

Weiwei Xing · Weibin Liu · Jun Wang
Shunli Zhang · Lihui Wang
Yuxiang Yang · Bowen Song

Visual Object Tracking from Correlation Filter to Deep Learning

Visual Object Tracking from Correlation Filter to Deep Learning


Weiwei Xing • Weibin Liu • Jun Wang •
Shunli Zhang • Lihui Wang • Yuxiang Yang •
Bowen Song

Visual Object Tracking from Correlation Filter to Deep Learning

 Springer

Weiwei Xing 
School of Software Engineering
Beijing Jiaotong University
Beijing, China

Weibin Liu 
Institute of Information Science
Beijing Jiaotong University
Beijing, China

Jun Wang 
College of Electronic Information
Engineering
Hebei University
Baoding, Hebei, China

Shunli Zhang 
School of Software Engineering
Beijing Jiaotong University
Beijing, China

Lihui Wang
Department of Information
and Communication
Army Academy of Armored Forces
Academy
Beijing, China

Yuxiang Yang 
School of Software Engineering
Beijing Jiaotong University
Beijing, China

Bowen Song
School of Software Engineering
Beijing Jiaotong University
Beijing, China

ISBN 978-981-16-6241-6 ISBN 978-981-16-6242-3 (eBook)
<https://doi.org/10.1007/978-981-16-6242-3>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

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

Preface

This book introduces some representative trackers through practical algorithm analysis and experimental evaluations. This book is intended for professionals and researchers interested in visual object tracking, and also can be used as a reference book by students. Readers will get comprehensive knowledge of tracking and can learn state-of-the-art methods through this content. In general, this book is organized as follows:

Chapter 1 introduces the wide application, existing challenges, and basic concepts in visual object tracking. Chapter 2 introduces some algorithm foundations from correlation filter basics, typical deep learning model, and performance evaluation. Chapter 3 presents the correlation filter-based visual object tracking. Chapter 4 mainly introduces correlation filter with deep feature for visual object tracking. Chapter 5 introduces deep learning-based visual object tracking. Finally, Chap. 6 summarizes our work and points out the potential future research directions for visual object tracking in appearance model construction and update.

Beijing, China
June 2021

Weiwei Xing

Acknowledgements

First of all, we would like to thank every one of the collaborators who worked tirelessly together and contributed to producing this monograph. We would also like to thank the organizers of visual tracking benchmarks for providing large-scale datasets to evaluate the trackers comprehensively. We would highly appreciate the help and assistance of the current and graduated students in our research group during the production of this book. In particular, we wish to thank Mr. Xinjie Wang, Mr. Menglei Jin, and Mr. Jiayi Yin for their time and efforts in providing materials and reviewing for revising the entire book and Mr. Hui Wang, Ms. Huaqing Hao, Mr. Yanhao Cheng, Ms. Yuxin Wang, and Mr. Yang Pei for improving the individual chapters of this monograph. We also want to especially thank the editor of Springer, Jing Dou, for her patience, support, guidance, and editorial assistance in the course of the preparation of this work. Part of the content in this book is based on the work supported by the National Natural Science Foundation of China under Nos. 61876018 and 61976017 and the Beijing Natural Science Foundation under No. 4212025.

This monograph, especially those proposed object tracking algorithms in Chaps. 3–5, is mainly prepared based on the following research papers of our group. We would like to give a special acknowledgement to all the authors for their contributions and to the related publishing organizations for their permission to use these materials. The referenced papers are as follows:

1. Jin, M.L., Liu, W.B., Xing, W.W.: A robust visual tracker based on DCF algorithm. *International Journal of Software Engineering and Knowledge Engineering*, 29(11n12):1819–1834 (2019)
2. Jin, M.L., Liu, W.B., Xing, W.W.: A robust Correlation Filter based tracker with rich representation and a relocation component. *KSII Transactions on Internet and Information Systems*, 13(10): 5161–5178 (2019)
3. Wang, J., Liu, W.B., Xing, W.W., Zhang, S.L.: Visual object tracking with multi-scale superpixels and color-feature guided Kernelized Correlation Filters. *Signal Processing: Image Communication*, 63: 44–62 (2018)

