

# Multiple Object Tracking: A Literature Review

Wenhan Luo<sup>1</sup> · Junliang Xing<sup>2</sup> · Xiaoqin Zhang<sup>3</sup> · Xiaowei Zhao<sup>1</sup> · Tae-Kyun Kim<sup>1</sup>

Received: date / Accepted: date

**Abstract** Multiple Object Tracking (MOT) is an important computer vision task which has gained increasing attention due to its academic and commercial potential. Although different kinds of approaches have been proposed to tackle this problem, there still exists many issues unsolved. For example, factors such as abrupt appearance changes and severe object occlusions pose great challenges for MOT. In order to help the readers understand this topic and start working on it, we contribute a systematic and comprehensive review. In the review, we inspect the recent advances in various aspects about this topic and propose some interesting directions for future research.

To our best knowledge, there has not been any review about this topic in the community. We endeavor to provide a thorough review on the development of this problem in the last decades. The main contributions of this review are fourfold: 1) Key aspects in a multiple object tracking system, including how to formulate MOT generally, how to categorize MOT algorithms, what needs to be considered when developing a MOT system and how to evaluate a MOT system, are discussed from the perspective of understanding a topic. We believe in that this could not only provide researchers, especially new comers to the topic of MOT, a general understanding of the state of the arts, but also help them to comprehend the essential components of a MOT system and the inter-component connection. 2) Instead of enumerating individual works, we discuss existing work according to the various aspects involved in a MOT system. In each aspect, methods are divided

into different groups and each group is discussed in details for the principles, advances and drawbacks. 3) We examine experiments of existing publications and give tables which list results on the popular data sets to provide convenient comparison. We also provide some interesting discoveries by analyzing these tables. 4) We offer some potential directions and respective discussions about MOT, which are still open issues and need more research efforts. This would be helpful for researchers to identify further interesting problems.

**Keywords** Multiple Object Tracking · Observation Model · Dynamic Model · Object Detection · Association · Tracklet · Survey

## 1 Introduction

Multiple Object Tracking (MOT), or Multiple Target Tracking (MTT), plays an important role in computer vision. The task of MOT is largely partitioned to locating multiple objects, maintaining their identities and yielding their individual trajectories given an input video. Objects to track can be, for example, pedestrians on the street (Yang et al. 2011; Pellegrini et al. 2009), vehicles (Koller et al. 1994; Betke et al. 2000), sport players in the court (Lu et al. 2013; Xing et al. 2011; Nillius et al. 2006), or a flock of animals (birds (Luo et al. 2014), bats (Betke et al. 2007), ants (Khan et al. 2004), fishes (Spampinato et al. 2008, 2012; Fontaine et al. 2007), cells (Meijering et al. 2009; Li et al. 2008), etc.). The multiple “objects” could also be different parts of a single object. In this review, we mainly focus on the research about pedestrian. There are three underlying reasons for this. Firstly, compared to other conventional objects in computer vision, pedestrians are typical non-rigid objects, which is an ideal example to study the

<sup>1</sup> Imperial College London, London, United Kingdom

<sup>2</sup> National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

<sup>3</sup> Institute of Intelligent System and Decision, Wenzhou University, Zhejiang, China

MOT problem. Secondly, video of pedestrians raises a huge number of practical applications which further results in great commercial potential. Thirdly, according to incomplete statistics collected by ourselves, at least 70% of current MOT research efforts are devoted to pedestrian (*cf.* Table 9 to Table 26).

As a mid-level task in computer vision, multiple object tracking grounds high-level tasks such as action recognition, behavior analysis, etc. It has numerous applications. Some of them are presented in the following.

- Visual Surveillance. The massive amount of videos, especially surveillance videos, requires automatic analysis to detect abnormal behaviors, which is based on analyses of objects’ actions, trajectories, etc. To obtain such information, we need to locate targets and track them, which is exactly the objective of multiple object tracking.
- Human Computer Interface (HCI). Visual information, such as expression, gesture, can be employed to achieve advanced HCI. Extraction of visual information requires visual tracking as the basis. When multiple objects appear in the scene, we need to consider interactions among them. In this case, MOT plays a crucial rule to make the HCI more natural and intelligent.
- Virtual Augment Reality (VAR). MOT also has an application for this problem. For instance, MOT can supply users with better experience in video conferences.
- Medical Image Processing. Some tasks of medical image processing require laborious manual labeling, for instance, labeling multiple cells in images. In this case, MOT can help to save a large amount of labeling cost.

The various applications above have sparked enormous interest in this topic. However, compared with Single Object Tracking (SOT) which primarily focuses on designing sophisticated appearance models or motion models to deal with challenging factors such as scale changes, out-of-plane rotation and illumination variations, multiple object tracking additionally requires maintaining the identities among multiple objects. Besides the common challenges in both SOT and MOT, the further key issues making MOT challenging include (but not limit to): 1) frequent occlusions, 2) initialization and termination of tracks, 3) small size of objects (Betke et al. 2007), 4) similar appearance among objects, and 5) interaction among multiple objects. In order to deal with the MOT problem, a wide range of solutions have been proposed in recent years. These solutions focus on different aspects of a MOT system, making it difficult for researchers especially new comers to MOT to gain a comprehensive understanding of this

problem. Thus we here provide a review to discuss the various aspects of multiple object tracking.

### 1.1 Differences from Other Related Reviews

To the best of our knowledge, there has not been any comprehensive literature review on the topic of multiple object tracking. However, there have been some other reviews related to multiple object tracking, which are listed in Table 1. We group these surveys into three sets and highlight the differences from ours as follows. The first set (Zhan et al. 2008; Hu et al. 2004; Kim et al. 2010; Candamo et al. 2010) also involves crowd, i.e., multiple objects but the focus is different from ours. We intend to review the primary aspects in developing a multi-object tracking system. In comparison, tracking is only discussed as an individual part (Zhan et al. 2008; Hu et al. 2004; Kim et al. 2010; Candamo et al. 2010; Wang 2013). More specifically, Zhan et al. (2008) focuses on crowd modeling, thus object tracking is only a step to obtain feature for crowd modeling. Hu et al. (2004) and Kim et al. (2010) discuss papers about building a surveillance system for high-level vision tasks, such as behavior understanding, so tracking is an intermediate step. Candamo et al. (2010) review publications about behavior recognition in a special scenario, i.e., transit scenes. In that review, object tracking is discussed as a core technology as well as motion detection and object classification. Multiple object tracking is also discussed as one module for video surveillance under multiple cameras (Wang 2013). The second set (Forsyth et al. 2006; Cannons 1991; Yilmaz et al. 2006; Li et al. 2013) is dedicated to general visual tracking techniques (Forsyth et al. 2006; Cannons 1991; Yilmaz et al. 2006) or some special issues such as appearance models in visual tracking (Li et al. 2013). Their scope is wider than our review while our review is more comprehensive and detailed in multiple object tracking. The third set (Wu et al. 2013a; Leal-Taixé et al. 2015) is about benchmark work about general visual tracking (Wu et al. 2013a) and specific multiple object tracking (Leal-Taixé et al. 2015). Their attention is paid on experimental study rather than literature review.

### 1.2 Contributions

The main contributions of this review are fourfold:

- Key aspects in a multiple object tracking system, including how to formulate MOT generally, how to categorize MOT algorithms, what needs to be considered when developing a MOT system and how

Table 1: Summary of other literature reviews

Reference	Topic	Year
Zhan et al. (2008)	Crowd Analysis	2008
Hu et al. (2004)	Object Motion and Behaviors	2004
Kim et al. (2010)	Intelligent Visual Surveillance	2010
Candamo et al. (2010)	Behavior Recognition in Transit Scenes	2010
Wang (2013)	Multi-Camera Video Surveillance	2013
Forsyth et al. (2006)	Human Motion Analysis	2006
Cannons (1991)	Visual Tracking	2008
Yilmaz et al. (2006)	Object Visual Tracking	2006
Li et al. (2013)	Appearance Models in Object Tracking	2013
Wu et al. (2013a)	Visual Tracking Benchmark	2013
Leal-Taixé et al. (2015)	MOT Benchmark	2015

to evaluate a MOT system, are discussed from the perspective of understanding a topic. We believe in that this could not only provide researchers, especially new comers to the topic of MOT, a general understanding of the state of the arts, but also help them to comprehend the essential components of a MOT system and the inter-component connection.

- Instead of enumerating individual works, we discuss existing work according to the various aspects involved in a MOT system. In each aspect, methods are divided into different groups and each group is discussed in details for the principles, advances and drawbacks.
- We examine experiments of existing publications and give tables which list results on the popular data sets to provide convenient comparison. We also provide some interesting discoveries by analyzing these tables.
- We offer some potential directions and respective discussions about MOT, which are still open issues and need more research efforts. This would be helpful for researchers to identify further interesting problems.

It is worthy noting that, this manuscript is dedicated to literately reviewing recent advances in multiple object tracking. As mentioned above, we also present experimental results in publicly available data sets excerpted from existing publications to provide a view of how the state-of-the-art MOT methods work. For more comprehensive survey of experimental study in multiple object tracking, readers may refer to the work by Leal-Taixé et al. (2015).

### 1.3 Organization of This Review

The goal of this review is to provide an overview of the major aspects which readers need to consider when developing a system for multiple object tracking. These aspects include what is the current state of research

about MOT, all the detailed issues requiring consideration in building a system, and how to evaluate a MOT system. Accordingly, the organization of this review is shown in Figure 1. Section 2 introduces some preliminary knowledge of MOT, including some important terminologies in this problem (Section 2.1) and the denotations used in this article (Section 2.2). Section 3 describes the MOT problem, including formula (Section 3.1) and categorization (Section 3.2). Section 4 contributes to the most important issues involved in multi-object tracking, specifically, appearance model (Section 4.1), motion model (Section 4.2), interaction model (Section 4.3), exclusion model (Section 4.4), occlusion handling model (Section 4.5) and estimation methods including both probabilistic inference (Section 4.6) and deterministic optimization (Section 4.7), respectively. Furthermore, issues concerning evaluations of a MOT system, including metrics (Section 5.1), public data sets (Section 5.2), public codes (Section 5.3) and benchmark results (Section 5.4) are discussed in Section 5. To this end, conclusions are drawn and some interesting directions are provided in Section 6.

## 2 Preliminaries

Before beginning the review, we provide some preliminary knowledge for going deeper into the MOT problem. We first introduce some terminologies that are frequently used in the MOT problem. To make the review easy to follow, we then list the denotations that are globally applied in the manuscript.

### 2.1 Terminologies

The terminologies listed in the following play an important role in the MOT research.

**Object.** In computer vision, an object is considered as a continuous closed area in an image which is distinct from its surroundings. The interest of multiple object

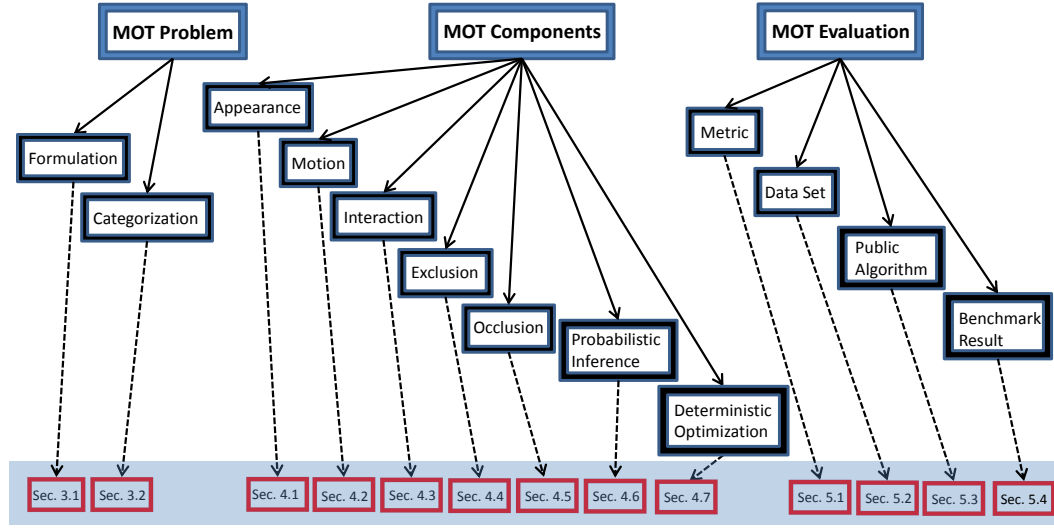


Fig. 1: Organization of this review

tracking often comes from one same type of objects, so an object should be identified additionally with an identity.

**Detection.** Detection is a computer vision task which localizes objects in images. It is usually conducted by training an object detector from a training dataset. In most situations, detection does not involve temporal information.

**Tracking.** Tracking is to localize an identical object in continuous frames. Thus tracking always involves in a video or image sequence with temporal information. In MOT, tracking means simultaneously localizing multiple objects and maintaining their identities.

**Detection response.** Detection responses are also known as detection observations or detection hypotheses, which are the outputs of an object detector trained for a specific kind of objects, e.g., human, vehicle, face or animal. They are configurations of objects including positions, sizes, etc, in an image sequence.

**Trajectory.** Trajectory is the output of a MOT system. One trajectory corresponds to one target, thus a trajectory is unique. At the same time, one trajectory is composed of multiple object responses of an identical target in an image sequence, each representing the location, size and some other information in one frame.

**Tracklet.** Tracklet is an intermediate level of output between detection responses and trajectories. It is composed of several detection responses which are believed to be from an identical target. As a fact, a detection response can be viewed as a tracklet composed of only one detection response. Tracklet is usually obtained by linking *confident* detection responses, thus it is shorter than trajectory regarding the time span. In some approaches, the final trajectories are obtained

by progressively linking detection responses into longer and longer tracklets and eventually forming trajectories. Figure 2 shows these three concepts.

**Data association.** Data association is a typical solution to multiple object tracking if we cast it as a paradigm of matching detection responses across frames based on object detection. The technique of data association figures out inter-frame correspondences between detection hypothesis.

## 2.2 Denotations

Throughout this manuscript, we represent scalar and vector by lowercase letter (e.g.  $x$ ) and bold lower case letter (e.g.  $\mathbf{x}$ ) individually. We use bold capital letter (e.g.  $\mathbf{X}$ ) to denote matrix or a set of vectors. Capital letters (e.g.  $X$ ) are adopted to denote specific functions or variables. Table 2 lists symbols utilized throughout this review. Except the symbols in the table, there might be some symbols for a specific reference. As these symbols are not commonly employed, we did not put them in the table. For these symbols, we will introduce their meanings in the context.

## 3 MOT Problem

The objective of MOT is to produce trajectories of objects as they move around the image plane. To elaborate this problem, in this section we firstly endeavor to give a general mathematical formulation of MOT, then we discuss its categorization from different aspects.

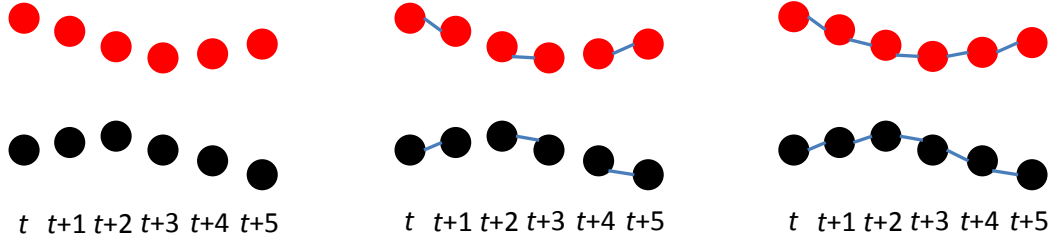


Fig. 2: Detection responses (left), tracklets (center), and trajectories (right) are shown in continuous 6 frames. Different colors encode different targets. Best viewed in color

Table 2: Denotations employed in this review

Symbol	Description	Symbol	Description	Symbol	Description	Symbol	Description
$P$	probability	$\mathbf{I}$	image	$\mathbf{p}$	position	$x, y$	position
$S$	similarity	$\mathbf{S}$	set of states	$\mathbf{v}$	velocity	$u, v$	speed
$C$	cost	$\mathbf{O}$	set of observations	$\mathbf{f}$	feature	$w, \alpha, \lambda$	weight
$N$	frame number	$\mathbf{T}$	trajectory/tracklet	$\mathbf{c}$	color	$t$	time index
$M$	object number	$\mathbf{M}$	feature matrix	$\mathbf{o}$	observation	$i, j, k$	general index
$G$	graph	$\mathbf{\Sigma}$	covariance matrix	$\mathbf{s}$	state	$\sigma$	variance
$V$	vertex set	$\mathbf{L}$	Laplacian matrix	$\mathbf{a}$	acceleration	$\epsilon$	noise
$E$	edge set	$\mathbf{Y}$	label set	$\mathbf{y}$	label	$s$	size
$D$	distance						
$L$	likelihood						
$F$	function						
$Z$	normalization factor						
$\mathcal{N}$	normal distribution						
$\mathcal{S}$	set						

### 3.1 Problem Formulation

Multiple object tracking can generally be formulated as a multi-variable estimation problem. Given an image sequence  $\{\mathbf{I}_1, \mathbf{I}_2, \dots, \mathbf{I}_t, \dots\}$  as input, we employ  $\mathbf{s}_t^i$  to denote the state of the  $i$ -th object in the  $t$ -th frame. We use  $\mathbf{S}_t = (\mathbf{s}_t^1, \mathbf{s}_t^2, \dots, \mathbf{s}_t^{M_t})$  to denote states of all the  $M_t$  objects in the  $t$ -th frame,  $\mathbf{s}_{1:t}^i = \{\mathbf{s}_1^i, \mathbf{s}_2^i, \dots, \mathbf{s}_t^i\}$  to denote the sequential states of the  $i$ -th object from the first frame to the  $t$ -th frame, and  $\mathbf{S}_{1:t} = \{\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_t\}$  to denote all the sequential states of all the objects from the first frame to the  $t$ -th frame. Note that the object number may vary from frame to frame.

To estimate the states of objects, we need to collect some observations from the image sequence. Correspondingly, we utilize  $\mathbf{o}_t^i$  to denote the collected observations for the  $i$ -th object in the  $t$ -th frame,  $\mathbf{O}_t = (\mathbf{o}_t^1, \mathbf{o}_t^2, \dots, \mathbf{o}_t^{M_t})$  to denote the collected observations for all the  $M_t$  objects in the  $t$ -th frame,  $\mathbf{o}_{1:t}^i = \{\mathbf{o}_1^i, \mathbf{o}_2^i, \dots, \mathbf{o}_t^i\}$  to denote the sequential observations collected from the first frame to the  $t$ -th frame, and  $\mathbf{O}_{1:t} = \{\mathbf{O}_1, \mathbf{O}_2, \dots, \mathbf{O}_t\}$  to denote all the collected sequential observations of all the objects from the first frame to the  $t$ -th frame.

The objective of multiple object tracking is to find the “optimal” sequential states of all the objects, which can be generally modeled by performing MAP (maxi-

mal a posteriori) estimation from the conditional distribution of the sequential states of all the objects given all the observations:

$$\hat{\mathbf{S}}_{1:t} = \arg \max_{\mathbf{S}_{1:t}} P(\mathbf{S}_{1:t} | \mathbf{O}_{1:t}). \quad (1)$$

The estimation can be performed using *probabilistic inference* algorithms based on a two-step iterative procedure (Liu et al. 2012; Breitenstein et al. 2009; Yang et al. 2009b; Mitzel and Leibe 2011; Rodriguez et al. 2011; Kratz and Nishino 2010; Reid 1979):

$$\begin{aligned} \text{Predict: } P(\mathbf{S}_t | \mathbf{O}_{1:t-1}) &= \int P(\mathbf{S}_t | \mathbf{S}_{t-1}) P(\mathbf{S}_{t-1} | \mathbf{O}_{1:t-1}) d\mathbf{S}_{t-1} \\ \text{Update: } P(\mathbf{S}_t | \mathbf{O}_{1:t}) &\propto P(\mathbf{O}_t | \mathbf{S}_t) P(\mathbf{S}_t | \mathbf{O}_{1:t-1}) \end{aligned}$$

In the formula above,  $P(\mathbf{S}_t | \mathbf{S}_{t-1})$  and  $P(\mathbf{O}_t | \mathbf{S}_t)$  are the *Dynamic Model* and the *Observation Model*, respectively. These two models play a very important role in a tracking algorithm. Since the distributions of these two models are usually unknown, sampling methods like Particle Filter (Jin and Mokhtarian 2007; Yang et al. 2005; Hess and Fern 2009; Han et al. 2007; Hu et al. 2012; Liu et al. 2012; Breitenstein et al. 2009; Yang et al. 2009b), MCMC (Khan et al. 2004, 2005, 2006), RJMCMC (Choi et al. 2013), etc are employed to perform the estimation.

