A review of several key techniques of target vision tracking in complex scenes

Zhengyu Wang¹, Ke Dong²

¹Community Education College, Anhui Radio and Television University, Hefei, 230022, China
²College of Economics and Management, Anhui Radio and Television University, Hefei, 230022, China

E-mail address: wangzy0604@163.com

Abstract. Target vision tracking is a hot research issue in the field of machine vision. It has been studied by many countries and widely used in many fields. Its main task is to detect, identify and locate moving targets in the video sequence, and obtain the parameters such as position, velocity, acceleration, direction, motion trajectory of the target, in order to provide the basis for further intelligent video analysis. Due to the complexity and unpredictability of the target itself and the external environment, robust real-time target tracking in complex scenes has become a key issue to be solved urgently. This paper aims to summarize key techniques such as motion detection, target classification and target tracking involved in video target tracking. And then the paper introduces the algorithms commonly used in several key techniques, compares and analyzes the advantages and disadvantages of various algorithms. Besides, the paper also summarizes the current research status and research issues, and prospects the future research trends.

1. Introduction

The main task of target vision tracking is to detect, identify and locate the moving target in the video sequence, and obtain the target's position, velocity, acceleration, direction, motion trajectory and other parameters, so as to provide the basis for further intelligent video analysis. Target vision tracking belongs to the category of image analysis and understanding, involving pattern recognition, computer vision, image processing, artificial intelligence, machine learning, probability theory and statistical analysis and other interdisciplinary frontier research topics. With the rapid increase of the application of video images in the economic and social fields, the research on target vision tracking has an extremely broad application prospect. Specifically speaking, it mainly has five aspects: (1) Intelligent video monitoring. The demand for intelligent video monitoring system comes mainly from those occasions that are sensitive to security requirements, such as airports, banks, shops. (2) Vision-based intelligent vehicle navigation. Intelligent vehicle navigation based on vision is an important part of Intelligent Transportation Systems, which is one of the research hotspots in the field of Intelligent Transportation in recent years. (3) Perceived interfaces. In the field of advanced user interface applications, it’s hoped that future machines will be able to communicate with us more easily and conveniently like humans, such as gesture-driven control, sign language translation. (4) Motion analysis. Segmenting the human body parts from the image, tracking and analyzing the joint motion of interest in the image sequence has a positive effect on establishing the geometric model of the human body, explaining the body’s motion and behavior mechanism, and thus improving the body’s motion function. The research conclusions can
be applied to guide sports, dance training and medical gait analysis. (5) Virtual reality. The realistic design of the shape, motion and behavior interaction of target animation in many computer games actually benefits from the target motion analysis in physical space, including the target skeleton model, the acquisition of the joint motion mechanism and the recovery of the posture. Vision tracking of targets primarily involves motion detection, target classification and target tracking and other key techniques.

2. Motion detection
The purpose of motion detection is to extract the changed regions from the background from the image sequence. Fast and efficient motion region segmentation is very important for target classification, tracking and behavior understanding, because the post-processing only needs to consider the pixels of the moving area in the image. However, the dynamic changes in background images (such as weather, lighting, shadow of moving targets and the influence of chaotic interference) present difficulties to motion detection. At present, several commonly used motion detection methods include:

2.1. Background subtraction.
Background subtraction method is the most commonly used method in motion segmentation at present. It is a method that uses the difference between the current image and the background image and thresholding to detect the motion region. Generally, it can provide the most complete feature data, but it is especially sensitive to the changes of dynamic scene, such as the interference of illumination and external things. The simplest background model is the time average image. At present, many researchers are working on good background models to reduce the impact of dynamic scene changes on accurate segmentation. Haritaoglu et al. used the minimum, maximum intensity value and maximum time difference score to conduct statistical modeling for each pixel in the scene, and carried out periodic background update. McKenna et al. used an adaptive background model combining pixel color and gradient information to solve the influence of shadow and unreliable color cues on segmentation. Stauffer and Grimson used adaptive background mixed gaussian model (every pixel was modeled by mixed gaussian distribution) and updated the model with online estimation, so as to effectively deal with the influence of light change and background chaotic motion. Lee introduced an adaptive learning factor on the method, which improves the convergence speed.

