median Flow (extracted from [29]).

The moving vector between two frames is calculated by OF. A set of points (100 points are selected randomly in the proposed method) $T^t = (x_1, x_2, \dots, x_n)$ is first tracked by OF and result in a set of forward points $T_f^{t+1} = (x_1, x_2, \dots, x_n)$ at frame

$t + 1$. However, unlike minimizing only forward error in OF, backward error is also considered in MF to exclude outlier pixels. Points $T_f^{t+1} = (x_1, x_2, \dots, x_n)$ at frame $t + 1$ are tracked backward to frame t and obtained backward points $T_b^t = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n)$. Euclidean distances between points in T^t and points in T_b^t are calculated

$$distance(x_i, \hat{x}_i) = ||x_i - \hat{x}_i|| \quad (3.1)$$

where x_i is the location of a point selected and $\hat{x}_i$ is the estimated location through forward and backward projection. If the distance between point location and estimated location is larger than a threshold (1 pixel), this point is considered as an outlier. Outliers are ruled out and those reliable points are kept to calculate the moving vector.

Thus the tracking result of MF is more reliable than OF. Even though, the short-term tracking result is not always accurate, it needs to be evaluated and adjusted. Therefore, in the proposed method, the tracking result after short-term tracking is evaluated by RF and a confidence score is assigned. Besides, patches in a local neighborhood of the tracking result are evaluated as well. Then patches with top ranking confidence are used to make adjustment to the location of target. Details of the adjustment is addressed in section 3.2.5

The tracker may fail even with location adjustment, especially when the target is under occlusion. Occlusion always causes abrupt appearance change for the tracking area of target. So when the confidence of tracking result is below the threshold and not convincing, the key-point based target detection will be performed by RF. By extracting key-points in a patch and assigning confidence scores to them, the patch will receive a confidence score for being the target based on the scores from key-points. Even if part of the target is occluded, as long as some key-points of the target are visible, the target still can be detected. Tracking in following frames can be performed based on the new detection. One example is shown in Figure 3.3.

3.2.3 Evaluation and On-line Learning of Key-points

Existing long-term tracking methods use the whole patch to represent the object. This causes the failure of object detection under occlusion. When object is partially occluded, the patch where object located will include part of the background. It will be different from sample patches which were used to train the classifier (Figure 3.4). Thus, the performance of detection will be undermined.

Figure 3.3: Detection after tracking failure. The tracker gradually drifts away from frame 157. It loses the target at frame 169. But the target is found by the detector at frame 176.

Figure 3.4: Patch comparison between target patch and patch with occlusion.

In the proposed framework, we aim to find a feature which can still be tracked when partial occlusion happens. Thus, key-points are used in the proposed framework. Since the appearance of the object as well as its neighborhood is changing, the object model should keep updating. An on-line classifier is needed to learn these key-points. In the proposed method, Random Ferns (RF) [35] is used to on-line update key-points. RF was originally used as an off-line classifier for key-point recognition, then Kalal used it as an online classifier to learn patches [11]. It has been proved to be an effective and efficient classifier for key-point recognition.

Figure 3.5: Transform from decision tree to fern.

Random Ferns is an variation of Random Trees [96]. As shown in Figure 3.5, a tree

can be transformed into a fern by performing the following steps. First, constrain the tree to systematically perform the same test across any given hierarchy level, which results in the same feature being evaluated independently of the path taken to get to a particular node. Second, remove away with the hierarchical structure and simply store the feature values at each level. It is applying a sequence of binary tests to the patch.

To evaluate key-points with RF, the local neighborhood patches of key-points are taken. In the proposed method, a local patch with $1/10$ of bounding box's size is extracted around the key-point. Then a set of random simple features can be extracted from the local patch to form the fern features, as shown in Figure 3.6. As fern feature is randomly chosen and compares pixel intensity of two points inside the local patch, it is illumination invariant.

Figure 3.6: Fern extracted from key-point.

For the fern feature extracted, the key-point is classified by maximizing a posteriori probability

$$\hat{c}_i = \arg \max_{c_i} P(C = c_i | f_1, f_2, \dots, f_N) \quad (3.2)$$

$$P(C = c_i | f_1, f_2, \dots, f_N) = \frac{P(f_1, f_2, \dots, f_N | C = c_i) P(C = c_i)}{P(f_1, f_2, \dots, f_N)} \quad (3.3)$$

where $\hat{c}_i$ is the class estimated. c_i represents the class from either object or background. f_i represents the fern feature and N represents the number of ferns used. With a uniform prior $P(C = c_i)$, and a scaling factor $P(f_1, f_2, \dots, f_N)$ which is

independent from the class, the class of key-point is decided by

$$\hat{c}_i = \arg \max_{c_i} P(f_1, f_2, \dots, f_N | C = c_i) \quad (3.4)$$

In the proposed method, RF is set to have 10 ferns and 11 bit features. Thus, using F_m to indicate the 11 bit feature with 2^{11} values. The number of ferns and bit features are set by experience. If the number is too small, the performance will be poor. On the other hand, the computation would be expensive if the number is too large. The probability for a specific value of F_m is

$$p_{k,c_i} = \frac{N_{k,c_i}}{N_k} \quad (3.5)$$

where k represents the value of a fern feature, N_{k,c_i} represents number of samples recorded with class label i and fern feature value k . N_k represents number of all samples with fern feature value k . Thus, we have $\sum_i p_{k,c_i} = 1$.

RF is on-line updated. When new samples come in, probabilities are updated as

$$p_{k,c_i} = \frac{N_{k,c_i} + N_r}{N_k + \beta \times N_r} \quad (3.6)$$

where N_r is number of points added with class label i and fern feature value k . β is the learning rate which determines how fast changes can be learned.

Final decision of one key-point is made by the average probability of ferns, which is called score of being an object key-point. Key-point with probability above 50% is considered belonging to the object, otherwise it is classified as background.

In each frame, points inside the bounding box and points assigned object label are selected as positive samples, points outside with high score and points assigned background label are selected as negative samples. These new samples will be used to update the RF to keep it adaptive.

3.2.4 Object Model and Surrounding Background Model

In the proposed method, object model and surrounding background model are combined together to improve the tracking accuracy and learning of RF. Most existing methods only consider the model of object, or take negative samples from the whole

background. While the surrounding background information is usually ignored. In this framework, object model and surrounding background model are used to store the object points and short-term surrounding background points, respectively. Both models are built by RF.

object model

Object model is used to detect and assign labels to object points. In each new frame, those points inside the tracking bounding box and matched with points in the object model are assigned object point labels. Points in very close neighborhood of tracking bounding box are also assigned object point labels if they receive high score matching with object points in the object model. For example, there is a key-point outside the bounding box in figure 3.7(a). By matching with points in the object model, it is recognized as a key-point belonging to the target, thus it is assigned object label even it is not in the target bounding box. When object location is determined, all the candidate points inside the target bounding box are matched with points stored in the object model. If a point has no match in the object model, it is considered as a new point and is added to the object model as a new point.

surrounding background model

Surrounding background model is used to detect and assign labels to surrounding background points. In this model, we keep all the points in the local neighborhood of object bounding box in the last 10 frames, as well as points which were assigned background labels. In the updating process of each new frame, we first match the points in local neighborhood with points in the surrounding background model. If the point receives matches in most of past 10 frames, it is considered as background point and assigned a background label. Take figure 3.7(b) as an example, some background is included in the bounding box. But with background model, those key-points from background can be recognized and assigned background labels.

usage of key-points models

In each new frame, if the tracking result is valid with the MF, we further perform position adjustment in a local range. A minor adjustment after MF tracker used can improve the performance of tracking. To achieve the adjustment, object points and background points are detected in a local neighbourhood of the tracking result. Those bounding boxes with high ratio of object points to background points are selected, because bounding box with a more accurate position should contain more object points and fewer background points.

Figure 3.7: Key-points stored in object model and background model. Red points are points with object label, while blue points are points with background label. In (a), one point is out of the bounding box, but is still recognized as object point. In (b), some background points in the bounding box are detected.

Object and surrounding background model are also used in the selection of positive and negative samples. When object is partially occluded, the bounding box of object may contain some background key-points (figure 3.7(b)), which will be detected and labelled by the surrounding background model. These background will undermine the object model in the RF if they are falsely used as positive samples. Therefore, points with background labels are ruled out of positive samples for updating the Random Ferns. Likewise, points with object sign are excluded from negative samples during the update of background model.

3.2.5 Position Adjustment After MF

MF is not always accurate, some minor shift happens during short-term tracking. Therefore, position adjustment in a local range is provided based on key-points models. After MF tracking result determined, bounding boxes in local range are searched to find candidate boxes which contain lots of object key-points and few background key-points. Then weighted average of bounding box positions is used to represent the final target location (Figure 3.8).

Figure 3.8: Position Adjustment. The first image shows the location after MF tracking, the second images shows candidates detected by key-points models, and the third image shows the target location determined by candidates and the tracking result of MF.

3.2.6 Adaptive Bounding Box to Scale Change

As scale of object changes during tracking, the size of bounding box is changing adaptively. During short-term tracking by MF, not only the moving vector is calculated, but also the scale change of bounding box. Besides, after tracking by MF, the result is combined with detection candidates from RF to determine the final target location. When detection is performed, scale change range from 0.8 to 1.2 times of the original size are considered. Thus, minor scale change can be dealt with in the proposed method.

3.3 Experimental Result and Discussion

To evaluate the performance of the proposed method, experiments are carried out on some videos with partial occlusion or full occlusion conditions.

3.3.1 Occlusion Evaluation

First, performance of the proposed method on different occlusion conditions is evaluated, to testify how good the proposed method can deal with occlusions.

We show the experiment results with different occlusion conditions (Figure 3.9). When object suffers minor occlusion, the tracking is not affected and the result is accurate. When half of the object is occluded, we can see that the target is still tracked accurately for long-term. Even if later “drift” problem happens (the third picture), detector can detect the target again and resume tracking (the fourth

Figure 3.9: Results in different occlusion conditions. First row is under minor occlusion, second row is under normal occlusion, the last row is under severe occlusion.

picture), so tracking is not failed. When object suffers severe occlusion, as we can observe, almost the whole object is occluded. Then our method fails to track the object, nor detect it (the third picture). But when part of the object appears again, our method can detect it again, and tracking will resume then (the fourth picture).

From the evaluation of different occlusion conditions, we can see that proposed method can deal with most of the occlusion conditions except severe occlusion. And even if severe occlusion happens, our method can detect the target when it reappears and achieve long-term tracking.

3.3.2 Comparison with Other Methods

The proposed method is also compared with some state-of-the-art works. Methods for comparison includes: TLD Tracker [11], Context Tracker [13] and MIL Tracker [10]. Some results of comparison are shown in Figure 3.10. When tracking target is partially occluded by background object, the tracker in TLD and Context Tracker stops at where occlusion happens rather than keep tracking of the target. In the following frames, because part of the target is occluded, the detector fails to detect the object as it is supposed to do. While our method succeeds to detect the target when it is partially occluded and keeps tracking of the occluded target. MIL Tracker is well known for its ability of adaptive tracking, but this also leads to the drift problem. As shown in Figure 3.10(b), MIL tracks well when object is occluded, but later drift happens and it loses the target. Our method tracks the object when it

Figure 3.10: Results of comparison with other methods. (a) and (e) are the results of TLD, (b) and (f) are results of Context Tracker, (c) and (g) are results of MIL, (d) and (h) are results of the proposed method.

Table 3.1: Numbers of correct tracking frames in test videos.

Video Sequence	Frames	TLD Tracker[11]	Context Tracker[13]	MIL Tracker[10]	Proposed Method
Person 1	256	108 lost@ 109	165 lost@ 166	172 lost@ 173	229
Person 2	279	140 lost@ 141	161 lost@ 162	161 lost@ 162	256
Kangaroo	273	112	115	159 lost@ 160	235

is occluded. When “drift” happens, it can detect the target and resume tracking later.

Table 3.1 shows the right tracking frames for the four methods in comparison, and where track failure happens. The testing videos are all involved with occlusion problems.

3.3.3 Discussion

As shown in previous sections, the proposed method can provide robust tracking under partial occlusions. However, the proposed method faces some challenges.

First, there is a resolution requirement for the proposed method. As the proposed method is based on key-points, the number of key points extracted may be insufficient for low resolution images. The performance will degrade without sufficient

key-points. One complementary method is to combine patch based model along with key-point based model.

Another problem faced is the expensive computation due to key-points extraction across the whole frame. Also, the extraction of bit features for all key-points is time-consuming. In order to reduce the computation, we can put a constraint on the spatial area searched for key-points. However, it may degrade the performance of the Random Ferns as it needs negative samples to train the classifier.

3.4 Concluding Remarks

Existing tracking methods combine detection and tracking can achieve long-term tracking, but they still face occlusion problem. In this chapter, we propose a method to solve the occlusion problem during object tracking. In order to find a feature which can be tracked even the object is partially occluded, we use key-points to represent the object. Key-points are on-line learned by Random Ferns. Both object and background models are used to store object points and surrounding background points, respectively. It can be further used to improve the tracking performance and reduce the false positive samples in learning of Random Ferns. Experiments show the proposed framework provides good performance in object tracking under partial occlusion. Comparisons with some of the state-of-art methods are provided.

Chapter 4

Salient Parts Based Multi-people Tracking

4.1 Introduction

Accurate tracking of multiple targets is crucial in many applications in surveillance, such as motion/scene analysis, abnormal event detections, detection and tracking of people in public places for safety reasons. Besides those challenges faced in single target tracking, extra challenges exist in multi-target tracking, such as mutual occlusion between targets, ID switch and so on.

There are mainly two kinds of method solving multi-people tracking problem. One is focusing on the frame to frame tracking [14, 15, 18]. Locations of people in current frame are decided by models constructed from past frames, no future information is needed. These methods usually combine tracking with detection. Models of targets are constructed for detection and keep updated during the tracking process. The other methods [16, 19, 20] consider multi-people tracking as an optimization problem based on detections over a period of time. With the assistant of human detection in each frame, data association methods are used to associate detections in different frames and form consecutive tracklets, aiming to reach the maximum probability or minimum cost.

In this chapter, a novel framework for multi-people tracking is proposed. A salient parts based model is constructed for each target. The way saliency used in the proposed method is different with existing saliency based tracking methods which include motion saliency for tracking, or select salient features from pool of local

features for tracking. In contrast, saliency is employed in the proposed method only to find local salient regions inside or around the target. In addition to the global appearance, regions that highly contrast to local neighborhood are included to represent the target effectively. These regions describe the distinctive parts of the target. Besides, context information can always provide useful information for tracking, which is not aware of in many existing methods. Therefore, salient regions in close neighborhood of the target are exploited in this research to assist tracking. These salient regions will be tracked simultaneously along with target, and provide relative spatial information with the whole body to determine the target location.