The estimation problem can also be coped with *deterministic optimization* approaches, e.g., directly max-



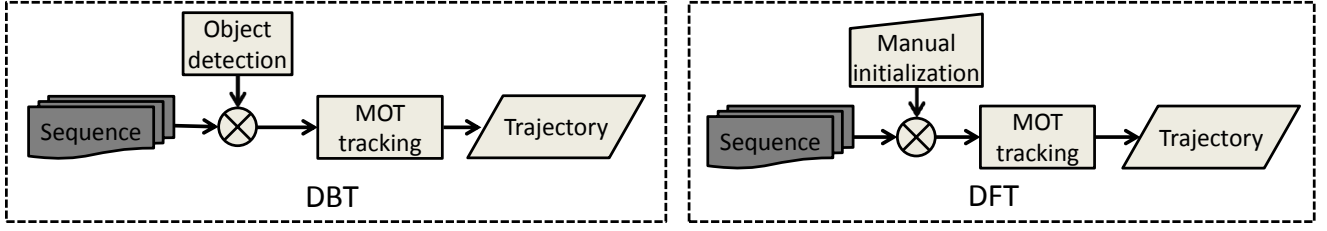


Fig. 3: Procedure flow of DBT (left) and DFT (right)

imizing the likelihood function  $P(\mathbf{O}_{1:t}|\mathbf{S}_{1:t})$  as a delegate of  $P(\mathbf{S}_{1:t}|\mathbf{O}_{1:t})$ :

$$\hat{\mathbf{S}}_{1:t} = \arg \max_{\mathbf{S}_{1:t}} P(\mathbf{S}_{1:t}|\mathbf{O}_{1:t}) = \arg \max_{\mathbf{S}_{1:t}} P(\mathbf{O}_{1:t}|\mathbf{S}_{1:t}), \quad (2)$$

or conversely minimizing an energy function

$$\begin{aligned} \hat{\mathbf{S}}_{1:t} &= \arg \max_{\mathbf{S}_{1:t}} P(\mathbf{S}_{1:t}|\mathbf{O}_{1:t}) \\ &= \arg \max_{\mathbf{S}_{1:t}} \frac{1}{Z} \exp(-C(\mathbf{S}_{1:t}|\mathbf{O}_{1:t})) \\ &= \arg \min_{\mathbf{S}_{1:t}} C(\mathbf{S}_{1:t}|\mathbf{O}_{1:t}), \end{aligned} \quad (3)$$

where  $Z$  is a normalization factor to make  $P(\mathbf{S}_{1:t}|\mathbf{O}_{1:t})$  a probability distribution.

The specific optimization approaches can range from bipartite graph Matching (Shu et al. 2012; Breitenstein et al. 2009; Wu and Nevatia 2007; Qin and Shelton 2012; Reilly et al. 2010; Perera et al. 2006; Xing et al. 2009; Huang et al. 2008), dynamic programming (Wolf et al. 1989; Jiang et al. 2007; Berclaz et al. 2009; Andriyenko and Schindler 2010), min-cost max-flow network flow (Zhang et al. 2008; Choi and Savarese 2012; Wu et al. 2012; Butt and Collins 2013; Pirsavash et al. 2011) to max weight independent set (Shafique et al. 2008; Brendel et al. 2011).

### 3.2 MOT Categorization

In general, it is difficult to get a universal classification of the MOT problem. Nevertheless, we can categorize MOT problem by different criteria to obtain a more comprehensive understanding of the problem. Existing work of visual tracking (Yilmaz et al. 2006; Cannons 1991) has provided some views for classification. For example, object shape representation is adopted by Yilmaz et al. (2006) to group existing work of visual tracking into subsets. To be specific, different types of object shape representations, such as point, primitive geometric shapes, object contours and region shapes are described individually. For the concerned MOT problem,

as object is represented by region shapes (bounding box or ellipse (Kuo and Nevatia 2011)) in most works, it is not necessary to discuss by object representations. Alternatively, we provide discussion according to the following aspects.

#### 3.2.1 Initialization Method

The first criterion is that how objects are initialized. According to this criterion, most of existing MOT work could be grouped into two sets (Yang and Nevatia 2012c): Detection Based Tracking (DBT) and Detection Free Tracking (DFT). DBT relies on object detection while DFT does not.

**DBT.** In DBT, objects are at first localized in each frame and then object hypotheses are linked into trajectories. Figure 3 (left) shows the flow of DBT. Given a sequence, type-specific object detection or motion detection (based on background modeling) (Bose et al. 2007; Song et al. 2010) is applied in each frame to obtain object hypotheses, then (sequential or batch) tracking is conducted to link detection hypotheses into trajectories. There are three issues worthy noting. First, in most cases object detection procedure is not the focus of DBT methods. The majority of DBT approaches builds upon a pre-trained object detector which produces object hypotheses as observations. Second, as mentioned above, since object detector is trained in advance, the majority of DBT focuses on specific kinds of targets, such as pedestrians, vehicles or faces. The underlying reason is that detection of these types of objects has gained great progress in recent years (Dalal and Triggs 2005; Felzenszwalb et al. 2010; Sun et al. 2006). Third, the performance of DBT depends on the performance of the employed model of object detection to a certain extent.

**DFT.** As shown in Figure 3 (right), DFT (Hu et al. 2012; Zhang and van der Maaten 2013, 2014; Yang et al. 2007) requires manual initialization of a fixed number of objects in the first frame (in the form of bounding boxes or other shape configurations), then localizes these fixed number of objects in the subsequent frames. It does not rely on object detector to provide object hypotheses.

Noting that, when the number of objects is one, DFT degrades as the classical visual tracking problem.

DBT is more popular for the fact that new objects are discovered and disappearing objects are terminated automatically. DFT requires manual initialization of each object to be tracked, thus it cannot deal with the case that objects appear. However, it is model-free, i.e., free of pre-trained object detectors. So it can deal with sequences of any type of objects. However, the setting of fixed number of objects limits its applications in practical systems. Table 3 lists the major differences between DBT and DFT.

### 3.2.2 Processing Mode

According to the way of processing data, MOT could be categorized into online tracking and offline tracking. The difference is whether the future frame observations are utilized when handling the current frame. Online tracking utilizes observations up to the current time instant to conduct the estimation, while offline tracking employs observations both in the past and in the future.

**Online tracking.** In online tracking, the image sequence is handled in a step-wise way, thus online tracking is also named as sequential tracking. As shown in Figure 4 (left), we present a toy example that there are four objects (different circles) in a video sequence with IDs a, b, c and d. The arrow attached to each object indicates its movement direction. The green arrows represent observations in the past. The results are represented by the object's location and its ID. Based on the up-to-time observations, trajectories are outputted on the fly.

**Offline tracking.** Offline tracking (Song et al. 2010; Qin and Shelton 2012; Yang and Nevatia 2012a,b; Brendel et al. 2011; Yang et al. 2011; Kuo et al. 2010; Henriques et al. 2011; Sugimura et al. 2009; Choi and Savarese 2010) utilizes a batch way to process the data therefore it is also called batch tracking. Figure 4 (right) illustrates how the batch tracking processes observations. Observations from all the frames are required to be obtained in advance and are investigated together to estimate the final output. Note that, due to computation ability, sometimes it is not possible to handle all the frames at one time. Alternatively, one solution is to divide the whole video into a set of segments or clips, handle these clips respectively, and infuse the results hierarchically.

In general, online tracking is appropriate in the case that video stream is obtained sequentially. Offline tracking typically deals with the data globally when all the frames are obtained. Theoretically offline tracking could obtain global optimal solution while it is not as practi-

cal as online tracking. To be clearer, we compare them in Table 4.

### 3.2.3 Mathematical Methodology

MOT could be classified into probabilistic tracking and deterministic tracking according to the adopted mathematical methodology. There are two differences between them. First, the approaches to estimating states of objects are different. In probabilistic tracking, the estimation is based on probabilistic inference, while in deterministic tracking the estimation is based on deterministic optimization. Second, the outputs are different. Output of probabilistic tracking may be different in different running trials while constant in deterministic tracking.

### 3.2.4 Discussion

The insights behind “online vs offline” and “DBT vs DFT” are related. The difference between DBT and DFT is whether a detection model is adopted (DBT) or not (DFT). The key to differentiate online and offline tracking is the way they process observations. Readers may question whether DFT is identical to online tracking because DFT always processes observations sequentially. That is true because DFT is free of (type-specific) object detection. It cannot attain future observations, thus it can only follow the sequential way. Another vagueness may rise between DBT and offline tracking, as in DBT tracklets or detection responses are usually associated in a batch way. Note that there are also sequential DBT which conducts association between previously obtained trajectories and new detection responses (Luo et al. 2014; Luo and Kim 2013; Xing et al. 2009).

At the same time, “online vs offline” and “probabilistic vs deterministic” are also related. In practice, online tracking usually adopts probabilistic inference for estimation while there indeed exists deterministic optimization based online tracking, such as online tracking by linking up-to-time trajectories and detections in the next frame based on Hungarian algorithm. On the other hand, offline tracking always employ deterministic optimization in the derivation of states of objects.

### 3.2.5 MOT Special Cases

Some publications cannot be simply grouped into traditional multi-target tracking as they exhibit different attributes. We represent three special cases as follows.

**MOT in sport scenarios.** Tracking of multiple targets in the sport court has wide applications such

Table 3: Comparison between DBT and DFT. Part of this table is from the work by [Yang and Nevatia \(2012c\)](#)

Item	DBT	DFT
Initialization	automatic, imperfect	manual, perfect
# of objects	varying	fixed
Applications	specific type of objects (in most cases)	any type of objects
Advantages	ability to handle varying number of objects	free of object detector
Drawbacks	performance depends on object detection	requires manual initialization

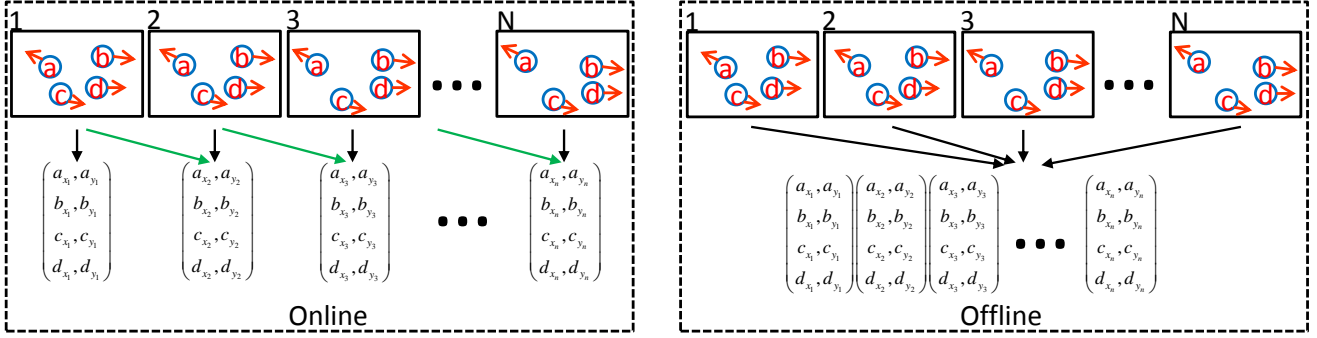


Fig. 4: Illustration of online and offline tracking. Best viewed in color

Table 4: Comparison between online and offline tracking

Items	Online tracking	Offline tracking
Required input	up-to-time observations	all observations
Methodology	gradually extend existing trajectories with current observations	link observations into trajectories
Advantages	suitable for online tasks	can obtain global optimal solution theoretically
Drawbacks	suffer from shortage of observation	Delay in outputting final results

as technical statistics in a match, automatic analysis of strategies. There are several differences between traditional MOT and this special problem. The first one is that the sport court is special. There are frequent shot switches and actions to zoom in and zoom out, which make it challenging. However, this scene also exhibits useful features. For example, regarding some kinds of sport such as basketball and football, there is a clear boundary between the court and the background. This kind of boundary could dismiss most of the confusion from the background once it is recognized. For instance, [Xing et al. \(2011\)](#) propose to segment the playfield from the non-playfield as a pre-processing step for MOT. The second difference is that targets in this problem probably dress uniform. This can lead to two effects. On one hand, as targets dress very similarly to each other, the appearance of targets is not as diverse as traditional MOT. Thus it may be more difficult to differentiate objects. On the other hand, as claimed before, targets of the same team wear clothes of the same color or pattern, while ones from different teams do not follow this. It could be helpful. For example, [Lu et al. \(2013\)](#) train a

logistic regression classifier which maps image patches to team labels using RGB color histograms. This step improves the precision while retaining the same level of recall rate.

**MOT in aerial scenes.** Multiple object tracking in aerial scenes is different from the normal MOT as it exhibits these features: 1) the size of targets is considerably small, thus the commonly used appearance information is not reliable, 2) the frame rate is quite low, 3) object density is high, e.g., the number of objects could be up to hundreds of. Therefore, the solution is also different from that to the normal multiple object tracking problem. [Reilly et al. \(2010\)](#) propose several techniques to tackle the difficulties mentioned above. To detect objects, motion detection is applied after background modeling. Obtaining objects in each frame, Hungarian algorithm is adapted to solve the assignment of detections to existing trajectories. When computing the affinity, the spatial proximity, velocity orientation and context are considered. Note that context is relatively important in this special case. [Shi et al. \(2013\)](#) deal with the multi-frame multi-target association problem



in the aerial scene as a rank-1 tensor approximation problem. High-order tensor is constructed based on all the candidate trajectories. Then the multi-dimensional assignment problem for MOT is transformed into the rank-1 approximation. In the end the approximation is solved by a  $l_1$  tensor power iteration.

**Generic MOT.** As we have discussed, the majority of MOT work focuses on some special kinds of objects, such as vehicles and pedestrians. The reason is that DBT solutions without manual initialization are more practical in real-life applications and object detection has achieved great progress, especially for pedestrians. However, these methods, although free of manual initialization, rely on specific kinds of object detectors to obtain detection observations. Typically, these object detectors have already been trained in advance. There arises two problems that, 1) these pre-trained detectors can only be applied to the image sequences of the same type of objects, 2) as these detectors are pre-trained, they are data-specific. Thus the performance is not optimal when they are applied to other image sequences. On the other hand, DFT approaches, although require manual initialization, they can be applied to type-free sequences.

Observing this, recently some researchers have proceeded a step to investigate generalization of the MOT problem to any kind of objects with minimum manual labeling and free of offline trained detectors. Zhao et al. (2012) propose to track multiple similar objects by requiring one instance in the first frame to be labeled. They firstly track this object, collect training samples, and train an object detector for this kind of objects in the first a few frames. Then they start from the first frame again, detect the top  $M$  (specified by the user) similar objects and track them in the subsequent frames. Compared with DFT, Zhao et al. (2012) saves much labeling labor. However, as the number of objects to track is still fixed, it is not comparable with DBT. To make the generalization of MOT more practical, a generic MOT problem is proposed by Luo and Kim (2013) that multiple similar objects are tracked with labeling of one example in the first frame. Object detection is progressively refined as the accumulation of training samples. A specialized tracker is initialized once an object is discovered and multiple trackers are learned together based on multiple task learning (Evgeniou and Pontil 2004; Caruana 1997). Dealing with the same problem, Luo et al. (2014) propose a bi-label propagation framework. The detection and tracking of multiple objects is cast as class and object label propagation in the spatial-temporal video tube. Multiple similar objects are tracked by Dicle et al. (2013), however, the detection responses are given as input to the al-

gorithm rather than by detection. Brostow and Cipolla (2006) deal with generic MOT without any supervision. Assuming that a pair of points that appears to move together is likely to be part of the same individual, feature points are tracked and motion clustering is conducted to discover and maintain identical entities.

## 4 MOT Components

As shown in Figure 5, MOT involves two primary components. One is observation model and the other one is dynamic model. Observation model measures similarity between object states and observations. To be more specific, an observation model includes modeling of appearance, motion, interaction, exclusion and occlusion. Dynamic model investigates states transition across frames. It can be classified into probabilistic inference and deterministic optimization. All these components will be present in this section.

### 4.1 Appearance Model

Appearance is an important cue for affinity computation in MOT. However, it is worthy noting that, different from single object tracking approach which primarily focuses on constructing a sophisticated appearance model to discriminate object from background, multiple object tracking does not mainly focus on appearance model, i.e., appearance cue is important but not the only cue to depend on. This is partly because that the multiple objects in MOT can hardly be discriminated by relying on only appearance information.

Technically, appearance model includes two components, i.e. *visual representation* and *statistical measuring*. Visual representation is closely related to features, but it is more than features. It is how to precisely describe the visual characteristics of object based on features, and in general it can be grouped into two sets, visual representation based on single cue and that based on multiple cues. Statistical measuring is the computation of similarity or dissimilarity between different observations when visual representation is ready. Eq. 4 gives an illustration of appearance modeling, where  $\mathbf{o}_i$  and  $\mathbf{o}_j$  are visual representation of different observations based on single cue or multiple cues, and  $F(\bullet, \bullet)$  is a function to measure the similarity  $S_{ij}$  between  $\mathbf{o}_i$  and  $\mathbf{o}_j$ . In the following, we firstly discuss the features/cues employed in MOT, and then describe appearance models based on single cue and multiple cues respectively.

$$S_{i,j} = F(\mathbf{o}_i, \mathbf{o}_j) \quad (4)$$

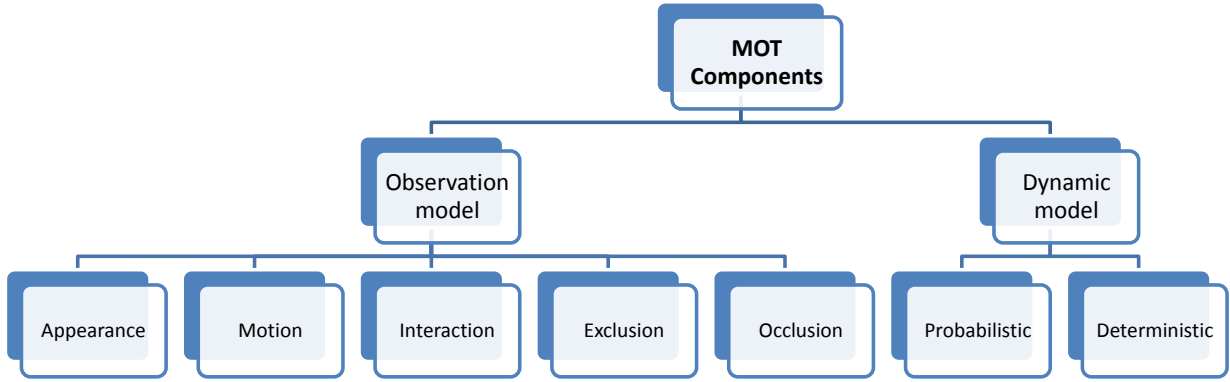


Fig. 5: Components of MOT

#### 4.1.1 Feature

Feature is indispensable for MOT. As shown in Figure 6, different kinds of features have been employed in MOT. We categorize features into the following subsets.

**Point features.** Point features are successful in single object tracking (Shi and Tomasi 1994). For MOT, point features can also be helpful. For instance, KLT tracker is employed to track feature points and generate a set of trajectories or short tracklets (Sugimura et al. 2009; Zhao et al. 2012). Choi and Savarese (2010) utilize the KLT features (Tomasi and Kanade 1991) as additional features to estimate the camera’s motion, greatly improving tracking performance. Similarly, KLT tracking is utilized by Benfold and Reid (2011) to estimate motion. Local feature points (Lowe 2004) are adopted along with the bag-of-word model by Yang et al. (2009b) to capture the texture characteristics of a region. Point features are also employed by Brostow and Cipolla (2006) for motion clustering.

**Color/intensity features.** This is the most popularly utilized feature for MOT. Usually the color or intensity features along with a measurement are employed to calculate the affinity between two counterparts (detection hypotheses, tracklets or short trajectories). The simple raw pixel template is employed by Yamaguchi et al. (2011) to compute the appearance affinity. Color histogram is used by Sugimura et al. (2009), Song et al. (2010), Mitzel et al. (2010), Izadinia et al. (2012), Okuma et al. (2004) and Mitzel and Leibe (2011).

**Optical flow.** The optical flow feature can be employed to conduct short-term visual tracking. Thus many solutions to MOT utilize optical flow to link detection responses from continuous frames into short tracklets for further data association processing (Rodriguez

et al. 2009) or directly use it for data association (Izadinia et al. 2012). Besides this, optical flow is also employed to complement HOG for observation model (Andriyenko and Schindler 2011). Additionally, optical flow is popular in extremely crowded scenarios for discovering crowd motion patterns (Ali and Shah 2008; Rodriguez et al. 2011).