2.2. Time-domain difference.
The time domain difference method uses pixel-based time difference and thresholding between adjacent frames in the image sequence to extract the motion regions in the image. Lipton detected the moving target from the actual video image with this method, and realized the classification and tracking of the target. This method has strong adaptability to the dynamic environment, but generally cannot fully extract all the relevant feature pixels, and cavitation is easy to occur in the moving entity.

2.3. The light flow.
The motion detection based on optical flow method adopts the optical flow characteristics of the moving target changing with time, and its advantage is that a moving target can be detected independently under the premise of camera motion. However, most of the optical flow computing methods are rather complex and have poor anti-noise performance, which cannot be applied to real-time processing of full-frame video streams without special hardware devices. Meyer et al. initialized the contour based on tracking algorithm by calculating the displacement vector optical flow field, thereby effectively extracting and tracking the articulated moving target.

2.4. Subspace analysis.
The motion detection method based on subspace analysis updates the low-dimensional feature space of the background through Online PCA, and performs foreground segmentation by calculating the distance between scene pixels and the background space. The advantage of this method is that it can effectively learn the time distribution information of background pixels and update the acquired feature
subspace model incrementally. The disadvantage is that the spatial distribution information of background pixels is lost, which makes the motion detection results more sensitive to noise.

Of course, there are other algorithms in the motion detection, such as Friedman et al. [13] used the extension of the (EM) Expectation Maximization algorithm to establish, the classification of gaussian mixture model established for each pixel. The model can be automatically updated and adaptively classify each pixel into the background, the shadow or sports outlook, so that it can perform the motion region segmentation well when the target motion speed is slow, in addition, it can also effectively eliminate the influence of the moving target shadow. VSAM [14] developed a hybrid algorithm combining adaptive background subtraction and three-frame difference, which can quickly and effectively detect moving targets. Wang et al. [15] proposed a dynamic conditional random field model to accomplish the task of foreground target segmentation and moving shadow detection. Candes et al. [16] proposed a background subtraction algorithm based on low-rank matrix restoration, which can deal with large background noise. Han and Davis [17] proposed a foreground detection algorithm based on SVM classification, which can integrate multiple features.

3. Target classification
In the monitoring scene, different motion regions may correspond to different moving targets. For example, the sequence images captured by the surveillance camera on the traffic road may contain moving objects such as pedestrians and vehicles, and the corresponding behaviors of different objects will vary greatly, the processing methods will also be quite different. The purpose of moving target classification is to classify the detected motion regions for further processing and behavior analysis. Target classification in monitoring scenario mainly includes two key technologies: feature extraction technology and classifier technology.

3.1. Feature extraction technique.
It mainly includes three types of extraction: shape feature, motion feature and abstract feature.

Shape features. For example, Lipton et al. [18] used dispersion and area information to classify two-dimensional motion regions, mainly to distinguish people, vehicles and chaotic disturbances, and the time consistency constraint made the classification more accurate. Kuno and Watanabe [19] used simple shape parameters of human contour mode to detect the motion of people from images. Literature [20] classified motor vehicles and pedestrians by using shape information such as the direction and axis length ratio of the envelope ellipse.

Motion characteristics. For example, Cutler and Davis [21] calculated the autocorrelation characteristics of the target with time by tracking the moving target of interest, and the periodic motion of people makes the autocorrelation periodic. Therefore, they identified people by analyzing whether the target has periodic motion characteristics by time-frequency method. Lipton [22] analyzed the rigidity and periodicity of the moving entity by calculating the Residual light Flow in the moving region, so as to distinguish the target. Javed et al. [23] used circular motion features to distinguish people from cars.