4. Yang, Y.X, Xing, W.W., Zhang, S.L., et al.: Visual tracking with long-short term based Correlation Filter. *IEEE Access*, 8, 20257–20269 (2020)
5. Wang, J., Liu, W.B., Xing, W.W.: Discriminative context-aware Correlation Filter network for visual tracking. In: *Intelligent Systems and Applications. Advances in Intelligent Systems and Computing*, 1250, pp.724–736 (2020)
6. Wang, J., Liu, W.B., Xing, W.W., Wang, L.Q., Zhang, S.L.: Attention shake siamese network with auxiliary relocation branch for visual object tracking. *Neurocomputing*, 400, 53–72 (2020)
7. Yang, Y.X, Xing, W.W., Zhang, S.L., et al.: A learning frequency-aware feature siamese network for real-time visual tracking. *Electronics*, 9(5):854 (2020)
8. Song, B.W., Lu, W., Xing, W.W., Xiang, W.: Real-time object tracking based on improved adversarial learning. In: *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3576–3581 (2020)
9. Yang, Y.X, Xing, W.W., Wang, D.D., Zhang, S.L., et al.: AEVRNet: Adaptive exploration network with variance reduced optimization for visual tracking. *Neurocomputing*, 449, 48–60 (2021)

Contents

1	Introduction	1
1.1	Motivation and Challenge	1
1.2	Basic Concepts and Features	7
1.3	Evolution of Visual Object Tracking Technology	8
1.4	Chapter Outline	10
	References	11
2	Algorithms Foundations	15
2.1	Correlation Filter Basics	15
2.1.1	MOSSE	16
2.1.2	Discriminative Correlation Filter	17
2.1.3	Kernel Correlation Filter	18
2.2	Typical Deep Learning Model for Tracking	22
2.2.1	Convolutional Neural Networks Based Model	22
2.2.2	Siamese Networks Based Model	24
2.2.3	Generative Adversarial Networks Based Model	26
2.2.4	Reinforcement Learning Based Model	28
2.3	Performance Evaluation	29
2.3.1	Performance Evaluation Criteria	29
2.3.2	Benchmark Datasets	35
2.4	Summary	41
	References	42
3	Correlation Filter Based Visual Object Tracking	45
3.1	Introduction	45
3.2	Correlation Filter Tracker with Context Aware Strategy	46
3.2.1	Context Aware Strategy	47
3.2.2	Adaptive Update Model	49
3.2.3	Framework and Procedure	50
3.2.4	Experimental Results and Discussions	51

3.3	Correlation Filter Tracker with Scale Pyramid	53
3.3.1	Scale Pyramid Filter	54
3.3.2	Rich Image Feature Representation	55
3.3.3	Framework and Procedure	56
3.3.4	Experimental Results and Discussions	58
3.4	Correlation Filter Tracker with Multi-Scale Superpixels	61
3.4.1	Multi-Scale Superpixels Segmentation	63
3.4.2	Structure Based Optimization Strategy	67
3.4.3	Framework and Procedure	70
3.4.4	Experimental Results and Discussions	72
3.5	Summary	81
	References	81
4	Correlation Filter with Deep Feature for Visual Object Tracking	85
4.1	Introduction	85
4.2	Long-Short Term Correlation Filter Based Visual Object Tracking	86
4.2.1	Fusion of Deep Features and Hand-Crafted Features	86
4.2.2	Correlation Filters with Long-Short Term Update	88
4.2.3	Framework and Procedure	92
4.2.4	Experimental Results and Discussions	92
4.3	Context-Aware Correlation Filter Network	102
4.3.1	Context-Aware Correlation Filter Network	103
4.3.2	Channel Attention Mechanism	106
4.3.3	Update with High Confidence Strategy	107
4.3.4	Framework and Procedure	108
4.3.5	Experimental Results and Discussions	108
4.4	Auxiliary Relocation in SiamFC Framework	114
4.4.1	Auxiliary Relocation with Correlation Filters	115
4.4.2	Switch Function for SiamFC Framework	118
4.4.3	Framework and Procedure	119
4.4.4	Experimental Results and Discussions	120
4.5	Summary	124
	References	126
5	Deep Learning Based Visual Object Tracking	129
5.1	Introduction	129
5.2	Attention Shake Siamese Based Visual Object Tracking	130
5.2.1	Attention Mechanisms in Siamese Network	130
5.2.2	Shake-Shake Mechanism in Siamese Network	133
5.2.3	Framework and Procedure	135
5.2.4	Experimental Results and Discussions	136
5.3	Frequency-Aware Siamese Network Based Visual Object Tracking	148
5.3.1	Frequency-Aware Siamese Network	148
5.3.2	Pre-training and Joint Update	149