4.2 Multi-people Tracking with Salient Parts

4.2.1 Overview

The proposed method follows tracking by detection approach. For each frame, the human detection is performed by Deformable Part Model (DPM) [44]. The detection results are further associated with trackers. Figure 4.1 is an overview of the proposed framework.

Tracker Initialization

A tracker initialization step is designed in the proposed method to remove potential false positive detections from DPM. New detections are first labelled as candidates before a tracker is initiated. Candidates are updated by data association in following frames. They are transferred to a tracker if the number of frames updated reaches the threshold. Salient parts are determined along with the initialization of tracker. They include regions inside the target and regions in local neighborhood around the target.

Tracking Process

During the tracking process, data association is first performed to find matched detections for the trackers. If matching fails for a tracker, short term tracking based on those salient parts is used to determine the target location. A supporting model is constructed to describe the spatial relationship between salient parts and the target.

Tracker Update

The tracker is updated if the target location is determined by either detection or short-term tracking. Contents to be updated include the appearance model, salient

Figure 4.1: The overview of the salient parts based multi-people tracking method.

parts and a supporting model. The appearance model is used for occlusion reasoning. The supporting model indicates the spatial relationship between salient parts and the target.

Tracker De-activation

Trackers which have not been updated for certain period of time will be deactivated and excluded from the tracking process, in order to keep the computing efficient. They will be re-activated if a new initiated tracker is identified to possess high

appearance similarity with them.

4.2.2 Tracking Model Description

The proposed multi-people tracking method aims to maximize the joint posterior probability of all people trackers χ based on the observation I from current and previous frames

$$\chi = \underset{\chi}{\operatorname{argmax}} P(\chi|I) = \underset{\chi}{\operatorname{argmax}} P(I|\chi)P(\chi) \quad (4.1)$$

During the tracking process, a model is constructed for each target. Use $\chi = \{\theta_i\}$ to represent a set of people trajectories in a video, where $\theta_i = \{x_{i,t}\}$ is one trajectory containing locations across frames. $x_{i,t} = (l_{i,t}, p_{i,t}, s_{i,t}, c_{i,t}, h_{i,t}, SP_{i,t})$ is the model of target i at frame t . It contains ID $l_{i,t}$, position of target $p_{i,t}$, scale of target $s_{i,t}$, color histogram of target $c_{i,t}$, head part $h_{i,t}$ and a set of salient parts $SP_{i,t} = \{sp_{i,t,k}\}$. The k in $SP_{i,t}$ represents the k th salient part. For head part $h_{i,t} = (p_{i,t}^h, s_{i,t}^h, c_{i,t}^h, r_{i,t}^h, rp_{i,t}^h)$ and salient parts $sp_{i,t,k} = (f_{i,t,k}^{SP}, p_{i,t,k}^{SP}, s_{i,t,k}^{SP}, c_{i,t,k}^{SP}, r_{i,t,k}^{SP}, rp_{i,t,k}^{SP})$, they all contain position $p_{i,t,k}$, scale $s_{i,t,k}$, color histogram (represents the appearance of the part) $c_{i,t,k}$, ratio of the part to the human patch $r_{i,t,k}$ and the relative position inside the human body $rp_{i,t,k}$, respectively. Besides, salient part contains feature of saliency $f_{i,t,k}$.

4.2.3 Tracking Process

In this section, the tracking process is described in detail. Considering the number of tracking targets in a video is varying, initialization and deactivation process are designed in the proposed method. During the tracking, data association and short-term tracking of salient parts are performed simultaneously to provide robust tracking. Occlusion is also detected during tracking to avoid the shift of tracker models.

Tracker Initialization

It is quite common in videos that people moving in or out of the scene. Therefore, a flexible way to handle the number of people tracker should be applied. As false positives exist in DPM detection results (Figure 4.2), tracker candidates are under observation before tracking targets are confirmed.

Tracker candidate is constructed when a new detection without associating to existing trackers is found. The candidate will be updated if matched detections appear

Figure 4.2: False Positive detections from DPM.

in following frames. Matching between existing candidates and detections in current frame is performed by data association based on spatial distance and appearance similarity (color histogram in this case). When the number of frames containing a specific candidate reaches the threshold, it is transited to a tracker. By setting the candidate, we can eliminate those transit false detections.

Along with the initialization of tracker, salient parts of the target are decided as well, as shown in Figure 4.3. For the candidate in each frame, we perform salient region extraction to locate salient regions in current frame. In our research, we exploit those salient parts that keep a relative static relationship with the target. Thus only salient parts with stable spatial relationship to the target are included in the target model. When candidate is transited to tracker, those salient parts across different frames are clustered by location inside the human patch. Regions in one cluster also must have similar appearance. Then those salient regions can be detected constantly during candidate period that is used in the tracker model. Salient regions that occasionally appear are discarded.

Data Association with Detections Across Frames

Following the tracking by detection approach, data association is used to link detection results with existing trajectories. Locations of human detections are provided by DPM. It is also used in the extension of candidate tracklets. To avoid the time delay caused by using a batch of frames, Hungarian Algorithm [49] is used in the proposed method to associate detections in current frame with locations of trackers in previous frame.

Spatial location p , scale of target s and appearance c are considered when performing data association. For one observant at current frame $y_{j,t} = (p_{j,t}, s_{j,t}, c_{j,t})$, the cost

Figure 4.3: Salient parts determination during the transition from Candidates to Tracker. The first four columns are from frames of candidates, while the last column is the first frame of the tracker. Only those salient parts that constantly appear in candidate frames are kept in the tracker model, while occasional appeared salient parts are discarded.

of linking this observant to a tracker location in last frame $x_{i,t-1}$ is calculated as

$$\begin{aligned}
 cost(y_{j,t}, x_{i,t-1}) = & cost_p(p_{j,t}^y, p_{i,t-1}^x) \\
 & + cost_s(s_{j,t}^y, s_{i,t-1}^x) \\
 & + cost_c(c_{j,t}^y, c_{i,t-1}^x)
 \end{aligned} \tag{4.2}$$

where $cost_p(p_{j,t}^y, p_{i,t-1}^x)$, $cost_s(s_{j,t}^y, s_{i,t-1}^x)$ and $cost_c(c_{j,t}^y, c_{i,t-1}^x)$ are costs of spatial location, cost of scale change and cost of appearance, respectively. They are calculated

as follows

$$cost_p(p_{j,t}^y, p_{i,t-1}^x) = 1 - \frac{Area(p_{j,t}^y \cap p_{i,t-1}^x)}{Area(p_{j,t}^y \cup p_{i,t-1}^x)} \quad (4.3)$$

$$cost_s(s_{j,t}^y, s_{i,t-1}^x) = 1 - \frac{\min(s_{j,t}^y, s_{i,t-1}^x)}{\max(s_{j,t}^y, s_{i,t-1}^x)} \quad (4.4)$$

$$cost_c(c_{j,t}^y, c_{i,t-1}^x) = 1 - \sum_b \min(c_{j,t,b}^y, c_{i,t-1,b}^x) \quad (4.5)$$

Observant receives higher cost are those who have smaller overlap area, larger size change and lower appearance similarity. Those costs larger than thresholds of $cost_p$, $cost_s$ or $cost_c$ are set as infinite.

Salient Parts Tracking

When the target tracking failed by data association, short-term tracking can be performed as an alternative based on head part and salient parts. Usually, short-term tracking of patch with smaller size is more stable than patch with larger size. Therefore, tracking a part of the human body instead of the whole body is preferred in short-term tracking (Figure 4.4).

Figure 4.4: Comparison of tracking parts and tracking the whole body. The yellow bounding box in the right frame is the tracking result of Median Flow on the whole body area, while the red bounding box is the target location indicated by the tracking result of green part.

As representative parts of the human body, salient parts can provide reliable information which indicates the target location. Salient parts around the target used as context information are helpful as well.

But not all salient parts are used in short-term tracking. Those salient parts with low appearance confidence or with occluded signs are excluded from short-term tracking, in case they deteriorate the performance of tracking. The assignment of occlusion signs will be addressed in section 4.2.7. Median Flow is used to track head part and salient parts from previous frame to current frame. By using Median Flow, minor scale changes can also be tracked. Then the location of the Tracker in current frame can be obtained by a supporting model. The supporting model will be addressed in detail in section 4.2.8.

Some examples of salient parts tracking are shown in Figure 4.5. As can be observed from the figure, some salient parts can be tracked reliably through period of time.

Figure 4.5: Examples of salient parts tracking.

Tracker Update

After the current location of a trajectory is determined by either detection association or salient parts tracking, trackers will be updated. Contents to be updated include the target location, target appearance represented by color histogram (a 48 bins histogram concatenated by histograms through RGB color plane), locations and appearance of head and salient parts.

Salient parts to be updated need to be detected or tracked validly, which means the similarity of appearance c and overlap of relative locations p must exceed thresholds

$$\sum_b \min(c_{i,t,k,b}^{SP}, c_{i,t-1,k,b}^x) > \tau_c \quad (4.6)$$

$$\frac{\text{Area}(p_{i,t,k}^{SP} \cap p_{i,t-1,k}^{SP})}{\text{Area}(p_{i,t,k}^{SP} \cup p_{i,t-1,k}^{SP})} > \tau_p \quad (4.7)$$

where τ_c and τ_p are thresholds of appearance similarity and spatial overlap, respectively.

And the supporting model is updated as well to reflect the spatial relationship between salient parts and target. The update of supporting model will be addressed in section 4.2.8.

Tracker De-activation

As the number of frames increases, the number of tracker increases as well. The computation could increase significantly if all trackers are involved in the tracking process. Therefore, those trackers have not been working for certain period of time (10 frames in the proposed method) are deactivated and not considered in the tracking process.

4.2.4 Salient Part Extraction

In this paper, saliency detection is employed to extract representative regions inside human body as well as context information around human body. Salient region detection was first introduced to computer vision by Itti [39]. From then onward, many techniques have been developed to detect salient regions/objects [97, 98, 99]. Saliency has been used in many areas in computer vision, such as object segmentation [100, 101] and people re-identification [83, 102].

It has been proved in Itti's work [39] that using the contrast with local neighborhood is an effective way to detect saliency with computational efficiency. And local contrast method also provides scale invariant and illumination invariant, which makes the extraction of salient parts more stable. Color and orientation are usually selected to detect salient regions.

As shown in Figure 4.6, a bottom up method similar to in [39] is used. The human patch is first smoothed by using Gaussian pyramids with 8 different scales. Then 4 contrast maps are obtained by subtracting patches smoothed with larger scales from patches smoothed with smaller scales. These contrast maps denote the difference between pixels and their neighborhood in different scales. They are combined to form the final saliency map. A sliding window with fixed size is applied to find areas with top saliency scores. Saliency score of a window is calculated as the average saliency value of pixels inside the window. Those windows with saliency score Sal_i exceed the threshold $\tau_{sal} = 0.8 \times Sal_{max}$ are kept. Among them, spatially close windows are combined together by weight to decide the location of this salient part. Weight of window is calculated as

$$Weight_i = \frac{Sal_i - \tau_{sal}}{\sum_j (Sal_j - \tau_{sal})} \quad (4.8)$$

Figure 4.6: Flow chart of the saliency detection.

Windows with higher salient score receives higher weight. Then the location of salient part is decided as

$$p^{SP} = \sum Weight_i \times p_i \quad (4.9)$$

where p_i is the location of a window.

In the selection of salient parts, we set different scales. Smaller scale parts can find areas with more contrast to the rest of body areas, while larger scale can find parts salient to the background in neighborhood.

Different with Itti's method, color and orientation salient maps are not combined together in the proposed method. Instead, color salient parts and orientation salient parts are detected separately (Figure 4.7 (b), (c)).

4.2.5 Single Layer Salient Parts

The first step of our research work was to develop salient parts based model with salient parts inside the human patch.

Different with existing methods using fixed body parts like torso, arms and legs assist multi-people tracking [20, 21, 60], in this proposed framework, those salient parts are extracted as representative body parts of the target. Salient parts inside the human body are considered to be important properties when constructing appearance model. Fixed body parts could be similar to the background or the rest of

Figure 4.7: Examples of inner salient parts detection. (a) is the original image. (b) and (c) are saliency maps formed by color and orientation respectively. (d) is the result of detected salient parts, where yellow boxes denote color salient part and green boxes denote orientation salient part.

the body parts, which degrade the performance of tracking. On the contrary, salient parts are highly contrasted to the local neighborhood. Therefore, they are able to represent the human appearance discriminatively and can be used effectively for short-term tracking. Besides, salient parts can also help to differentiate the target with other targets.

In the inner layer, clothes colors with high local contrast or items carried by people are usually extracted as color salient parts. Besides, patterns on clothes or items carried usually cause texture saliency and are extracted as orientation salient parts. These parts are representative and can describe the human appearance well.

For example, in Figure 4.7, the lady is carrying a white bag with a white glove. With Izadinia's method [20], the body parts used would be the arms and legs. For the torso area, the clothes are in uniform color, which make the short-term tracking less reliable. By tracking the salient parts instead, we find the glove and the bag area. They are highly contrast with the local neighborhood and can be tracked more reliable.

4.2.6 Double Layer Salient Parts

Some problems still exist by employing salient parts inside the target patch only. As in the video, mutual occlusion or occlusion by background is common. Sometimes, the visible part of partial occluded target may contain no salient part. Some salient

parts of the target may remain out of the target bounding box since the movement of the human body is flexible.

Thus, salient parts around the target are considered to be included as context information. For example, if a man is riding a bike, the bike parts could be included as context information of the target and used to indicate the target location. The tracking of context information can also provide crucial information for the location of target. Besides, considering the targets in our method are not rigid, bounding box is not able to include all the body part all the time. The detection of salient parts in areas around the bounding box could help to find those body parts outside the bounding box. In the proposed method, the outer layer is define as the bounding box outside the target bounding box, with extra 0.6 times higher and 2 times wider than the target.

In the outer layer, salient parts are usually extracted from items moving along the people, such as bikes, or bags carried by people. Since these objects may not always inside the target bounding box, to detect salient parts around the human body area can include these objects moving along with the target. Besides, these objects are usual different from the background. Therefore, the saliency is guaranteed (black box in Figure 4.8(d)).

Figure 4.8: Examples of two layers salient parts detection. (a) is the original image. (b) and (c) are saliency maps formed by color and orientation respectively. (d) is the result of detected salient parts. Red box shows the inner layer and black box shows the outer layer. Yellow boxes denote color salient parts and green boxes denote orientation salient parts.