Abstract features. Such as Haar features [24], SIFT features with scale-invariant characteristics [25], The Principle Component Analysis (PCA) is adopted to obtain the mathematical characteristics [26].

3.2. Classifier technique.
In addition to selecting appropriate features, selecting appropriate classification method is also the key part of achieving the correct classification. Target classification in monitoring scenarios has special performance requirements for classifiers. Compared with general pattern classification problems, it requires faster classification speed and higher classification accuracy, so special classification mechanism must be designed. In recent years, the classification mechanism based on monitoring scenario focuses on the following two aspects:

A classification mechanism based on a single classifier. For example, a single Support Vector Machine (SVM) is adopted [27,28,29,30], Artificial neural network is adopted [31]. SVM has strong theoretical support and has been well verified in other applications. Artificial neural network has strong
nonlinear partitioning ability and can adapt to large deformation.

A Classification mechanism based on combined classifier. The combined classification organizes multiple single classifiers according to a certain structure, and comprehensively judges the category of the target to be detected. The typical representative is Adaboost [32]. According to the organization of single classifier, it can be divided into serial arrangement [33], parallel arrangement [34] and cascade combination [35,36]. The performance of the combined classifier is greatly influenced by the organization of single classifier [37].

4. Target tracking

The tracking problem is equivalent to creating the corresponding matching problem of related features based on position, speed, shape, texture and color between successive image frames. Common mathematical tools include Kalman Filtering [38], Condensation algorithm [39,40], Mean Shift [41], Dynamic Bayesian Network [42, 43] and geodesic line [44, 45]. There are two ways to deal with this problem: a top-down processing method and bottom-up processing method.

Top-down processing method can also be called the model driven method. This method first establishes a target model, and then turns a tracking problem into a matching problem or the maximum likelihood problem, in order to meet the target appearance changes and the change of scene. The target model often needs to update or online learning, this method has a solid mathematical theoretical basis and is the mainstream method to study visual tracking problems.

The bottom-up processing method can also be called a data-driven method, which directly obtains the target area and its motion information through methods such as background clipping. This method can quickly detect and track the target, but such method is more suitable for the tracking of the camera at rest.

According to the algorithms used in tracking, the vision tracking methods are divided into five categories: region-based tracking, feature-based tracking, active contour tracking, model-based tracking, and multi-strategy fusion tracking methods.

(1) Region-based tracking [46,47]: region-based tracking method has been in wide use at present, which can be useful to track the changes of the corresponding moving target region in the image. In this type of method, the background is obtained by dynamic modeling, and the foreground is obtained by background clipping.

(2) Feature-based tracking [48,49,50]: it includes feature extracting and feature matching. Feature-based tracking can be divided into three categories according to the characteristics of its selected features: the method based on global features, the method based on local features, and the method combining global and local features. Global features include the center, diameter, and color of the target. Methods based on local features include line segmentation, curve segmentation, and corner information.

(3) Tracking based on deformation templates: deformation templates are panels or curves whose textures or edges can be deformed under certain restrictions. The commonly used deformation template is the active contour model proposed by Kass in 1987, also known as Snake model [51]. This model combined with kalman filter can better track [52]. For multi-target tracking, the active contour model based on the level set method is more used [53].

(4) Model-based tracking: when tracking the human body, there are usually four forms of models, namely line graph model [54], 2D model [55], 3D model [56] and subspace model [57].

(5) Multi-strategy fusion tracking method [58,59,60]: in order to enhance the robustness and adaptability of the algorithm, most methods tend to fuse different features or tracking methods, and group them into the same framework by switching or weighting.

In addition, vision tracking methods can be divided into: point tracking, core tracking, and silhouette tracking. Point tracking can be divided into deterministic method and probabilistic statistical method. The kernel tracking methods include template-based tracking and multi-angle tracking. Silhouette tracking can be divided into contour evolution method and shape matching [61].

