Advances in Computer Vision and Pattern Recognition

Founding Editor

Sameer Singh

Series Editor

Srinivasa Narasimhan, Carnegie Mellon University, Pittsburgh, PA, USA

Advisory Editors

Richard Bowden, University of Surrey, Guildford, UK

Sven Dickinson, University of Toronto, Toronto, ON, Canada

Jiaya Jia, The Chinese University of Hong Kong, Shatin, Hong Kong

Zhouchen Lin, Peking University, Beijing, China

Bernt Schiele, Max Planck Institute for Informatics, Saarbrücken, Germany

The field of computer vision and pattern recognition has a rich history of nearly 50 years. In the past decade, however, the field has experienced remarkable advances in scene understanding and image generation. This advancement is driven by three key factors: (a) the availability of large, diverse datasets, (b) the accessibility of cloud and personal computing, and (c) the open release of advanced neural network architectures and models. These breakthroughs have led to significant successes across numerous application domains, including intelligent transportation, augmented reality, healthcare, agriculture, oceanography, and more. The ACVPR book series aims to introduce, analyze, and synthesize recent and foundational research, offering valuable references to both beginners and expert practitioners. The series covers timely topics such as:

- Deep Learning for Vision
- Large-scale Foundational Models
- Generative Methods
- Multimodal Learning (vision, audio, language, action, etc.)
- Neural Fields for Vision
- 3D Computer Vision
- Computational Photography, Display and Illumination
- Video Understanding and Synthesis
- Virtual, Mixed and Augmented Reality
- Biological and Human Vision
- Physics-based Vision
- Vision for Graphics
- Ethics in Computer Vision
- Applications (Robotics, Agriculture, Health, Intelligent Transportation, Oceanography, Safety/Security, etc.)

This series includes monographs, introductory and advanced textbooks, and state-of-the-art collections. Furthermore, it supports Open Access publication mode.

Xin Zhao · Shiyu Hu · Xu-Cheng Yin

Visual Object Tracking

An Evaluation Perspective

Springer

Xin Zhao
School of Computer and Communication Engineering
University of Science and Technology Beijing
Beijing, China

Xu-Cheng Yin
School of Computer and Communication Engineering
University of Science and Technology Beijing
Beijing, China

Shiyu Hu
School of Physical and Mathematical Sciences
Nanyang Technological University
Nanyang, Singapore

ISSN 2191-6586 ISSN 2191-6594 (electronic)
Advances in Computer Vision and Pattern Recognition
ISBN 978-981-96-4557-2 ISBN 978-981-96-4558-9 (eBook)
https://doi.org/10.1007/978-981-96-4558-9

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

If disposing of this product, please recycle the paper.

Preface

Pursuing dynamic visual intelligence represents one of the most significant challenges in artificial intelligence, requiring the synthesis of perceptual acuity, cognitive reasoning, and real-time adaptability. With applications spanning autonomous systems, augmented reality, and surveillance, the ability to emulate human-like visual intelligence has become an essential benchmark for advancing AI. Despite considerable progress in algorithmic innovations and multimodal integration, achieving seamless interaction between human adaptability and machine consistency remains elusive. This book addresses these challenges, positioning itself at the intersection of cognitive science, computer vision, and machine learning.

Unlike traditional texts focusing on algorithmic performance or isolated datasets, this book offers a comprehensive framework that bridges theoretical principles with practical applications. It emphasizes a cross-disciplinary approach, integrating neuroscientific insights, novel benchmarking techniques, and state-of-the-art algorithms. Readers will find an analysis of existing methods and guidance on how to innovate, evaluate, and deploy advanced systems capable of dynamic and human-like visual reasoning.

The rapid proliferation of large vision-language models, multimodal algorithms, and human-machine collaborative systems underscores the need for a unified understanding of dynamic visual tracking. As AI systems transition from static, isolated tasks to dynamic, real-world scenarios, there is an urgent demand for robust evaluation frameworks and adaptive methodologies. This book is one of the first to systematically explore these emerging areas, providing actionable insights for researchers and practitioners to navigate this rapidly evolving field.