**Gradient/pixel-comparison features.** There are some features based on gradient or pixel comparison. Mitzel et al. (2010) utilize a variation of the level-set formula, which integrates three terms penalizing the deviation between foreground and background, a embedding function from a signed distance function and the length of the contour to track objects in continuous frames. Besides the success in human detection, HOG (Dalal and Triggs 2005) plays a vital role in the multiple pedestrian tracking problem as well. For instance, HOG is employed (Izadinia et al. 2012; Kuo et al. 2010; Breitenstein et al. 2009; Choi and Savarese 2012; Yu et al. 2008) to detect objects and/or compute similarity between pedestrian detections for data association.

**Region covariance matrix features.** Region covariance matrix (Porikli et al. 2006; Tuzel et al. 2006) features are robust to issues such as illumination changes, scale variations, etc. Therefore, it is also employed for the MOT problem. The region covariance matrix based dissimilarity is used to compare appearance for data association (Henriques et al. 2011). Covariance matrices along with other features constitute the feature pool for appearance learning by Kuo et al. (2010). Hu et al. (2012) utilize the covariance matrix to represent object for both single and multiple object tracking.

**Depth.** Depth information is employed for various computer vision tasks. With regard to MOT, Mitzel et al. (2010) utilize depth information to correct bounding box of detection response and re-initialize the bounding box for level-set tracking in their work. Depth in-

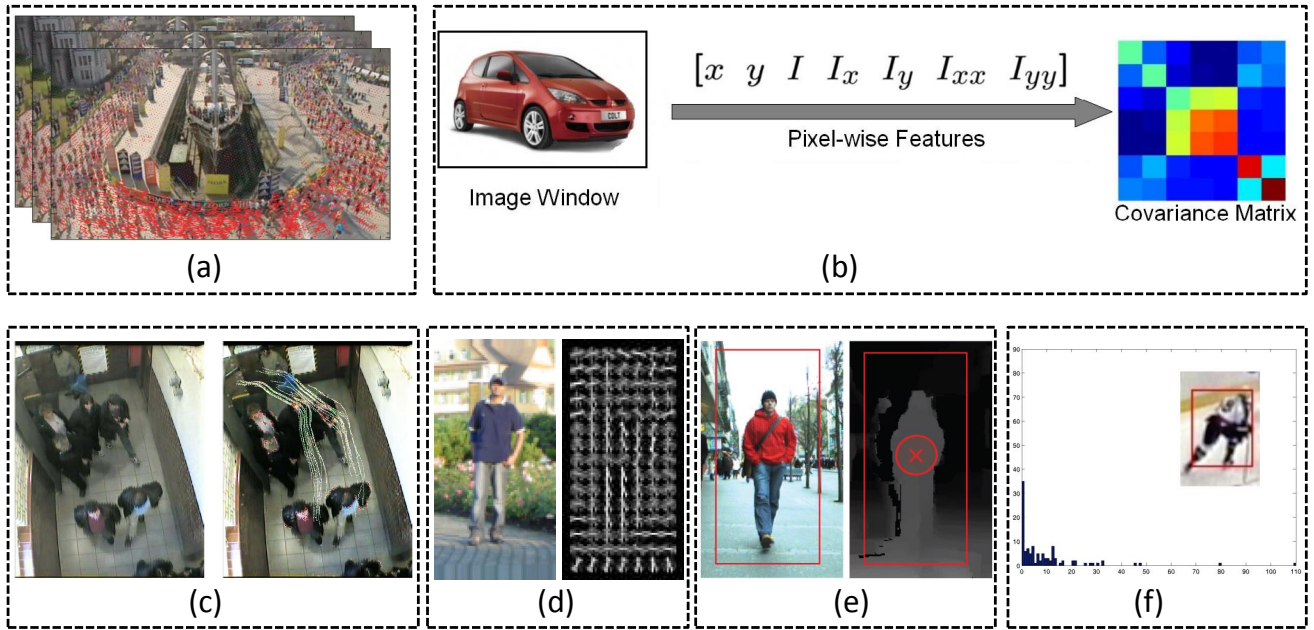


Fig. 6: Some exemplar features. (a) Image showing optical flow (Ali and Shah 2008), (b) image showing covariance matrix, (c) image showing point feature (Brostow and Cipolla 2006), (d) image showing gradient based features (Dalal and Triggs 2005), (e) image showing depth (Mitzel et al. 2010) and (f) image showing color feature (Okuma et al. 2004). Best viewed in color

formation is integrated into the framework (Ess et al. 2009, 2007) to augment detection hypotheses with a depth flag 0 or 1, which further refines the detection responses. Similarly, Ess et al. (2008) employ depth information to obtain more accurate object detections in a mobile vision system and then use the detection result for multiple object tracking. The stereo depth is taken into account by Giebel et al. (2004) to estimate weight of a particle in the proposed Bayesian framework for multiple 3D object tracking. Gavrila and Munder (2007) integrate depth to generate detections and consequently verify them for multiple object tracking from a moving car.

**Others.** Some other features, which are not so popular, are utilized to conduct multiple object tracking as well. For instance, gait features in the frequency domain, which are unique for every person, are employed by Sugimura et al. (2009) to maximize the discrimination between the tracked individuals. Given a trajectory, a line fitting via linear regression is conducted to extract the periodic component of the trajectory. Then the Fast Fourier Transform (FFT) is applied to the residual periodic signal to obtain the amplitude spectra and phase of the trajectory, which are utilized to compute the dissimilarity between a trajectory and other trajectories. The Probabilistic Occupancy Map (POM) (Fleuret et al. 2008; Berclaz et al. 2011) is employed

to estimate how probable an object would occur in a specific grid under the multi-camera settings. It creates detection hypotheses on top of background modeling as the input for MOT.

Generally speaking, most of the features are efficient. At the same time, they also have shortcomings. For instance, color histogram has well studied similarity measures, but it ignores the spatial layout of the object region. Point features are efficient, but sensitive to issues like occlusion and out-of-plane rotation. Gradient based features like HOG can describe the shape of object and robust to issues such as illumination changes, but it cannot handle occlusion and deformation well. Region covariance matrix features are more robust as they take more information in account, but this benefit is obtained at the cost of more computation. Depth features make the computation of affinity more accurate, but they require multiple views of the same scenery and/or additional algorithm (Felzenszwalb and Huttenlocher 2006) to obtain depth.

#### 4.1.2 Single cue based appearance model

In most cases, appearance model in MOT is kind of simplicity and efficiency, thus a single cue is a popular option. In the following we present appearance models employing a single cue in five aspects.

**Raw pixel template representation.** Pixel is the most granular element of images and videos, thus pixel is the foundation of computer vision problems. Beside its importance, it is also popular for its simplicity. The raw pixel template representation is the raw pixel intensity or color of a region. It can encode the spatial information since the comparison is element wise when matching two templates. Yamaguchi et al. (2011) employ the Normalized Cross Correlation (NCC) to evaluate the predicted position of object. This appearance model is very simple, but helpful. The appearance affinity is calculated as the NCC between the target template and a candidate bounding box (Ali and Shah 2008). Note that the target template is progressively updated at each time instant. Wu et al. (2012) build a network-flow approach to handle multiple target tracking. When they compute the transitional cost on the arcs of the network as flows, the normalized cross correlation between the upper one-fourth bounding boxes of the corresponding two detection observations is used. A simple patch-based tracker is implemented to calculate the data likelihood with regard to appearance (Pellegrini et al. 2009). The similarity is computed between the initial bounding box and a candidate bounding box as the squared exponential  $P_{data}(\mathbf{p}) \propto \exp\left(-\left(NCC(\mathbf{p}, \mathbf{p}^0) - 1\right)^2\right)$ , where  $\mathbf{p}$  and  $\mathbf{p}^0$  represent the candidate bounding box and the initial bounding box. Despite of efficiency, this kind of representation easily suffers from the change of illumination, occlusion or some other issues.

**Color histogram representation.** Color histogram is the most popular representation for appearance modeling in MOT approaches as a result of its effectiveness to capture the statistical information of target region. For example, Kratz and Nishino (2010) employ the color histogram model (Pérez et al. 2002) to calculate the likelihood in terms of appearance, and they use an exponential function to transform the histogram distance into probability. Similarly, to capture the dissimilarity, Sugimura et al. (2009) use the Bhattacharyya distance between hue-saturation color histograms when constructing a graph. A Mean Shift tracker employing color histogram is utilized to sequentially seize the object (Choi and Savarese 2010). Appearance model is defined as the RGB color histogram of the trajectory by Leibe et al. (2008). It is initialized as the first detection response’s color histogram and evolves as a weighted mean of all the detection responses which belong to this trajectory. The likelihood considering appearance is proportional to the Bhattacharyya coefficient of two histograms. Affinity regarding appearance is dealt with by calculating the Bhattacharyya distance between the average HSV color histograms of the concerned track-

lets (Qin and Shelton 2012). Xing et al. (2009) represent target (pedestrian body) as a constitution of three overlapped part areas, i.e. the Full Body (FB), the Head-Shoulder (HS) and the Head-Torso (HT). To link tracklets for data association, they consider the color histogram of each part in the detection response. The affinity regarding appearance to link two tracklets  $\mathbf{T}_i$  and  $\mathbf{T}_j$  is calculated as in Eq. 5,

$$S(\mathbf{T}_i, \mathbf{T}_j) = \exp(-D(\mathbf{c}_i, \mathbf{c}_j)), \quad (5)$$

where  $D(\mathbf{c}_i, \mathbf{c}_j) = \text{mean}\{D(\mathbf{c}_i^k, \mathbf{c}_j^k) | k = FB, HT, HS\}$ ,  $\mathbf{c}$  means color histogram and  $D(\bullet, \bullet)$  is the Bhattacharyya distance measure. The link affinity of two detection responses regarding appearance is calculated based on RGB histograms (Zhang et al. 2008). Given two detection observations  $\mathbf{o}_i$  and  $\mathbf{o}_j$ , the corresponding color histograms  $\mathbf{c}_i$  and  $\mathbf{c}_j$  are extracted. Then the Bhattacharyya distance  $D_{ij}$  between  $\mathbf{c}_i$  and  $\mathbf{c}_j$  is obtained, and the probability of linking  $\mathbf{o}_i$  and  $\mathbf{o}_j$  is

$$P(\mathbf{o}_i | \mathbf{o}_j) = \frac{\mathcal{N}(D_{ij}; D_s, \sigma_s^2)}{\mathcal{N}(D_{ij}; D_s, \sigma_s^2) + \mathcal{N}(D_{ij}; D_d, \sigma_d^2)}, \quad (6)$$

where  $\mathcal{N}(\bullet; D_s, \sigma_s^2)$  and  $\mathcal{N}(\bullet; D_d, \sigma_d^2)$  are two Gaussian distributions of appearance dissimilarity between the same object and different objects respectively. The parameters of these two distributions are learned from the training data. Noting that, besides its advances, the color histogram representation has the drawback of losing spatial information.

**Covariance matrix representation.** Covariance matrix is robust to illumination change, rotation, etc. The covariance matrix descriptor is employed to represent the appearance of an object by Henriques et al. (2011). The likelihood concerning appearance to link two detection responses is modeled as

$$P_{link}(\mathbf{T}_i, \mathbf{T}_j) \propto \mathcal{N}(D_{ij}; \bar{D}, \Sigma), \quad (7)$$

where  $\mathcal{N}$  is the Gaussian distribution,  $D_{ij}$  is the appearance dissimilarity between  $\mathbf{T}_i$  and  $\mathbf{T}_j$ ,  $\bar{D}$  and  $\Sigma$  are parameters estimated from training data. A block-division appearance model which divides the object region into blocks is proposed by Hu et al. (2012). Within each block, the covariance matrix is extracted as the region descriptor to characterize the block. At the same time, likelihood of each block is computed with regard to the corresponding block of the target, and likelihood of the whole region is the product of the likelihood of all blocks.



**Pixel comparison representation.** Nothing could be simpler than giving a binary result of comparison between two pixels, and this is the advance of this type of representation over other kinds of representation. Zhao et al. (2012) adopt Random Ferns in tracking (Kalal et al. 2012). It encodes the results of comparisons between pairs of pixels and vote the comparison based on training data. The probability of a patch being positive is calculated based on how many positive samples and negative samples have been recorded by that leaf.

**Bag of words representation.** Fast dense SIFT-like features (Lowe 2004) are computed by Yang et al. (2009b) and encoded based on the bag-of-word model. To incorporate spatial information, the spatial pyramid matching (SPM) method (Lazebnik et al. 2006) is adapted. This is used as an observation model for appearance modeling.

#### 4.1.3 Multi-cue based appearance model

Different kinds of cues could compensate each other, which would make appearance model robust. However, there arises an issue that how to fuse the information from multiple cues. Regarding this, we present multi-cue based appearance models according to five kinds of fusion strategies, *Boosting*, *Concatenating*, *Summation*, *Product* and *Cascading* (also see Table 5).

**Boosting.** The strategy of Boosting usually selects a portion of features from a feature pool sequentially via a Boosting based algorithm (e.g. Adaboost by Kuo et al. (2010) and RealBoost by Yang and Nevatia (2012c)). Features are selected according to their discrimination power. A discriminative appearance model is proposed by Kuo et al. (2010) to assign high similarity to tracklets which are of the same object, but low affinity to tracklets of different objects. Specifically, color histogram in RGB space, HOG and covariance matrix descriptor are employed as features. They choose 15 regions so that they have 45 cues in total in the feature pool. Collecting positive and negative training pairs according to the so-called spatial-temporal constraints, they employ Adaboost to choose the most representative features to discriminate pairs of tracklets belonging to the same object from those belonging to different objects. Similarly, Yang and Nevatia (2012c) adopt features (Kuo et al. 2010) and employ the standard RealBoost algorithm to learn the feature weights from training sample set, which is composed of correctly linked pairs (as positive samples) and incorrectly matched pairs (as negative samples). A HybridBoost algorithm is proposed by Li et al. (2009) to automatically select features with maximum discrimination. This algorithm employs a hybrid loss function composed of a classification term and

a ranking term. The ranking term ranks correct tracklet associations higher than their counterparts and the classification term dismisses wrong associations.

**Concatenating.** A SVM model classifier is trained to distinguish a specific target from targets in its temporal window. Color, HOG and optical flow are concatenated and further processed with PCA projection for dimension reduction to describe the detection response (Brendel et al. 2011). The similarity  $S$  between two detection responses is

$$S = \exp \left( - (\mathbf{f} - \mathbf{f}')^T \mathbf{M} (\mathbf{f} - \mathbf{f}') \right), \quad (8)$$

where  $\mathbf{f}$  and  $\mathbf{f}'$  are features corresponding to the two detection responses,  $\mathbf{M}$  is a distance metric matrix learned online.

**Summation.** Mitzel et al. (2010) simultaneously segment and track multiple objects. As pedestrians and backgrounds often contain the same color, color information alone yields rather unreliable segmentation for pedestrians. To address it, they integrate color information with depth information. To be specific, they firstly compute an expected depth of foreground in the following frames according to the current depth and a maximum velocity. Then each depth of foreground in the following frame could be assigned a probability based on a Gaussian distribution centered at the expected depth. The probability based on color representation is the accordance of color histogram calculated in the Lab space with the learned appearance model (Bibby and Reid 2008). Then these two probabilities (computed from color and depth cues) are weighted by a parameter  $\alpha$  as in Eq. 9. The similar weighting strategy is adopted by Liu et al. (2012) to balance two cues of raw pixel intensity and silhouette.

$$P_i = (1 - \alpha) P_{i,color} + \alpha P_{i,depth}, i \in \{fg, bg\}. \quad (9)$$

Besides similarity, distance could also be added together to conduct multiple object tracking. Distance in terms of different appearance cues, including RGB color histogram, correlogram and LBP is computed as a summation for matching (Takala and Pietikainen 2007).

**Product.** Yang et al. (2009b) integrate multiple cues including color, shapes and bags of local features (Lowe 2004; Lazebnik et al. 2006) to calculate the likelihood of linking a detection response with an existing trajectory. Assuming these three cues are  $\mathbf{f}^1$ ,  $\mathbf{f}^2$  and  $\mathbf{f}^3$ , the likelihood linking a detection response with a trajectory (state  $\mathbf{s}$ ) is modeled as  $P(\mathbf{f}^1, \mathbf{f}^2, \mathbf{f}^3 | \mathbf{s})$ . These



Table 5: An overview of typical appearance models employing multiple cues

Strategy	Employed Cue	Representative Reference
Boosting	Color, HOG, covariance matrix, shapes, etc.	Kuo et al. (2010) Li et al. (2009) Yang and Nevatia (2012c)
Concatenating	Color, HOG, optical flow, etc.	Brendel et al. (2011)
Summation	Color, depth, correlogram, LBP, etc.	Mitzel et al. (2010) Liu et al. (2012) Takala and Pietikainen (2007)
Product	Color, shapes, bags of local features, etc.	Yang et al. (2009b) Song et al. (2010) Giebel et al. (2004) Berclaz et al. (2006)
Cascading	Depth, shape and texture, etc.	Gavrila and Munder (2007) Izadinia et al. (2012)

cues are assumed to be independent from each other, thus the likelihood is

$$P(\mathbf{f}^1, \mathbf{f}^2, \mathbf{f}^3 | \mathbf{s}) = \prod_{i=1}^3 P(\mathbf{f}^i | \mathbf{s}). \quad (10)$$

A similar formula is adopted by Song et al. (2010). The likelihood considering color histogram is multiplied with the likelihood regarding foreground response as the final likelihood in the observation model. In this work, these two cues are assumed to be independent. Likelihoods in terms of three cues, shape, texture and depth, are multiplied to compute the weight of a particle in a Bayesian framework (Giebel et al. 2004). Dividing the scene under multiple cameras into multiple grids, appearance model is constructed based on color model and ground plane occupancy estimation (Berclaz et al. 2006). Similarity concerning these two cues are multiplied in the MAP formula.

**Cascading.** Gavrila and Munder (2007) utilize cues of depth, shape and texture in a cascade manner to narrow the search space for multiple object detection and tracking. Specifically, depth is used to generate hypotheses, and shape and texture cues are employed sequentially. Finally depth is utilized to verify detections. Real-time performance is achieved by doing so. Appearance information is considered in two stages (Izadinia et al. 2012). In the first stage, detection responses are linked into short tracklets. The affinity to link two detection responses is calculated according to color histogram, a HOG feature descriptor and a motion feature descriptor with optical flow. In the second stage, a pedestrian-specific appearance model considering multiple parts of object.

## 4.2 Motion Model

Object motion model describes how an object moves. It is important for multiple object tracking since it can predict the potential position of objects in the future frames, reducing search space. In general, objects are assumed to move smoothly in the image scene (*cf.* the abrupt motion is a special case). Popular motion models employed in multiple object tracking are divided into the following two classes.