5. Conclusion
For target motion tracking and behavior understanding, the main development trends and research difficulties can be summarized as follows:

(1) Motion segmentation with adaptive background model. Fast and accurate motion segmentation is an important but difficult problem. This is because the images captured in the dynamic environment are affected by many aspects, such as the change of weather, the change of lighting conditions, the chaotic interference in the background, the shadow of the moving targets, and even the motion of the camera itself, all of which bring difficulties to the accurate and effective motion segmentation. Although the background subtraction method is mainly used in image motion segmentation, it is still an urgent problem to establish a reliable adaptive background model for dynamic changes of any complex environment.

(2) Effectively feature extraction and accurate recognition of moving targets. The effective extraction of moving target features in complex scenes is the premise of accurate target recognition. In video monitoring system, especially in outdoor video monitoring, usually the target contains few pixels, the target is ambiguous, the target form changes greatly, coupled with the projection deformation and geometric distortion, it is difficult to classify the target. How to choose the appropriate features to make the target classification to best resist the influence of these objective conditions is a difficult part of video monitoring technology.

(3) Occlusion processing in the tracking process. At present, most of the target motion analysis systems are unable to solve the problem of mutual occlusion and self-occlusion between targets, especially in the process of multi-target detection and tracking in the crowded state. In the case of occlusion, only part of the target is visible, and this process is untrainable. Therefore, the technology that simply relies on background subtraction to segment motion will no longer be reliable, and a better model must be able to deal with the correspondence between features and parts of the target in the case of occlusion. However, the general system cannot provide when to stop and restart the tracking of the target, i.e. the tracking initialization before and after occlusion lacks bootstrap method. The current method of interest is to use statistical characteristics to predict the target motion from the available image information.

In short, the target’s motion tracking, as a research field with important scientific significance and great application prospects in the field of computer vision, has attracted extensive attention from domestic industry and foreign academic circles, and has carried out a lot of research work, but there are still many theoretical and practical issues that need to be resolved, and further in-depth research work is needed.

Acknowledgement
This research is supported by the Key Program of Anhui Natural Science, China (No. KY2018A0687); Anhui Radio and Television University Young Teacher Research Fund Project (No. qn15-30).

Reference
[1] Pentland A. Looking at people: sensing for ubiquitous and wearable computing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 22(1):107-119.
[2] Mohan A, Pajageorgiou C, Poggio T. Example-based object detection in images by components[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(4):349-361.
[3] Haritaoglu I, Harwood D, Davis L S. W4: Real-Time Surveillance of People and Their Activities[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 22(8):809-830.
[4] Mckenna S J, Jabri S, Duric Z, et al. Tracking Groups of People[J]. Computer Vision & Image Understanding, 2000, 80(1):42-56.
[5] Stauffer C, Grimson W E L. Adaptive background mixture models for real-time tracking[C]//Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149). IEEE, 1999, 2: 246-252.
[6] Lee D S. Effective Gaussian Mixture Learning for Video Background Subtraction[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(5):827-832.
[7] Lipton A J, Fujiyoshi H, Patil R S. Moving target classification and tracking from real-time video[C]//Proceedings Fourth IEEE Workshop on Applications of Computer Vision. WACV’98. IEEE, 1998: 8-14.

[8] Meyer D. Model based extraction of articulated objects in image sequences for gait analysis[C]//International Conference on Image Processing. IEEE Computer Society, 1997.

[9] Barron J L, Fleet D J, Beauchemin S S. Performance of optical flow techniques[J]. International Journal of Computer Vision, 1994, 12(1):43-77.

[10] Thonnat M. Image understanding for visual surveillance applications[C]//Proc. 3rd Int. Workshop on Cooperative Distributed Vision, Nov. 1999. 51-82.

[11] Torre F D L, Black M J. A Framework for Robust Subspace Learning[J]. International Journal of Computer Vision, 2003, 54(1-3):117-142.