For example, in Figure 4.8, the lady is walking with a bike. By tracking fixed body parts, they are in uniform colors and not good for tracking. By tracking the salient parts instead, we find scarf salient both in color and orientation. Besides, with salient detection in the local neighborhood, context information on the bike is included as salient parts. They can provide context information for the tracking.

4.2.7 Occlusion Reasoning

Figure 4.9: Examples of mutual occlusion.

Mutual occlusion is a common problem in multi-people tracking (Figure 4.9). To prevent the drift of the salient parts based model, a Bayesian based method is used in the proposed method for occlusion reasoning. When overlapping is detected between two targets, occlusion reasoning is performed (Figure 4.10). The mutual occlusion relationship between two targets is decided based on posterior probability

Figure 4.10: Occlusion reasoning. For each target model, appearance histogram in head, torso and leg areas are recorded. The overlapped area is divided into different parts, occlusion reasoning is performed based on posterior probability.

$$l_o = \underset{l_i}{\operatorname{argmax}} P(c_i | I_o) \quad (4.10)$$

where l_o represents target ID which the overlapped area belongs to. I_o represents appearance of the overlapped area. l_i and c_i represent the ID and appearance of the

target i .

Suppose we have mutual occlusion between target A and B. If the overlapped area is assigned to target A, for the target B, this area is considered as occluded. Any salient parts in the model of target B located at this area is assigned a occlusion sign, which is excluded from tracker update in current frame.

4.2.8 Supporting Model

The determination of salient parts during the tracker initialization is not only based on saliency, but also considered stable spatial relationship to the target. It demands all salient parts in the model keep relative static, or move almost in the same way the target moves. Therefore, a supporting model $P_t(x_i|SP_i)$ indicating the spatial relationship between salient parts SP_i and the target x_i can be constructed. By using the support model, the target location can be predicted based on the observation of available salient parts. As shown in Figure 4.11

Figure 4.11: Supporting model. Target location is indicated by tracking results of salient parts when detection association fails.

The supporting model is on-line updated once the target location is determined. The exponential forgetting principle is used to achieve on-line updating

$$P_t(x_i|SP_i) \propto \alpha P_{t-1}(x_i|SP_i) + (1 - \alpha)p(x_{i,t}|SP_{i,t}) \quad (4.11)$$

where $p(x_{i,t}|SP_{i,t})$ indicates the spatial relation between salient parts and the target in current frame. α is a forgetting factor controlling the learning rate.

4.2.9 ID Re-identification

Another problem faces in multi-people tracking is how to link different tracklets of the same target together. As mutual occlusion is common, some times the target could be occluded completely. Tracking usually fail after complete occlusion. Unlike optimization methods [16, 19, 20] which can employ future information to link different tracklets and fill the gap, tracking by detection method can not keep tracking during the occlusion. Therefore, when the target appears again, it is usually recognized as a new target and assigned new ID. In this section, an appearance based method is used to perform ID re-identification which could link the new appeared target with previous failed trackers.

Structural Local Sparse Appearance Model

A sparse appearance model is constructed for each target by using Structural Local Sparse Appearance Model (SLSAM) [103]. Instead of performing sparse coding on the whole target patch, SLSAM exploits local sparse information and spatial relationship between different parts of the patch (Figure 4.12).

Figure 4.12: Example of Structural Local Sparse Appearance Model. The human patch is divided into 3 by 3 blocks with overlap, sparse code is generated in each block and concatenated to form the sparse code of the target.

For one target, several templates are used to construct the sparse appearance model during the tracker initialization step. Then one template is segmented into 3 by 3 local patches with 50% overlapping. For each patch, a sparse based representation is used

$$y_i = Dt_i + e_i \quad (4.12)$$

where t_i is the coefficient vector and the e_i is the approximation error. Dictionary

D is formed with local patches of several templates. By solving

$$\begin{aligned} & \min_{\alpha} \|t\|_0 \\ & s.t. \|e\|_2^2 < \varepsilon \\ & e = y - Dt \end{aligned} \quad (4.13)$$

a sparse representation of the template $T = [t_1, t_2, \dots, t_N]$ is obtained, where t_i represents the sparse code of local patch i . Then t_i from all templates are weighted combined to represent the local patch i . Weights are calculated as

$$v_i = \frac{1}{C} \sum_{k=1}^n t_i^{(k)}, \quad i = 1, 2, \dots, N \quad (4.14)$$

where $t_i^{(k)}$ is a $N \times 1$ vector, it represents the sparse code of patch i from template k . C is a normalization term. For more details of the SLSAM, please refer to [103].

ID Re-identification

When the target comes out from the occlusion or detected again by the detector, it should be tracked with the same ID as origin. In the proposed method, a simple ID re-identification step is conducted when a new tracker is initiated. The appearance model of the new tracker is compared with those de-activated trackers within reasonable temporal range

$$e = y^{(n)} - D^{(n)}v(d) \quad (4.15)$$

where $y^{(n)}$ represents the appearance from new initiated tracker, and $D^{(n)}$ represents the dictionary obtained from templates of the new initiated tracker. $v(d)$ represents the sparse code from de-activated tracker. If the distance between the two sparse appearances $\|e\|_2^2$ is below the threshold, the new tracker is assigned with the same ID as the de-activated trackers. All models and information saved in the de-activated tracker are transferred to the new tracker. In this way, target ID can be conserved when the tracking is lost temporarily.

4.3 Datasets and Evaluation Criterion

To evaluate the performance of the proposed framework, experiments are carried out on public datasets. The tracking performance of the framework with both single

layer salient parts and double layer salient parts are compared. Comparison with state-of-the-art works is provided as well. To evaluate the overall performance of multi-people tracking, a measurement similar to F-measure is proposed.

4.3.1 Datasets

Public datasets used in the experiment include Town Center [16], PETS09 [104] and TUD Stadtmitte [14]. Example frames of these video sequences are shown as in Figure 4.13.

Figure 4.13: Example frames of sequences.

Town Center dataset [16] contains only one video. It captures the view of a street in a town center. The length of the video is five minutes, but the ground truth covers only the first three minutes. 25 frames per second and 4500 frames in total are included. This video contains mid-level people density with some severe mutual occlusion conditions.

Three sequences from PETS09 dataset [104] are used in the experiment S2L1, S2L2 and S2L3. All sequences were captured outdoor with an elevated viewpoint. Sequence S2L1 contains 795 frames. People appear in the camera performs usual walking as well as some irregular non-linear motions. The sequence also includes occlusions by object and mutual occlusions between people. Sequence S2L2 contains 435 frames. People with large density were walking randomly in this sequence,

which caused many intersections of trajectory and severe occlusions. Sequence S2L3 contains 239 frames. In this sequence, a group of people were walking toward the same direction simultaneously. Heavy occlusion conditions exist in this video.

TUD Stadtmitte [14] was captured in a horizontal level and contains 179 frames. People captured in the video were walking on the street. They occupied large areas of the frame and mutual occlusion happens.

4.3.2 Evaluation Criterion

Precision, Recall and F-measure are used in the evaluation as they are common used in the tracking problems. Besides, as common used in multi-people tracking methods, the CLEAR MOT [105] is adopted in this research to evaluate multi-people tracking performance of the proposed method. By comparing the tracking result with Ground Truth provided by datasets, two important metrics are obtained. The Multiple Object Tracking Accuracy (MOTA) evaluates three types of errors: False Positive (denoted as fp_t), False Negative (called miss detected target and denoted as m_t) and ID switch (called mismatches and denoted as mme_t).

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \quad (4.16)$$

where g_t is the number of ground truth at frame t .

The Multiple Object Tracking Precision (MOTP) evaluates the ability of the tracker to estimate precise object positions to the Ground Truth. Higher score means smaller distances between tracking results and ground truth.

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \quad (4.17)$$

where c_t is the number of matches at frame t . And d_t^i is the distance between matches, which can be calculated as overlap of bounding boxes.

Similar to F-measure, Multiple Object Tracking F-measure (MOTF) is proposed to measure the overall performance on the multi-people tracking. It is calculated as

$$MOTF = \frac{2MOTP \times MOTA}{MOTP + MOTA} \quad (4.18)$$

4.4 Experimental Results and Discussion

In this section, parameter settings in experiment of the proposed method are first introduced. Followed by evaluation on ID re-identification, qualitative analysis on experiments of different datasets and comparison of the proposed method with other methods on different datasets.

4.4.1 Parameter Settings

Meta parameter settings in the experiment are introduced here. The threshold of frames before tracking candidate transited to a tracker is set to be 5. For salient parts determination when candidate is transited to tracker, as 5 frame candidates is used, salient parts appeared in at least 3 frames are included in the salient parts based model. The thresholds used in data association $cost_p$, $cost_s$, $cost_c$ are set to be 0.4, 0.2, 0.3, respectively. Threshold of position overlapping and appearance similarity used in salient parts association during tracker update are set to be 0.4 and 0.8, respectively. During salient parts detection, three scales of window sizes are used in the proposed method. They are set to be 0.3, 0.4 and 0.5 of the target patch size, respectively. The learning rate α in supporting model is set to be e^{-1} .

4.4.2 Evaluation on ID Re-identification

Those methods pursuing global optimization can make use of future information and therefore can link tracklets even if there are gaps between them. However, only information provided from previous and current frames are used in the proposed method. Sometimes, the failure of tracker is inevitable due to severe occlusion or dramatic appearance change. If the target is detected again after the tracking failure, it is usually assigned a new label and become a new trajectory. To deal with this problem, sparse based appearance model (SLSAM) is added to perform ID re-identification in a certain temporal range.

Figure 4.14 shows some examples of successful ID re-identification, comparing with the similar tracking conditions without performing ID re-identification. Results of proposed method without ID re-identification (2L_SPT) are shown at the left side. When the target is detected again after the tracking failure, a new ID is assigned and the former tracker is abandoned. While in the proposed method with ID re-identification (2L_SPT + IDRI), after the target is detected and initiated a tracker

Figure 4.14: Examples of ID re-identification of the proposed method. Left: results from method without ID re-identification (2L_SPT). Right: results of method with ID re-identification (2L_SPT+IDRI). First columns show the tracking before track failure, IDs are shown along with the target. The second columns show the track failure and third columns show the detection of the candidate after track failure. The last columns show the result when the target is tracked again. In the method with re-identification, it can be observed that the same IDs (in the first column) can be resumed in the last column.

again, the new initiated tracker is compared with those de-activated trackers on the sparse appearance. As shown in the right side of Figure 4.14, these new initiated trackers are successfully re-identified as a former tracker and resumed the tracking under the previous ID assignment.

To evaluate the re-identification performance of the proposed method with ID re-identify process, the total numbers of IDs assigned in each experiment are compared. Ground truth of the target numbers in all videos are provided for comparison. As shown in Table 4.1, the proposed method with double layers and ID re-identify (which is performed by sparse appearance) 2L_SPT +IDRI is compared with the proposed method without ID re-identify 2L_SPT and the single layer salient parts tracking (1L_SPT). 2L_SPT can improve the overall performance on multi-people tracking compared with 1L_SPT. However, the number of trackers increases. In most videos, with the help of sparse appearance model, the number of trackers can be

Table 4.1: Comparison on tracker numbers. The closest numbers to the ground truth are shown in bold.

Methods	1L_SPT	2L_SPT	2L_SPT+IDRI	Ground Truth
Town Center	MOTF = 72.9 F-measure = 87.6 ID_no=369	MOTF = 75.7 F-measure = 89.6 ID_no=420	MOTF = 75.4 F-measure = 89.6 ID_no= 368	ID_no=231
PETS09 S2L1	MOTF = 82.4 F-measure = 96.2 ID_no=66	MOTF = 82.6 F-measure = 96.2 ID_no=53	MOTF = 82.5 F-measure = 95.8 ID_no= 45	ID_no=19
PETS09 S2L2	MOTF = 51.3 F-measure = 59.9 ID_no=184	MOTF = 58.9 F-measure = 74.3 ID_no=231	MOTF = 58.1 F-measure = 72.2 ID_no= 139	ID_no=43
PETS09 S2L3	MOTF = 42.6 F-measure = 50.6 ID_no=68	MOTF = 49.2 F-measure = 63.4 ID_no= 48	MOTF = 46.9 F-measure = 58.8 ID_no=38	ID_no=44
TUD Stadt- mitte	MOTF = 74.3 F-measure = 86.2 ID_no= 13	MOTF = 77.7 F-measure = 89.0 ID_no=15	MOTF = 76.9 F-measure = 88.1 ID_no=15	ID_no=10

reduced without affecting much on the performance. 2L_SPT +IDRI can effectively re-identify targets after track failure.

4.4.3 Qualitative Analysis

Failure of detection and mutual occlusion are very common in multi-people tracking. To solve these problems, salient parts are used in the proposed method. The inner salient parts are representative. They can be used for tracking when detection fails or the target is partially occluded. The outer layer of salient parts can be considered as context information and therefore indicate the location of target.

As shown in Figure 4.15, the proposed method is compared with DPM detection and tracking using inner layer salient parts only (1L_SPT) [106]. The first column

Figure 4.15: Comparison between DPM detection, single layer salient parts tracking (1L_SPT) and double layer salient parts tracking (2L_SPT). The first row are the results from DPM detection, second row are the results from 1L_SPT, while the last row are the results from the proposed method.

shows results from frame 182 of Town Center. Due to the occlusion of the woman in white, DPM fails to detect the man with a bag and white hat on the top of the frame, while 1L_SPT and the proposed method successfully tracked the target with the assistance of the salient part on the bag. The second column shows results from frame 585 of Town Center. The woman is not detected from the crowd by DPM. However, by tracking the scarf area, 1L_SPT and the proposed method can track the woman successfully. The third column shows results from frame 794 of Town Center. The woman with yawn is not detected by DPM and tracking by 1L_SPT is poor. But the proposed method can employ the context information from the yawn and therefore tracks the target accurately. The forth column shows results from frame 1598 of Town Center. Both DPM and 1L_SPT failed on tracking the woman with yawn, while the proposed method tracked the target with the assistance of context information from the yawn.

4.4.4 Comparison on Tracking Performance Evaluation with Other Methods

In this section, comparison of the proposed methods (including Single Layer Salient Parts Tracking and Double Layer Salient Parts Tracking method) with existing

multi-people tracking methods on public datasets is provided. Evaluation data of other methods come from their publications.