### 4.2.1 Constant velocity motion models/linear motion models

As the name indicates, objects following these models are assumed to move with constant velocity. This is the most popular model (Shafique et al. 2008; Yu et al. 2007). The velocity of object in the next time is the same as the current velocity (added by process noise independently drawn from some types of distributions). For example, Breitenstein et al. (2009) employ a constant velocity motion model to propagate particles like this:

$$\begin{aligned} (x, y)_t &= (x, y)_{t-1} + (u, v)_{t-1} \bullet \Delta t + \epsilon_{x,y}, \\ (u, v)_t &= (u, v)_{t-1} + \epsilon_{u,v}, \end{aligned} \quad (11)$$

where  $(x, y)$  and  $(u, v)$  represent 2D image position and speed respectively,  $\epsilon_{x,y}$  and  $\epsilon_{u,v}$  are noise variables drawn from Gaussian distributions with the mean of zero and variances to consider the previous states of object. Specifically, the position variance  $\sigma_{x,y}^2$  varies with the size of object, and the speed variance  $\sigma_{u,v}^2$  is inversely proportional to the number of successfully tracked frames. The more number of the successfully tracked frames, the smaller the variance, i.e., the less particles to spread. Motion model by Andriyenko and Schindler (2011) and Milan et al. (2014) is also a constant velocity model. To

be specific, a term which considers differences between the velocities of one object in different time instants is formulated in Eq. 12 as,

$$C_{dyn} = \sum_{t=1}^{N-2} \sum_{i=1}^M \|\mathbf{v}_i^t - \mathbf{v}_i^{t+1}\|^2, \quad (12)$$

where  $\mathbf{v}_i^t$  is the velocity of target  $i$  at time  $t$ . It is computed as the displacement between object positions in two continuous frames. The first summation takes all the  $N$  frames into account and the second summation counts all the  $M$  trajectories/objects. Intuitively, this term penalizes the difference between velocities and forces trajectories to be smooth. A constant velocity model simultaneously considering the forward velocity and the backward velocity is proposed by Xing et al. (2009) to compute the affinity linking two tracklets in terms of motion. Given two tracklets  $\mathbf{T}_i$  and  $\mathbf{T}_j$ , let us assume there is a temporal gap between the tail of  $\mathbf{T}_i$  and the head of  $\mathbf{T}_j$ . The forward-direction motion is described by a Gaussian distribution centered in  $\mathbf{p}_j^{head}$ , the position of the head response in  $\mathbf{T}_j$ , with variance  $\Sigma_j^B$ . It estimates the probability of the position of tail response in  $\mathbf{T}_i$  plus forward displacement as  $\mathbf{p}_i^{tail} + \mathbf{v}_i^F \Delta t$ . The backward-direction motion is also represented as a Gaussian distribution, with the difference that motion is calculated backwardly from the position of head response in  $\mathbf{T}_j$  to the position of tail response in  $\mathbf{T}_i$ . The model is given in Eq. 13 as

$$P_m(\mathbf{T}_i, \mathbf{T}_j) = \mathcal{N}(\mathbf{p}_i^{tail} + \mathbf{v}_i^F \Delta t; \mathbf{p}_j^{head}, \Sigma_j^B) * \mathcal{N}(\mathbf{p}_j^{head} + \mathbf{v}_j^B \Delta t; \mathbf{p}_i^{tail}, \Sigma_i^F). \quad (13)$$

Different from previous graph based MOT approaches which treat each node as an individual observation (e.g, one detection response), Yang and Nevatia (2012b) employ the Conditional Random Field (CRF) model, treating the node as a pair of tracklets. The label of each node indicates whether the two tracklets corresponding to this node can be associated or not. This is addressed by the unary term in the CRF model, considering both the appearance and motion information. The probability in terms of motion is calculated based on the displacement between the estimated positions via a linear motion model and the observed positions. Figure 7(a) illustrates this model clear. Given two tracklets  $\mathbf{T}_1$  and  $\mathbf{T}_2$ , assuming that  $\mathbf{T}_1$  is the front one along the time axis compared with  $\mathbf{T}_2$ , there is a time gap  $\Delta t$  between tail of  $\mathbf{T}_1$  and head of  $\mathbf{T}_2$ . The probability of linking  $\mathbf{T}_1$  and  $\mathbf{T}_2$  depends on two terms. One is from the displacement between the observed position and the estimated position of tail of  $\mathbf{T}_1$ , which is defined as  $\Delta \mathbf{p}_1 = \mathbf{p}^{head} - \mathbf{v}^{head} \Delta t - \mathbf{p}^{tail}$ . The other one

is from the displacement between the observed position and the estimated position of head of  $\mathbf{T}_2$ , defined as  $\Delta \mathbf{p}_2 = \mathbf{p}^{tail} + \mathbf{v}^{tail} \Delta t - \mathbf{p}^{head}$ . The probability is

$$P_m(\mathbf{T}_1, \mathbf{T}_2) = \mathcal{N}(\Delta \mathbf{p}_1; \mathbf{0}, \Sigma_p) \mathcal{N}(\Delta \mathbf{p}_2; \mathbf{0}, \Sigma_p). \quad (14)$$

This motion model is essentially the same as the one by Xing et al. (2009). It is quite popular (Kuo et al. 2010; Kuo and Nevatia 2011; Yang et al. 2011; Qin and Shelton 2012; Nillius et al. 2006). However, it only considers the pair of tracklets itself. A motion model between two pairs of tracklets are also taken into consideration (Yang and Nevatia 2012b). Figure 7(b) shows the pairwise motion model. Considering two pairs of tracklets,  $(\mathbf{T}_1, \mathbf{T}_3)$  and  $(\mathbf{T}_2, \mathbf{T}_4)$ , they suppose  $\mathbf{T}_1$  and  $\mathbf{T}_2$  are tail-close tracklets. Firstly the earlier tail time when  $\mathbf{T}_1$  and  $\mathbf{T}_2$  end is defined as  $t_x = \min\{t_1^e, t_2^e\}$ , where  $t_i^e$  means the ending time of tracklet  $\mathbf{T}_i$ . Similarly, the later time when  $\mathbf{T}_3$  and  $\mathbf{T}_4$  start is represented as  $t_y = \max\{t_3^s, t_4^s\}$ . Obviously  $t_y > t_x$ . Then they compute the relative distance between the estimated positions of  $\mathbf{T}_1$  and  $\mathbf{T}_2$  at frame  $t_y$  as  $\Delta \mathbf{p}_1 = (\mathbf{p}_1^{t_1^e} + \mathbf{v}_1^{tail}(t_y - t_1^e)) - (\mathbf{p}_2^{t_2^e} + \mathbf{v}_2^{tail}(t_y - t_2^e))$ , where  $\mathbf{v}_1^{tail}$  and  $\mathbf{v}_2^{tail}$  are the tail velocity of  $\mathbf{T}_1$  and  $\mathbf{T}_2$ . On the other hand, the real relative distance between  $\mathbf{T}_1$  and  $\mathbf{T}_2$  at frame  $t_y$  is  $\Delta \mathbf{p}_2 = \mathbf{p}_3^{t_y} - \mathbf{p}_4^{t_y}$ . Similar to the motion model in the unary term, they employ a zero-mean Gaussian function as  $\mathcal{N}(\Delta \mathbf{p}_1 - \Delta \mathbf{p}_2; \mathbf{0}, \Sigma_p)$  to estimate the linkable probability. The insight behind this pairwise motion model is that if the difference between  $\Delta \mathbf{p}_1$  and  $\Delta \mathbf{p}_2$  is small and  $\mathbf{T}_1$  is associated with  $\mathbf{T}_3$  (i.e., the label of the node corresponding to  $(\mathbf{T}_1, \mathbf{T}_3)$  is 1), then the probability to associate  $\mathbf{T}_2$  and  $\mathbf{T}_4$  is high.

Besides considering position and velocity, Kuo and Nevatia (2011) also take the accelerate rate into consideration. The probability concerning motion of a state  $\{\hat{\mathbf{s}}_k\}$  ( $k$  is the frame index) given observation tracklet  $\{\mathbf{o}_k\}$  is modeled as,

$$P(\{\hat{\mathbf{s}}_k\} | \{\mathbf{o}_k\}) = \prod_k \mathcal{N}(\mathbf{x}_k - \hat{\mathbf{x}}_k; \mathbf{0}, \Sigma_p) \prod_k \mathcal{N}(\mathbf{v}_k; \mathbf{0}, \Sigma_v) \prod_k \mathcal{N}(\mathbf{a}_k; \mathbf{0}, \Sigma_a), \quad (15)$$

where  $\mathbf{v}_k = \frac{\hat{\mathbf{x}}_{k+1} - \hat{\mathbf{x}}_k}{t_{k+1} - t_k}$  is the velocity,  $\mathbf{a}_k = \frac{\mathbf{v}_k - \mathbf{v}_{k-1}}{0.5(t_{k+1} - t_{k-1})}$  is the acceleration, and  $\mathcal{N}$  is a zero-mean Gaussian distribution.

#### 4.2.2 Non-linear motion model

Linear motion model is commonly used to explain object's movement. However, there are some cases which

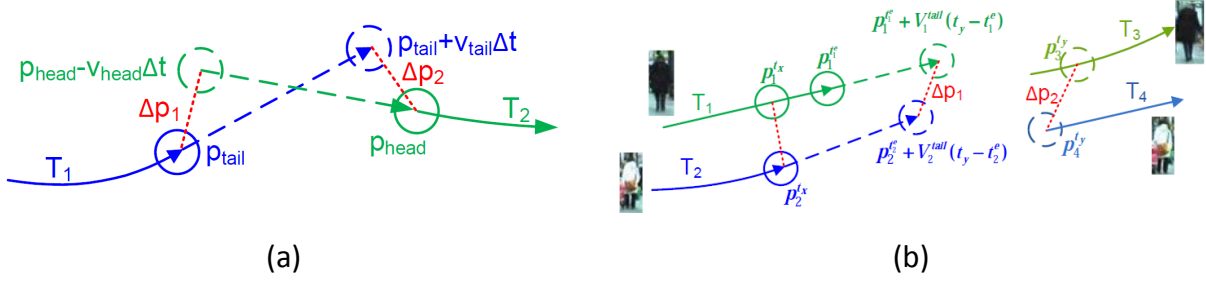


Fig. 7: The unary motion model (a) and pairwise motion model (b) (Yang and Nevatia 2012b)

the linear motion model cannot deal with. To this end, non-linear motion models are proposed to produce more accurate motion affinity between tracklets. For instance, Yang and Nevatia (2012a) employ a non-linear motion model to handle the situation that targets may move freely. Given two tracklets  $\mathbf{T}_1$  and  $\mathbf{T}_2$  which belong to the same target in Figure 8(a), the linear motion model (Yang and Nevatia 2012b) would produce low probability to link them, which is not consistent with the truth. Alternatively, employing the nonlinear motion model, which is composed of a set of pattern tracklets, the gap between tail of tracklet  $\mathbf{T}_1$  and head of tracklet  $\mathbf{T}_2$  could be reasonably explained by a tracklet  $\mathbf{T}_0 \in \mathcal{S}$ . As shown in Figure 8(b), the tracklet  $\mathbf{T}_0$  is a support tracklet to explain  $\mathbf{T}_1$  and  $\mathbf{T}_2$  because there exist elements  $\{(\mathbf{p}_i, s_i, \mathbf{v}_i)\}$  in  $\mathbf{T}_0$  which are matched with the tail of  $\mathbf{T}_1$  and the head of  $\mathbf{T}_2$ , where  $\mathbf{p}$ ,  $s$  and  $\mathbf{v}$  are position, size and velocity, respectively. Then the real path to bridge  $\mathbf{T}_1$  and  $\mathbf{T}_2$  is estimated based on  $\mathbf{T}_0$ , and the similar way as the linear motion model is employed to calculate the affinity between  $\mathbf{T}_1$  and  $\mathbf{T}_2$ , but based on the non-linear motion positions.

#### 4.3 Interaction Model

Interaction model, also known as mutual motion model, captures the influence of an object to other objects. This is a distinct issue of multiple object tracking compared with single object tracking. In the crowd scenery, obviously an object would consider some “force” from others. For instance, when a pedestrian is walking on the street, he would consider his speed, direction and destination, in order to avoid collision with others. Another example is that when a crowd of people walk across a street, each of them follows others and guides others at the same time, i.e., they form a motion pattern and every one follows this pattern. In fact, these are examples of two typical interaction models known as the *social force models* (Helbing and Molnar 1995) and the *crowd motion pattern models* (Hu et al. 2008).

There are some representative work of these models, which are illustrated as follows.

##### 4.3.1 Social force models

Social force models are also known as group models. In these models, each object is considered to be dependent from other objects and environmental factors. This type of information could alleviate performance deterioration in crowded scenes. In social force models, targets are considered as agencies which determine their speed, velocity, destination based on observations of other objects and the around environment. More specifically, in social force models, targets behavior is modeled based on two aspects, *individual force* and *group force*.

**Individual force.** For each individual in the scenario of multiple objects, two types of force are considered: 1) *fidelity*, which means one should not change his desired destination, 2) *constancy*, which means one should not suddenly change his velocity, including speed and direction.

**Group force.** For a whole group, three types of force are considered: 1) *attraction*, which means individuals moving together as a group should stay close, 2) *repulsion*, which means that individuals moving together as a group should keep some distance away from others to make all members comfortable and 3) *coherence*, which means individuals moving together as a group should be with similar velocity.

The majority of existing publications with social force model follows these two types of force. For instance, assuming that each pedestrian in a social group always adjusts its trajectory at an early stage in order to avoid possible collisions, Pellegrini et al. (2009) maximize the minimum distance which makes targets comfortable. In their model, each subject is represented as  $\mathbf{s}_i = \{\mathbf{p}_i^t, \mathbf{v}_i^t\}$ , where  $\mathbf{p}_i^t$  and  $\mathbf{v}_i^t$  are position and velocity at time  $t$ . Regarding subject  $i$ , the minimum distance between it and another subject  $j$  to make them comfortable is  $D_{ij}(\mathbf{v}_i)$ . The corresponding energy term is  $C_{ij}$ . To model this property among multiple objects

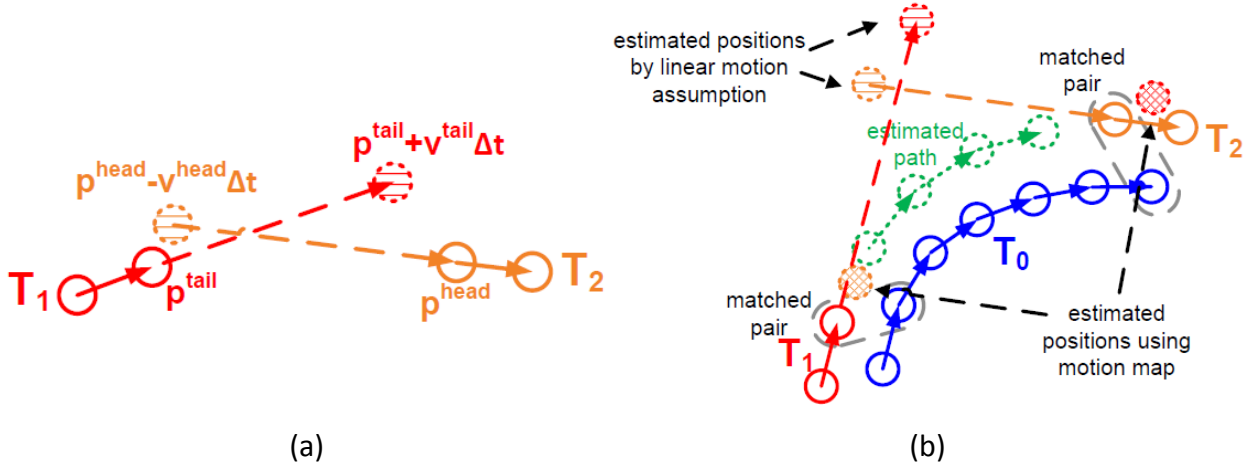


Fig. 8: An image comparing the linear motion model (a) with the non-linear motion model (b) (Yang and Nevatia 2012a). Best viewed in color

considering subject  $i$ , they balance each pair-wise energy term by a weight. Thus, the final energy between subject  $i$  and other objects considering interaction is

$$C_i^{inter}(\mathbf{v}_i) = \sum_{j \neq i} w_{ij} C_{ij}(\mathbf{v}_i), \quad (16)$$

where  $w_{ij}$  is the weight assigned to subject  $j$  considering subject  $i$ . Furthermore, they assume subject  $i$  walks to a destination  $\tilde{\mathbf{p}}_i$  with a desired speed  $u_i$ . Therefore they have two more energy terms as  $C_i(\mathbf{v}_i)^{ssc} = (u_i - \|\mathbf{v}_i\|)^2$  penalizing sudden speed change and  $C_i^{dfd}(\mathbf{v}_i) = -\frac{(\tilde{\mathbf{p}}_i - \mathbf{p}_i) \cdot \mathbf{v}_i}{\|\tilde{\mathbf{p}}_i - \mathbf{p}_i\| \|\mathbf{v}_i\|}$  penalizing the drift from destination. Then the complete energy objective is

$$C_i(\mathbf{v}_i) = C_i^{inter}(\mathbf{v}_i) + \lambda_1 C_i^{ssc}(\mathbf{v}_i) + \lambda_2 C_i^{dfd}(\mathbf{v}_i), \quad (17)$$

where  $\lambda_1$  and  $\lambda_2$  are parameters to balance these terms. By minimizing the above energy function for subject  $i$ , the search space of its destination could be largely reduced, and the data association procedure is further simplified.

Yamaguchi et al. (2011) also assume objects are agencies of a social force model. The destination of an object is determined by considering the so-called personal, social and environmental factors, which are formulated as terms in a cost function. In their model, each object state is represented as  $\mathbf{s}_i = \{\mathbf{p}_i, \mathbf{v}_i, u_i, \tilde{\mathbf{p}}_i, \mathcal{S}_i\}$  where  $\mathbf{p}_i$  is position,  $\mathbf{v}_i$  is velocity,  $u_i$  is the desired speed,  $\tilde{\mathbf{p}}_i$  is the desired destination and  $\mathcal{S}_i$  is the group of objects including object  $i$ . The behavior model to navigate objects has 6 terms, which are 1) a damping term penalizing the sudden change of velocity as  $C_{damping}(\mathbf{v}; \mathbf{s}_i) = \|\mathbf{v} - \mathbf{v}_i\|$ , 2) a speed term giving a cost as  $C_{speed}(\mathbf{v}; \mathbf{s}_i) = (u_i - \|\mathbf{v}_i\|)^2$  if the speed varies

from the desired speed, 3) a direction term as

$C_{direction}(\mathbf{v}; \mathbf{s}_i) = -\frac{\tilde{\mathbf{p}}_i - \mathbf{p}_i}{\|\tilde{\mathbf{p}}_i - \mathbf{p}_i\|} \cdot \frac{\mathbf{v}}{\|\mathbf{v}\|}$  to penalize the case that object does not follow the desired direction, 4) an attraction term

$$C_{attraction}(\mathbf{v}; \mathbf{s}_i, \mathcal{S}_i) = \sum_{j \in \{\mathcal{S}_i\}} \left( \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|} \cdot \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} \right) \left( \frac{\Delta \mathbf{p}_{ij}}{\|\Delta \mathbf{p}_{ij}\|} \cdot \frac{\mathbf{v}}{\|\mathbf{v}\|} \right)$$

assuming that people tend to stay close when they move together, 5) a group term  $C_{group}(\mathbf{v}; \mathbf{s}_i, \mathcal{S}_i) = \|\mathbf{v} - \bar{\mathbf{v}}_{\mathcal{S}_i}\|$  penalizing the variance of velocity in a group ( $\bar{\mathbf{v}}_{\mathcal{S}_i}$  is the group average velocity) and 6) a collision term (Pellegrini et al. 2009) (see Eq. 16 for details).

Social grouping behavior is considered to improve data association for MOT (Qin and Shelton 2012). To be specific, they assume people form  $K$  groups, where  $K$  can be learned optimally, and every tracklet assigned to the same group should be consistent with the group mean trajectory. Thus they have an extra cost term which takes the distance between a concerned tracklet and its assigned group trajectories into consideration.

Two factors, repulsion and group motion, are considered by Choi and Savarese (2010). The repulsion factor tries to separate objects if they are too close to each other. Given two targets  $i$  and  $j$  at time  $t$ , the potential concerning repulsion is  $S_{repulsion}(i, j) = \exp(-1/(c_r D_{ij}))$ , where  $D_{ij}$  is the distance between the two targets in the 3D space and  $c_r$  is a controlling parameter. It is obvious that if two objects are too close, then the potential between them is very small. Group motion factor assumes the relative distance between two objects in continuous two frames should keep unchanged. This also means that velocities of the pair of objects should be similar to each other. Thus the group motion is modeled as a potential term as  $S_{group}(i, j) = \exp(-c_g \bullet \|\mathbf{v}_i - \mathbf{v}_j\| / (1 + \exp(s_g(D_{ij} - D_g))))$ , where  $\mathbf{v}$

is velocity and  $c_g$  is a parameter to control this factor. The first factor is a soft step function to model how we consider two objects to be in the same group.  $D_g$  is a threshold distance and  $s_g$  is a parameter to balance the slope. From the group motion factor, if two objects are close enough and with similar velocity to be considered in the same group, then the potential is high.

A social force model which has four components is proposed by Scovanner and Tappen (2009) to learn dynamics of pedestrians in the real world. These four components contribute four energy terms as 1)  $C_{LM}$  which constraints the movement of a target to avoid jump in the space grid, 2)  $C_{CV}$  which maintains target's constant velocity, 3)  $C_{Dest}$  which guides a target to reach a destination, 4)  $C_{AV}$  which takes other targets into consideration, producing repulsion to avoid possible collisions. These four energy terms are weighted to form an energy objective which are then minimized to predict the movement of a target, generating tracks.

The data association problem and the group relationship mining problem are jointly estimated (Pellegrini et al. 2010). They model the trajectory assignment problem based on the motion and appearance information, and mine the group relationship by assuming targets belonging to the same group keep the distance constant and have the same direction. A three-order CRF model is proposed and an energy function with regarding to these two problems is constructed. By inferring the most probable estimation, these two problems are solved.