[12] Li Y. On incremental and robust subspace learning[J]. Pattern Recognition: The Journal of the Pattern Recognition Society, 2004, 37(7):1509-1518.

[13] Friedman N, Russell S. Image Segmentation in Video Sequences: A Probabilistic Approach[J]. 2013.

[14] R. T. Collins et al, “A system for video surveillance and monitoring”, Technical Report, CMU-RI-TR-00-12, Carnegie Mellon University, 2000.

[15] Wang Y, Loe K F, Wu J K. A dynamic conditional random field model for foreground and shadow segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2005, 28(2): 279-289.

[16] Emmanuel J. Candès, Li X, Ma Y, et al. Robust principal component analysis?[J]. Journal of the Acm, 2011, 58(3):1-37.

[17] Han B, Davis L S. Density-Based Multifeature Background Subtraction with Support Vector Machine[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(5):1017-1023.

[18] Lipton A J, Fujiyoshi H, Patil R S. Moving target classification and tracking from real-time video[C]//Proceedings Fourth IEEE Workshop on Applications of Computer Vision. WACV’98. IEEE, 1998: 8-14.

[19] Kuno Y, Watanabe T, Shimosakoda Y, et al. Automated Detection of Human for Visual Surveillance System[C]// International Conference on Pattern Recognition. IEEE, 1996.

[20] Haritaoglu I, Harwood D, David L S. W4: Real-time surveillance of people and their activities[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 22(8):809-830.

[21] Cutler R, Davis L S. Robust Real-Time Periodic Motion Detection, Analysis, and Applications[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1999, 22(8):781-796.

[22] Lipton A J. Local application of optic flow to analyse rigid versus non-rigid motion[M]. Carnegie Mellon University, The Robotics Institute, 1999.

[23] Javed O, Shah M. Tracking and object classification for automated surveillance[C]//European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2002: 343-357.

[24] Viola P, Jones M J, Snow D. Detecting pedestrians using patterns of motion and appearance[J]. International Journal of Computer Vision, 2005, 63(2): 153-161.

[25] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[26] Munder S, Gavrila D M. An experimental study on pedestrian classification[J]. IEEE transactions on pattern analysis and machine intelligence, 2006, 28(11): 1863-1868.

[27] Cheng H, Zheng N, Qin J. Pedestrian detection using sparse Gabor filter and support vector machine[C]//IEEE Proceedings. Intelligent Vehicles Symposium, 2005. IEEE, 2005: 583-587.

[28] Tian Q, Sun H, Luo Y, et al. Nighttime Pedestrian Detection with a Normal Camera Using SVM Classifier[J]. Lecture Notes in Computer Science, 2005.

[29] Grubb G, Zelinsky A, Nilsson L, et al. 3D vision sensing for improved pedestrian safety[C]//IEEE Intelligent Vehicles Symposium, 2004. IEEE, 2004: 19-24.
[30] Schauland S, Kummert A. Implementation and optimization of wavelet and symmetry features for vision-based pedestrian detection[M]. ACTA Press, 2007.

[31] Munder S, Gavrila D M. An experimental study on pedestrian classification[J]. IEEE transactions on pattern analysis and machine intelligence, 2006, 28(11): 1863-1868.

[32] Viola P, Jones M J, Snow D. Detecting pedestrians using patterns of motion and appearance[J]. International Journal of Computer Vision, 2005, 63(2): 153-161.

[33] Shashua A. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance[C]// Proc. Intelligent Vehicle Symposium, 2004. IEEE, 2004.

[34] Grubb G, Zelinsky A, Nilsson L, et al. 3D vision sensing for improved pedestrian safety[C]//IEEE Intelligent Vehicles Symposium, 2004. IEEE, 2004: 19-24.

[35] Cao X B, Qiao H, Keane J. A low-cost pedestrian-detection system with a single optical camera[J]. IEEE Transactions on Intelligent Transportation Systems, 2008, 9(1): 58-67.

