Monocular and Binocular People Tracking

Chong Luo and Wenjun Zeng
Microsoft Research Asia, Beijing, China

Synonyms

Person following

Related Concepts

- ▶ Articulated People Tracking
- ▶ Multiple Object Tracking (MOT)
- ▶ People Detection
- ▶ Person Re-identification
- ▶ RGB-D People Tracking
- ▶ Visual Object Tracking (VOT)

Definition

People tracking is the process of estimating and recording the locations of target people in 2D image sequences (monocular computer vision) or 3D spaces over time (binocular computer vision). It may also refer to the process of estimating and recording the pose (or joint locations) of target people.

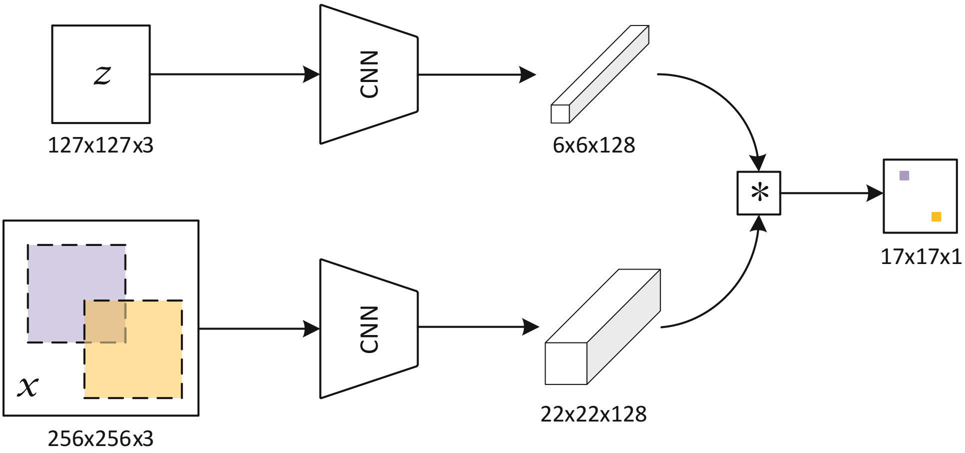
Background

Visual object tracking is a fundamental problem in computer vision. As a person is the most important type of object, people tracking has received tremendous interest from both academia and industry. While generic object tracking methods can be directly applied to people tracking, there are also many algorithms and schemes that are tailored for people tracking.

Monocular People Tracking

Monocular people tracking finds applications in surveillance, video indexing, and many other video analytic applications. When a specific person is considered, people tracking can be treated as a generic visual object tracking problem. The most well-known tracking algorithm is the Kanade-Lucas-Tomasi (KLT) feature tracking algorithm [1]. The basic idea is to find a bunch of good features to track and then estimate the object changes based on the points movement. It was the dominant tracking algorithm before the MOSSE tracker [2] was proposed. MOSSE tracker is a correlation filter (CF)-based tracker designed for fast object tracking. There have been several modifications to this method, including kernelization [3] and scale adaptation [4]. In recent years, the Siamese fully convolutional network (SiamFC)-based trackers have attracted much attention [5, 6]. These trackers also perform fairly well on human

K. Ikeuchi, *Computer Vision*,
https://doi.org/10.1007/978-3-030-03243-2_872-1



Monocular and Binocular People Tracking, Fig. 1 Fully convolutional Siamese network for visual object tracking [5], where z is the target object and x is the search image. CNN denotes the convolutional neural network for feature extraction, and * denotes the cross-correlation operation. The output is a scalar-valued score map, where the location of the maximal value indicates the location of the object in the search image

objects. Figure 1 shows the architecture of the original SiamFC tracker.

When multi-person tracking is considered, it can be similarly treated as a generic multiple object tracking (MOT) problem. In MOT research, it is often assumed that object detection results are given in each frame as input, and the tracking task is to associate the detections to find object trajectories. Existing works solve the data association problem based on the Hungarian algorithm, Markov decision process (MDP) [7], or more complex graph models [8]. There are a handful of efforts that treat object detection and tracking as a coupled optimization problem [9].