#### 4.3.2 Crowd motion pattern models

Inspired by the crowd simulation literature (Zhan et al. 2008), motion patterns are introduced to alleviate the difficulty of tracking an individual object in crowd. In general, this type of models is usually applied in the over-crowded scenario, i.e., the density of targets is considerably high. In the highly-crowded scenery, objects are usually quite small, and cues such as appearance and individual motion are ambiguous. In this case, motion from the crowd is a comparably reliable cue for the problem.

There have been some work in this direction. For example, an assumption is made that the behavior of an individual is determined by the scene layout and the surrounding objects (Ali and Shah 2008). In order to model the influence from others and the scene structure, three kinds of force from the floor fields are proposed. These fields are Static Floor Fields (SFF), Boundary Floor Field (BFF) and Dynamic Floor Field (DFF). SEF considers the scene structure, including the favorite path a crowd takes and the sinking point (exit)

of the scene. BFF takes the barriers in the scene into consideration by recognizing the physical and virtual barriers in the scene. DFF captures the motion of a crowd around the object being tracked to determine the future positions of objects in the crowd.

Zhao et al. (2012) deal with the multiple object tracking problem in the structured crowd scene by observing that crowd exhibits clear motion patterns in this case, and these patterns could benefit the tracking problem. Motion patterns are discovered by the ND tensor voting (Mordohai and Medioni 2010) of the tracklet points which are obtained by KLT tracker. These motion patterns are represented as a set of 4-D points  $\{\mathbf{p}_i = (x_i, y_i, u_{x_i}, v_{y_i}), i = 1, \dots, n\}$ , where  $x, y$  are spatial positions, and  $u, v$  are speed along the two dimensions of the image. Then these motion patterns are employed to 1) estimate the probability of a candidate for detection and 2) predict velocity of an object in a special position for tracking.

Observing that a group of pedestrians exhibits collective spatio-temporal structure, movement of an object within any local space-time location of a video are learned by training a set of Hidden Markov Models (HMM) (Kratz and Nishino 2010, 2012). The entire video is viewed as 3D volume, and it is divided into local spatio-temporal cubes. The motion pattern of a specific spatio-temporal cube is represented as a 3D Gaussian distribution considering the 3D gradients of all the pixels in the cube. This motion pattern is assumed to vary through time and exhibit the Markov property. Thus the future motion pattern could be predicted based on the previous states (motion patterns), and this predict motion pattern could be employed to constraint tracking of the object in this spatio-temporal location.

The motion pattern models described above make an assumption that the objects move coherently in a common direction. This may hold in the case of the so-called structured crowd scenarios, but does not comply with the unstructured crowd which exhibits various modalities of motion. To address it, Correlated Topic Model (CTM) is adopted by Rodriguez et al. (2009) to learn various motion behaviors in the scene. A tracker which can predict a rough displacement based on scene codebook from all the moving pixel in the scene, along with the learned high-level behavior, are weighted to track objects in the unstructured scenes. Similar to image retrieval, motion pattern could also be retrieved (Rodriguez et al. 2011). Motion patterns are firstly learned in an unsupervised and offline manner from a database composed of a large number of videos. Then given a test video, a set of space-time patches are matched to explain the test video. After that, motion priors in the retrieved video patches are transferred to



the test video as a prior to assist object tracking along with a Kalman filter based tracker.

#### 4.4 Exclusion Model

Exclusion is a constraint when seeking a solution to the MOT problem due to physical collisions. Given multiple detection responses and multiple trajectory hypotheses, generally there are two constraints to be considered. The first one is the so-called *detection-level exclusion* (Milan et al. 2013), i.e., two different detection responses in the same frame cannot be assigned to an identical trajectory hypothesis. The second one is the so-called *trajectory-level exclusion*, i.e., two trajectories cannot occupy an identical detection response. Modeling of them is presented as follows.

##### 4.4.1 Detection-level exclusion modeling

The detection-level exclusion is explicitly modeled by defining a cost term to penalize the case if two simultaneous detection response  $\mathbf{o}_i^t$  and  $\mathbf{o}_j^t$  at time  $t$  are assigned the same label of trajectory with a cost if they are distant (Milan et al. 2013).

KC and De Vleeschouwer (2013) employ label propagation for multiple object tracking. To model exclusion, a special exclusion graph is constructed to capture the constraint that detection responses with the same time stamp (occurring at the same time) should have different labels. Given all the detection responses, they define a graph where nodes represent detection responses. These nodes are not fully connected. Alternatively, each node (one detection) is connected only to nodes (other detections) happening at the same time as the node concerned. The connecting weight of each edge between two nodes are uniform as  $w_{ij} = 1/n$ , where  $n$  is the number of detection responses happening at the same time as these two nodes. After constructing this graph, the Laplacian matrix of this exclusion graph could be computed as  $\mathbf{L}$  and label error regarding exclusion is maximized as,

$$\arg \max_{\mathbf{Y} \in \mathcal{S}} Tr(\mathbf{YLY}), \quad (18)$$

where  $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_{|V|})$  is the label assignment of all the  $|V|$  nodes in the graph,  $\mathcal{S}$  is the set of all row-stochastic matrices of size  $|V| \times |V|$  and  $Tr(\bullet)$  is the trace norm of a matrix.

A topic model based on Dirichlet Process Mixture Model (DPMM) is employed for multi-object tracking

(Luo et al. 2015). To model the detection-level exclusion, the so-called cannot links are introduced to imitate that if two tracklets have overlap in their time span, then they cannot be assigned into one cluster, i.e., trajectory.

##### 4.4.2 Trajectory-level exclusion modeling

A term in Eq. 19 to model exclusion is accounted (Andriyenko and Schindler 2011) in an energy function as

$$C_{exc} \propto \sum_{t=1}^N \sum_{i \neq j} \frac{1}{\|\mathbf{p}_i^t - \mathbf{p}_j^t\|^2}. \quad (19)$$

The first summation accounts for all the  $N$  frames in the sequence, and the second summation accounts for all the detection responses in the same frame.  $\mathbf{p}_i^t$  is the ground plane coordinate of object  $i$  at time  $t$ . If two detection responses are too close, it will lead to considerably large cost of the energy function.

To model the trajectory level exclusion, Milan et al. (2013) penalize the case that two close trajectories  $\mathbf{T}_i$  and  $\mathbf{T}_j$  have different labels. The penalty is proportional to the spatial-temporal overlap between  $\mathbf{T}_i$  and  $\mathbf{T}_j$ . The closer the two trajectories, the higher penalty it is.

Similarly, mutual exclusion is modeled as an additional cost term to penalize the case that two trajectories are very close to each other. The cost is reversely proportional to the minimum distance between the trajectories in their temporal overlap (Andriyenko et al. 2012). By doing so, one of the trajectory would be abandoned to avoid the collision.

Exclusion is modeled as an extra constraint in the objective function of network flow (Butt and Collins 2013). Let the detection observations at frame  $k$  as  $\mathbf{O}_k = \{\mathbf{o}_1^k, \dots, \mathbf{o}_{M_k}^k\}$ . Given detection responses in two consecutive frames as  $\mathbf{O}_k$  and  $\mathbf{O}_{k+1}$ , one detection from  $\mathbf{O}_k$  and the other detection from  $\mathbf{O}_{k+1}$  can form a match. Based on all matches between these two frames, a graph is constructed as  $G = (V, E)$ , where each node in  $G$  is a pair of detections and each edge belonging to  $E$  represents flow in the graph, where flow 1 means linkable and 0 means not. Conflict edges are represented as  $E_{conflict}$ . Recalling the constraint that one detection should only be occupied by no more than one trajectory, the flow through edge in  $E_{conflict}$  is constrained to be at most 1.

#### 4.5 Occlusion Handling

Occlusion is a fatal issue in multiple object tracking. It could lead to ID switch or fragmentation of trajectories.

In order to handle occlusion, various kinds of strategies have been proposed. We here summarize the popular ones in the following aspects.

**Part-to-whole.** This strategy is the most popular one for occlusion handling. It is built on the assumption that, part of the object is still visible when occlusion happens. This assumption holds in most cases because even the complete occlusion still begins with partial occlusion. Based on this assumption, this strategy observe and utilize the visible part to infer state of the whole object.

Hu et al. (2012) propose a block-division model to deal with occlusion along with the task of recovering occlusion relationship among objects. In this model, object is divided into multiple un-overlapped blocks and for each block an appearance model based on subspace learning is constructed. Likelihood is computed according to reconstruction error in the subspace corresponding to each block. This model benefits the tracking problem under occlusion in two aspects. Firstly, spatial information is considered as likelihood of an observation is the product of likelihood of all its blocks. Secondly, an occlusion map could be obtained according to reconstruction errors of all blocks, and this occlusion map can be utilized to reason the occlusion relationship among objects. The relationship could be further utilized to selectively update appearance model.

Part based appearance model is learned to discriminate an object from other objects around and the background (Yang and Nevatia 2012c). To explicitly deal with occlusion, object is represented as 15 parts. Given a tracklet  $\mathbf{T}_k$ , its appearance model consists of a set of features  $\{\mathbf{f}_k^1, \mathbf{f}_k^2, \dots, \mathbf{f}_k^n\}$  and the corresponding weights  $\{w^1, w^2, \dots, w^n\}$  from its head observation and its tail observation. The link probability considering appearance between two tracklets is calculated as the similarity between the tail of one tracklet and the head of the other tracklet, which is  $\sum_i w_i F(\mathbf{f}_j^i, \mathbf{f}_k^i)$ , where  $\mathbf{f}_j^i$  and  $\mathbf{f}_k^i$  are from the  $i$ th part of the two tracklets and  $F(\bullet, \bullet)$  is a similarity evaluation function. The meaning of part model is that once a part is found occluded, all the features from that part are assumed to be invalid. Learning of the appearance model is conducted via a boosting algorithm.

Part based model is also applied by Izadinia et al. (2012) as a multi-person multi-part tracker. Beginning from a state-of-the-art part-based human detector (Felzenszwalb et al. 2010), they track the whole human body and individual body parts, and the final trajectory estimation is obtained by jointly considering association between the whole human body and the individual human body parts. Figure 9 shows how the part based model handles occlusion. The pedestrian is occluded

from frame 47 to frame 134. During this period, the whole-body human detector would be confused. However, thanks to the part detector, the visible parts are detected. Based on these parts, trajectories of visible parts are estimated. Furthermore, along with the trajectory of the whole body, the complete trajectory is recovered. A similar part based model for occlusion handling is also proposed by Shu et al. (2012).

In general, tracking based on appearance information may fail when occlusion happens. Analogous to part based model which can still observe some parts in case of occlusion, feature point clustering based tracking, which assumes feature points with similar motion should belong to the same object, is also applicable to address occlusion. As long as some parts of an object are visible, the clustering of feature point trajectories will work. There are some examples (Sugimura et al. 2009; Brostow and Cipolla 2006; Fragkiadaki et al. 2012).

**Hypothesize-and-test.** This strategy sidesteps challenges from occlusion by hypothesizing proposals and testing the proposals according to observations at hand.

An Explicit Occlusion Model (EOM) is proposed by Zhang et al. (2008) and integrated into the cost-flow framework to better handle long-term occlusion in data association for MOT. In classical data association, two tracklets are assumed to be linkable only when the temporal gap between them is small. This would somehow result in fragments in the final association estimation. Increasing the temporal gap threshold to make more tracklets linkable could ease the situation to some extent, but would possibly yield more errant associations. To deal with this, occlusion hypotheses are generated based on the occlusion constraints. If the distance and scale difference of two observations are small enough, then they are occludable. Assuming  $\mathbf{o}_i$  is occluded by  $\mathbf{o}_j$ , a corresponding occlusion hypothesis is  $\tilde{\mathbf{o}}_i^j = (\mathbf{p}_j, s_i, \mathbf{f}_i, t_j)$ , where  $\mathbf{p}_j$  and  $t_j$  are the position and time stamp of  $\mathbf{o}_j$ , and  $s_i$  and  $\mathbf{f}_i$  are the size and appearance feature of  $\mathbf{o}_i$ . Along with the original observations (tracklets), all the observations are given as input to the cost-flow framework and MAP is conducted to obtain the optimal solution.

The model adopted by Tang et al. (2013) and Tang et al. (2014) is also a hypothesize-and-test fashion to handle occlusion. Different from the traditional detector which treats occlusion as distraction, occlusion is employed to help detection by observing that occlusion yields typical appearance patterns. Specifically, a double-person detector is built to be aware of different levels of occlusion between two people. They train the double-person detector based on instances generated by synthetically combining two objects with different levels of occlusion, thus the resulting detector can be



Fig. 9: An image from illustrating how the part based model deals with occlusion (Izadinia et al. 2012)



Fig. 10: Training samples for the double-person detector (Tang et al. 2014). From left to right, the level of occlusion increases

occlusion aware. Figure 10 shows exemplar training instances for the double-person detector. Along with the traditional single person detector, this multi-person detector could be employed as the basis of multiple object tracking.

**Buffer-and-recover.** This strategy buffers observations when occlusion happens and remember states of objects before occlusion. When occlusion ends, object states are recovered based on the buffered observations and the stored states before occlusion.

Mitzel et al. (2010) combine a level-set tracker based on image segmentation and a high-level tracker based on detection for MOT. In their approach, the high-level tracker is employed to initialize new tracks from detection response and the level-set tracker is used to tackle the frame-to-frame data association. When occlusion occurs, the level-set tracker would fail. To tackle this, the high-level tracker keeps a trajectory alive for up to 15 frames when occlusion happens, and extrapolates the position to grow the dormant trajectory through occlusion. In case the object reappears, the track is fired again and the identity is maintained. The similar idea is also used by Mitzel and Leibe (2011).

Ryoo and Aggarwal (2008) propose an “observe-and-explain” strategy to handle the inter-object occlusion and scene-object occlusion. Their strategy could save computation cost as an observation mode is activated when the state of tracking is not clear due to occlusion. When they get enough observations, expla-

nations are generated to correspond to the observations. This could also be treated as “buffer-and-recover” strategy.

**Others.** The strategies described above do not cover all the tactics in the community. On one hand, in practice there exists a method which addresses occlusion based on overlap between detection bounding boxes. It is simple but works in some cases. On the other hand, the three types of strategies are not strictly parallel. Sometimes they are combined and used simultaneously.

#### 4.6 Probabilistic Inference

Approaches based on probabilistic inference framework typically represent states of objects (like size, position and velocity) with probabilistic distribution. The goal of algorithm is to estimate the probabilistic distribution of target status by a variety of probability reasoning methods based on existing observations. Through the probabilistic distribution, object states can be estimated. As this kind of approaches requires only the existing observations, they are especially appropriate for the task of online tracking. As only the existing observations are employed for estimation, it is naturally to impose the assumption of Markov property in the objects state sequence. This assumption includes two aspects and let us illustrate them by recalling the formula in Section 3.1.

First, the current object state only depends on the previous states. Further, it only depends on the very last state if the first-order Markov property is imposed. Formally,  $P(\mathbf{S}_t|\mathbf{S}_{1:t-1}) = P(\mathbf{S}_t|\mathbf{S}_{t-1})$ .

Second, observation of the object is only related to its state corresponding to the observation. Formally,  $P(\mathbf{O}_{1:t}|\mathbf{S}_{1:t}) = \prod_{i=1}^t P(\mathbf{O}_i|\mathbf{S}_i)$ .

Based on the two aspects, the estimation in Eq. 1 could be conducted by two steps of predict and update, which are:

$$\begin{aligned} \text{Predict: } P(\mathbf{S}_t|\mathbf{O}_{1:t-1}) &= \int P(\mathbf{S}_t|\mathbf{S}_{t-1})P(\mathbf{S}_{t-1}|\mathbf{O}_{1:t-1})d\mathbf{S}_{t-1} \\ \text{Update: } P(\mathbf{S}_t|\mathbf{O}_{1:t}) &\propto P(\mathbf{O}_t|\mathbf{S}_t)P(\mathbf{S}_t|\mathbf{O}_{1:t-1}) \end{aligned}$$

In the formula above,  $P(\mathbf{S}_t|\mathbf{S}_{t-1})$  and  $P(\mathbf{O}_t|\mathbf{S}_t)$  are the *Dynamic Model* and the *Observation Model* individually. The dynamic model corresponds to tracking strategy and the observation model provides observation measurement concerning object states. The *predict* step is to estimate the current state based on all the previous observations. More specifically, the posterior probability distribution of the current state is estimated by integrating in the space of the last object state via the dynamic model. The *update* step is to update the posterior probability distribution of states of objects based on the observation model.

According to the formula, states of objects can be estimated by iteratively conducting the predict and update steps. However, in practice, the object state distribution cannot be represented by simple distribution such as multivariate normal distribution, thus there is no analytical solution to the integral procedure. Additionally, for multiple objects, the dimension of the set of states is very large, which makes the integral and derivation of approximate solution more difficult.

Various kinds of probabilistic inference models have been applied to multi-object tracking (Kratz and Nishino 2010; Fortmann et al. 1983; Giebel et al. 2004), such as Kalman filter (Rodriguez et al. 2011; Reid 1979), Extended Kalman filter (Mitzel and Leibe 2011) and Particle filter (Jin and Mokhtarian 2007; Yang et al. 2005; Hess and Fern 2009; Han et al. 2007; Hu et al. 2012; Liu et al. 2012; Breitenstein et al. 2009; Yang et al. 2009b).

**Kalman filter.** In the case of linear system and Gaussian-distribution object states, Kalman filter is proved to be the optimal estimator. It has been applied (Rodriguez et al. 2011; Reid 1979).

**Extended Kalman filter.** For the nonlinear case, extended Kalman filter is a solution. It approximates the nonlinear system by Taylor Expansion (Mitzel and Leibe 2011).

**Particle filter.** Monte Carlo sampling based models become popular in tracking, especially after the introduction of Particle filter (Jin and Mokhtarian 2007;

Yang et al. 2005; Hess and Fern 2009; Han et al. 2007; Hu et al. 2012; Liu et al. 2012; Breitenstein et al. 2009; Yang et al. 2009b; Khan et al. 2004). Typically, the strategy of Maximum A Posteriori (MAP) (Liu et al. 2012; Breitenstein et al. 2009; Yang et al. 2009b; Mitzel and Leibe 2011; Rodriguez et al. 2011; Kratz and Nishino 2010; Reid 1979) is adopted to derive a state with the maximum probability.

#### 4.7 Deterministic Optimization

Approaches based on deterministic optimization framework at first obtain observations from each of the frame in the image sequence. Defining similarity among observations, the tracking problem is cast as a special optimization problem. Then by seeking the optimal solution of the optimization problem, final estimation of tracking results is obtained. Approaches within this framework are suitable for the task of offline tracking because observations from all the frames or at least a time window are required to be ready in advance. Given observations (usually detection hypotheses) from all the frames, these types of methods endeavor to globally associate observations belonging to an identical object into a trajectory. The key issue is how to seek the optimum association. Usually, the global optimum of all trajectories is formulated as a particular type of optimization problem. Some popular and well-studied approaches are detailed in the following.

**Bipartite Graph Matching.** By modeling the MOT problem as Bipartite Graph Matching, two disjoint sets of graph nodes could be existing trajectories and new detections in sequential tracking or two sets of tracklets in batch tracking. Weights among nodes are modeled as affinities among trajectories and detections. Then greedy bipartite assignment algorithm (Shu et al. 2012; Breitenstein et al. 2009; Wu and Nevatia 2007) or Hungarian algorithm (Qin and Shelton 2012; Reilly et al. 2010; Perera et al. 2006; Xing et al. 2009; Huang et al. 2008) are employed to derive the matching between nodes in the two sets.

**Dynamic Programming.** Extend Dynamic Programming (Wolf et al. 1989), Linear Programming (Jiang et al. 2007; Berclaz et al. 2009; Andriyenko and Schindler 2010), Quadratic Boolean Programming (Leibe et al. 2007), K-shortest path (Berclaz et al. 2011; Choi and Savarese 2012) and set cover (Wu et al. 2011) are adopted to solve the association problem among detections or tracklets.

**Min-cost Max-flow Network Flow.** Network flow is known as the transportation network. It is a directed graph where each edge has capacity. In the MOT problem, nodes in the graph for network flow are usually



low-level observations, which could be detection responses or tracklets. Usually the flow is modeled as an indicator to link two nodes (flow is 1) or not (flow is 0). To meet the flow balance requirement, a source node corresponding to the start of a trajectory and a sink node corresponding to the end of a trajectory are added to the original graph (see Figure 11). One trajectory corresponds to one flow path in the graph. The flow transited from the source node to the sink node equals to the number of trajectories or the objects in the video, and the cost to transit the flow from the source node to the sink node is the neg-likelihood of all the association hypotheses. Some examples (Zhang et al. 2008; Choi and Savarese 2012; Wu et al. 2012; Butt and Collins 2013; Pirsiavash et al. 2011) adopt this for MOT problem.