[36] Wei C X, Cao X B, Xu Y W, et al. The treelike assembly classifier for pedestrian detection[C]//Pacific-Asia Workshop on Intelligence and Security Informatics. Springer, Berlin, Heidelberg, 2007: 232-237.

[37] Fujimura, Kikuo. Pedestrian detection and tracking with night vision[J]. IEEE Transactions on Intelligent Transportation Systems, 2005, 6(1):63-71.

[38] Links I K F. An Introduction to the Kalman Filter[J]. 1995.

[39] Isard M, Blake A. CONDENSATION—Conditional Density Propagation for Visual Tracking[J]. International Journal of Computer Vision, 1998, 29(1):5-28.

[40] Sidenbladh H, Black M J, Fleet D J. Stochastic tracking of 3D human figures using 2D image motion[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2000: 702-718.

[41] Comaniciu D, Ramesh V, Meer P. Kernel-based object tracking[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2003, 25(5):564-575.

[42] Pavlovic V, Rehg J M, Cham T J, et al. A dynamic Bayesian network approach to figure tracking using learned dynamic models[C]//Proceedings of the seventh IEEE international conference on computer vision. IEEE, 1999, 1: 94-101.

[43] Nillius P, Sullivan J, Carlsson S. Multi-target tracking-linking identities using bayesian network inference[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06). IEEE, 2006, 2: 2187-2194.

[44] Paragios N, Deriche R. Geodesic active regions: A new framework to deal with frame partition problems in computer vision[J]. Journal of Visual Communication and Image Representation, 2002, 13(1-2): 249-268.

[45] Paragios N, Deriche R. Geodesic active contours and level sets for the detection and tracking of moving objects[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2000, 22(3):266-280.

[46] Matthews I, Ishikawa T, Baker S. The Template Update Problem[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2004, 26(6):810-815.

[47] Lim H, Camps O I, Szaier M, et al. Dynamic Appearance Modeling for Human Tracking[C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2006.

[48] Comaniciu D, Ramesh V, Meer P. Kernel-based object tracking[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2003, 25(5):564-575.

[49] Tissainayagam P, Suter D. Object tracking in image sequences using point features[J]. Pattern Recognition, 2005, 38(1):105-113.

[50] Nickels K, Hutchinson S. Estimating uncertainty in SSD-based feature tracking[J]. Image and Vision Computing, 2002, 20(1):47-58.

[51] Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models[J]. International Journal of Computer Vision, 1988, 1(4):321-331.

[52] Terzopoulos D, Szeliski R. Tracking with Kalman snakes[M]// Active vision. 1993.
[53] Yilmaz A, Li X, Shah M. Contour-based object tracking with occlusion handling in video acquired using mobile cameras[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(11):0-1536.
[54] Karaulova I A, Hall P M, Marshall A D. A Hierarchical Model of Dynamics for Tracking People with a Single Video Camera[C]//BMVC. 2000: 1-10.
[55] Ju S X, Black M J, Yacoob Y. Cardboard people: A parameterized model of articulated image motion[C]//Proceedings of the Second International Conference on Automatic Face and Gesture Recognition. IEEE, 1996: 38-44.
[56] Balan A O, Black M J. An adaptive appearance model approach for model-based articulated object tracking[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006, 1: 758-765.
[57] Nguyen H T, Ji Q, Smeulders A W M. Spatio-Temporal Context for Robust Multitarget Tracking[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(1):52.
[58] Morenonoguer F, Sanfeliu A, Samaras D. Dependent multiple cue integration for robust tracking[J]. IEEE Trans Pattern Anal Mach Intell, 2008, 30(4):670-685.
[59] Lee H S, Kim D. Robust face tracking by integration of two separate trackers: Skin color and facial shape[M]. Elsevier Science Inc. 2007.
[60] Yilmaz A, Javed O, Shah M. Object tracking: A survey[J]. Acm computing surveys (CSUR), 2006, 38(4): 13-57.