- 5.3.3 Framework and Procedure..... 152
- 5.3.4 Experimental Results and Discussions 152
- 5.4 Improved Generative Adversarial Network Based Visual Object Tracking 157
 - 5.4.1 Improved Adversarial Learning Strategy 158
 - 5.4.2 Precise ROI Pooling for Faster Feature Extraction..... 159
 - 5.4.3 Framework and Procedure..... 161
 - 5.4.4 Experimental Results and Discussions 163
- 5.5 Improved Policy-Based Reinforcement Learning Based Visual Object Tracking 167
 - 5.5.1 Non-Convex Optimized Variance Reduced Backward Propagation..... 169
 - 5.5.2 ϵ -Greedy Strategy for Action Space Exploration 171
 - 5.5.3 Regression Based Reward Function 173
 - 5.5.4 Framework and Procedure..... 174
 - 5.5.5 Experimental Results and Discussions 175
- 5.6 Summary 185
- References 186
- 6 Summary and Future Work 191**
 - 6.1 Summary 191
 - 6.2 Future Work 192

About the Authors

Weiwei Xing received the B.S. degree in Computer Science and Technology and the Ph.D. degree in Signal and Information Processing from the Beijing Jiaotong University, Beijing, China, in 2001 and 2006, respectively. She was a visiting scholar at the University of Pennsylvania, PA, USA, during 2011–2012. She is currently a professor at the School of Software Engineering, Beijing Jiaotong University and leads the research group on Intelligent Computing and Big Data. Her research interests include visual object tracking, computer vision, pattern recognition, and deep learning.

Weibin Liu received the Ph.D. degree in Signal and Information Processing from the Institute of Information Science at Beijing Jiaotong University, China, in 2001. During 2001–2005, he was a researcher in the Information Technology Division at Fujitsu Research and Development Center Co., LTD. Since 2005, he has been with the Institute of Information Science at Beijing Jiaotong University, where currently he is a professor in Digital Media Research Group. He was also a visiting researcher in the Center for Human Modeling and Simulation at University of Pennsylvania, PA, USA, during 2009–2010. His research interests include computer vision, computer graphics, image processing, virtual human and virtual environment, and pattern recognition.

Jun Wang received the M.S. degree in Pattern Recognition and Intelligent Systems from the Hebei University, China, in 2015. He received the Ph.D. degree in Signal and Information Processing from the Institute of Information Science at Beijing Jiaotong University, China, in 2020. He was also a visiting researcher in Visual Object Tracking at the University of Central Florida, USA, during 2018–2019. Currently, he is an associate professor at the College of Electronic Information Engineering, Hebei University. His research interests include image processing, computer vision, visual object tracking, and pattern recognition.

Shunli Zhang received the B.S. and M.S. degrees in Electronics and Information Engineering from the Shandong University, Jinan, China, in 2008 and 2011, respectively, and the Ph.D. degree in Signal and Information Processing from the Tsinghua University in 2016. He was a visiting scholar at the Carnegie Mellon University, Pittsburgh, from 2018 to 2019. He is currently a faculty member at the School of Software Engineering, Beijing Jiaotong University. His research interests include pattern recognition, computer vision, and image processing.