**Conditional Random Field.** The Conditional Random Field model is adopted to handle the multiple object tracking problem (Yang and Nevatia 2012b; Yang et al. 2011; Milan et al. 2013). Defining a graph  $G = (V, E)$  where  $V$  is the set of vertexes and  $E$  is the set of edges between vertexes, low level tracklets are given as input to the graph. Each node in the graph is defined as a pair of tracklets, and a label is predicted to indicate whether this pair of tracklets can be linked (label is 1) or not (label is 0). These labels compose the label map which corresponds to the optimal association of the tracklets for the MOT problem.

**MWIS.** The maximum-weight independent set (MWIS) is the heaviest subset of non-adjacent nodes of an attributed graph. Concerning the MOT problem, nodes in the attribute graph represent pair of tracklets in continuous frames, weights of nodes represent the affinity of the pair of tracklets, and the edge is connect if two tracklets share the same detection. Given this graph, the data association problem is modeled as the MWIS problem (Shafique et al. 2008; Brendel et al. 2011).

In practice, deterministic optimization usually outperforms approaches based on probabilistic inference, especially in the case of occlusion among objects. With the help of global information, occlusion is addressed better than its counterparts. However, global information simultaneously results in more consumption of time and space for the optimization. Additionally, the requirement of access to all frames in advance limits their applications in online tracking scenarios.

## 5 MOT Evaluation

Given a MOT algorithm developed, metrics, data sets are required to evaluate its performance. At the same time, one also need to compare with appropriate public algorithms to verify a developed algorithm. In the

following, we list metrics, publicly available data sets, codes and benchmark results.

### 5.1 Metrics

Evaluation metrics of MOT approaches are crucial as they provide standard for fair quantitative comparison. A brief review on different MOT evaluation metrics is presented in this section. As many approaches to MOT employ the tracking-by-detection strategy, they often measure detection performance as well as tracking performance. Metrics for object detection are therefore employed in MOT approaches. Based on this, MOT metrics can be largely categorized into two sets evaluating detection and tracking respectively, as listed in Table 6.

#### 5.1.1 Metrics for detection

We further group metrics for detection into two subsets. One set measures accuracy, and the other one measures precision.

**Accuracy.** The commonly used Recall and Precision metrics as well as the average False Alarms per Frame (FAF) rate are employed as MOT metrics (Yang et al. 2011). Choi and Savarese (2010) use the False Positive Per Image (FPPI) to evaluate detection performance in MOT. A comprehensive metric called Multiple Object Detection Accuracy (MODA), which considers the relative number of false positives and miss detections is utilized by Kasturi et al. (2009).

**Precision.** The Multiple Object Detection Precision (MODP) metric measures the quality of alignment between true detections and the ground truth (Kasturi et al. 2009).

#### 5.1.2 Metrics for tracking

Metrics for tracking are classified into four subsets by different attributes as the following.

**Accuracy.** This kind of metrics measures how accurately an algorithm could track target objects. The metric of ID switches (IDs) (Yamaguchi et al. 2011) counts how many times a MOT algorithm switches to wrong objects. Multiple Object Tracking Accuracy (MOTA) metric (Keni and Rainer 2008) combines the false positive rate, false negative rate and mismatch rate for MOT.

**Precision.** Two metrics, Multiple Object Tracking Precision (MOTP) (Keni and Rainer 2008) and Tracking Distance Error (TDE) (Kratz and Nishino 2010) belong to this subset. They describe how precisely the objects are tracked from the view of overlap and distance.



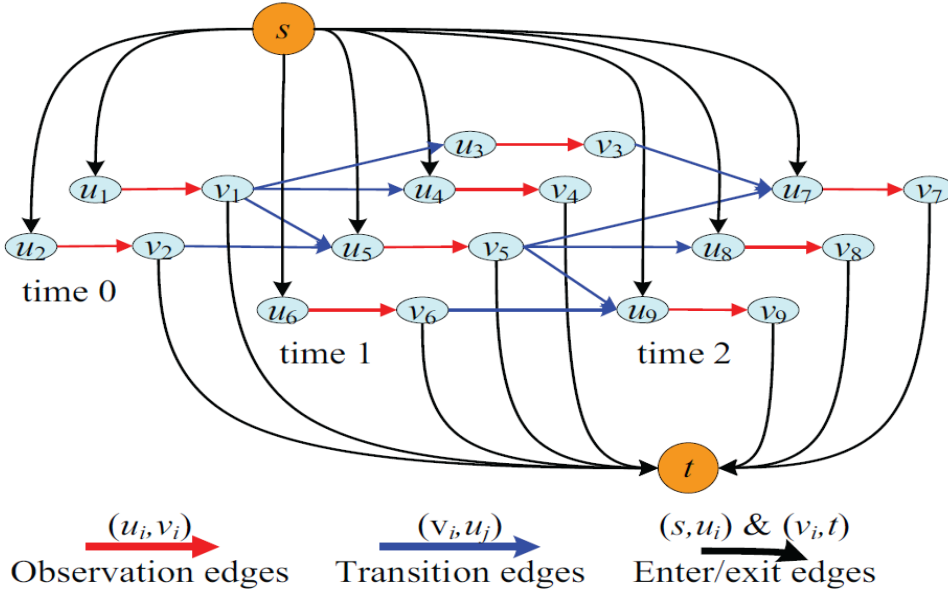


Fig. 11: An example of the cost-flow network with 3 timesteps and 9 observations (Zhang et al. 2008)

**Completeness.** Metrics for completeness indicate how completely the ground truth trajectories are tracked. Metrics of Mostly Tracked (MT), Partly Tracked (PT), Mostly Lost (ML) and Fragmentation (FM) (Li et al. 2009) could be grouped to this set.

**Robustness.** To assess the ability of a MOT algorithm to recover from occlusion, metrics of Recover from Short-term occlusion (RS) and Recover from Long-term occlusion (RL) are considered (Song et al. 2010).

## 5.2 Data Sets

To compare with various state-of-the-art MOT methods, publicly available datasets are employed to evaluate the proposed methods in individual publications. We here summarize the popular data sets in the literature, to give a clear view in Table 7.

## 5.3 Public Algorithms

We examine the literature and list algorithms with which the associated source codes are publicly available to make further comparisons convenient in Table 8.

## 5.4 Benchmark Results

We list public results on the data sets mentioned above to get a direct comparison among different approaches and provide convenience for future comparison in Table

9 to Table 26. Please note that this kind of direct comparison among different approaches on the same data set may not be fair. We list some issues which may result in unfairness in case of direct comparison:

- Different methodologies. For example, some publications belong to offline methods while others belong to online ones. Due to the difference described in Section 3.2.2, it is unfair to directly compare them.
- Different detection hypotheses. Different approaches adopt various detectors to obtain detection hypotheses. One approach based on different detection hypotheses would output different results, let alone different approaches.
- Some approaches utilize observations from multiple views while some approaches adopt information from a single view. This makes the comparison between them difficult.
- Prior information, such as scene structure and the number of pedestrians, are employed by some approaches. Direct comparison between these approaches and others is not so convincing.

In order to make direct and fair comparison, one needs to fix all the other components while vary the concerned component. For instance, adopting different data association models while keeping all other parts the same could directly compare performance of different data association methods. This is what another survey (Leal-Taixé et al. 2015) specifically focusing on evaluation of multiple object tracking provides. For intensive experimental comparison among different MOT

Table 6: An overview of evaluation metrics for MOT. The up arrow (*resp.* down arrow) indicates that the performance is better if the quantity is greater (*resp.* smaller)

Type	Concern	Metric	Description	Note
Detection	Accuracy	Recall	correctly matched detections over ground-truth detections	↑
		Precision	correctly matched detections over result detections	↑
		FAF/FPPI	number of false alarms averaged over a sequence	↓
		MODA	take the miss detection, false positive rate into account	↑
	Precision	MODP	the overlap between true positives and ground truth	↑
Tracking	Accuracy	MOTA	take the false negative, false positive and mismatch rate into account	↑
		IDS	the number of times that a tracked trajectory changes its matched ground-truth identity	↓
	Precision	MOTP	overlap between the estimated positions and the ground truth averaged over the matches	↑
		TDE	difference between the ground-truth annotation and the tracking result	↓
	Completeness	MT	percentage of ground-truth trajectories which are covered by tracker output for more than 80% in length	↑
		ML	percentage of ground-truth trajectories which are covered by tracker output for less than 20% in length	↓
		PT	1.0 - MT - ML	-
		FM	the number of times that a ground-truth trajectory is interrupted in tracking result	↓
	Robustness	RS	the ratio of tracks which are correctly recovered from short occlusion	↑
		RL	the ratio of tracks which are correctly recovered from long occlusion	↑

solutions, one may refer to (Leal-Taixé et al. 2015) for interest. In spite of the issues mentioned above, it is still worthy to list all the public results on the same data set due to the following reasons.

- By listing the public results, it at least provides intuitive comparison among different methods on the same data set and convenience for future comparison if one evaluates his developed MOT algorithm.
- Although comparison among individual methods may not be fair, intuitive comparison between different types of methods such as that between offline methods and online methods could tell us how these types of methods work in public data sets.
- Additionally, we could observe how the research of MOT progresses along years by comparing performance of methods in different years.

We list public results on the popular data sets, specifically, PETS2009S2L1 (Table 9), PETS2009S2L2 (Table 10), PETS2009S2L3 (Table 11), PETS2009S1L1-2 (Table 12), PETS2009S1L1-1 (Table 13), PETS2009S3 MF1 (Table 14), CAVIAR (Table 15), Airport (Table 16), Town Center (Table 17), i-LIDS (Table 18),

TRECVID2008 (Table 19), ETH Central (Table 20), Hockey (Table 21), Parking Lot (Table 22), TUD Crossing (Table 23), TUD Stadtmitte (Table 24), ETHMS (Table 25) and TUD Campus (Table 26). For each data set, we report the results in terms of the MOTA, MOTP, IDS, Precision, Recall, MT, PT, ML, FM and F1 metrics. At the same time, we organize the results in two groups, which correspond to offline and online methods respectively. Please note that, 1) in each table, the top rows and bottom rows (separated by a line) are results of offline and online methods correspondingly, 2) in some tables there is no separation of rows, which means all the methods in this table belongs to offline methods, 3) cells in tables may be null, which means that we did not find the corresponding value neither from the original publication nor from other publications which cite it, 4) in some cases, there could be different results for an unique publication (for example, result from the original publication versus result from another publication which compares with it). This might because different configurations are adopt (e.g. different detection hypotheses). In this case, we quote either the most popularly cited or the latest results here.

Table 7: An overview of publicly available data sets. The tick means ground truth is available while the cross means not available

Data set	Multi-view	Ground truth	Web link
PETS 2009	✓	✓	<a href="http://www.cvg.rdg.ac.uk/PETS2009/a.html">www.cvg.rdg.ac.uk/PETS2009/a.html</a>
PETS 2006	✓	✗	<a href="http://www.cvg.rdg.ac.uk/PETS2006/data.html">www.cvg.rdg.ac.uk/PETS2006/data.html</a>
PETS 2007	✓	✓	<a href="http://www.pets2007.net/">www.pets2007.net/</a>
CAVIAR	✓	✓	<a href="http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/">http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/</a>
Trecvid 2008	✓	✗	<a href="http://www-nlpir.nist.gov/projects/tv2008/">www-nlpir.nist.gov/projects/tv2008/</a>
TUD	✗	✓	<a href="http://www.d2.mpi-inf.mpg.de/datasets">www.d2.mpi-inf.mpg.de/datasets</a>
Caltech Pedestrian	✗	✓	<a href="http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/">www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/</a>
UBC Hockey	✗	✗	<a href="http://www.cs.ubc.ca/~okumak/research.html">www.cs.ubc.ca/~okumak/research.html</a>
Lids AVSS 2007	✗	✓	<a href="http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html">www.eecs.qmul.ac.uk/~andrea/avss2007_d.html</a>
ETH pedestrian	✓	✓	<a href="http://www.vision.ee.ethz.ch/~aess/dataset/">www.vision.ee.ethz.ch/~aess/dataset/</a>
ETHZ Central	✗	✓	<a href="http://www.vision.ee.ethz.ch/datasets/">www.vision.ee.ethz.ch/datasets/</a>
Town Centre	✗	✓	<a href="http://www.robots.ox.ac.uk/ActiveVision/Research/Projects/2009bбенfold_headpose/project.html#datasets">www.robots.ox.ac.uk/ActiveVision/Research/Projects/2009bбенfold_headpose/project.html#datasets</a>
Zara	✗	✗	<a href="https://graphics.cs.ucy.ac.cy/research/downloads/crowd-data">https://graphics.cs.ucy.ac.cy/research/downloads/crowd-data</a>
UCSD	✗	✗	<a href="http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm">http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm</a>
UCF Crowds	✗	✗	<a href="http://www.crcv.ucf.edu/data/crowd.php">www.crcv.ucf.edu/data/crowd.php</a>

Table 8: List of publicly available program codes

Reference	Web Link
Choi and Savarese (2010)	<a href="https://www.eecs.umich.edu/vision/mttproject.html">https://www.eecs.umich.edu/vision/mttproject.html</a>
Jiang et al. (2007)	<a href="http://www.cs.bc.edu/~hjiang/details/tracking/index.html">http://www.cs.bc.edu/~hjiang/details/tracking/index.html</a>
Milan et al. (2014)	<a href="http://research.milanton.de/contracking/">http://research.milanton.de/contracking/</a>
Andriyenko et al. (2012)	<a href="http://research.milanton.de/dctracking/">http://research.milanton.de/dctracking/</a>
Milan et al. (2013)	<a href="https://bitbucket.org/amilan/dctracking">https://bitbucket.org/amilan/dctracking</a>
Zamir et al. (2012)	<a href="http://crcv.ucf.edu/projects/GMCP-Tracker/">http://crcv.ucf.edu/projects/GMCP-Tracker/</a>
Berclaz et al. (2011)	<a href="http://cvlab.epfl.ch/software/ksp">http://cvlab.epfl.ch/software/ksp</a>
Okuma et al. (2004)	<a href="http://www.cs.ubc.ca/~okumak/research.html">http://www.cs.ubc.ca/~okumak/research.html</a>
Zhang and van der Maaten (2013) Zhang and van der Maaten (2014)	<a href="http://visionlab.tudelft.nl/spot">http://visionlab.tudelft.nl/spot</a>
Pirsiavash et al. (2011)	<a href="http://www.ics.uci.edu/~dramanan/">http://www.ics.uci.edu/~dramanan/</a>
Rodriguez et al. (2009)	<a href="http://www.mikelrodriguez.com/crowd-tracking-matlab-application">http://www.mikelrodriguez.com/crowd-tracking-matlab-application</a>
Possegger et al. (2014)	<a href="https://lrs.icg.tugraz.at/download.php#motog">https://lrs.icg.tugraz.at/download.php#motog</a>

We conduct analysis of benchmark results on the Town Center and the CAVIAR data set to investigate the comparison between offline methods and online methods. The reason we choose these two data sets is that they are popularly utilized for evaluation. We average values of each metric across each type of methods, and report the mean value and the standard deviation in Table 27 and Table 28. To make the comparison more intuitive, we additionally show the comparative results in the form of bar. For each table, IDS and FM are shown separately as their value range is different from other metrics.

Some interesting points could be achieved based on the analysis. As shown in Figure 12 and Figure 13, offline methods generally outperform online ones regarding most of the metrics. The same conclusion could also

be draw from Figure 14 and Figure 15. It coincides with the fact that offline methods employ globally temporal information for the estimation.

Additionally, we conduct analysis of the year-wise results on the PETS2009S2L1 data set. To be specific, we calculate metric values of methods in years ranging from 2009 to 2015, and report the mean and standard deviation value in Table 29. These results are also represented in Figure 16 and Figure 17. In general, the performance improves across years. We suspect that contributors such as better models and progress in object detection could be employed to explain the achieved progress.

Table 9: Benchmark results on the PETS2009S2L1 data set

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Berclaz et al. (2011)	0.803	0.720	13	0.963	0.838	0.739	0.174	0.087	22	0.896	2011	Wen et al. (2014)
Ben Shitrit et al. (2011)	0.815	0.584	19	0.907	0.908	-	-	-	-	0.907	2011	Zamir et al. (2012)
Andriyenko and Schindler (2011)	0.863	0.787	38	0.976	0.895	0.783	0.174	0.043	21	0.934	2011	Wen et al. (2014)
Henriques et al. (2011)	0.848	0.687	10	0.924	0.940	-	-	-	-	0.932	2011	Zamir et al. (2012)
Izadinia et al. (2012)	0.907	0.760	-	0.968	0.952	-	-	-	-	0.960	2012	Izadinia et al. (2012)
Zamir et al. (2012)	0.903	0.690	8	0.936	0.965	-	-	-	-	0.950	2012	Zamir et al. (2012)
Segal and Reid (2013)	0.900	0.750	6	-	-	0.890	-	-	-	-	2013	Dehghan et al. (2015)
KC and De Vleeschouwer (2013)	0.910	0.700	5	-	-	-	-	-	-	-	2013	Dehghan et al. (2015)
Milan et al. (2014)	0.906	0.802	11	0.984	0.924	0.913	0.043	0.043	-	0.953	2014	Wen et al. (2014)
Andriyenko et al. (2012)	0.883	0.796	18	0.987	0.900	0.826	0.174	0.000	14	0.941	2012	Wen et al. (2014)
Pirsiavash et al. (2011)	0.774	0.743	57	0.972	0.812	0.609	0.347	0.043	62	0.885	2011	Wen et al. (2014)
Wen et al. (2014)	0.927	0.729	5	0.984	0.944	0.957	0.043	0.000	10	0.964	2014	Wen et al. (2014)
Kuo and Nevatia (2011)	-	-	1	0.996	0.895	0.789	0.211	0.000	23	0.943	2011	Yang and Nevatia (2012a)
Yang and Nevatia (2012a)	-	-	0	0.990	0.918	0.895	0.105	0.000	9	0.953	2012	Yang and Nevatia (2012a)
Hofmann et al. (2013b)	0.980	0.828	10	-	-	1.000	0.000	0.000	11	-	2013	Possegger et al. (2014)
Yang and Nevatia (2012c)	-	-	0	0.948	0.978	0.950	0.050	0.000	2	0.963	2012	Zhang et al. (2015)
Zhang et al. (2015)	0.956	0.916	0	0.986	0.970	0.950	0.050	0.000	4	0.978	2015	Zhang et al. (2015)
Milan et al. (2013)	0.903	0.743	22	-	-	0.783	0.217	0.000	15	-	2013	Milan et al. (2013)
Andriyenko et al. (2011)	0.917	0.745	11	-	-	-	-	-	-	-	2011	Zhang et al. (2012)
Leal-Taixé et al. (2011)	0.670	-	-	-	-	-	-	-	-	-	2011	Zhang et al. (2012)
Berclaz et al. (2009)	0.830	0.520	-	0.820	0.530	-	-	-	-	0.644	2009	Izadinia et al. (2012)
Hofmann et al. (2013a)	0.978	0.753	8	0.991	0.990	1.000	0.000	0.000	8	0.990	2013	Hofmann et al. (2013a)
Leal-Taixé et al. (2012)	0.760	0.600	-	-	-	-	-	-	-	-	2012	Possegger et al. (2014)
Shi et al. (2013)	0.927	0.818	7	0.982	0.960	0.947	0.053	0.000	11	0.971	2013	Shi et al. (2014)
Shi et al. (2014)	0.961	0.818	4	0.989	0.977	0.947	0.053	0.000	6	0.983	2014	Shi et al. (2014)
Dehghan et al. (2015)	0.904	0.631	3	-	-	0.950	0.050	0.000	-	-	2015	Dehghan et al. (2015)
Breitenstein et al. (2011)	0.797	0.563	-	-	-	-	-	-	-	-	2011	Possegger et al. (2014)
Yang et al. (2009a)	0.759	0.538	-	-	-	-	-	-	-	-	2009	Possegger et al. (2014)
Bae and Yoon (2014)	0.830	0.696	4	-	-	1.000	0.000	0.000	4	-	2014	Bae and Yoon (2014)
Zhang et al. (2012)	0.933	0.682	19	-	-	-	-	-	-	-	2012	Zhang et al. (2012)
Conte et al. (2010)	0.810	0.570	-	0.850	0.580	-	-	-	-	0.690	2010	Izadinia et al. (2012)
Alahi et al. (2009)	0.830	0.520	-	0.690	0.530	-	-	-	-	0.600	2009	Izadinia et al. (2012)
Wu et al. (2013b)	0.928	0.743	8	-	-	1.000	0.000	0.000	11	-	2013	Wu et al. (2013b)
Possegger et al. (2014)	0.981	0.805	9	-	-	1.000	0.000	0.000	16	-	2014	Possegger et al. (2014)