This book provides a systematic exploration of visual object tracking (VOT), focusing on its core principles, algorithmic techniques, evaluation methodologies, and future development trends. Adopting a "from fundamentals to frontier" approach, the book is organized into five logical modules: Foundations of Visual Tracking (Part I), Experimental Environment (Part II), Task Executors (Part III), Evaluation Techniques (Part IV), and Future Trends and Outlook (Part V). Beginning with short-term, long-term, and global instance tracking tasks, the book introduces their definitions, challenges, and key techniques, offering practical applications across

diverse scenarios. By integrating traditional methods, deep learning paradigms, and insights from human vision theories, this book bridges foundational concepts and cutting-edge research in VOT.

Part I: Foundations of Visual Tracking

Chapter 1: **Task Introduction**
This chapter introduces the fundamental concepts of visual object tracking, focusing on its core principles and application scenarios. It explores the distinctions and connections among short-term tracking (STT), long-term tracking (LTT), and global instance tracking (GIT), while highlighting major challenges in dynamic tracking environments, such as occlusion, lighting variations, and scale changes.

Part II: Experimental Environment

Chapter 2: **Experimental Environment**
This chapter examines the experimental environments crucial for evaluating tracking algorithms. It categorizes datasets into general, specialized, and competition-specific datasets, analyzing their attributes and suitability for various tasks. Emphasis is placed on the role of diverse datasets in assessing algorithmic performance under realistic conditions.

Part III: Task Executors

Chapter 3: **Algorithms**
This chapter delves into object tracking algorithms, discussing the evolution from traditional methods to advanced techniques. Topics include correlation filter-based approaches, deep learning paradigms such as Siamese networks and transformers, and emerging large vision models (LVMs). The discussion highlights key trade-offs in precision, robustness, and adaptability.

Chapter 4: **Humans**
This chapter connects VOT to human vision theories, such as the Feature Integration Theory and Recognition-by-Components Theory. It examines human visual capabilities, particularly in static and dynamic contexts, and discusses how understanding human perception can inspire more effective tracking systems.

Part IV: Evaluation Techniques

Chapter 5: **Machine-To-Machine Comparisons**
This chapter reviews traditional evaluation techniques for VOT, such as One-Pass Evaluation (OPE) and Temporal Robustness Evaluation (TRE), while introducing enhanced metrics to address robustness, spatial accuracy, and fairness in algorithmic performance assessments.

Chapter 6: **Visual Turing Test: Human-To-Machine Comparisons**
This chapter proposes the Visual Turing Test (VTT) as a framework for evaluating machine vision systems. It bridges traditional evaluation approaches with the requirements of advanced general intelligence (AGI), particularly in complex scenarios demanding human-like robustness.

Part V: Conclusions and Future Works

Chapter 7: **More Human-Like Task Design**
This chapter explores designing tasks inspired by human visual capabilities. It emphasizes multimodal integration and adaptive memory mechanisms to enhance performance in complex and dynamic scenarios.

Chapter 8: **More Realistic Data Environment**
This chapter advocates for more realistic datasets and testing environments, emphasizing dataset diversity, annotation quality, and generalization to real-world conditions.

Chapter 9: **More Human-Like Executors**
This chapter focuses on creating human-like executors, incorporating perceptual strategies and adaptive behaviors to push the boundaries of current tracking technologies.

Chapter 10: **More Intelligent Evaluation**
The final chapter summarizes trends in intelligent evaluation, advocating for bridging human and machine capabilities through advanced AI methodologies. It highlights future directions in enhancing evaluation systems for dynamic and complex visual tasks.

One of the core contributions of this book lies in its exploration of human-like adaptability. By leveraging cognitive principles such as memory-driven reasoning and real-time contextual adaptation, the book establishes a foundation for developing systems that go beyond traditional AI. Furthermore, its focus on human-machine collaboration and multimodal tracking sets a new standard for addressing the complexities of real-world scenarios.