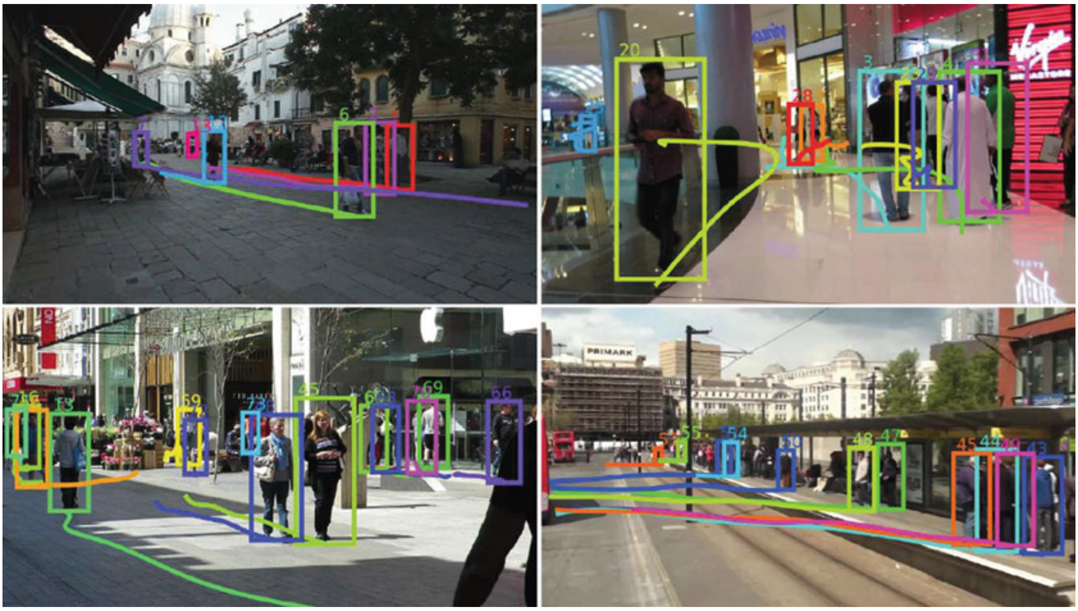
For the human object, there is a stronger demand to jointly optimize detection and tracking. It is observed that people detectors are able to locate pedestrians even in complex street scenes, but false positives have remained frequent. Tracking methods are able to find a particular individual in image sequences but are severely challenged by real-world scenarios such as crowded street scenes. Therefore, the advantages of detection and tracking should be combined in a single framework [10]. Breitenstein et al. [11] propose a tracking-by-detection algorithm which uses the continuous confidence of pedestrian detectors and online trained, instance-specific classifiers as a graded observation model. Tang et al. [12] propose training people detectors explicitly on failure cases of the overall tracker.

In addition to the tracking algorithms which treat a human in its entirety, there are approaches which utilize the body parts or articulations. Rather than simply determining the position and scale of a person, Andriluka et al. also extract the 2D articulation [10] or 3D pose [13] for the tracking task. Shu et al. [14] extend part-based human detection methods to the tracking task. Henschel et al. [15] use body parts, such as the head and/or shoulder, to facilitate person tracking in very crowded scenes.

Note that, although person reidentification has been a separate research topic, it is a highly related concept. Traditionally, person re-ID addresses the association of people across nonoverlapping cameras, but it can be used for linking the detected persons across the whole video after long-term occlusion in multi-person tracking [16]. Some of the results achieved by this work are shown in Fig. 2.

Binocular People Tracking

People detection and tracking are key capabilities for a mobile robot acting in populated environ-



Monocular and Binocular People Tracking, Fig. 2 Some tracking results achieved by [16] on the MOT16 benchmark. The line under each bounding box indicates the lifetime of the track

ments. It is often addressed in the context of binocular vision because a robot can be equipped with two cameras. Conversely, robot control and human-robot interaction are the most important applications of binocular people tracking.

Around a decade ago, when the mainstream tracking algorithms were feature-based tracking, Chen et al. [17] proposed a binocular person following algorithm which uses Lucas-Kanade feature detection and matching to determine the location of the person in the image and thereby control the robot. Matching was performed between two images of a stereo pair, as well as between successive video frames. It is one of the earliest works that do not rely on clothing colors for tracking.

Usually, depth information can be estimated from a pair of stereo images. The depth information complements the appearance model and allows tracking algorithms to achieve good results in challenging scenarios. The majority of binocular people tracking algorithms are directly based on RGB-D data, but the depth information can be used in different ways. For example, Ess et al. [18] consider multi-person tracking in busy pedestrian zones. The depth information is used to verify people candidates obtained by a people detector. Luber et al. [19] combine a multi-cue person detector for RGB-D data with an online detector that learns individual target models. In the 3D tracking method proposed by Cao et al. [20], depth information obtained from a moving binocular camera is used to detect and recover from occlusions.