Table 10: Benchmark results on the PETS2009S2L2 data set

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Milan et al. (2014)	0.569	0.594	99	0.898	0.655	0.378	0.459	0.162	73	0.757	2014	Wen et al. (2014)
Berclaz et al. (2011)	0.242	0.609	22	0.921	0.268	0.095	0.364	0.541	38	0.415	2011	Wen et al. (2014)
Andriyenko and Schindler (2011)	0.485	0.620	152	0.937	0.539	0.203	0.608	0.189	128	0.684	2011	Wen et al. (2014)
Andriyenko et al. (2012)	0.480	0.616	143	0.947	0.526	0.203	0.649	0.149	125	0.676	2012	Wen et al. (2014)
Pirsiavash et al. (2011)	0.450	0.641	137	0.954	0.490	0.095	0.676	0.230	216	0.647	2011	Wen et al. (2014)
Wen et al. (2014)	0.621	0.527	125	0.903	0.712	0.365	0.595	0.041	175	0.796	2014	Wen et al. (2014)
Zhang et al. (2015)	0.556	0.671	112	0.920	0.621	0.284	0.554	0.162	91	0.742	2015	Zhang et al. (2015)
Milan et al. (2013)	0.460	0.598	126	-	-	0.338	0.554	0.108	105	-	2013	Milan et al. (2013)
Andriyenko et al. (2011)	0.602	0.630	104	-	-	-	-	-	-	-	2011	Zhang et al. (2012)
Hofmann et al. (2013a)	0.571	0.564	67	0.921	0.638	0.395	0.421	0.184	59	0.754	2013	Hofmann et al. (2013a)
Hofmann et al. (2013b)	0.758	0.721	234	-	-	0.651	0.349	0.000	252	-	2013	Possegger et al. (2014)
Breitenstein et al. (2011)	0.500	0.513	-	-	-	-	-	-	-	-	2011	Bae and Yoon (2014)
Bae and Yoon (2014)	0.701	0.539	45	0.857	0.850	0.716	0.270	0.014	52	0.853	2014	Bae and Yoon (2014)
Zhang et al. (2012)	0.667	0.582	215	-	-	-	-	-	-	-	2012	Zhang et al. (2012)
Wu et al. (2013b)	0.733	0.732	122	-	-	0.689	0.270	0.041	113	-	2013	Possegger et al. (2014)
Possegger et al. (2014)	0.660	0.648	181	-	-	0.415	0.585	0.000	315	-	2014	Possegger et al. (2014)

Table 11: Benchmark results on the PETS2009S2L3 dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Milan et al. (2014)	0.454	0.646	38	0.909	0.518	0.205	0.386	0.409	27	0.660	2014	Wen et al. (2014)
Berclaz et al. (2011)	0.288	0.618	7	0.957	0.304	0.114	0.182	0.705	-	0.461	2011	Wen et al. (2014)
Andriyenko and Schindler (2011)	0.512	0.542	82	0.929	0.581	0.159	0.614	0.227	-	0.715	2011	Wen et al. (2014)
Andriyenko et al. (2012)	0.469	0.578	73	0.961	0.513	0.159	0.432	0.409	-	0.669	2012	Wen et al. (2014)
Pirsiavash et al. (2011)	0.430	0.630	52	0.970	0.460	0.114	0.477	0.409	-	0.624	2011	Wen et al. (2014)
Wen et al. (2014)	0.553	0.532	36	0.930	0.610	0.273	0.522	0.205	-	0.737	2014	Wen et al. (2014)
Zhang et al. (2015)	0.426	0.639	13	0.964	0.433	0.136	0.341	0.523	8	0.598	2015	Zhang et al. (2015)
Milan et al. (2013)	0.398	0.650	27	-	-	0.182	0.386	0.432	22	-	2013	Milan et al. (2013)
Andriyenko et al. (2011)	0.434	0.600	23	-	-	-	-	-	-	-	2011	Zhang et al. (2012)
Hofmann et al. (2013a)	0.415	0.650	46	-	-	0.341	0.341	0.318	67	-	2013	Possegger et al. (2014)
Hofmann et al. (2013b)	0.628	0.705	225	-	-	0.545	0.341	0.114	217	-	2013	Possegger et al. (2014)
Zhang et al. (2012)	0.404	0.564	80	-	-	-	-	-	-	-	2012	Zhang et al. (2012)
Wu et al. (2013b)	0.583	0.697	41	-	-	0.477	0.341	0.182	39	-	2013	Possegger et al. (2014)
Possegger et al. (2014)	0.625	0.626	59	-	-	0.318	0.546	0.136	98	-	2014	Possegger et al. (2014)

## 6 Conclusion and Future Directions

This paper has presented a comprehensive review of Multiple Object Tracking (MOT). The review has illustrated the current state, the key issues, and evaluation of this topic by classification, discussion and case study. Although great progress in MOT has been recently achieved, there are still issues remaining to be tackled.

- *MOT with video adaptation.* As aforementioned, the majority of current MOT methods requires an of-

fine trained object detector. There arises a problem that the detection result for a specific video is not optimal since the object detector is not trained for the given video. This would decrease the performance of multiple object tracking. The customization of the object detector is necessary to improve MOT performance. One solution is proposed by [Shu et al. \(2013\)](#), which adapts a generic pedestrian detector to a specific video by progressively refining the generic pedestrian detector. This is an important direction to improve the pre-procedure for MOT.

Table 12: Benchmark results on the PETS2009S1L1-2 dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Milan et al. (2014)	0.579	0.597	21	0.918	0.645	0.528	0.167	0.306	13	0.758	2014	Wen et al. (2014)
Berclaz et al. (2011)	0.515	0.648	4	0.936	0.555	0.444	0.167	0.389	8	0.697	2011	Wen et al. (2014)
Andriyenko and Schindler (2011)	0.480	0.645	17	0.974	0.500	0.250	0.417	0.333	12	0.661	2011	Wen et al. (2014)
Andriyenko et al. (2012)	0.544	0.643	24	0.965	0.574	0.417	0.278	0.306	17	0.720	2012	Wen et al. (2014)
Pirsiavash et al. (2011)	0.454	0.668	38	0.995	0.471	0.250	0.361	0.389	32	0.639	2011	Wen et al. (2014)
Wen et al. (2014)	0.571	0.548	4	0.978	0.586	0.500	0.278	0.222	7	0.733	2014	Wen et al. (2014)
Milan et al. (2013)	0.600	0.619	22	-	-	0.583	0.111	0.306	19	-	2013	Milan et al. (2013)

Table 13: Benchmark results on the PETS2009S1L1-1 dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Andriyenko and Schindler (2011)	0.400	0.694	25	0.979	0.415	0.196	0.370	0.435	18	0.583	2011	Wen et al. (2014)
Andriyenko et al. (2012)	0.376	0.658	44	0.968	0.401	0.196	0.391	0.413	36	0.567	2012	Wen et al. (2014)
Pirsiavash et al. (2011)	0.328	0.765	35	0.978	0.345	0.152	0.370	0.478	42	0.510	2011	Wen et al. (2014)
Wen et al. (2014)	0.411	0.719	11	0.997	0.415	0.239	0.348	0.413	10	0.586	2014	Wen et al. (2014)
Milan et al. (2014)	0.308	0.490	61	0.864	0.385	0.163	0.372	0.465	35	0.533	2014	Milan et al. (2014)
Berclaz et al. (2011)	0.195	0.606	7	0.926	0.214	0.093	0.233	0.674	11	0.348	2011	Milan et al. (2014)
Milan et al. (2013)	0.265	0.602	34	-	-	0.130	0.370	0.500	27	-	2013	Milan et al. (2013)

Table 14: Benchmark results on the PETS2009S3 MF1 dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Milan et al. (2014)	0.967	0.827	0	0.990	0.977	1.000	0.000	0.000	-	0.983	2014	Milan et al. (2014)
Berclaz et al. (2011)	0.837	0.778	0	0.954	0.879	0.857	0.143	0.000	0	0.915	2011	Milan et al. (2014)
Andriyenko et al. (2011)	0.963	0.841	-	-	-	1.000	0.000	0.000	-	-	2011	Hofmann et al. (2013a)
Hofmann et al. (2013a)	0.992	0.777	0	0.991	1.000	1.000	0.000	0.000	0	0.995	2013	Hofmann et al. (2013a)

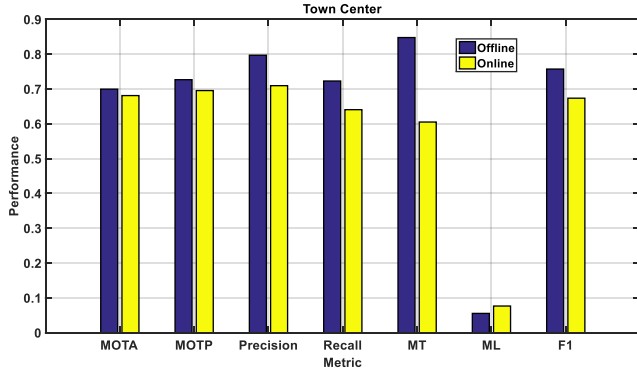


Fig. 12: Comparison between offline methods and on-line methods on the Town Center data set in terms of MOTA, MOTP, Precision, Recall, MT, ML and F1 metrics

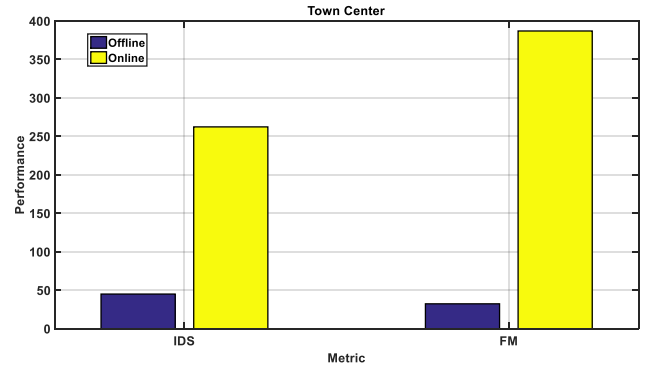


Fig. 13: Comparison between offline methods and online methods on the Town Center data set in terms of IDS and FM metrics

- *Balance between crowd density and completeness of object representation.* In general, the denser the crowd is, the less ratio of object's body is visible. In the under-crowd scenery, usually the whole body of an object can be recovered easily. However, in the over-crowd scenery, often only the head of an object is observable. With the minimum appearance information, motion pattern of the crowd could be, alternatively, helpful for MOT since targets exhibit coherent motion patterns.
- *MOT under multiple cameras.* It is obvious that MOT would benefit from multi-camera settings (Hofmann et al. 2013b; Chang et al. 2000). There are two kinds of configurations of multiple cameras. The first one is that multiple cameras record the same scene, i.e., multiple views. How to fuse information

from multiple cameras is a key issue in this configuration. The second one is that each camera of multiple camera records a different scene, i.e., a multi-camera network. One issue in data association of multiple cameras is the object re-identification problem.

- *Multiple 3D object tracking.* Most of the current approaches focus on multiple object tracking in 2D scenery, even in the case of multiple cameras. 3D tracking (Park et al. 2008), which could provide more accurate position, size estimation and effective occlusion handling for high-level computer vision tasks, is potentially more useful. However, 3D tracking requires more parameters to estimate and more computation cost compared with 2D tracking. Meanwhile, 3D model acquisition is another issue exclusive to 2D MOT.



Table 15: Benchmark results on the CAVIAR dataset

Reference	MOTA $\uparrow$	MOTP $\uparrow$	IDS $\downarrow$	Precision $\uparrow$	Recall $\uparrow$	MT $\uparrow$	PT	ML $\downarrow$	FM $\downarrow$	F1 $\uparrow$	Year	Source
Huang et al. (2008)	-	-	12	-	0.863	0.783	0.147	0.070	54	-	2008	Kuo et al. (2010)
Zhang et al. (2008)	-	-	15	-	0.764	0.857	0.107	0.036	20	-	2008	Zhang et al. (2008)
Chen et al. (2014)	-	-	5	-	-	0.907	0.066	0.027	6	-	2014	Chen et al. (2014)
Qin and Shelton (2012)	-	-	5	-	-	0.893	0.080	0.027	7	-	2012	Chen et al. (2014)
Kuo et al. (2010)	-	-	11	0.969	0.894	0.846	0.147	0.007	18	0.930	2010	Yang and Nevatia (2012a)
Li et al. (2009)	-	-	11	0.941	0.890	0.846	0.140	0.014	17	0.915	2009	Zhang et al. (2015)
Kuo and Nevatia (2011)	-	-	4	0.966	0.881	0.860	0.133	0.007	17	0.922	2011	Yang and Nevatia (2012a)
Yang and Nevatia (2012a)	-	-	5	0.961	0.902	0.891	0.102	0.007	11	0.931	2012	Yang and Nevatia (2012a)
Zhang et al. (2015)	0.872	0.763	6	0.978	0.892	0.853	0.132	0.015	32	0.933	2015	Zhang et al. (2015)
Yang et al. (2009b)	0.593	-	-	0.819	0.754	-	-	-	-	0.783	2009	Yang et al. (2009b)
Wu and Nevatia (2007)	-	-	17	-	0.752	0.757	0.179	0.064	35	-	2007	Xing et al. (2009)
Wu and Nevatia (2006)	-	-	17	-	-	0.757	0.179	0.064	35	-	2006	Zhang et al. (2008)
Zhao and Nevatia (2004)	-	-	14	-	-	0.646	0.440	0.049	57	-	2004	Wu and Nevatia (2006)
Xing et al. (2009)	-	-	14	-	0.818	0.843	0.121	0.036	24	-	2009	Xing et al. (2009)
Bae and Yoon (2014)	0.865	0.872	9	0.990	0.886	0.895	0.105	0.000	8	0.935	2014	Bae and Yoon (2014)

Table 16: Benchmark results on the Airport dataset

Reference	MOTA $\uparrow$	MOTP $\uparrow$	IDS $\downarrow$	Precision $\uparrow$	Recall $\uparrow$	MT $\uparrow$	PT	ML $\downarrow$	FM $\downarrow$	F1 $\uparrow$	Year	Source
Izadinia et al. (2012)	0.466	0.679	-	0.899	0.554	-	-	-	-	0.686	2012	Izadinia et al. (2012)
Shu et al. (2012)	0.522	0.672	-	0.674	0.536	-	-	-	-	0.597	2012	Shu et al. (2012)
Pirsiavash et al. (2011)	0.327	0.677	-	0.765	0.549	-	-	-	-	0.639	2011	Izadinia et al. (2012)
Yang et al. (2009b)	-	-	-	0.831	0.632	-	-	-	-	0.718	2009	Yang et al. (2009b)

Table 17: Benchmark results on the Town Center dataset

Reference	MOTA $\uparrow$	MOTP $\uparrow$	IDS $\downarrow$	Precision $\uparrow$	Recall $\uparrow$	MT $\uparrow$	PT	ML $\downarrow$	FM $\downarrow$	F1 $\uparrow$	Year	Source
Benfold and Reid (2011)	0.648	0.804	-	0.805	0.649	-	-	-	-	0.719	2011	Shu et al. (2012)
Zhang et al. (2008)	0.657	0.715	-	0.715	0.661	-	-	-	-	0.687	2008	Shu et al. (2012)
Leal-Taixé et al. (2011)	0.673	0.715	86	-	-	-	-	-	-	-	2011	Zhang et al. (2012)
Izadinia et al. (2012)	0.757	0.716	-	0.936	0.818	-	-	-	-	0.873	2012	Izadinia et al. (2012)
Zamir et al. (2012)	0.756	0.719	-	-	-	-	-	-	-	-	2012	Zamir et al. (2012)
McLaughlin et al. (2013)	0.742	0.724	-	0.904	0.833	-	-	-	-	0.867	2013	McLaughlin et al. (2013)
Chen et al. (2014)	-	-	36	-	-	0.855	0.086	0.059	26	-	2014	Chen et al. (2014)
Qin and Shelton (2012)	-	-	39	-	-	0.832	0.109	0.059	28	-	2012	Qin and Shelton (2012)
Pellegrini et al. (2010)	0.634	0.707	-	0.708	0.641	-	-	-	-	0.673	2010	Zamir et al. (2012)
Shu et al. (2012)	0.729	0.713	-	0.714	0.735	-	-	-	-	0.724	2012	Shu et al. (2012)
Yan et al. (2014)	-	-	19	-	-	0.856	0.096	0.048	43	-	2014	Yan et al. (2014)
Yamaguchi et al. (2011)	0.633	0.709	-	0.711	0.640	-	-	-	-	0.674	2011	Izadinia et al. (2012)
Zhang et al. (2012)	0.736	0.688	421	-	-	-	-	-	-	-	2012	Zhang et al. (2012)
Pellegrini et al. (2009)	0.634	0.707	-	0.708	0.641	-	-	-	-	0.673	2009	Izadinia et al. (2012)
Wu et al. (2013b)	0.695	0.687	209	-	-	0.647	0.274	0.079	453	-	2013	Possegger et al. (2014)
Possegger et al. (2014)	0.707	0.686	157	-	-	0.563	0.363	0.074	321	-	2014	Possegger et al. (2014)

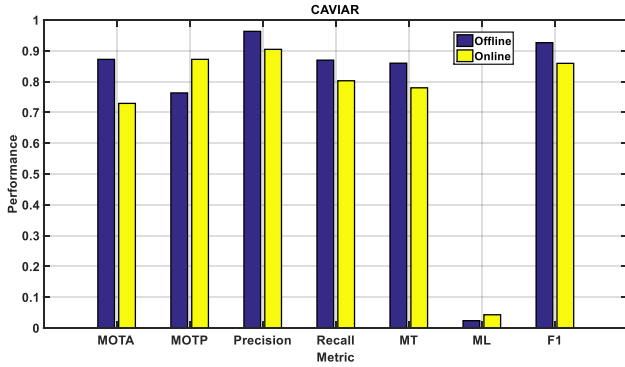


Fig. 14: Comparison between offline methods and online methods on the CAVIAR data set in terms of MOTA, MOTP, Precision, Recall, MT, ML and F1 metrics

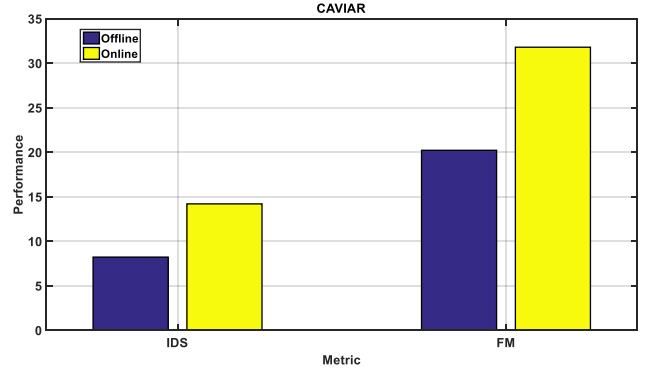


Fig. 15: Comparison between offline methods and online methods on the CAVIAR data set in terms of IDS and FM metrics

- *MOT with scene understanding.* Even more attention is paid to the over-crowded scenarios such as underground station, train station and so on. In this kind of scenario, most of the objects are very small and/or largely occluded. To this end, scene understanding, including crowd analysis (Rodriguez et al. 2011, 2009; Zhou et al. 2012b,a) for the objects themselves and scene structure recognition (exit, en-

trance, etc) (Zhou et al. 2011) is being investigated to help multiple object tracking in this scene.