Table 4.2: Comparison on Town Center dataset. The best is shown in bold, the second best is underlined and the third best is in italic

Methods	MOTP	MOTA	MOTF	Precision	Recall	F-measure
Benfold [16]	80.3	61.3	69.5	82.0	79.0	80.5
Leal-Taixe [107]	71.5	67.3	69.3	71.6	67.6	69.5
Pirsiavash [57]	68.8	63.5	66.0	84.9	78.9	81.8
Yamaguchi [108]	70.9	63.3	66.9	71.1	64.0	67.4
Izadinia [20]	71.6	<i>75.7</i>	<i>73.6</i>	<i>93.6</i>	81.8	<i>87.3</i>
Zhang [50]	68.8	73.6	71.1	84.5	74.2	79.0
1L_SPT [106]	<i>71.7</i>	74.1	72.9	87.1	88.2	<u>87.6</u>
2L_SPT	<u>71.9</u>	<u>79.9</u>	75.7	<u>95.2</u>	<u>84.7</u>	89.6
2L_SPT+IDRI	71.3	80.0	<u>75.4</u>	95.3	<i>84.6</i>	89.6

Table 4.2 shows the evaluation of methods on Town Center dataset. Since the crowd density in the video is medium and some people are carrying objects like bike and pram, the proposed method can make good use of outer layer salient parts. They are used as context information and provide helpful information and make target tracking more stable even they are partially occluded. Therefore, 2L_SPT shows superior performance compared with 1L_SPT, it also outperforms other methods on this dataset. Compared with 2L_SPT, 2L_SPT+IDRI reduces tracelet breakage without sacrificing too much performance. As shown in the table, the proposed method 2L_SPT receives the second best performance on four evaluations of MOTP, MOTA, Precision and Recall. While with the assistance of ID re-identify, 2L_SPT+IDRI obtains the best score in MOTA and Precision, third best score in Recall. For the overall performance, the proposed method 2L_SPT shows superior performance on both MOTF and F-measure than all other methods. In addition, 2L_SPT+IDRI achieves the best performance on F-measure and the second best performance on MOTF, which is close to the best performance. 2L_SPT shows improvement on the performance compared with 1L_SPT.

Table 4.3 to Table 4.5 show the evaluation of methods on PETS09 dataset. Table 4.3 lists performance of methods on S2L1 sequence. People in this video does not carry anything along, thus context information can hardly be employed. Therefore, on this

Table 4.3: Comparison on PETS09 S2L1. The best is shown in bold, the second best is underlined and the third best is in italic

Methods	MOTP	MOTA	MOTF	Precision	Recall	F-measure
Berclaz [58]	72.0	80.3	75.9	96.3	83.8	89.6
Breitenstein [15]	56.3	79.7	66.0	89.0	60.0	71.7
Pirsiavash [57]	73.7	84.6	78.8	96.8	93.2	95.0
Izadinia [20]	<u>76.0</u>	90.7	<u>82.7</u>	96.8	<i>95.2</i>	<i>96.0</i>
Wu [53]	70.0	94.0	80.2	-	-	-
Yang [21]	-	-	-	94.8	97.8	96.3
Zhang [50]	68.2	<u>93.3</u>	78.8	94.8	82.2	88.1
Milan [109]	74.3	90.3	81.5	-	-	-
Segal [18]	75.0	<i>92.0</i>	<i>82.6</i>	-	-	-
Shitrit [59]	-	81.0	-	-	-	-
Bae [54]	69.6	83.0	75.7	<u>98.8</u>	80.6	88.8
Milan [19]	80.2	90.6	85.1	<i>98.4</i>	92.4	95.3
1L_SPT [106]	75.4	90.8	82.4	96.8	<u>95.6</u>	<u>96.2</u>
2L_SPT	75.3	91.4	<i>82.6</i>	98.9	93.7	<u>96.2</u>
2L_SPT+IDRI	<i>75.6</i>	90.7	82.5	<u>98.8</u>	92.9	95.8

sequence, 2L_SPT obtains comparable performance with 1L_SPT. Another challenge is the post in the middle of video, people may stopped at there and chat, where severe occlusion happens. The proposed method 2L_SPT receives the best performance on Precision, while 2L_SPT + IDRI achieves the second best performance on Precision and the third best performance on MOTP. On the overall performance, the proposed method 2L_SPT achieves the second best performance on F-measure and close to the performance with the second best on MOTF. 2L_SPT + IDRI's overall performance is slightly inferior to 2L_SPT.

Table 4.4 shows the performance of methods on S2L2 sequence. The proposed method 2L_SPT + IDRI has the second best performance on MOTP and Precision. However, it does not achieve good performance on MOTA and Recall in this sequence. 2L_SPT achieves the third best on MOTP, Precision and Recall. On the overall performance, 2L_SPT receives the third best on MOTF and F-measure.

Table 4.4: Comparison on PETS09 S2L2. The best is shown in bold, the second best is underlined and the third best is in italic

Methods	MOTP	MOTA	MOTF	Precision	Recall	F-measure
Berclaz [58]	60.9	24.2	34.6	92.1	26.8	41.5
Breitenstein [15]	51.3	50.0	50.6	-	-	-
Zhang [50]	58.7	<u>64.1</u>	61.3	90.5	51.6	65.7
Milan [109]	59.8	46.0	52.0	-	-	-
Bae [54]	53.9	70.1	<u>60.9</u>	85.0	85.7	85.3
Milan [19]	59.7	<i>57.2</i>	58.4	88.0	<u>67.9</u>	<u>76.7</u>
1L_SPT [106]	70.4	40.3	51.3	95.2	43.7	59.9
2L_SPT	<i>64.3</i>	54.3	<i>58.9</i>	<i>92.3</i>	<i>62.2</i>	<i>74.3</i>
2L_SPT+IDRI	<u>64.7</u>	52.7	58.1	<u>92.6</u>	59.1	72.2

While Zhang’s method shows the best performance on MOTF, Bae’s method shows the second best performance on MOTF and the best performance on F-measure. 2L_SPT shows improvement on 1L_SPT. Table 4.5 shows the performance of methods on S2L3 sequence. The proposed method 2L_SPT achieves the third best performance on the overall performance of F-measure. Overall, Milan’s method shows the best performance. 2L_SPT receives superior performance than 1L_SPT.

Table 4.5: Comparison on PETS09 S2L3. The best is shown in bold, the second best is underlined and the third best is in italic

Methods	MOTP	MOTA	MOTF	Precision	Recall	F-measure
Berclaz [58]	61.8	28.8	39.3	95.7	30.4	46.1
Breitenstein [15]	52.1	67.5	58.8	-	-	-
Zhang [50]	56.4	40.4	47.1	90.5	<u>51.6</u>	<u>65.7</u>
Milan [109]	<u>65.0</u>	39.8	<i>49.4</i>	-	-	-
Milan [19]	<i>64.6</i>	<u>45.4</u>	<u>53.3</u>	90.9	51.8	66.0
1L_SPT [106]	67.4	31.1	42.6	<u>93.3</u>	34.7	50.6
2L_SPT	59.2	<i>42.1</i>	49.2	90.8	<i>48.7</i>	<i>63.4</i>
2L_SPT+IDRI	59.9	38.6	46.9	<i>92.9</i>	43.0	58.8

As shown in Table 4.4-4.5, the proposed method does not show superior performance to other methods on all video sequences, this is because large population density exists in PETS09 S2L2 and PETS09 S2L3, which may cause the inferior performance of the proposed method. As the model is constructed during the initialization of the proposed method, the model could be deteriorated if information from other targets is involved due to occlusion. And large population density also caused the low detection rate, which is adverse to the proposed method. Furthermore, in some sequences, not so many people carrying objects along with them. In this case, context information could not be exploited, thus tracking with two layers of salient parts does not show superior performance than tracking with the inner layer salient parts only.

Table 4.6 shows the evaluation of methods on Stadtmitte sequence of TUD dataset. Compared to previous datasets, the camera is set close to targets. Thus salient parts can be well extracted and provided robust assistance to tracking. The proposed method 2L_SPT receives the best performance on MOTA and Precision, second best on MOTP and third best on Recall. While 2L_SPT + IDRI receives the second best performance on MOTA and Precision, as well as the third best performance on MOTP. Overall, 2L_SPT has the best performance on both MOTF and F-measure, while 2L_SPT + IDRI receives slight inferior performance.

Table 4.6: Comparison on TUD Stadtmitte. The best is shown in bold, the second best is underlined and the third best is in italic

Methods	MOTP	MOTA	MOTF	Precision	Recall	F-measure
Berclaz [58]	56.7	45.8	50.7	79.2	63.1	70.2
Milan [109]	61.6	56.2	58.8	-	-	-
Segal [18]	71.0	73.0	72.0	-	-	-
Milan [19]	65.0	<i>73.4</i>	68.9	<i>87.9</i>	85.6	<i>86.7</i>
1L_SPT [106]	77.3	71.6	<i>74.3</i>	87.8	<u>84.6</u>	86.2
2L_SPT	<u>77.0</u>	78.5	77.7	98.2	<i>81.3</i>	89.0
2L_SPT+IDRI	<i>76.8</i>	<u>77.0</u>	<u>76.9</u>	<u>96.8</u>	80.9	<u>88.1</u>

4.5 Concluding Remarks

In this paper, we proposed a Salient Parts Model for multi-people tracking. Inner layer and outer layer salient parts are extracted to represent the target and context information, respectively. Following the tracking-by-detection approach, salient parts tracking can be a good complementary to data association. Supporting models are constructed and on-line updated to indicate the spatial relations between salient parts and target locations. Appearance model is added to perform occlusion reasoning, as well as performing ID re-identification to link tracklets after tracking failure.

As shown in experimental results, tracklets can be linked by ID re-identification and reduce tracklets breakage without sacrificing too much tracking performance. Besides, experimental results validate the improvement of performance by adding two layers of salient parts tracking. 2L_SPT always performs better than 1L_SPT when context information is available. Furthermore, it shows the comparable performance of the proposed method with state-of-the-art methods. The proposed method shows superior performance when targets are in high resolution and the density is not too large, or targets are carrying objects along.

Chapter 5

People Descriptor for Re-identification

5.1 Introduction

One challenge of people tracking is the construction of people appearance model. To successfully identify those people, the key is to find robust and discriminative features. In this chapter, we propose a framework to extract people descriptor in videos. The descriptor consists of a set of semantic features, which includes soft biometric traits and appearance attributes. Compared to traditional low level features used in object detection and tracking, like color, gradient and so on. These semantic features involve with semantic meaning and are more invariant to illumination changes, people pose changes and so on. They can provide robust information about an individual, though they are not able to individually authenticate the subject because they lack distinctiveness and permanence [88]. There are many semantic features including color of hair and eyes, color of skin, beard, body characteristics like height or weight as well as gender and age. Semantic features are good complementary to biometric identifiers like fingerprint, face and hand-geometry. Though a single semantic features is not distinguishable, a set of them can well describe and help to identify a person.

The semantic features used in the proposed method not only consider soft-biometric traits from human body, but also features adhere to human characteristics in a period of time like clothes and accessories. As simple methods are developed for the extraction of each semantic feature, the proposed method is able to achieve real-time processing on videos. Since the proposed method is designed to be applied on

real-time videos, in which training sample data of people appeared is not available. Training beforehand is not needed in the proposed method. Unlike other methods extracting descriptors from well cropped and resized images, the proposed method can extract descriptors from frames with different scales, even with some rotations. Descriptors of partially visible people can also be extracted, which is implausible in other state-of-the-art methods.

5.2 A Framework for People Descriptor Extraction

Though human descriptor can be useful in applications like multi-people tracking, most current researches focus on people re-identification with static image datasets. These methods require images to be well aligned beforehand. However, there are disadvantages by using static image based datasets. For applications like multi-people tracking in practice, real-time processing on the video is expected. That demands methods can directly process on videos, which means images of people are not well segmented and they are in different scales. Also, people in camera may be tilted. Furthermore, learning beforehand for any two cameras may not be available. Therefore, in this research, a framework that can automatically detect people in surveillance cameras and model the people with semantic features is proposed. Image alignment and learning procedure are not involved beforehand.

5.2.1 Overview

The overview of the proposed framework is shown in Figure 5.1. Basically it consists of 4 steps.

Pre-processing

Before processing to the extraction of semantic features on people detected, some pre-processing works need to be done. In the proposed framework, background subtraction [110] and head detection method [111] are employed to detect people in the video. After each people detected, an ID is assigned to link the same people in consecutive frames. Body crop and rotation correction are included in pre-processing as well.

Body Segmentation

Figure 5.1: Overview of the proposed framework

For each people detected in a frame, the body will be segmented into head, torso and leg areas. Symmetry and color information are both used for segmentation, which will be detailed in section 5.2.3.

Semantic Feature Extraction

There are many semantic features could be used. However, for demonstration purpose, 2 to 4 soft-biometric traits in each body part, 8 in total are used in the simulation of the proposed framework to construct the human appearance model. In the head area, soft-biometric traits like hair length and skin tone are extracted. In the torso area, traits describing the clothes are considered, including clothes color, clothes pattern and the length of sleeves. The detection of bag is performed as well.

In the leg area, pant color and pant length are extracted.

Combination

In applications like People Re-identification, all features are combined by weight to form an overall descriptor for a person after each semantic feature extracted. Re-identification work is further done by comparing the similarity of two sets of descriptors and assigning IDs to subjects in the probe set. During the calculation of similarities, only those features extracted are considered.

5.2.2 Pre-processing

Usually the video is transformed into frames for further processing. All frames used in the proposed framework is resized to the same size.

People Detection

Before constructing the human appearance model with semantic features, one important preprocessing task is to detect people in the surveillance video. Existing methods in people detection include Histograms of Oriented Gradient (HOG) [3], DPM [44] and background subtraction based methods [110, 112]. HOG is a popular feature used for people detection in images, it describes the local shape information. Supervised learning methods like SVM are used to train HOG features to obtain a classifier. Background subtraction based methods are commonly used for mobile object detection in videos. A background model is constructed, and the foreground object in a frame with the same background can be detected by subtracting the background model.

HOG and DPM have been proven to be efficient in the detection of people. However, in the database used for evaluation, some drastic illumination conditions and affine transformation exist and therefore downgrade the performance of HOG or DPM. Therefore, background subtraction based method [110] is adopted along with head detection method [111] in this framework. Besides, by using background subtraction, the moving part can be extracted, thus dealing with scale change is plausible. Some of the background subtraction results are shown in Figure 5.2. For other datasets, human detection methods like HOG and DPM could be used.

Body Crop

After the location of head is detected, a morphological operation is employed to crop the foreground mask of the human body. With the foreground mask given by background subtraction and the location given by head detection, the foreground

Figure 5.2: Examples of the preprocessing work. Original images are shown in row 1. Row 2 are results of background subtraction, the target is shown in red bounding box. Row 3 are masks of cropped bodies. Rows 4 and 5 are masks and images of body after rotation correction

mask area connected with head is cropped and considered as the mask of full body of people detected, as shown in Figure 5.2. Further processing like segmentation and extraction of semantic features are based on the cropped images. By employing the foreground mask, background information is excluded from the extraction of traits, therefore improving the accuracy in descriptor construction.