Lihui Wang received the Ph.D. degree in Signal and Information Processing from the Beijing Jiaotong University, Beijing, China, in 2011. She is currently a lecturer in the Department of Information and Communication, Army Academy of Armored Forces. Her main research interests include computer application, big data analysis, and three-dimensional reconstruction.

Yuxiang Yang received the B.S. degree in Computer Science and Technology from the Northeastern University of China, Liaoning, China, in 2014. He is currently pursuing the Ph.D. degree with the School of Software Engineering, Beijing Jiaotong University. His research interests include image processing, deep learning, reinforcement learning, and object tracking.

Bowen Song received the B.S. degree in Computer Science and Technology from the School of Computer Science and Technology, Heilongjiang University, China, in 2018. She received the M.S.E. degree in Software Engineering from the School of Software Engineering, Beijing Jiaotong University, China, in 2021. Her research interests include visual tracking and deep learning.

Chapter 1

Introduction



Visual tracking is a rapidly evolving field of computer vision that has been attracting increasing attention in the vision community. One reason is that visual tracking offers many challenges as a scientific problem. Moreover, it is a part of many high-level problems of computer vision, such as motion analysis, event detection, and activity understanding. In this chapter, we give a detailed introduction to visual tracking which includes basic components of tracking algorithms, difficulties in tracking, datasets used to evaluate trackers, and evaluation metrics.

1.1 Motivation and Challenge

How to make computers have vision and analyze the information in videos has always been a desire of human. In recent years, Artificial Intelligence (AI) gradually applied in various industries, along with the continuous improvement of machine learning and deep learning research [1], such as automatic driving [2], speech recognition [3], face recognition [4], VR games [5], etc. Visual object tracking, which provides the trajectory characteristics for behavior analysis by predicting the state of the object in the video, is one of the important components in Computer Vision (CV). It has been widely applied in intelligent monitoring [6], human-computer interaction [7], automatic driving [8, 9], virtual reality [10, 11], crime projections [12], surgical navigation [13], aerospace [14, 15] and so on.

Figure 1.1 shows some common application scenarios for visual object tracking. In smart traffic, visual object tracking can judge whether there is a violation by monitoring the tracking the vehicles, such as illegal U-turn, speeding, etc. In human-computer interaction, computers can determine the instructions of the person and make corresponding actions without pressing the buttons, by tracking and calculating the states of human body parts, such as hands, legs, head, eyes, etc. In automatic driving, visual object tracking can perceive the change and motion of



Fig. 1.1 Some applications of visual object tracking

the objects around the vehicle to provide a certain reference for vehicle computer. In virtual reality, visual object tracking combined with object segmentation algorithm can calculate the location and shape of the objects. For example, in the application of virtual changing, the shape of cloth can be automatically adjusted to make it more suitable for the outline of the human body. In crime prediction, monitoring and tracking the sudden aggregation and dispersion of people or other objects in video could predict the abnormal and possible emergencies and help the police find out illegal crimes and improve the social environment. In surgery navigation, the success rate of surgery can be improved by tracking the position and posture of the scalpel and probe. Visual object tracking also has important applications in military fields. In missile navigation and military reconnaissance, the objects are often moving and the cameras on the missiles are also jittering, visual object tracking can be used to determine the position of the object and adjust the attitude of the missile to improve the guidance accuracy.