We believe this book will not only serve as a valuable reference for seasoned researchers but also inspire the next generation of AI practitioners to reimagine what is possible in dynamic visual intelligence. Whether you are a student, an academic, or an industry professional, this book offers the tools and perspectives needed to advance your understanding and impact in this transformative domain.

Beijing, China — Xin Zhao
Nanyang, Singapore — Shiyu Hu
Beijing, China — Xu-Cheng Yin
January 2025

Acknowledgements We would like to extend our deepest gratitude to the many individuals and institutions that have supported the completion of this book. Their contributions, whether academic, technical, or personal, have been invaluable in shaping the ideas and content presented here.

First and foremost, we are grateful to Beijing Natural Science Foundation (No. JQ24024) and National Key Laboratory of Science and Technology on Space-Born Intelligent Information Processing Project (No. TJ-04-23-02) for their generous financial support, which enabled the development of the research underlying this book. This financial backing provided the resources necessary to conduct extensive experiments, develop innovative methodologies, and produce high-quality benchmarks, which form the foundation of this book.

We are also immensely thankful to the editorial team at Springer, particularly Ramesh Kumaran, for their patience, guidance, and professionalism throughout the publication process. Their meticulous attention to detail, constructive feedback, and unwavering support have significantly enhanced the quality and readability of this book. Their

Additionally, we are grateful to the broader academic community for the datasets, tools, and open-source resources that have facilitated our work. Without these contributions, it would have been impossible to conduct the extensive evaluations and comparisons included in this book.

Some subsections of this book are based on previously published papers. We wish to thank all the co-authors of these works, whose collaboration and expertise have been indispensable in developing the core ideas. The key papers include:

- Huang, L., Zhao, X., & Huang, K. (2019). Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE transactions on pattern analysis and machine intelligence, 43(5), 1562–1577.
- Zhao, X., Hu, S., Wang, Y., Zhang, J., Hu, Y., Liu, R., ... & Li, J. (2024). Biodrone: A bionic drone-based single object tracking benchmark for robust vision. International Journal of Computer Vision, 132(5), 1659–1684.
- Hu, S., Zhao, X., Huang, L., & Huang, K. (2022). Global instance tracking: Locating target more like humans. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1), 576–592.
- Hu, S., Zhao, X., & Huang, K. (2024). SOTVerse: A user-defined task space of single object tracking. International Journal of Computer Vision, 132(3), 872–930.
- Hu, S., Zhang, D., Feng, X., Li, X., Zhao, X., & Huang, K. (2024). A multimodal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship. Advances in Neural Information Processing Systems, 36.
- Zhang, D., Hu, S., Feng, X., Li, X., Wu, M., Zhang, J., & Huang, K. Beyond accuracy: Tracking more like human via visual search. In The Thirty-eighth Annual Conference on Neural Information Processing Systems.
- Feng, X., Li, X., Hu, S., Zhang, D., Wu, M., Zhang, J., ... & Huang, K. (2024). MemVLT: Vision-Language Tracking with Adaptive Memory-based Prompts. In The Thirty-eighth Annual Conference on Neural Information Processing Systems.

- Huang, L., Zhao, X., & Huang, K. (2020). Globaltrack: A simple and strong baseline for long-term tracking. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 11037–11044).
- Huang, L., Zhao, X., & Huang, K. (2019). Bridging the gap between detection and tracking: A unified approach. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3999–4009).
- Li, X., Feng, X., Hu, S., Wu, M., Zhang, D., Zhang, J., & Huang, K. (2024). DTLLM-VLT: Diverse Text Generation for Visual Language Tracking Based on LLM. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7283–7292).

Finally, we would like to express our heartfelt appreciation to our families and friends for their unwavering support and encouragement throughout this endeavor. Their patience and understanding during the long hours of research and writing have been a constant source of strength.

Competing Interests The authors have no competing interests to declare that are relevant to the content of this manuscript.