Rotation Correction

Images used usually come from surveillance systems in practice. Distortion may occur, but it can be approached or estimated by affine transformation. (Figure 5.2). As the people appeared in image are not always in upright position, a rotation correction is conducted if needed. The orientation of the principle axis is first calculated based on the mask cropped by Principle Component Analysis (PCA) [113]. The first component is calculated as

$$v_1 = \operatorname{argmax}\left\{\frac{v^T X^T X v}{v^T v}\right\} \quad (5.1)$$

where $X = [x^T \ y^T]$ is the matrix storing location information for foreground pixels. v represents an eigenvector. Then the orientation θ can be computed with the first component $v_1 = [v_x \ v_y]$

$$\theta = \arctan\left(\frac{v_y}{v_x}\right) \quad (5.2)$$

Then rotation transformation is applied to get the image of upright aligned person.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(-\theta) & \sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (5.3)$$

where (x, y) is the original coordinate and (x', y') is the coordinate after rotation correction.

5.2.3 Body Segmentation

After the cropped image of human body obtained, the human body is to be divided into three areas for further extraction of semantic features. The conventional method is to segment the body into several horizontal stripes [23, 62], or segment the body into 3 parts – head, torso and leg [22, 67]. In the proposed framework, the latter segmentation is followed. A spatial distribution and color histogram combined method similar to [67] is used for the segmentation.

The segmentation is based on image pixels with the foreground mask which obtained from pre-processing. Segmentation between head and torso relies on the spatial distribution of mask area. For i_{th} row, the width of foreground area is denoted as w_i . The spatial distribution sd_i of two areas segmented by row i is calculated as

$$sd_i = var(W_h) + var(W_t) \quad (5.4)$$

where $W_h = [w_1, \dots, w_i]$ and $W_t = [w_{i+1}, \dots, w_n]$ are the sets of foreground width from the starting rows to the ending rows in two areas above and below the row i . Then the row which receives the minimum spatial distribution is searched. This row is taken as the border of head area and torso area r_{ht} .

$$r_{ht} = \underset{i}{\operatorname{argmin}}(sd_i) \quad (5.5)$$

where the search is constrained to the upper area of mask image.

For the segmentation between torso and leg, color histograms of 48 bins concatenated by three color planes are used. Usually, the clothes color is not the same with the

color of pants. Therefore, the distance between color histograms of upper area and lower area of human body is checked. For row i , the distance d_i is calculated as

$$d_i = \|H_t - H_l\| \quad (5.6)$$

where $\|\cdot\|$ represents Euclidean distance. H_t and H_l represent the color histograms in two areas, respectively. The row with maximum distance is considered as the border of torso and leg areas r_{tl} .

$$r_{tl} = \underset{i}{\operatorname{argmax}}(d_i) \quad (5.7)$$

To speed up the calculation, Integral Histogram [114] is used when color histogram is constructed. Some results of segmentation are shown in Figure 5.3.

Figure 5.3: Examples of body segmentation. First row are human body masks, while second row are upper and lower body segmentation results.

Figure 5.4: Examples of body segmentation with half body visible.

Unlike the segmentation method used in [67] which can only conduct the segmentation on the full body, the method used in the proposed framework can be used to segment human body partially appears by considering the vertical body size. For

example, if only the head and torso can be seen from the frame, as the body size is way smaller than normal, the leg area is considered not visible. Thus the body can be segmented without the leg area (Figure 5.4).

5.2.4 Semantic Feature Extraction

As introduced earlier, semantic features are those traits temporally adhere to human characteristic. They are with semantic meanings created by human to distinguish different people. Normally, one semantic feature is not discriminative, but a set of them is proved to have good performance on people re-identification [23, 24]. In the proposed framework, some robust semantic features are chosen in simulation. The detection methods of these semantic features are simple yet fast, and some of them are scale invariant or illumination invariant. Details are given as follows.

Head Area

In the head area, there are many traits can be involved, such as hair, skin tone, wearing glasses and hat, etc. In the proposed framework, due to low video resolution used, only hair length and skin tone are chosen as the traits to describe the head area. Because they occupy larger areas than glasses and facial hairs and are detectable even in low resolution images. Of course, if high resolution video is available, more traits can be included.

Two categories are defined for the detection of hair length – short hair and long hair. Hair length is extracted in the head patch, using areas with foreground mask. Blob analysis based method [115] is used to detect the hair length (Figure 5.5). When the connected hair area occupies 80% of the head length, it is considered as long hair. Otherwise it is classified as short hair.

Figure 5.5: Examples of hair segmentation in head area.

Skin tone also has two categories – bright skin and dark skin. In the head patch, areas excluded from hair area and with foreground mask are considered as face area. Color histogram is extracted from face area and classified into bright skin and dark skin. Note that the extraction of skin tone is highly depend on the segmentation

of face area, therefore the detection rate is relatively low. Future work may exploit other segmentation method and other color spaces for better performance. Some examples of the traits in head area are shown in Figure 5.6.

Figure 5.6: Examples of soft-biometric traits in head area.

Torso Area

More traits are extracted from the torso area because usually this area provides more information than the other two areas. Four traits are used in the torso area in the proposed framework, including clothes color, clothes pattern, sleeve length and bag carrying.

Color is the most common used feature in computer vision due to its good ability to differentiate. However, images in disjointed surveillance cameras suffer from two aspects: (1) different illumination conditions; (2) cameras with different specifications or brands. The same color may appear differently in different cameras. To make the detection more robust, predefined color categories is used instead of color histograms in the conventional way. Colors of clothes are classified into nine common categories – black, gray, white, red, green, blue, cyan, magenta and yellow. Pixels in the torso area with foreground mask are considered belonging to the body. A 9-bin color histogram is formed in HSV color space (Figure 5.7) from these pixels. Color category with the maximum vote is considered as the dominant color and is used to represent the color of the clothes.

As often used by human when describing a people from distance, clothes pattern is used as a semantic feature in the proposed framework. There are many categories to describe the pattern of clothes, but it is not practical to use all the categories in the framework. Thus the most common three patterns – plain color, horizontal stripe and vertical stripe are chosen, the rest is categorized as complicate pattern. Therefore, there are four categories in total. Histograms of magnitude and orientation of gradient are extracted from pixels with foreground masks. Plain color is decided only by the histogram of magnitude of gradient when pixels in the bin of

Figure 5.7: HSV color space.

low intensities are larger than 80% of the histogram. The rest categories are decided by both of the two histograms. Horizontal stripe is determined when pixels of horizontal orientation are more than 60% of the orientation histogram, similarly it is classified as vertical stripe when pixels of vertical orientation are more than 60%.

Another semantic feature used is the length of sleeves. The length of sleeve is extracted indirectly by detecting nude arms. Just two categories are defined – short sleeve and long sleeve. This is decided by whether large area of skin can be detected in the torso area. Skin can be detected by color since the skin color usually stays in a specific range in HSV color space [116]. Some examples of skin detection result are shown in Figure 5.8.

Figure 5.8: Examples of skin detection.

Bag carrying can also be considered as a semantic feature. There are some works exist to detect the bag carrying in pictures [117, 118, 119]. In this framework, a simple method detecting the vertical Hough lines on the edge of the torso is used, since straps of bag usually have different color with clothes and form vertical lines.

Some examples of the traits in torso area are shown in Figure 5.9.

Leg area

In the leg area, pant color and pant length are extracted as semantic features. For the extraction of pant color, similar method is used as in the extraction of clothes color. For the extraction of pant length, short pant and long pant are the defined categories, and similar method is used as in the extraction of sleeve length.

Some examples of the traits in leg area are shown in Figure 5.9.

Figure 5.9: Examples of traits extraction in torso and leg areas. For the torso area, first row of texts denotes the clothes color, second row denotes sleeve length, third row denotes clothes pattern, forth row denotes whether carrying bag. For the leg area, first row of texts denotes pant color, second row denotes pant length.

5.2.5 Combination of Semantic Features and Re-identification

A descriptor for a person f is a combination of semantic features extracted from three body parts

$$f = [v_1, \dots, v_t]^T \quad (5.8)$$

where v denotes a semantic feature and t denotes the number of features extracted. Most of existing datasets on people re-identification are consist of static images. That means the descriptor can only be extracted from a single image. However, it is highly possible that the image includes some unexpected conditions which would downgrade the re-identification performance. The descriptor f_{is} extracted from single frame can be represented as

$$f_{is} = f_i + \epsilon \quad (5.9)$$

where ϵ represents unpredicted error and f_i is descriptor from the real model. Descriptor f_{im} extracted from a set of frames can be represented as

$$f_{im} = f_i + \frac{1}{n} \sum_{k=1}^n \epsilon_k \quad (5.10)$$

where n denotes the number of frames used to construct the descriptor. It can be an arbitrary number which depends on different conditions. It is set to 5 in the experiment part of the proposed method. Use σ^2 to denote the variance of error. Suppose error ϵ follows normal distribution. The variance of error σ_{im}^2 in multi-frame based descriptor is reduced, compared with variance of error σ_{is}^2 in single frame based descriptor

$$\sigma_{im}^2 = \frac{1}{n}\sigma_{is}^2 = \frac{1}{n}\sigma^2 \quad (5.11)$$

Therefore, the descriptor constructed is based on the video instead of one static image. The video is first converted to a set of frames, then a set of descriptors is extracted based on frames. For a video, the final descriptor for the people is obtained by weighted combination of all descriptors from frames. In multi-frame based descriptor, each trait is constructed as

$$v_k = \operatorname{argmax}(v_{kc}) \quad (5.12)$$

where k represents the k_{th} semantic feature, and c is the number of categories of this feature predefined.

Nearest Neighbor (NN) matching strategy is used in people re-identification experiments when assigning IDs to targets in the probe set. The similarity of two descriptors is calculated as

$$s(f_i, f_j) = \sum_{k=1}^t \alpha_k h(v_{ik}, v_{jk}) \quad (5.13)$$

$h(v_{ik}, v_{jk})$ equals to one if categories of the trait from two descriptors are the same, otherwise it equals to zero. The weight α_k is calculated as the entropy of trait category distributions in real-life. In the experiment, we simulate the trait category distributions with part of the dataset to calculate α_k for each trait. Then for a target in the probe set, the ID is determined by the person in the gallery set with closest distance:

$$ID_j = \operatorname{argmax}_{i \in 1, \dots, N} s(g_i, p_j) \quad (5.14)$$

5.3 Datasets and Evaluation Criterion

5.3.1 Datasets

To evaluate the performance of the proposed framework, experiments are carried out on public datasets. There are several public datasets exist. Currently, the most commonly used datasets are VIPeR [66], iLIDS [120] and ETHZ [121], which consist of static image pairs for people. VIPeR and iLIDS contain static image pairs which are resized to 128×64 . Challenging conditions like illumination variations and occlusion are included. ETHZ dataset contains images of people cropped from videos taken by one moving camera. Multiple images of one person are taken from the same video sequence.

Datasets mentioned above only contain static images for people. However, these static image based datasets have their disadvantages. Considering unpredictable conditions often occur in a single image, such as sudden illumination change, occlusion by background object, or occlusion caused by movement of limbs, these will distract the description from the real model. On the other hand, constructing the model with a set of frames in a video can reduce the error. Therefore, the experiment of the proposed method is carried out on SAIVT-SoftBio database [122], a video based dataset recently published. This database consists of videos from eight cameras distributed inside a hotel. Scenarios of cameras are shown in Figure 5.10. For each person walking across the hotel, he or she may appear in one or several cameras. 152 people in total are included in this database.

Figure 5.10: Scenarios of surveillance cameras in SAIVT-SoftBio database.

5.3.2 Evaluation Criterion

To evaluate the re-identification performance, Cumulative Matching Characteristic (CMC) curve and Synthetic Recognition Rate (SRR) curve are employed. CMC curve records the probability that the right match can be found within the rank score. It is a cumulative curve that increases as the rank score increases. SRR curve records the probability that finds the right match at the first rank. This curve decreases as the number of targets increases. Area under the Curve (AUC) is provided for CMC curve as well.

In the open ID experiment, besides CMC curve, Ratio of Detection rate to False Positive rate (RDFP) curves and Ratio of Detection rate to False Acceptance Rate (RDFAR) curves are used. In this two curves, False Positive rate and Mis-Matches rate are considered. False Positive denotes outliers from probe which were falsely matched as one of subjects in gallery, while Mis-Matches denote non-outliers from probe which were assigned a false ID. False Acceptance Rate includes both False Positive rate and Mis-Match rate. Ratio of Detection rate to False Positive rate (RDFP) is defined as

$$RDFP = \frac{DetectionRate}{DetectionRate + FalsePositiveRate} \quad (5.15)$$

Ratio of Detection rate to False Acceptance Rate (RDFAR) is defined as

$$RDFAR = \frac{DetectionRate}{DetectionRate + FalsePositiveRate + MisMatchRate} \quad (5.16)$$

5.4 Experimental Results

There are five parts in the experiment. Comparison with some other methods on static image datasets is addressed in the first part. In the second part, the effects of descriptors constructed with multi-frames and single frame are investigated. In the third part, the performance of the proposed method on various viewing conditions is evaluated, followed by the evaluation of the proposed method without considering viewing angles. The last part discusses an open ID recognition problem.

5.4.1 Performance Evaluation on Static Image Datasets

First, we evaluate the performance of the proposed framework on some static image datasets and compare with other methods.

Comparison with Other Methods on VIPeR Dataset

In this part, the proposed method is tested on VIPeR dataset and compared with Symmetry Driven Accumulation of Local Features (SDALF) [67], Relative Distance Comparison method (RDC) [82] and Saliency Distance Comparison method (SDC) [83]. The codes for the comparing methods are provided by their authors. For RDC and SDC, 316 pairs of images are randomly selected as training images, the rest are used as testing images. While for SDALF and the proposed method, 316 pairs of images are randomly selected for the re-identification experiment. Figure 5.11 presents the CMC curves and Figure 5.12 presents the SRR curves of the proposed method along with other methods. As shown in the CMC curves, the proposed method achieves comparable performance with RDC, and outperforms SDALF and SDC. The proposed method also shows the best overall performance on the SRR curve.

Figure 5.11: Comparison on CMC curves between the proposed method and other methods on VIPeR dataset.

AUCs are list in Table 5.1. RDC receives the best performance, while the proposed method receives a close score with RDC. Some statistics at lower rank scores of CMC curves are shown in Table 5.2. SDC receives best performance on rank 1, 5 and 10, while the proposed method achieves best performance on rank 25 and 50.

Figure 5.12: Comparison on SRR curves between the proposed method and other methods on VIPeR dataset.

Table 5.1: Comparison of AUCs (in percentage) on VIPeR dataset, the best is shown in bold.