Several countries and institutions have established major projects around visual object tracking. In the early 1997s of American, Project Video Surveillance And Monitoring (VSAM) was co-founded by Carnegie Mellon University and the David Sarno Research Center, with the funding from the Defense Advanced Research Projects Agency (DARPA) [16]. This project aims to track people and cars in complex environments continuously with multiply video sensors, and develops visual surveillance systems. DARPA then funded project Human Identification at a Distance (HID) in 2000, which is led by University of Maryland [17]. The Framework 5 on EU information technologies also created Annotated Digital Video for Surveillance and Optimized Retrieval (ADVISOR) project in 1999 which used to manage the urban traffic systems and analyse pedestrian behaviour. At the same time, Hiroshima University in Japan hosted the Cooperative Distributed Vision Project (CDVP) project of intelligent monitoring from 1996 to 1999, in order to build community-oriented monitoring systems. The Institute of Automation,

the Chinese Academy of Science and Technology also presided over the Visual Surveillance Star (VStar) project for urban traffic monitoring and management.

In addition to the visual object tracking based projects, many top publications and conferences in the world also continue to promote the progress and innovation of visual object tracking algorithms. Visual object tracking, which is the basic research direction of video processing, occupies a certain proportion in the top journals and conferences in Computer Vision (CV) every year. The top conferences of visual object tracking direction are mainly IEEE International Conference on Computer Vision (ICCV), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), European Conference on Computer Vision (ECCV) and International Joint Conference on Artificial Intelligence (IJCAI), etc. While, the top journals are mainly IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), IEEE Transactions on Signal Processing (TSP), IEEE Transactions on Image Processing (TIP) and IEEE Transactions on Multimedia (TMM), etc. It can be seen that the research of visual object tracking is of great theoretical value and extensive application background.

With the continuous progress of science and technology, the demand of people for visual object tracking is also increasing. The objects to be tracked changes from rigid objects to non-grid objects. The background of video is from simple to complex, such as occlusion, motion blur and some other complex situations. Besides, the demand for tracking time also grows. The essence of visual object tracking is an online learning problem with the small size of non-annotated samples: How to construct a robust representation model, fast motion model and effective update model is the mainly challenge in visual object tracking. In addition, tracking tasks are more and more close to the real life. There are multiple tracking challenges in one sequence. Thus, how to design a universal tracing algorithm which could deal with the multiple challenges in one sequence is one of the key problems in the research of visual object tracking. Wu et al. [18] divided the challenges existing in visual object tracking into 11 categories. The details of the 11 challenges are as follows:

Illumination Variation, IV IV refers to the situation in which the illumination of the object region changes significantly as the video plays, as shown in Fig. 1.2. This situation could lead to a sharp change in color and gray based features of the object. Thus, the color or gray feature based appearance model could not represent the object well and lead to the tracking failure.

Scale Variation, SV SV refers to the situation in which the ratio of the bounding box in the initial frame to the bounding box in current frame exceeds a specific



Fig. 1.2 Example of illumination variation



Fig. 1.3 Example of scale variation



Fig. 1.4 Example of occlusion

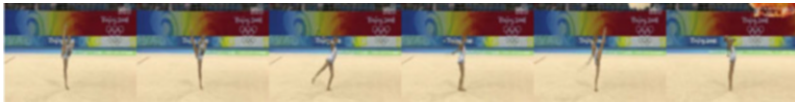


Fig. 1.5 Example of deformation

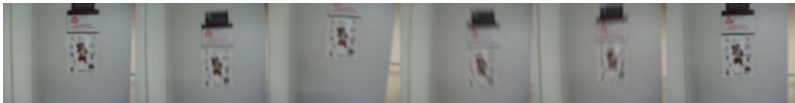


Fig. 1.6 Example of motion blur

threshold $ts > 1$, usually, ts is set to be 2, as shown in Fig. 1.3. The scale change of the object will lead to the change of the number of pixels, which is a great challenge to construct the appearance model. Some tracking methods could find the center of the object, but fail to estimate the size. These methods may show a high accuracy score but a low success rate.

Occlusion, OCC OCC refers to the situation in which the object is partially or completely occluded as the video plays, as shown in Fig. 1.4. The occlusion challenge will change the statistical and structural features of the object, which brings challenges to the tracking methods. Thus, how to re-detect and continue to track the object when occlusion occurs is one of the main problems that the tracking methods should be solved.