Contents

Part I Foundations of Visual Tracking

1 Task Introduction 3
1.1 Defining Single Object Tracking: Core Principles and Applications 3
1.1.1 Short-Term Tracking: Core Attributes and Foundational Role 5
1.1.2 Long-Term Tracking: Bridging Continuity and Real-World Complexity 6
1.1.3 Global Instance Tracking: Expanding Spatial and Temporal Horizons 8
1.2 Key Challenges in Single Object Tracking 9
1.2.1 Shot-Cuts and Disrupted Continuity in Tracking 11
1.2.2 Handling Occlusion and Target Disappearance in Dynamic Tracking 12
1.2.3 Addressing Challenges of Lighting Variations in Single Object Tracking 13
1.2.4 Overcoming Background Clutter in Dynamic Tracking Environments 14
1.2.5 Addressing Challenges of Fast Motion in Dynamic Tracking 15
1.2.6 Addressing Challenges of Extreme Scales and Aspect Ratios in Object Tracking 16
1.2.7 Challenges and Solutions for Scale and Aspect Ratio Variations in SOT 17
1.3 Conclusion 18
References 19

Part II Experimental Environment

2 Experimental Environment 25
2.1 General Datasets 25
2.1.1 Small-Scale Datasets 26
2.1.2 Large-Scale Datasets 29
2.2 Specialized Datasets 40
2.2.1 Special Object 40
2.2.2 Special Scenario 42
2.3 Competition Datasets 46
2.4 Conclusion 49
References 50

Part III Task Executors

3 Algorithms 55
3.1 Traditional Trackers 55
3.1.1 Motion Modeling: Balancing Precision and Efficiency in Dynamic Environments 55
3.1.2 Feature Representation: Balancing Global Simplicity and Local Adaptability in SOT 57
3.1.3 Appearance Modeling: Balancing Reconstruction and Discrimination in Single Object Tracking 58
3.1.4 Adaptive Model Update Strategies in Single Object Tracking 59
3.2 Advancements in Correlation Filter-Based Trackers 60
3.2.1 Kernelized Correlation Filter: Efficiency and Adaptability 62
3.2.2 ECO: Optimizing Efficiency and Robustness in Correlation Filters 63
3.2.3 UPDT: Adaptive Feature Fusion for Robust Tracking 64
3.3 Siamese Neural Networks: A Paradigm Shift in SOT 65
3.3.1 SiamFC: A Foundational Framework for SNN-Based Tracking 66
3.3.2 Advancing Object Tracking with SiamRPN: Integration of Detection and Tracking 68
3.3.3 DaSiamRPN: Enhancing Robustness and Distractor Awareness in Long-Term Object Tracking 70
3.3.4 SiamRPN++: Advancing Robustness with Deep Backbones 71
3.3.5 TransT: Transforming Tracking with Attention Mechanisms 72
3.3.6 SwinTrack: Advancing Tracking with Hierarchical Transformers 73

3.4 One-Stream Trackers: A Unified Approach to SOT 74
3.4.1 OSTrack: Unifying Feature Learning and Relation Modeling 75
3.4.2 MixFormer: Iterative Mixed Attention for Unified Tracking 76
3.5 LVM-Based Trackers: Leveraging Large Vision Models for Object Tracking 78
3.5.1 Track Anything Model (TAM): Bridging Segmentation and Temporal Tracking 79
3.5.2 SAM-Track: A Multimodal Framework for Interactive Video Segmentation and Tracking 80
3.5.3 TrackGPT: Instruction-Driven Video Tracking Framework 81
3.6 Conclusion 82
References 84

4 Humans 89
4.1 Vision Theory 89
4.1.1 Feature Integration Theory 90
4.1.2 Visual Computation Theory 91
4.1.3 Recognition-By-Components Theory 92
4.2 Vision Abilities 94
4.2.1 Static Visual Ability 95
4.2.2 Dynamic Visual Ability 96
4.3 Conclusion 98
References 99