Open Problems

People tracking faces challenges when there are appearance changes due to illumination, pose changes, or cluttered background. But the biggest challenge arises from the person interactions and long-term occlusions in crowded real-world scenes. These may cause identity switches between multiple tracked people. Incorporating long-range person re-ID into the tracking algorithms could be a feasible solution. Besides, joint detection and tracking is a promising approach to improve the overall system performance.

References

1. Shi J, Tomasi C (1993) Good features to track. Technical report, Cornell University
2. Bolme DS, Beveridge JR, Draper BA, Lui YM (2010) Visual object tracking using adaptive correlation filters. In: 2010 IEEE computer society conference on computer vision and pattern recognition, pp 2544–2550. IEEE
3. Henriques J, Caseiro R, Martins P, Batista J (2015) High-speed tracking with kernelized correlation filters. IEEE Trans Pattern Anal Mach Intell 37(3):583–596
4. Danelljan M, Häger G, Khan FS, Felsberg M (2017) Discriminative scale space tracking. IEEE Trans Pattern Anal Mach Intell 39(8):1561–1575
5. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PHS (2016) Fully-convolutional siamese networks for object tracking. In: European conference on computer vision, pp 850–865
6. Li B, Yan J, Wu W, Zhu Z, Hu X (2018) High performance visual tracking with siamese region proposal network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8971–8980
7. Xiang Y, Alahi A, Savarese S (2015) Learning to track: Online multi-object tracking by decision making. In: Proceedings of the IEEE international conference on computer vision, pp 4705–4713
8. Berclaz J, Fleuret F, Turetken E, Fua P (2011) Multiple object tracking using k-shortest paths optimization. IEEE Trans Pattern Anal Mach Intell 33(9):1806–1819
9. Leibe B, Schindler K, Van Gool L (2007) Coupled detection and trajectory estimation for multi-object tracking. In: 2007 IEEE 11th international conference on computer vision, pp 1–8. IEEE
10. Andriluka M, Roth S, Schiele B (2008) People-tracking-by-detection and people-detection-by-tracking. In: 2008 IEEE conference on computer vision and pattern recognition, pp 1–8. IEEE
11. Breitenstein MD, Reichlin F, Leibe B, Koller-Meier E, Van Gool L (2009) Robust tracking-by-detection using a detector confidence particle filter. In: 2009 IEEE 12th international conference on computer vision, pp 1515–1522. IEEE
12. Tang S, Andriluka M, Milan A, Schindler K, Roth S, Schiele B (2013) Learning people detectors for tracking in crowded scenes. In: Proceedings of the IEEE international conference on computer vision, pp 1049–1056
13. Andriluka M, Roth S, Schiele B (2010) Monocular 3D pose estimation and tracking by detection. In: 2010 IEEE computer society conference on computer vision and pattern recognition, pp 623–630. IEEE
14. Shu G, Dehghan A, Oreifej O, Hand E, Shah M (2012) Part-based multiple-person tracking with partial occlusion handling. In: 2012 IEEE conference on computer vision and pattern recognition, pp 1815–1821. IEEE
15. Henschel R, Leal-Taixé L, Cremers D, Rosenhahn B (2018) Fusion of head and full-body detectors for multi-object tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1428–1437
16. Tang S, Andriluka M, Andres B, Schiele B (2017) Multiple people tracking by lifted multicut and person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3539–3548
17. Chen Z, Birchfield ST (2007) Person following with a mobile robot using binocular feature-based tracking. In: 2007 IEEE/RSJ international conference on intelligent robots and systems, pp 815–820. IEEE
18. Ess A, Leibe B, Schindler K, Van Gool L (2009) Robust multiperson tracking from a mobile platform. IEEE Trans Pattern Anal Mach Intell 31(10):1831–1846
19. Luber M, Spinello L, Arras KO (2011) People tracking in rgb-d data with on-line boosted target models. In: 2011 IEEE/RSJ international conference on intelligent robots and systems, pp 3844–3849. IEEE
20. Cao L, Wang C, Li J (2015) Robust depth-based object tracking from a moving binocular camera. Signal Process 112:154–161