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
AUCs	87.85	91.97	88.39	91.83

Table 5.2: Comparison of recognition rate (in percentage) in lower rank scores on VIPeR dataset, the best is shown in bold.

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
Rank 1	13.54	17.32	24.05	20.68
Rank 5	32.82	35.69	42.72	38.64
Rank 10	45.95	45.39	50.63	49.15
Rank 25	66.87	62.68	69.62	70.17
Rank 50	84.05	76.78	79.75	86.78

Comparison with Other Methods on iLIDS Dataset

In this part, the proposed method is tested on iLIDS dataset and compared with SDALF [67], RDC [82] and SDC [83]. For RDC and SDC, images from half of 119 subjects are randomly selected as training images, the rest are used as testing images. For SDALF and the proposed method, images from randomly selected 60

subjects are used in the re-identification experiment. Figure 5.13 presents the CMC curves and Figure 5.14 presents the SRR curves of the proposed method and other methods. As shown in the both figures, SDC receives the best performance.

Figure 5.13: Comparison on CMC curves between the proposed method and other methods on iLIDS dataset.

Figure 5.14: Comparison on SRR curves between the proposed method and other methods on iLIDS dataset.

AUCs are list in Table 5.3. SDC receives the best performance. The proposed method outperforms SDALF and RDC on iLIDS dataset. Some statistics at lower rank scores of CMC curves are shown in Table 5.4. Overall, SDC receives best performance.

Table 5.3: Comparison of AUCs (in percentage) on iLIDS dataset, the best is shown in bold.

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
AUCs	81.27	83.07	87.28	85.83

Table 5.4: Comparison of recognition rate (in percentage) in lower rank scores on iLIDS dataset, the best is shown in bold.

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
Rank 1	30.51	24.33	42.17	34.24
Rank 5	55.59	51.67	62.33	58.47
Rank 10	68.14	63.17	75.33	71.02
Rank 20	85.25	79.17	84.50	83.56

As shown on the experimental results of the two static image datasets, the proposed method does not achieve overwhelming performance on these dataset. On the other hand, methods with supervised learning show their advantages on images pairs with similar viewing angles like in iLIDS dataset.

5.4.2 Single Frame vs Multi-frames for Human Appearance Model Construction

In this section, the performances of descriptors formed by using multi-frames and single frame are discussed. Videos from all cameras are used. For single frame based descriptors of each people, a random frame from a random camera which is available is chosen. For multi-frame based descriptors of each people, 5 frames are randomly selected from one video and a single descriptor is formed by combining descriptors from these 5 frames. The CMC curves are shown in Figure 5.15. Multi-frames based descriptors show superior performance than single frame based descriptors in people re-identification. The curves are obtained with 10 times of repeated experiments. AUC is shown in Table 5.5, it also proves that multi-frames based descriptor receives better performance. The SRR curves are shown in Figure 5.16. It is obtained from a 100 times repeated experiment. From the SRR curves we can observe, the performance of multi-frames based descriptors is slightly better than single frame

Figure 5.15: CMC curves of single frame based descriptor and multi-frame based descriptor.

Figure 5.16: SRR curves of single frame based descriptor and multi-frame based descriptor.

based descriptors. Therefore, we can conclude that using multiple frames based descriptors can improve the people re-identification performance.

Table 5.5: AUC of single frame based descriptor and multi-frame based descriptor.

Method	Single Frame	Multi-frame
AUCs	78.62	87.18

5.4.3 Evaluation of the Performance on Different View Angles

In this section, the performance of the proposed framework on different view angles is evaluated. Comparison with single features and other state-of-the-art methods is provided as well. The SAIVT-SoftBio database [122] provides videos from 8 cameras distributed in the hotel, which perfectly reflect view angle conditions in reality. Three conditions are analyzed in this part: cameras with similar views, cameras with opposite views and cameras with dissimilar views. Table 5.6 shows the view differences between each two cameras. In all these cameras, camera 2 captured too few subjects which are insufficient to be evaluated. Camera 4 suffered severe illumination condition, and camera 5 also has some illumination problem (Figure 5.17).

Table 5.6: View differences between each pair of cameras in SAIVT-SoftBio database, where s represents similar views, o represents opposite views and d represents dissimilar views.

Cam No.	1	2	3	4	5	6	7	8
1	-	d	o	d	d	d	d	o
2	d	-	d	d	o	o	o	d
3	o	d	-	d	d	d	d	s
4	d	d	d	-	d	d	d	d
5	d	o	d	d	-	s	s	d
6	d	o	d	d	s	-	s	d
7	d	o	d	d	s	s	-	d
8	o	d	s	d	d	d	d	-

For each condition, the performances of semantic features combined descriptor with descriptors using single feature are first compared. For descriptors using single semantic feature, continuous value without quantization is employed. Skin tone and Bag carrying are not included in the comparison. Bag carrying is a binary trait and not able to be represented by continuous value. The successful extraction of skin tone is quite low compared with other traits. Methods used for comparison includes SDALF [67], RDC [82] and SDC [83]. For existing methods, we use codes provided by authors for evaluation.

Figure 5.17: Severe illumination conditions. The first two images are from one target and the last two are from another target. As can be seen, target in camera 5 suffers bright lighting condition, while target in camera 4 suffers darkness.

Condition 1: Similar Views

Firstly, the effect of similar viewing angles is evaluated. In cameras with similar viewing angles, people captured usually have close angles of view, but other variations like scale and illumination may exist. Among the 8 cameras, there are two sets of similar view cameras. One set contains camera 3 and 8, the other set contains camera 5, 6 and 7.

Figure 5.18: CMC curves of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views (Camera 3 vs Camera 8).

The performances of multiple semantic features combined descriptor with single feature based descriptor are first compared. Comparison on the CMC and SRR curves between camera 3 and camera 8 are shown in Figure 5.18 and Figure 5.19, respectively. The AUCs are listed in Table 5.7. Experiments are carried out on videos from a pair of two cameras with similar views. As can be seen in Figure 5.18, the combined descriptor achieves the best performance in CMC curves. In SRR

Figure 5.19: SRR curves of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views (Camera 3 vs Camera 8).

Table 5.7: AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views (Camera 3 vs Camera 8).

Features	AUCs(cam 3 and 8)
Combined	84.16
Hair Length	53.05
Pant color	62.67
Clothes color	65.60
Pant length	55.13
Sleeve length	60.10
Clothes Pattern	73.38

curves (Figure 5.18), the combined descriptor is performing better than most of single feature based descriptors. Only the descriptor of hair length achieves higher recognition rate on the SRR curve, this shows the discriminative of using hair length when targets are few, it will decrease drastically as the number of targets increases which leads to the poor performance in CMC curve. Overall, multiple semantic features combined descriptor outperforms descriptors extracted from single feature. This supports the view that, though single semantic feature is not discriminative, a set of these features can do well in identification. **For experimental result**

between camera 5 and camera 7, please refer to Appendix A on page 116-117, Figures A.1-A.2 and Tables A.1.

Figure 5.20: CMC curves of proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).

Figure 5.21: SRR curves of proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).

The comparison of the proposed method with existing methods between camera 3 and camera 8 are shown in Figure 5.20 and Figure 5.21. The AUCs are listed in Table 5.8. Note that, the number of targets used in SRR curve might be less than 25 for RDC and SDC, this is because training is needed for RDC and SDC. Samples

Table 5.8: AUCs of proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
AUCs(cam 3 vs cam 8)	76.90	83.81	77.20	84.16

are divided into two groups, and half of the samples are used for training. Thus samples being used for test are less than 25. Overall, the result of the proposed method has superior performance over the other three methods.

For experimental result between camera 5 and camera 7, please refer to Appendix A on page 117-118, Figures A.3-A.4 and Tables A.2. The performance of experiment on camera 5 and 7 is inferior to experiment on camera 3 and 8. This is due to the poor illumination condition in camera 5, which deteriorates the re-identification performance.

Condition 2: Opposite Views

Figure 5.22: CMC curves of multiple soft-biometric traits based descriptor and single feature based descriptors with opposite views (Camera 1 vs Camera 3,8).

There are two sets of opposite cameras: (1) camera 1 vs camera 3 and 8; (2) camera 2 vs camera 5, 6, 7. However, due to insufficient subjects collected from camera 2, only the set camera 1 vs camera 3 and 8 is tested in the experiment. With opposite views, people re-identification may face more problems than in similar

viewing angles. Because even for the same person, some of the traits appear in the front view and the back view could be totally different. For example, there is a people wearing a white T-shirt and carrying a black backpack. From the front view, the clothes color is extracted as white, but from the back view, black may be extracted instead as the clothes color.

Figure 5.23: SRR curves of multiple soft-biometric traits based descriptor and single feature based descriptors with opposite views (Camera 1 vs Camera 3,8).

Table 5.9: AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with opposite views (Camera 1 vs Camera 3,8).

Features	AUCs(cam 1 vs cam 3 and 8)
Combined	78.96
Hair Length	57.80
Pant color	58.01
Clothes color	60.22
Pant length	59.73
Sleeve length	68.21
Clothes Pattern	66.65

Similar with the previous condition, performances of semantic features combined descriptor and descriptors of single semantic feature are first compared. The CMC

Figure 5.24: CMC curves of proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).

Figure 5.25: SRR curves of proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).

Table 5.10: AUCs of proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
AUCs(cam 1 vs cam 3 and 8)	75.45	73.32	71.44	78.96

and SRR curves are shown in Figure 5.22 and Figure 5.23, respectively. The AUCs are listed in Table 5.9. From the figures and table, it can be concluded that the combined descriptor outperforms descriptors of single feature in people re-identification of opposite views.

Then the proposed method is compared with existing methods on opposite viewing angles. The comparison on the CMC and SRR curves are shown in Figure 5.24 and Figure 5.25, while the AUCs are listed in Table 5.10. It can be observed that the proposed method is performing better. Compared with the performance of all methods under similar views (camera 3 and 8), the AUCs of all methods declined under opposite views. This denotes that people re-identification is more challenging in opposite views than in similar views.

Condition 3: Dissimilar Views

There are more combinations for cameras with dissimilar views. Considering the views of angle are different, the performance could be inferior to that of similar views.

Figure 5.26: CMC curves of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views (Camera 3,8 vs Camera 5,6,7).

As the same with previous conditions, semantic features combined descriptor is compared with single feature based descriptors. The CMC and SRR curves between cameras 3,8 and cameras 5,6,7 are shown in Figure 5.26 and Figure 5.27, respectively. The AUCs are listed in Table 5.11. **For more experimental results on dissimilar views, please refer to Appendix A on page 119-120, Figures A.5-A.4 and Tables A.3.**

Figure 5.27: SRR curves of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views (Camera 3,8 vs Camera 5,6,7).

Table 5.11: AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views (Camera 3,8 vs Camera 5,6,7).

Features	AUCs(cam 3,8 vs cam 5,6,7)
Combined	79.10
Hair Length	55.73
Pant color	58.69
Clothes color	64.80
Pant length	57.64
Sleeve length	61.36
Clothes Pattern	71.77

Table 5.12: AUCs of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 3,8 vs Camera 5,6,7).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
AUCs(cam 3,8 vs cam 5,6,7)	72.07	68.54	71.94	79.10

Figure 5.28: CMC curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 3,8 vs Camera 5,6,7).

Figure 5.29: SRR curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 3,8 vs Camera 5,6,7).

Then the proposed method is compared with existing methods on dissimilar viewing angles between cameras 3,8 and cameras 5,6,7. The comparisons on the CMC and SRR curves are shown in Figure 5.28 and Figure 5.29. And the AUCs are listed in Table 5.12. Again, as can be observed, the proposed method is performing better. **For more experimental results in comparison with other methods, please refer to Appendix A on page 120-121, Figures A.7-A.8 and Tables A.4.**

Discussion

Comparing the performance in three different viewing conditions, it is evident that experiments on similar views have the best re-identification performance. In experiments on opposite views and dissimilar views, the performance of all methods declines. It can be concluded that the viewing angle of camera is very important in people re-identification. Large variation in views has great impact on the performance. Among the four methods, RDC declines the most from similar views to other viewing conditions. RDC and SDC are receiving inferior performance on opposite views and dissimilar views. This may indicate that learning based methods performance largely depends on the viewing angles, especially when the sample data are insufficient.

5.4.4 Comparison with Other Methods on Multi-views

People re-identifications in previous section are based on images from two cameras. What if the images in the gallery and probe are both from different viewing angles? In this section, evaluation of the performance of the proposed method under this condition is provided and compared with some existing methods. In this section, images of gallery and probe sets are randomly chosen from several cameras, without considering the viewing angles. From Figure 5.30, Figure 5.31 and Table 5.13, it is observed that with more view angles involved, the performance of the proposed method is still better than the other three methods.

Figure 5.30: CMC curves of proposed method, SDALF, RDC and SDC with multiple views.

Figure 5.31: SRR curves of proposed method, SDALF, RDC and SDC with multiple views.

Table 5.13: AUCs of proposed method, SDALF, RDC and SDC with multiple views.

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
AUCs(all cameras)	73.45	64.03	69.48	78.59

5.4.5 Evaluation on Open ID Recognition

Currently, most existing people re-identification experiments are based on gallery and probe sets containing exactly the same subjects. That means for each image in the probe set, there must be at least one image in the gallery set belonging to the same ID. This is defined as closed set experiment [123]. However, what faced in reality is an open set problem [123], this means some people appear in the probe set are never included in the gallery set, vice versa.

The way to evaluate the performance in open set experiment is different with the closed set experiment. For the closed set experiment, the matching accuracy is the only index matters. However, for open set problems, besides matching accuracy, we have to care about the detection of outliers, which denotes the rate of mismatching for those images appear in only probe set.

Usually in the closed set experiment, the distance of two descriptors is measured to represent the similarity of two subjects. For one probe subject, the subject in gallery with the closest distance is picked as the matching ID. This method is not

suitable for open set problem. Because in an open set problem, for a subject appears only in probe set, there is no right matching in the gallery set. A way is needed to judge whether the subject is an outlier or not. In open set experiment, a threshold could be set for the distance of two descriptors to decide outliers. The matching is confirmed only if the similarity of two subjects is larger than the threshold τ

$$s(g_i, p_j) > \tau \quad (5.17)$$

In this section, performances of the proposed method with existing methods SDALF [67], RDC [82] and SDC [83] are compared on open ID experiment. The CMC curves are shown in Figures 5.32, 5.35 and 5.38, Ratio of Detection rate to False Positive rate (RDFP) curves are shown in Figures 5.33, 5.36 and 5.39, and Ratio of Detection rate to False Acceptance Rate (RDFAR) curves are shown in Figures 5.34, 5.37 and 5.40.