Part IV Evaluation Techniques

5 Machine-To-Machine Comparisons 105
5.1 Machine-To-Machine Evaluation 105
5.2 Evaluation Mechanisms 106
5.2.1 One-Pass Evaluation (OPE): A Comprehensive Framework for Tracking Assessment 106
5.2.2 Temporal Robustness Evaluation (TRE): A Comprehensive Temporal Assessment 108
5.2.3 Spatial Robustness Evaluation (SRE): Assessing Spatial Adaptability 109
5.2.4 Restart Mechanism in OTB: Enhancing Recovery in Tracking Evaluations 110
5.2.5 Anchor-Based Restart Mechanism in VOT Challenge: Ensuring Fairness and Robustness 111
5.2.6 Enhanced Restart Mechanism in VideoCube Benchmark 111

5.3 Evaluation Metrics 112
5.3.1 Precision (PRE): Assessing Spatial Accuracy in Tracking 113
5.3.2 Normalized Precision (N-PRE): Addressing Biases in Tracking Accuracy 114
5.3.3 Success Rate (SR): A Comprehensive Metric for Overlap Evaluation 116
5.3.4 Robustness: Evaluating Stability and Resilience 117
5.3.5 Expected Average Overlap (EAO): A Unified Metric for Tracking Performance 118
5.3.6 F-Score: A Balanced Metric for Precision and Recall in Single-Target Tracking 119
5.4 Results and Analyses About Representative Trackers 120
5.4.1 Performance Advancements in the OTB100 Benchmark 120
5.4.2 Advancements in the TrackingNet Benchmark 122
5.5 Advancements in the GOT-10k Benchmark 122
5.6 Conclusion 125
References 125

6 Visual Turing Test: Human-To-Machine Comparisons 129
6.1 Human-To-Machine Evaluation 129
6.2 From the Turing Test to Modern AI Evaluation: Legacy and Beyond 130
6.3 Visual Turing Test: A Holistic Framework for Machine Vision Assessment 132
6.3.1 Case Study 1: Assessing Classification Robustness Through the Visual Turing Test 133
6.3.2 Case Study 2: Evaluating Deep Learning Robustness Under Ambiguous Visual Conditions 135
6.3.3 Case Study 3: Bridging the Gap in Global Instance Tracking Through the Visual Turing Test 138
6.3.4 Summary: A Structured Methodology for the Visual Turing Test 141
6.4 Conclusion 143
References 144

Part V Conclusions and Future Works

7 More Human-Like Task Design 147
7.1 From Perception to Cognition: Advancing Dynamic Vision Task Design 147
7.2 Integrating Vision and Language: A Paradigm Shift in Object Tracking 149

7.3 MGIT: Redefining Global Instance Tracking with Multimodal Frameworks 151
7.4 Conclusion 155
References 156

8 More Realistic Data Environment 159
8.1 From Controlled Benchmarks to Real-World Dynamics: Evolution of Datasets for Dynamic Vision Tasks 159
8.2 From Model Refinement to Data Optimization: Redefining the AI Development Paradigm 161
8.2.1 Step 1: Systematic Design of Realistic Datasets for AI: Addressing Complexity and Diversity 162
8.2.2 Step 2: Refining Annotations for AI: Merging Human Expertise with Automation 164
8.2.3 Step 3: Advanced Testing Strategies: Combating Shortcut Learning and Promoting Generalization 167
8.3 Revolutionizing Single Object Tracking: The Dynamic and Modular Framework of SOTVerse 168
8.4 DTVLT: Redefining Visual-Language Tracking with Semantic Depth and Temporal Granularity 171
8.5 Conclusion 173
References 174

9 More Human-Like Executors 177
9.1 Foundations of Human-Like Visual Intelligence 177
9.2 Enhancing Visual Object Tracking with Human-Inspired Perceptual Strategies 178
9.3 Advancing Visual Object Tracking with Human-Inspired CPDTrack 180
9.4 Human-Like Visual-Language Tracking: Adaptive Memory Modeling 184
9.5 Conclusion 187
References 188

10 More Intelligent Evaluation 191
10.1 Evaluating Dynamic Visual Intelligence: Bridging Human and Machine Capabilities 191
10.2 Unifying Perspectives in Dynamic Visual Intelligence Evaluation 193
10.3 Advancing Future Directions in Dynamic Visual Tracking 196
10.4 Conclusion 197
References 198