Deformation, DEF DEF refers to the situation in which the shape of the object change significantly compared to the initial frame. This challenge is primarily aimed at non-rigid objects, as shown in Fig. 1.5. Although deformation challenge does not change the statistical features of the object, such as gray histograms, it changes the target structural characteristics. This makes it difficult to determine the boundary of the objects, and the bounding box may not cover the object properly.

Motion Blur, MB MB refers to the situation that the object image is blurred due to the movement of the object or the shake of camera, as shown in Fig. 1.6. The



Fig. 1.7 Example of fast motion



Fig. 1.8 Example of in-plane rotation

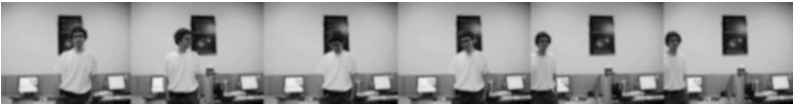


Fig. 1.9 Example of out-of-plane rotation

blurring of the object image not only loses the detail information of the object, but also changes the structure information of the object and makes the boundary of the object become blurred. These makes boundary information changes. Thus, MB makes the trackers difficult to locate the boundary of the object and determine the object size accurately.

Fast Motion, FM FM refers to the situation that the distance of the objects between two adjacent frames is over a certain threshold tm . Generally, tm is set to be 20, $tm = 20$, as shown in Fig. 1.7. Although FM does not change the appearance of the object itself, it requires a higher search ability of the motion model. Usually, trackers assume that the motion of the object is smooth, which means the change of the object state follows the Gaussian distribution. Gaussian distribution based motion model could reduce the number of candidate samples and speed up the running time, but limits the search ability of the trackers. These make the trackers lose the objects, and result in tracking failure.

In-Plane Rotation, IPR IPR refers to the rotation of the object in image plane when the video plays, as shown in Fig. 1.8. IPR can be regarded as a rotation movement of the whole pixels of the object centered at a certain position. It does not introduce new information and new pixels of the objects, but requires higher rotation invariance of appearance model.

Out-of-Plane Rotation, OPR OPR refers to the rotation of the object out of the image plane when the video plays, as shown in Fig. 1.9. OPR brings the change of the object appearance and puts forward higher requirements for the update ability of tracking methods and adaptability of appearance model.



Fig. 1.10 Example of out-of-view



Fig. 1.11 Example of background clutters



Fig. 1.12 Example of low resolution

Out-of-View, OV OV refers to the situation that the object appears at the boundary of the frame and some parts of the object jump out of this frame, as shown in Fig. 1.10. Similar to occlusion, some information of the object is lost under OV challenge. Since the object is located at the boundary of the frame, when the object moves out of frame, some background information is also lost, which increases the difficulty of some context based trackers. This challenge also affects the estimation of the size and center of the object, as well as the extraction of context information around the object.

Background Clutters, BC BC refers to the situation that there are some similar objects around the object to be tracked. These objects and objects to be tracked are similar in color and shape, which may mislead the tracking algorithm, as shown in Fig. 1.11. BC challenges the discrimination ability of the appearance model. Some trackers with a weak discriminative appearance model may tend to track the wrong object and ignore the real object to be tracked, especially when the objects are close to each other or even overlaps.

Low Resolution, LR LR refers to the case that the number of the pixels in object bounding box does not exceed a certain threshold tr . Usually, tr is set to be 400 pixels, as shown in Fig. 1.12. The lack of pixels of the object makes the construction of the appearance model difficult, and the loss of detail information limits the discriminant ability of the appearance model.

It is worth noting that the 11 challenges described above do not appear alone. Usually, there exist multiple tracking challenges at the same time in one video sequence. The videos which are obtained from real life and contain multiple challenges in the complex scenes in visual object tracking. Thus, how to design a tracking algorithm with good tracking effect for every challenge is still a problem to be solved in visual object tracking.