Three scenarios—similar views, opposite views and dissimilar views are considered. As can be seen from CMC curves (Figure 5.32, 5.35 and 5.38), compared with closed ID experiment, the performances of all methods drop in open ID experiment. This is due to the distraction of unmatched samples. In open ID experiment the detection of outliers is more cared about, which is not reflected in CMC curve. Therefore, the RDFP and RDFAR are used. Since a fixed threshold is not desirable for different camera combinations, a varying threshold is used. In the experiment (Figures 5.33, 5.34, 5.36, 5.37, 5.39 and 5.40), the threshold varies from the minimum value to the maximum value of similarity scores. 100 values are set for the varying threshold.

Table 5.14: Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
RDFP(cam 5 vs cam 7)	50.00	20.00	14.51	22.22

Table 5.15: Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
RDFAR(cam 5 vs cam 7)	6.18	6.06	7.22	12.80

Figure 5.32: CMC curves of proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).

Figure 5.33: RDFP curves of proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).

Table 5.16: Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
RDFP(cam 1 vs cam 3 and 8)	19.05	57.14	20.53	37.50

Figure 5.34: RDFAR curves of proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).

Figure 5.35: CMC curves of proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).

Table 5.17: Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
RDFAR(cam 1 vs cam 3 and 8)	4.88	15.15	6.81	8.26

Figure 5.36: RDFP curves of proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).

Figure 5.37: RDFAR curves of proposed method, SDALF, RDC and SDC with opposite views (Camera 1 vs Camera 3,8).

Table 5.18: Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
RDFP(cam 4 vs cam 5,6,7)	16.47	50.00	14.00	45.45

Figure 5.38: CMC curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).

Figure 5.39: RDFP curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).

Table 5.19: Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
RDFAR(cam 4 vs cam 5,6,7)	3.98	9.26	2.17	8.13

Figure 5.40: RDFAR curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).

For more experimental results on Open ID recognition, please refer to Appendix B on page 122-121, Figures B.1-B.4 and Tables B.1-B.4.

As can be observed from Figures 5.32-5.40, RDC does not have good performance on CMC curve. However, its performance on RDFP and RDFAR is the best most of the time. This denotes RDC has good performance on the detection of outliers. RDC is a supervised learning based method. With learning progress, it gives low similarity score to outliers. Therefore the false positive, which is not preferred in open ID experiment, is kept low even though the detection rate of RDC is low. Meanwhile, SDC does not receive good performance on RDFP and RDFAR, the reason is that SDC depends on finding the salient features of the target. However, in SAIVT-SoftBio database, appearances of most subjects are not salient to others. This causes the low detection rate for SDC even training is involved.

Another reason contributes to the superior performance of RDC in RDFP and RDFAR is that, subjects used in testing are less than subjects used in SDALF and the proposed method, due to the training data needed. With less subjects involved in experiment, the ratio of detection rate is supposed to be higher. Therefore, the proposed method is more proper to compare with SDALF, which is under the same experiment condition. As shown in Figures 5.32-5.40 and Tables 5.14-5.19, the proposed method performs better than SDALF. Even though, the proposed method still faces the problem of excluding outliers when performing re-identification in an open ID experiment.

5.5 Concluding Remarks

In this paper, a framework is proposed to extract semantic features based people descriptors from videos. This framework is designed to solve people re-identification problem in disjoint camera based surveillance systems. People detection, body segmentation and extraction of soft-biometric traits are included in the framework. The performance of the extracted descriptors is evaluated on several aspects and comparisons with state-of-the-art methods are provided. The experimental results show the good performances of the proposed method under conditions of different view angles, illumination change and pose variations. However, in open ID experiment, it faces the problem of excluding outliers, which should be considered and solved in future work.

Chapter 6

Conclusion and Recommendations for Future Work

6.1 Conclusion

In this thesis, we have presented our research on multi-people tracking and people description, especially under cluttered environment. In this section, main contributions in our research are summarized as follows:

1. A key-points based tracking method is proposed to perform long-term tracking, aiming to deal with partial occlusion problems. In the framework we proposed, SURF key-points are used to represent the object. The evaluation of the tracking result and detection are based on recognizing key-points inside. Key-points are evaluated and on-line learned by Random Ferns. Object model and surrounding background model are proposed to store key-points from object and surrounding background, respectively. By tracing object points along with points in surrounding background, more accurate position can be obtained from short-term tracking and the learning of Random Ferns can be improved.
2. In this research, regions that highly contrast to local neighborhood are included in addition to the global appearance in the construction of target model. These salient regions will be tracked simultaneously along with target, and provide relative spatial information with the whole body when determining the location of people. Salient regions inside the target patch can be considered as representative to the target, thus can be used for short-term tracking. Besides, context information can always provide useful information for tracking, which is not aware of in many existing methods.

Supporting model is constructed to indicate the spatial relationship between salient parts and target. The supporting model is on-line updated with a forgetting factor. Occlusion reasoning is performed to detect those salient parts being occluded. Parts with occlusion labels are excluded in update to prevent drift of the model.

3. A method extracting semantic people description from surveillance cameras is exploited. By segmenting the human body and extracting a set of appearance based traits from each segmentation area, a combined people descriptor can be formed. With semantic feature, the descriptor is more invariant compared with low level features. No training or camera calibration is needed beforehand. As simple methods are developed for the extraction of soft-biometric traits, the proposed method is able to achieve real-time processing on videos. In addition, it is also able to handle the scale change and incomplete appearance. Unlike other methods extracting descriptors from well cropped and resized images, the proposed method can extract descriptors from images with different scales, even with some rotations. Furthermore, descriptors of partially visible people can also be extracted, which is implausible in other state-of-the-art methods. In the experiment of People Re-identification, the semantic feature based descriptor shows promising performance under conditions of different view angles, illumination change and pose variations.

6.2 Recommendations for Future Works

Some potential extensions for future work are discussed in this section.

6.2.1 Exploit Advanced Method for Salient Part Extraction

It has been proven in Chapter 4 that salient parts can be used to assist people tracking. Therefore, the extraction of salient parts would be important since these parts could be more representative if they are carefully selected. In the multi-people tracking method proposed so far, we only exploit color and orientation as features for saliency detection and local contrast method is used to calculate the saliency score. It is a relative easy and intuitive way to find salient regions. As decades of development, there are already some advanced salient region detection methods proposed. For example, Hou [124] exploited the visual saliency in the spectral domain instead of feature domains in other methods. Gopalakrishnan [125] considered the distribution of colors and orientations in both local and global scales for saliency detection. A hierarchical model is proposed in [126] to handle multi-scale

problem in saliency detection. Cheng [127] evaluated the global contrast differences with histogram based contrast and spatial coherence for each region segments. In our future work, more feature spaces or feature combinations can be exploited for better performance on salient part extraction.

6.2.2 Salient Part Extraction with Adaptive Shape

Salient regions used in this research are square boxes with fixed size. But in real scenes, the salient part could be any shapes. Using only square parts might not be the optimal choice for short-term tracking. Besides, using pre-defined windows size may not fit the salient part well. In Luo's work [128], a bounding box of Maximum Saliency Density (MSD) can be provided that fit the shape and size of the salient object. An efficient sub-window searching method is employed by considering both the size and saliency density of the bounding box. In our future work, techniques like MSD will be exploited to fit the irregular shapes of salient parts with adaptive window size.

6.2.3 Extension of Semantic Feature Extraction

The proposed people descriptor faces the problem of excluding outliers in open ID experiment which should be considered and solved in future work. In future, more semantic features should be included into the framework to achieve more accurate description. As listed in Figure 2.22, many semantic features included but not limited to bag carrying, skin tone and body shape, these could be considered in the proposed framework. Besides, extension work on existing semantic features can be done in our future work, more detailed categories can be defined for more accurate description.

6.2.4 Multi-layer People Descriptor

In our future work, a multi-layer descriptor can be exploited. In addition to existing semantic features, other semantic features requiring good resolution (wearing of cap, wearing of glasses and facial hair, etc.) can be extracted when resolution condition is satisfied. Besides, semantic feature based descriptor can be used for rough re-identification, while low level features can be included to provide more specific description. With the help of multi-layer description, it may improve the performance of outlier detecting, thus a better performance can be expected in open ID experiment.

6.2.5 Application of People Descriptor

People Descriptor proposed in this research work could be used in applications for people tracking in future work. For instance, it can be used for ID re-identification as in section 4.2.9, which is actually an open ID problem with few variation in view angles. Besides, it can be used in people tracking in surveillance system consists of disjoint cameras. The people descriptor could play a key-role in re-identifying the target if the target across areas covered by different cameras without overlap. With the assistant of people descriptor, the trajectories of the same target in different areas can be linked together to provide more consistent information.

Appendix A

More Experimental Results on Different View Angles in Section 5.4.3

Similar View (Camera 5 vs Camera 7)

Figure A.1: CMC curves of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views (Camera 5 vs Camera 7).

Figure A.2: SRR curves of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views (Camera 5 vs Camera 7).

Table A.1: AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views (Camera 5 vs Camera 7).

Features	AUCs(cam 5 and 7)
Combined	79.93
Hair Length	64.38
Pant color	56.23
Clothes color	62.39
Pant length	54.65
Sleeve length	61.24
Clothes Pattern	76.72

Table A.2: AUCs of proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
AUCs(cam 5 vs cam 7)	69.14	72.09	73.25	79.93

Figure A.3: CMC curves of proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).

Figure A.4: SRR curves of proposed method, SDALF, RDC and SDC with similar views (Camera 5 vs Camera 7).

Dissimilar View (Camera 4 vs Camera 5,6,7)

Figure A.5: CMC curves of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views (Camera 4 vs Camera 5,6,7).

Figure A.6: SRR curves of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views (Camera 4 vs Camera 5,6,7).

Table A.3: AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views (Camera 4 vs Camera 5,6,7).

Features	AUCs(cam 4 vs cam 5,6,7)
Combined	70.85
Hair Length	53.17
Pant color	52.69
Clothes color	55.36
Pant length	51.07
Sleeve length	59.76
Clothes Pattern	59.39

Figure A.7: CMC curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).

Table A.4: AUCs of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
AUCs(cam 4 vs 5,6,7)	62.17	61.50	58.18	70.85

Figure A.8: SRR curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 4 vs Camera 5,6,7).

Appendix B

More Experimental Results on Open ID in Section 5.4.5

Similar View (Camera 3 vs Camera 8)

Figure B.1: CMC curves of proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).

Table B.1: Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
RDFP(cam 3 vs cam 8)	50.00	100.00	9.79	33.33

Figure B.2: RDFP curves of proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).

Figure B.3: RDFAR curves of proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).

Table B.2: Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with similar views (Camera 3 vs Camera 8).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
RDFAR(cam 3 vs cam 8)	7.25	12.50	7.71	9.09

Dissimilar View (Camera 1 vs Camera 5,6,7)

Figure B.4: CMC curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 1 vs Camera 5,6,7).

Figure B.5: RDFP curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 1 vs Camera 5,6,7).

Figure B.6: RDFAR curves of proposed method, SDALF, RDC and SDC with dissimilar views (Camera 1 vs Camera 5,6,7).

Table B.3: Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with dissimilar views (Camera 1 vs Camera 5,6,7).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
RDFP(cam 1 vs 5,6,7)	16.47	50.00	14.00	45.45

Table B.4: Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with dissimilar views (Camera 1 vs Camera 5,6,7).

Method	SDALF[67]	RDC[82]	SDC[83]	Proposed Method
RDFAR(cam 1 vs 5,6,7)	3.98	9.26	2.17	8.13

Author's Publications

Journal Paper:

- [1] Z. Zhou, Y. Wang and E. K. Teoh. "A framework for soft-biometric based people description in videos", in *Image and Visual Computing (IVC)*, vol. 46, pp. 29-46, Feb 2016.

Journal Paper Manuscript:

- [1] Z. Zhou, Y. Wang and E. K. Teoh. "Salient Parts Model for Multi-people Tracking", *Submitted to International Journal of Pattern Recognition and Artificial Intelligence* in February 2016.

Conference Papers:

- [1] Z. Zhou, Y. Wang and E. K. Teoh, "Robust Object Tracking Using Bi-model", *IEEE International Conference on Image Processing (ICIP)*, Melbourne, Australia, 15-18 Sep 2013, pp. 3103-3107.
- [2] Z. Zhou, Y. Wang and E. K. Teoh, "People apparel model for human appearance description", *IEEE 9th International Conference on Information, Communications and Signal Processing (ICICSP)*, Tainan, Taiwan, Dec 10-13, 2013, pp. 1-5.
- [3] Y. Wang, Z. Zhou, and E. K. Teoh, "Human hair segmentation and length detection for human appearance model", *IEEE 22th International Conference on Pattern Recognition, (ICPR)*, Stockholm, Sweden, 2014, pp. 450-454.

- [4] Z. Zhou, H. T. Ong and E. K. Teoh, "Soft-biometric detection based on supervised learning", *13th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Singapore, 10-12 Dec 2014, pp. 234-238.
- [5] Z. Zhou, Y. Wang and E. K. Teoh, "People re-identification based on bags of semantic features", *Workshop on Human Identification for Surveillance in 12th Asian Conference on Computer Vision (ACCV 2014)*, Singapore, 1-4 Nov 2014.
- [6] Z. Zhou, Y. Wang and E. K. Teoh, "Salient parts based multi-people tracking", *10th International Conference on Computer Vision Theory and Applications (VISAPP 2015)*, Berlin, Germany, 10-14 Mar 2015.
- [7] Z. Zhou, Y. Wang and E. K. Teoh, "Double Layer Salient Parts based Multi-people Tracking", *IEEE International Conference on Image Processing (ICIP)*, Quebec City, Canada, 27-30 Sep 2015.

Bibliography

- [1] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 1999, pp. 1150–1157.
- [2] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 404–417, 2006.
- [3] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893.
- [4] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.
- [5] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2001, pp. I–511.
- [6] M. J. Black and A. D. Jepson, “Eigentracking: Robust matching and tracking of articulated objects using a view-based representation,” *International Journal of Computer Vision*, vol. 26, no. 1, pp. 63–84, 1998.
- [7] J. Kang, I. Cohen, and G. Medioni, “Object reacquisition using invariant appearance model,” in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, 2004, pp. 759–762.
- [8] S. Avidan, “Ensemble tracking,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 261–271, 2007.