1.2 Basic Concepts and Features

Visual object tracking is a classic research direction in the field of computer vision. In 1982, Man et al. [19] constructed the framework of computer vision and demonstrated that the Fourier transform of spatial frequency sensitive data can derive the retinal receptive field geometry. That is, we can detect edges and contours by Laplace or the second derivative method to find zero intersection in image intensity gradient. The excitatory and inhibitory receptive fields can be constructed by Difference Of two Gaussians (DOG) function. Visual systems can use two-dimensional convolutions and Gaussian filters as operators to optimize the bandwidth of the optical distribution. It provides the theoretical basis for computer vision and visual object tracking. In 2014, Arnold W. M. Smeulders et al. [20] gave a specific definition of visual object tracking in TPAMI, one of the top journals in computer vision fields. That is “tracking is the analysis of video sequences for the purpose of establishing the location of the target over a sequence of frames (time), starting from the bounding box given in the first frame.”

Arnold W. M. Smeulders et al. [20] not only gave the definition of the visual object tracking, but also summarized the principle and process of some existing visual object tracking algorithms. The visual object tracking is divided into five parts: object area, appearance model, motion model, tracking algorithm and update model. General video object tracking ideas and processes are shown in Fig. 1.13. At the beginning of tracking, the object region that needs to be tracked should be selected. Then, the appearance model is constructed based on the object image, and the motion model is constructed according to the relationship of the object states between two adjacent frames. Tracking algorithms, which are based on different principles, such as similarity matching and optimization, calculate the state of the object in current frame through appearance model and motion model. Update model is used to update the appearance model and motion model of the object, in order to make the appearance and motion model adapt to the change of shape and appearance change of the object, caused by deformation and occlusion of the object. The position and state of the object in each frame could be predicted through iteration. In the five parts in Fig. 1.13, the object area refers to the image of the object that needs to be tracked. Normally, the object area is constructed by a bounding box. Such as NCC tracker [21]. Some object areas are constructed in an elliptical

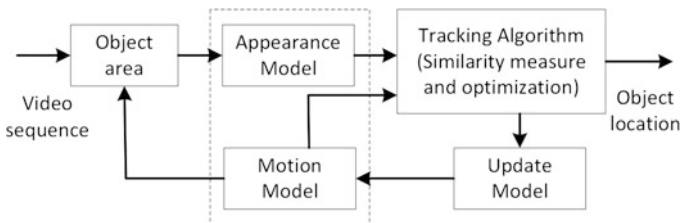


Fig. 1.13 Basic ideas of visual object tracking

way, such as MST tracker [22]. The appearance model mainly extracts the feature representation of the object image. According to the features extracted from objects, the appearance model can be divided into two dimension image array [21, 23], one dimension histogram [24–26], and the feature vector [27–29]. The motion model is to model the motion pattern of the object. Usually, it is assumed that the motion of the object in the video is smooth. That is the center of the object in the next frame is around the center of the object in current frame. Therefore, the motion model is mainly based on the Gaussian distribution [30, 31]. In some trackers, the detection methods and the tracking algorithms are combined to construct the motion model of the object, such as TLD tracker[32]. Tracking algorithms apply the appearance model and motion model to predict the position and state of the object. Tracking algorithms are divided into two kinds: matching based [33] and classification based methods [34] which are also known as generative tracking algorithms and discriminative tracking algorithms. Along with the heating up of deep learning and artificial intelligence, deep learning based tracking algorithms also got a rapid development [35, 36]. There are two ways to update the appearance model. One way is to use the object template calculated in the current frame to partially update the whole object appearance model [37, 38], such as the weighted sum. The other way is to reconstruct the appearance model entirely based on the object template computed in the current frame [39].

1.3 Evolution of Visual Object Tracking Technology

The development of the visual object tracking algorithm shows a trend from traditional tracking methods [40–42] to deep learning based tracking methods [43–45], and from generative methods [30, 46–48] to discriminative methods [23, 34, 49]. The evolution of visual object tracking technology can be simply summarized by Fig. 1.14.