- [9] H. Grabner and H. Bischof, “On-line boosting and vision,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2006, pp. 260–267.
- [10] B. Babenko, M. H. Yang, and S. Belongie, “Visual tracking with online multiple instance learning,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, 2009.
- [11] Z. Kalal, J. Matas, and K. Mikolajczyk, “Pn learning: Bootstrapping binary classifiers by structural constraints,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 49–56.
- [12] H. Grabner, J. Matas, L. Van Gool, and P. Cattin, “Tracking the invisible: Learning where the object might be,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1285–1292.
- [13] T. B. Dinh, N. Vo, and G. Medioni, “Context tracker: Exploring supporters and distracters in unconstrained environments,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1177–1184.
- [14] M. Andriluka, S. Roth, and B. Schiele, “People-tracking-by-detection and people-detection-by-tracking,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.
- [15] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, “On-line multiperson tracking-by-detection from a single, uncalibrated camera,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820–1833, 2011.
- [16] B. Benfold and I. Reid, “Stable multi-target tracking in real-time surveillance video,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3457–3464.
- [17] A. Andriyenko and K. Schindler, “Multi-target tracking by continuous energy minimization,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1265–1272.
- [18] A. Segal and I. Reid, “Latent data association: Bayesian model selection for multi-target tracking,” in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 2904–2911.

- [19] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multi-target tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, 2014.
- [20] H. Izadinia, I. Saleemi, W. Li, and M. Shah, "2t: Multiple people multiple parts tracker," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2012, pp. 100–114.
- [21] B. Yang and R. Nevatia, "Online learned discriminative part-based appearance models for multi-human tracking," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2012, pp. 484–498.
- [22] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *Workshop on Applications of Computer Vision (WACV)*, 2009, pp. 1–8.
- [23] R. Layne, T. M. Hospedales, S. Gong *et al.*, "Person re-identification by attributes," in *Proc. British Machine Vision Conference (BMVC)*, vol. 2, 2012, p. 3.
- [24] R. Layne, T. M. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," in *ECCV Workshops and Demonstrations*, 2012, pp. 402–412.
- [25] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys (CSUR)*, vol. 38, no. 4, p. 13, 2006.
- [26] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981.
- [27] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2000, pp. 142–149.
- [28] A. Wedel, D. Cremers, T. Pock, and H. Bischof, "Structure-and motion-adaptive regularization for high accuracy optic flow," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 1663–1668.
- [29] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *Proceedings of 20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 2756–2759.

- [30] V. Lepetit, P. Lager, and P. Fua, “Randomized trees for real-time keypoint recognition,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2005, pp. 775–781.
- [31] D. Ramanan, D. A. Forsyth, and A. Zisserman, “Strike a pose: Tracking people by finding stylized poses,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 271–278.
- [32] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” in *Computational Learning Theory*, 1995, pp. 23–37.
- [33] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society, Series B (Methodological)*, pp. 1–38, 1977.
- [34] G. Nebehay and R. Pflugfelder, “Consensus-based matching and tracking of keypoints for object tracking,” in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014, pp. 862–869.
- [35] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua, “Fast keypoint recognition using random ferns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 448–61, 2010.
- [36] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [37] S. Hare, A. Saffari, and P. H. Torr, “Efficient online structured output learning for keypoint-based object tracking,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1894–1901.
- [38] S. Yang, J. Xu, Y. Chen, and M. Wang, “On-road vehicle tracking using keypoint-based representation and online co-training,” *Multimedia Tools and Applications*, vol. 72, no. 2, pp. 1561–1583, 2014.
- [39] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [40] H. Wu, G. Li, and X. Luo, “Weighted attentional blocks for probabilistic object tracking,” *The Visual Computer*, vol. 30, no. 2, pp. 229–243, 2014.

- [41] M. N. Ali, M. Abdullah-Al-Wadud, and S.-L. Lee, "Multiple object tracking with partial occlusion handling using salient feature points," *Information Sciences*, vol. 278, pp. 448–465, 2014.
- [42] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1007–1013.
- [43] K. M. Yi, H. Jeong, B. Heo, H. J. Chang, and J. Y. Choi, "Initialization-insensitive visual tracking through voting with salient local features," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 2912–2919.
- [44] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.
- [45] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems*, 2002, pp. 561–568.
- [46] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, 2008.
- [47] H. Jiang, S. Fels, and J. Little, "A linear programming approach for multiple object tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–8.
- [48] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.
- [49] H. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
- [50] J. Zhang, L. Presti, and S. Sclaroff, "Online multi-person tracking by tracker hierarchy," in *AVSS*, 2012, pp. 379–385.
- [51] W. Ge and R. Collins, "Multi-target data association by tracklets with unsupervised parameter estimation." in *Proc. British Machine Vision Conference (BMVC)*, vol. 2, 2008, p. 5.

- [52] C. Tomasi and T. Kanade, *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.
- [53] Z. Wu, A. Thangali, S. Sclaroff, and M. Betke, “Coupling detection and data association for multiple object tracking,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1948–1955.
- [54] S. Bae and K. Yoon, “Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1218–1225.
- [55] C. Kuo, C. Huang, and R. Nevatia, “Multi-target tracking by on-line learned discriminative appearance models,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 685–692.
- [56] B. Yang and R. Nevatia, “An online learned crf model for multi-target tracking,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2034–2041.
- [57] H. Pirsiavash, D. Ramanan, and C. Fowlkes, “Globally-optimal greedy algorithms for tracking a variable number of objects,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1201–1208.
- [58] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, “Multiple object tracking using k-shortest paths optimization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1806–1819, 2011.
- [59] B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua, “Multi-commodity network flow for tracking multiple people,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1614–1627, 2013.
- [60] B. Wu and R. Nevatia, “Tracking of multiple, partially occluded humans based on static body part detection,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2006, pp. 951–958.
- [61] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, “Tracking across multiple cameras with disjoint views,” in *Proceedings of Ninth IEEE International Conference on Computer Vision (ICCV)*, 2003, pp. 952–957.

- [62] N. D. Bird, O. Masoud, N. P. Papanikolopoulos, and A. Isaacs, "Detection of loitering individuals in public transportation areas," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 167–177, 2005.
- [63] J.-H. Kao, C.-Y. Lin, W.-H. Wang, and Y.-T. Wu, "A unified hierarchical appearance model for people re-identification using multi-view vision sensors," in *Advances in Multimedia Information Processing-PCM*, 2008, pp. 553–562.
- [64] C. Madden, E. D. Cheng, and M. Piccardi, "Tracking people across disjoint camera views by an illumination-tolerant appearance representation," *Machine Vision and Applications*, vol. 18, no. 3-4, pp. 233–247, 2007.
- [65] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions." in *Proc. British Machine Vision Conference (BMVC)*, vol. 8, 2008, pp. 164–1.
- [66] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *IEEE International workshop on performance evaluation of tracking and surveillance*, 2007.
- [67] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2360–2367.
- [68] G. Berdugo, O. Soceanu, Y. Moshe, D. Rudoy, and I. Dvir, "Object reidentification in real world scenarios across multiple non-overlapping cameras," in *Proc. Euro. Sig. Proc. Conf*, 2010, pp. 1806–1810.
- [69] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *ECCV Workshops and Demonstrations*, 2012, pp. 413–422.
- [70] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013.
- [71] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 1528–1535.

- [72] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, “Shape and appearance context modeling,” in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2007.
- [73] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, “Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences,” in *Second ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, 2008, pp. 1–6.
- [74] W.-S. Zheng, S. Gong, and T. Xiang, “Associating groups of people,” in *Proc. British Machine Vision Conference (BMVC)*, 2009.
- [75] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, “Person re-identification using haar-based and dcd-based signature,” in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2010, pp. 1–8.
- [76] —, “Boosted human re-identification using riemannian manifolds,” *Image and Vision Computing*, vol. 30, no. 6, pp. 443–452, 2012.
- [77] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in *Proc. European Conference on Computer Vision (ECCV)*, 2008, pp. 262–275.
- [78] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, “Person re-identification by support vector ranking,” in *Proc. British Machine Vision Conference (BMVC)*, vol. 1, no. 3, 2010, p. 5.
- [79] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch, “Learning implicit transfer for person re-identification,” in *ECCV Workshops and Demonstrations*, 2012, pp. 381–390.
- [80] Y. Brand, T. Avraham, and M. Lindenbaum, “Transitive re-identification,” in *Proc. British Machine Vision Conference (BMVC)*, 2013.
- [81] S. Wang, M. Lewandowski, J. Annesley, and J. Orwell, “Re-identification of pedestrians with variable occlusion and scale,” in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 1876–1882.
- [82] W.-S. Zheng, S. Gong, and T. Xiang, “Reidentification by relative distance comparison,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, 2013.

- [83] R. Zhao, W. Ouyang, and X. Wang, “Unsupervised salience learning for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3586–3593.
- [84] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 152–159.
- [85] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, “Local fisher discriminant analysis for pedestrian re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3318–3325.
- [86] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, “Person re-identification by regularized smoothing kiss metric learning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 10, pp. 1675–1685, 2013.
- [87] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, “Large scale metric learning from equivalence constraints,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2288–2295.
- [88] A. K. Jain, S. C. Dass, and K. Nandakumar, “Soft biometric traits for personal recognition systems,” in *Biometric Authentication*, 2004, pp. 731–738.
- [89] —, “Can soft biometric traits assist user recognition?” in *Defense and Security*. International Society for Optics and Photonics, 2004, pp. 561–572.
- [90] A. Dantcheva, C. Velardo, A. Dangelo, and J.-L. Dugelay, “Bag of soft biometrics for person identification,” *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 739–777, 2011.
- [91] S. Samangooei, B. Guo, and M. S. Nixon, “The use of semantic human description as a soft biometric,” in *2nd IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2008, pp. 1–7.
- [92] G. Shakhnarovich, P. A. Viola, and B. Moghaddam, “A unified learning framework for real time face detection and classification,” in *Proceedings. Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002, pp. 14–21.
- [93] S. Denman, C. Fookes, A. Bialkowski, and S. Sridharan, “Soft-biometrics: unconstrained authentication in a surveillance environment,” in *Digital Image Computing: Techniques and Applications (DICTA)*, 2009, pp. 196–203.

- [94] A. Dantcheva and J.-L. Dugelay, “Frontal-to-side face re-identification based on hair, skin and clothes patches,” in *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2011, pp. 309–313.
- [95] V. Lepetit and P. Fua, “Keypoint recognition using randomized trees,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume=28, number=9, pages=1465–1479, year=2006.
- [96] Y. Amit and D. Geman, “Shape quantization and recognition with randomized trees,” *Neural Computation*, vol. 9, no. 7, pp. 1545–1588, 1997.
- [97] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, “Global contrast based salient region detection,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 409–416.
- [98] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, “Saliency filters: Contrast based filtering for salient region detection,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 733–740.
- [99] H. Yeh, K. Liu, and C. Chen, “Salient object detection via local saliency estimation and global homogeneity refinement,” *Pattern Recognition*, vol. 47, no. 4, pp. 1740–1750, 2014.
- [100] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille, “The secrets of salient object segmentation.” *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [101] Y. Tian, J. Li, S. Yu, and T. Huang, “Learning complementary saliency priors for foreground object segmentation in complex scenes,” *International Journal of Computer Vision*, pp. 1–18, 2014.
- [102] S. Iodice and A. Petrosino, “Person re-identification based on enriched symmetry salient features and graph matching,” in *Pattern Recognition*, 2013, pp. 155–164.
- [103] X. Jia, H.-C. Lu, and M.-H. Yang, “Visual tracking via adaptive structural local sparse appearance model,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1822–1829.
- [104] J. Ferryman and A. Shahrokni, “Pets2009: Dataset and challenge,” in *Winter-PETS*, 2009.

- [105] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, “Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 319–336, 2009.
- [106] Z. Zhou, Y. Wang, and E. Teoh, “Salient parts based multi-people tracking,” in *VISAPP*, 2015.
- [107] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn, “Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker,” in *ICCV Workshops*, 2011, pp. 120–127.
- [108] K. Yamaguchi, A. Berg, L. Ortiz, and T. Berg, “Who are you with and where are you going?” in *CVPR*, 2011, pp. 1345–1352.
- [109] A. Milan, K. Schindler, and S. Roth, “Detection-and trajectory-level exclusion in multiple object tracking,” in *CVPR*, 2013, pp. 3682–3689.
- [110] L. Li, W. Huang, I. Y. Gu, and Q. Tian, “Foreground object detection from videos containing complex background,” in *Proceedings of the eleventh ACM international conference on Multimedia*. ACM, 2003, pp. 2–10.
- [111] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li, “Learning multi-scale block local binary patterns for face recognition,” in *Advances in Biometrics*, 2007, pp. 828–837.
- [112] H. Wang and D. Suter, “A consensus-based method for tracking: Modelling background scenario and foreground appearance,” *Pattern Recognition*, vol. 40, no. 3, pp. 1091–1105, 2007.
- [113] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.
- [114] F. Porikli, “Integral histogram: A fast way to extract histograms in cartesian spaces,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, 2005, pp. 829–836.
- [115] Y. Wang, Z. Zhou, and E. K. Teoh, “Human hair segmentation and length detection for human appearance model,” in *Proceedings of International Conference on Pattern Recognition (ICPR)*, 2014.

- [116] K. Sobottka and I. Pitas, “A novel method for automatic face segmentation, facial feature extraction and tracking,” *Signal processing: Image communication*, vol. 12, no. 3, pp. 263–281, 1998.
- [117] I. Haritaoglu, R. Cutler, D. Harwood, and L. S. Davis, “Backpack: Detection of people carrying objects using silhouettes,” in *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 1, 1999, pp. 102–107.
- [118] D. Damen and D. Hogg, “Detecting carried objects from sequences of walking pedestrians,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1056–1067, 2012.
- [119] C.-S. Lee and A. Elgammal, “Carrying object detection using pose preserving dynamic shape models,” in *Articulated Motion and Deformable Objects*, 2006, pp. 315–325.
- [120] U. H. Office, *i-LIDS multiple camera tracking scenario definition*, 2008.
- [121] A. Ess, B. Leibe, and L. Van Gool, “Depth and appearance for mobile scene analysis,” in *IEEE International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.
- [122] A. Bialkowski, S. Denman, P. Lucey, S. Sridharan, and C. B. Fookes, “A database for person re-identification in multi-camera surveillance networks,” in *Proceedings of the 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*, 2012, pp. 1–8.
- [123] A. Bedagkar-Gala and S. K. Shah, “Part-based spatio-temporal model for multi-person re-identification,” *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1908–1915, 2012.
- [124] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.
- [125] V. Gopalakrishnan, Y. Hu, and D. Rajan, “Salient region detection by modeling distributions of color and orientation,” *IEEE Transactions on Multimedia*, vol. 11, no. 5, pp. 892–905, 2009.
- [126] Q. Yan, L. Xu, J. Shi, and J. Jia, “Hierarchical saliency detection,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1155–1162.

-
- [127] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu, “Global contrast based salient region detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.
- [128] Y. Luo, J. Yuan, P. Xue, and Q. Tian, “Saliency density maximization for efficient visual objects discovery,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 12, pp. 1822–1834, 2011.