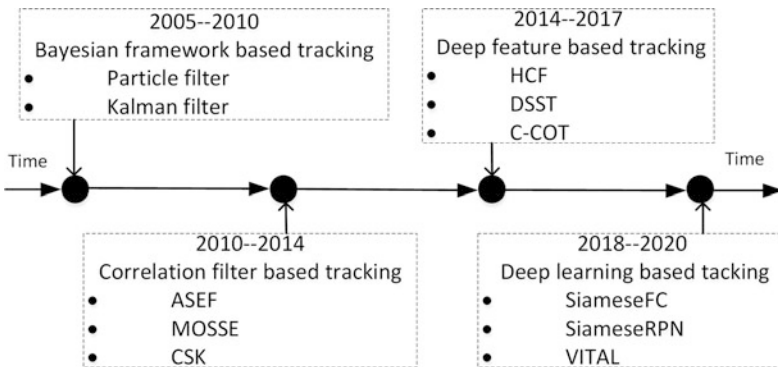


Fig. 1.14 The evolution of visual object tracking

From 2005 to 2010, visual object tracking methods are mainly based on the Bayesian framework, Particle filter, and Kalman filter, which are generative methods. During this period, visual object tracking is mainly considered as a template matching problem. Some manual designed features are used to construct the appearance model for comparison. Gaussian distribution is applied as a motion model to provide candidate objects and the final state of the object can be calculated by finding the candidate with the maximum similarity. Abdel-Hadi et al. [46] and Han et al. [47] proposed visual object trackers based on the Kalman filter and particle filter respectively. Yang et al. [48] extracted the superpixel feature to construct the appearance model for comparison.

From 2010 to 2014, the correlation filtering based trackers with kernel methods which belongs to the discriminative methods are widely studied by researchers [50–54]. The purpose of correlation filtering based trackers is to train a correlation filter which could make the object center locate at the peak value in the response map after the correlation filtering operation. Bolme et al. [50] applied the correlation filter to determine the location of eye and proposed the ASEF filter. Then, Bolme et al. [51] further improved the ASEF filter and applied the correlation filter to visual object tracking, and proposed the MOSSE tracker which is also the first correlation filter based tracker. Henriques et al. [52] proposed the CSK tracker which tries to solve the correlation filter with a linear classifier. In 2014, Henriques et al. [53] viewed tracking as a ridge regression problem and applies a circulant matrix to collect the positive and negative samples around the object for training the correlation filter. Aiming at the problems of scale variation in the KCF tracker, Danelljan et al. [54] applied two correlation filters: translation filter and scale filter. Translation filter is used to detect the center location of the object and the scale filter is used to estimate the scale change of the object.

Because of the powerful representation ability of deep feature, deep feature is merged into the correlation filtering based trackers during 2015 and 2017 [55–57]. Well pre-trained networks are used as feature extractors. Ma et al. [55] applied a pretrained deep network to extract the deep feature of the object and combined multi-features from feature maps of different layers in deep network to construct the proposed HCF tracker. In addition, the characteristics of the feature map from different layers in the deep network are also discussed in [55]. Hong et al. [56] proposed a learnable saliency map based on CNN, and combined the saliency map with SVM based classifier to construct the appearance model. Danelljan et al. [57] proposed the continuous convolution operators to integrate multiple resolution feature maps and achieve accurate sub-pixel location.

From 2018 to 2020, along with the development of deep network, deep learning based tracking methods also get rapid developments, especially the Siamese network which obtained a remarkable performance in visual object tracking [36, 58, 59]. Bertinetto et al. [58] combined the Siamese network and correlation filter to propose the SiameseFC tracker. Li et al. [36] introduced the region proposal network into the Siamese network to provide candidates of the object. The region proposal network in [36] can be viewed as a motion model in visual object tracking. Wang et al. [59] combined visual object tracking with instance segmentation