- *MOT with other computer vision tasks.* Though multiple object tracking is in serve of other high-level computer vision tasks, there is a trend to solve multi-object tracking as a mid-level computer vision task and some other computer vision tasks at the same time as they are beneficial to each other. For example, Zhang et al. (2007) tackle multiple object track-

Table 18: Benchmark results on the i-LIDS dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Huang et al. (2008)	0.684	-	-	-	-	-	-	-	-	-	2008	Brendel et al. (2011)
Brendel et al. (2011)	0.786	0.700	1	-	-	-	-	-	-	-	2011	Brendel et al. (2011)
Benfold and Reid (2011)	0.599	0.736	-	0.803	0.820	-	-	-	-	0.811	2011	Benfold and Reid (2011)
Wu and Nevatia (2007)	0.553	-	-	-	-	-	-	-	-	-	2007	Brendel et al. (2011)
Breitenstein et al. (2009)	0.760	0.660	2	-	-	-	-	-	-	-	2009	Brendel et al. (2011)
Stalder et al. (2010)	-	-	-	0.894	0.533	-	-	-	-	0.668	2010	Benfold and Reid (2011)

Table 19: Benchmark results on the TRECVID2008 dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Kuo et al. (2010)	-	-	224	0.861	0.804	0.761	0.193	0.046	322	0.832	2010	Yang and Nevatia (2012a)
Huang et al. (2008)	-	-	278	0.808	0.716	0.570	0.281	0.149	487	0.759	2008	Kuo et al. (2010)
Li et al. (2009)	-	-	288	0.835	0.800	0.775	0.176	0.049	310	0.817	2009	Kuo et al. (2010)
Yang et al. (2011)	-	-	253	0.858	0.792	0.782	0.169	0.049	319	0.824	2011	Yang and Nevatia (2012a)
Kuo and Nevatia (2011)	-	-	171	0.868	0.792	0.770	0.177	0.052	283	0.828	2011	Yang and Nevatia (2012a)
Yang and Nevatia (2012a)	-	-	153	0.875	0.802	0.769	0.176	0.055	242	0.837	2012	Yang and Nevatia (2012a)
Yang and Nevatia (2012b)	-	-	147	0.878	0.798	0.755	0.187	0.058	240	0.836	2012	Yang and Nevatia (2012b)

Table 20: Benchmark results on the ETH Central dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Brendel et al. (2011)	0.742	0.720	0	-	-	-	-	-	-	-	2011	Brendel et al. (2011)
Breitenstein et al. (2009)	0.729	0.700	0	-	-	-	-	-	-	-	2009	Brendel et al. (2011)
Leibe et al. (2007)	0.338	0.660	5	-	-	-	-	-	-	-	2007	Brendel et al. (2011)
Breitenstein et al. (2011)	0.729	0.700	0	-	-	-	-	-	-	-	2011	Yan et al. (2012)
Yan et al. (2012)	0.754	0.715	0	-	-	-	-	-	-	-	2012	Yan et al. (2012)

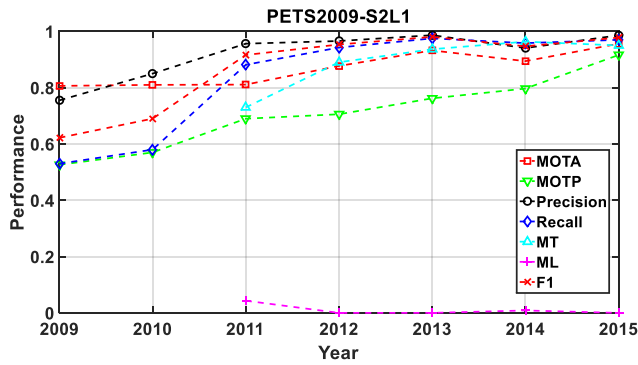


Fig. 16: Statistics of results in different years on the PETS2009-S2L1 data set in terms of MOTA, MOTP, Precision, Recall, MT, ML and F1 metrics

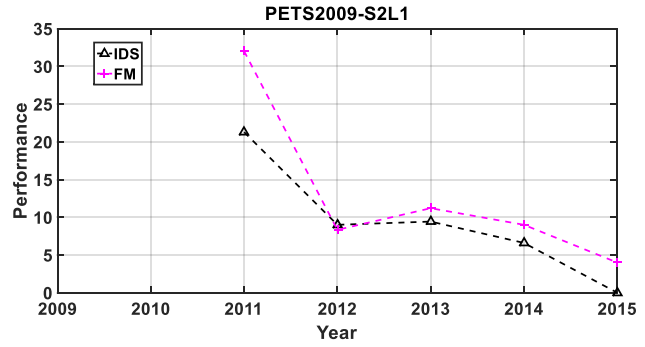


Fig. 17: Statistics of results in different years on the PETS2009-S2L1 data set in terms of IDS and FM metrics

ing and pose estimation simultaneously. Detection and tracking are combined and better results are achieved by Andriluka et al. (2008). Ishiguro et al. (2008) learn the dynamics among multiple objects and track them at the same time. Multiple object tracking is also addressed along with human body pose estimation by Gammeter et al. (2008). Another example is addressing MOT and action recognition in a unified framework (Choi and Savarese 2012).

## References

- Alahi A, Jacques L, Boursier Y, Vanderghenst P (2009) Sparsity-driven people localization algorithm: Evaluation in crowded scenes environments. In: Proc. IEEE Int. Conf. Advanced Video Signal-Based Surveillance, pp 1–8
- Ali S, Shah M (2008) Floor fields for tracking in high density crowd scenes. In: Proc. Eur. Conf. Comput. Vis., pp 1–14
- Andriluka M, Roth S, Schiele B (2008) People-tracking-by-detection and people-detection-by-tracking. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1–8
- Andriyenko A, Schindler K (2010) Globally optimal multi-target tracking on a hexagonal lattice. In: Proc. Eur. Conf. Comput. Vis., pp 466–479
- Andriyenko A, Schindler K (2011) Multi-target tracking by continuous energy minimization. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1265–1272
- Andriyenko A, Roth S, Schindler K (2011) An analytical formulation of global occlusion reasoning for multi-target tracking. In: Proc. IEEE Int. Conf. Comput. Vis. Workshops, pp 1839–1846
- Andriyenko A, Schindler K, Roth S (2012) Discrete-continuous optimization for multi-target tracking. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1926–1933
- Bae SH, Yoon KJ (2014) Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1218–1225

Table 21: Benchmark results on the Hockey dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Brendel et al. (2011)	0.797	0.600	0	-	-	-	-	-	-	-	2011	Brendel et al. (2011)
Breitenstein et al. (2009)	0.765	0.570	0	-	-	-	-	-	-	-	2009	Brendel et al. (2011)
Okuma et al. (2004)	0.678	0.510	11	-	-	-	-	-	-	-	2004	Brendel et al. (2011)
Breitenstein et al. (2011)	0.765	0.570	0	-	-	-	-	-	-	-	2011	Yan et al. (2012)
Yan et al. (2012)	0.918	0.716	0	-	-	-	-	-	-	-	2012	Yan et al. (2012)

Table 22: Benchmark results on the Parking Lot dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Shu et al. (2012)	0.741	0.793	-	0.913	0.817	-	-	-	-	0.862	2012	Wen et al. (2014)
Izadinia et al. (2012)	0.889	0.775	-	0.936	0.965	-	-	-	-	0.950	2012	Izadinia et al. (2012)
Zamir et al. (2012)	0.904	0.741	-	0.982	0.853	-	-	-	-	0.913	2012	Wen et al. (2014)
Tang et al. (2013)	0.893	0.777	-	-	-	-	-	-	-	-	2013	Dehghan et al. (2015)
Wen et al. (2014)	0.884	0.819	21	0.983	0.908	0.786	0.214	0.000	23	0.944	2014	Wen et al. (2014)
Dehghan et al. (2015)	0.907	0.693	3	-	0.860	0.140	0.000	-	-	-	2015	Dehghan et al. (2015)
Andriyenko and Schindler (2011)	0.600	0.707	68	0.913	0.693	0.214	0.714	0.071	97	0.788	2011	Wen et al. (2014)
Andriyenko et al. (2012)	0.731	0.765	83	0.894	0.978	0.786	0.214	0.000	70	0.934	2012	Wen et al. (2014)
Pirsiavash et al. (2011)	0.657	0.753	52	0.868	0.694	0.071	0.857	0.071	60	0.771	2011	Wen et al. (2014)
Tang et al. (2015)	0.938	0.783	1	0.970	0.969	0.929	0.071	0.000	6	0.969	2015	Tang et al. (2015)

Table 23: Benchmark results on the TUD Crossing dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Brendel et al. (2011)	0.859	0.730	2	-	-	-	-	-	-	-	2011	Brendel et al. (2011)
Zamir et al. (2012)	0.916	0.756	0	0.986	0.928	-	-	-	-	0.956	2012	Zamir et al. (2012)
Dehghan et al. (2015)	0.929	0.692	1	-	-	-	-	-	-	-	2015	Dehghan et al. (2015)
Pirsiavash et al. (2011)	0.655	0.768	50	0.950	0.739	0.462	0.538	0.000	42	0.831	2011	Tang et al. (2015)
Segal and Reid (2013)	0.740	0.760	2	-	-	-	-	-	12	-	2013	Tang et al. (2015)
Tang et al. (2013)	0.760	0.786	-	0.939	0.827	-	-	-	-	0.879	2013	Tang et al. (2015)
Tang et al. (2015)	0.809	0.780	1	0.988	0.820	0.615	0.231	0.154	1	0.896	2015	Tang et al. (2015)
Breitenstein et al. (2009)	0.843	0.710	2	0.851	0.986	-	-	-	-	0.914	2009	Zamir et al. (2012)
Breitenstein et al. (2011)	0.843	0.710	2	-	-	-	-	-	-	-	2011	Yan et al. (2012)
Yan et al. (2012)	0.894	0.708	2	-	-	-	-	-	-	-	2012	Yan et al. (2012)
Wu et al. (2013b)	0.906	0.769	8	-	-	0.846	0.154	0.000	5	-	2013	Wu et al. (2013b)
Zhang et al. (2012)	0.713	0.675	11	-	-	0.538	0.462	0.000	15	-	2012	Wu et al. (2013b)

Ben Shitrit H, Berclaz J, Fleuret F, Fua P (2011) Tracking multiple people under global appearance constraints. In: Proc. IEEE Int. Conf. Comput. Vis., pp 137–144

Benfold B, Reid I (2011) Stable multi-target tracking in real-time surveillance video. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 3457–3464

Berclaz J, Fleuret F, Fua P (2006) Robust people tracking with global trajectory optimization. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 744–750

Berclaz J, Fleuret F, Fua P (2009) Multiple object tracking using flow linear programming. In: Proc. IEEE Int. Workshop Perform. Eval. Track. Surveillance, pp 1–8

Berclaz J, Fleuret F, Turetken E, Fua P (2011) Multiple object tracking using k-shortest paths optimization. IEEE Trans Pattern Anal Mach Intel 33(9):1806–1819

Betke M, Haritaoglu E, Davis LS (2000) Real-time multiple vehicle detection and tracking from a moving vehicle. Mach Vis Appl 12(2):69–83

Betke M, Hirsh DE, Bagchi A, Hristov NI, Makris NC, Kunz TH (2007) Tracking large variable numbers of objects in clutter. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1–8

Bibby C, Reid I (2008) Robust real-time visual tracking using pixel-wise posteriors. In: Proc. Eur. Conf. Comput. Vis., pp 831–844

Bose B, Wang X, Grimson E (2007) Multi-class object tracking algorithm that handles fragmentation and grouping. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1–8

Breitenstein MD, Reichlin F, Leibe B, Koller-Meier E, Van Gool L (2009) Robust tracking-by-detection using a detector confidence particle filter. In: Proc. IEEE Int. Conf. Comput. Vis., pp 1515–1522

Breitenstein MD, Reichlin F, Leibe B, Koller-Meier E, Van Gool L (2011) Online multiperson tracking-by-detection from a single, uncalibrated camera. IEEE Trans

Pattern Anal Mach Intel 33(9):1820–1833

Brendel W, Amer M, Todorovic S (2011) Multiobject tracking as maximum weight independent set. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1273–1280

Brostow G, Cipolla R (2006) Unsupervised bayesian detection of independent motion in crowds. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 594–601

Butt A, Collins R (2013) Multi-target tracking by lagrangian relaxation to min-cost network flow. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1846–1853

Candamo J, Shreve M, Goldgof DB, Sapper DB, Kasturi R (2010) Understanding transit scenes: A survey on human behavior-recognition algorithms. IEEE Trans Intell Transp Syst 11(1):206–224

Cannons K (1991) A review of visual tracking. Tech. Rep. CSE-2008-07, Dept. Comput. Sci. Eng., York Univ.

Caruana R (1997) Multitask learning. Mach Learn 28(1):41–75

Chang TH, Gong S, Ong EJ (2000) Tracking multiple people under occlusion using multiple cameras. In: Proc. Brit. Mach. Vis. Conf., pp 1–10

Chen X, Qin Z, An L, Bhanu B (2014) An online learned elementary grouping model for multi-target tracking. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1242–1249

Choi W, Savarese S (2010) Multiple target tracking in world coordinate with single, minimally calibrated camera. In: Proc. Eur. Conf. Comput. Vis., pp 553–567

Choi W, Savarese S (2012) A unified framework for multi-target tracking and collective activity recognition. In: Proc. Eur. Conf. Comput. Vis., pp 215–230

Choi W, Pantofaru C, Savarese S (2013) A general framework for tracking multiple people from a moving camera. IEEE Trans Pattern Anal Mach Intel 35(7):1577–1591

Conte D, Foggia P, Percannella G, Vento M (2010) Performance evaluation of a people tracking system on pets2009

Table 24: Benchmark results on the TUD Stadtmitte dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Zamir et al. (2012)	0.777	0.634	0	0.956	0.814	-	-	-	-	0.879	2012	Zamir et al. (2012)
Andriyenko and Schindler (2011)	0.605	0.658	7	-	-	0.670	0.330	0.000	4	-	2011	Andriyenko and Schindler (2011)
Kuo and Nevatia (2011)	-	-	1	0.995	0.810	0.600	0.300	0.100	0	0.893	2011	Yang and Nevatia (2012b)
Yang and Nevatia (2012b)	-	-	0	0.967	0.870	0.700	0.300	0.000	1	0.916	2012	Yang and Nevatia (2012b)
Zhang et al. (2015)	0.842	0.865	1	0.981	0.858	0.800	0.200	0.000	2	0.915	2015	Zhang et al. (2015)
Milan et al. (2014)	0.711	0.655	4	0.867	0.847	0.778	0.222	0.000	3	0.857	2014	Milan et al. (2014)
Berclaz et al. (2011)	0.458	0.567	5	0.792	0.631	0.111	0.778	0.111	15	0.702	2011	Berclaz et al. (2011)
Milan et al. (2013)	0.562	0.616	15	-	-	0.444	0.555	0.000	13	-	2013	Milan et al. (2013)
Andriyenko et al. (2012)	0.618	0.632	4	-	-	0.600	0.400	0.000	1	-	2012	Andriyenko et al. (2012)
Andriyenko et al. (2011)	0.686	0.640	-	-	-	0.556	0.000	0.444	-	-	2011	Hofmann et al. (2013a)
Hofmann et al. (2013a)	0.724	0.720	8	0.835	0.928	0.900	0.000	0.100	10	0.879	2013	Hofmann et al. (2013a)
Pirsiavash et al. (2011)	0.759	0.826	8	0.965	0.838	0.800	0.200	0.000	10	0.897	2011	Shi et al. (2014)
Shi et al. (2013)	0.804	0.877	3	0.988	0.839	0.700	0.300	0.000	10	0.907	2013	Shi et al. (2014)
Shi et al. (2014)	0.825	0.893	0	0.999	0.840	0.800	0.200	0.000	1	0.913	2014	Shi et al. (2014)
Segal and Reid (2013)	0.730	0.710	2	-	-	-	-	-	1	-	2013	Segal and Reid (2013)
Yan et al. (2014)	-	-	1	-	-	0.700	0.300	0.000	2	-	2014	Yan et al. (2014)
Wu et al. (2013b)	0.754	0.7	3	-	-	0.900	0.100	0.000	2	-	2013	Wu et al. (2013b)
Zhang et al. (2012)	0.75	0.598	2	-	-	0.600	0.400	0.000	1	-	2012	Zhang et al. (2012)

Table 25: Benchmark results on the ETHMS dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Kuo and Nevatia (2011)	-	-	11	0.866	0.768	0.584	0.336	0.080	23	0.814	2011	Milan et al. (2013)
Yang and Nevatia (2012b)	-	-	11	0.904	0.790	0.680	0.248	0.072	19	0.843	2012	Milan et al. (2013)
Pirsiavash et al. (2011)	-	-	4	0.914	0.674	0.502	0.399	0.099	143	0.776	2011	Milan et al. (2013)
Milan et al. (2013)	-	-	57	0.872	0.773	0.664	0.254	0.082	69	0.820	2013	Milan et al. (2013)
Poiesi et al. (2013)	-	-	45	0.855	0.787	0.624	0.296	0.080	69	0.820	2013	Poiesi et al. (2013)
Bae and Yoon (2014)	0.720	0.640	18	0.958	0.764	0.738	0.238	0.024	38	0.850	2014	Bae and Yoon (2014)

Table 26: Benchmark results on the TUD Campus dataset

Reference	MOTA ↑	MOTP ↑	IDS ↓	Precision ↑	Recall ↑	MT ↑	PT	ML ↓	FM ↓	F1 ↑	Year	Source
Pirsiavash et al. (2011)	0.606	0.782	10	0.955	0.666	0.375	0.500	0.125	13	0.785	2011	Tang et al. (2015)
Segal and Reid (2013)	0.820	0.740	0	-	-	0.625	-	-	3	-	2013	Tang et al. (2015)
Tang et al. (2015)	0.833	0.769	0	0.993	0.838	0.625	0.250	0.125	1	0.909	2015	Tang et al. (2015)
Breitenstein et al. (2011)	0.733	0.670	2	-	-	-	-	-	-	-	2011	Tang et al. (2015)
Yan et al. (2012)	0.848	0.678	0	1.000	0.848	-	-	-	-	0.918	2012	Yan et al. (2012)
Wu et al. (2013b)	0.685	0.713	5	-	-	0.500	0.500	0.000	5	-	2013	Wu et al. (2013b)
Zhang et al. (2012)	0.747	0.680	3	-	-	0.667	0.333	0.000	4	-	2012	Wu et al. (2013b)

- database. In: Proc. IEEE Int. Conf. Advanced Video Signal-Based Surveillance, pp 119–126
- Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 886–893
- Dehghan A, Tian Y, Torr PH, Shah M (2015) Target identity-aware network flow for online multiple target tracking. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1146–1154
- Dicle C, Sznaiar M, Camps O (2013) The way they move: Tracking multiple targets with similar appearance. In: Proc. IEEE Int. Conf. Comput. Vis., pp 2304–2311
- Ess A, Leibe B, Van Gool L (2007) Depth and appearance for mobile scene analysis. In: Proc. IEEE Int. Conf. Comput. Vis., pp 1–8
- Ess A, Leibe B, Schindler K, Van Gool L (2008) A mobile vision system for robust multi-person tracking. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1–8
- Ess A, Leibe B, Schindler K, Van Gool L (2009) Robust multi-person tracking from a mobile platform. IEEE Trans Pattern Anal Mach Intel 31(10):1831–1846
- Evgeniou T, Pontil M (2004) Regularized multi-task learning. In: Proc. ACM Int. Conf. Knowl. Discov. Data Min, pp 109–117
- Felzenszwalb P, Girshick R, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intel 32(9):1627–1645
- Felzenszwalb PF, Huttenlocher DP (2006) Efficient belief propagation for early vis. Int J Comput Vis 70(1):41–54
- Fleuret F, Berclaz J, Lengagne R, Fua P (2008) Multicamera people tracking with a probabilistic occupancy map. IEEE Trans Pattern Anal Mach Intel 30(2):267–282
- Fontaine E, Barr AH, Burdick JW (2007) Model-based tracking of multiple worms and fish. In: Proc. IEEE Int. Conf. Comput. Vis. Workshops, pp 1–13
- Forsyth DA, Arikan O, Ikemoto L, O’Brien J, Ramanan D, et al. (2006) Computational studies of human motion: Part 1, tracking and motion synthesis. Found Trends Comput Graph Vis 1(2-3):77–254
- Fortmann TE, Bar-Shalom Y, Scheffe M (1983) Sonar tracking of multiple targets using joint probabilistic data association. IEEE J Ocean Eng 8(3):173–184
- Fragkiadaki K, Zhang W, Zhang G, Shi J (2012) Two-granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions. In: Proc. Eur. Conf. Comput. Vis., pp 552–565
- Gammeter S, Ess A, Jäggli T, Schindler K, Leibe B, Van Gool L (2008) Articulated multi-body tracking under egomotion. In: Proc. Eur. Conf. Comput. Vis., pp 816–830
- Gavrila DM, Munder S (2007) Multi-cue pedestrian detection and tracking from a moving vehicle. Int J Comput Vis 73(1):41–59
- Giebel J, Gavrila DM, Schnörr C (2004) A bayesian framework for multi-cue 3d object tracking. In: Proc. Eur. Conf. Comput. Vis., pp 241–252
- Han B, Joo SW, Davis LS (2007) Probabilistic fusion tracking using mixture kernel-based bayesian filtering. In: Proc. IEEE Int. Conf. Comput. Vis., pp 1–8
- Helbing D, Molnar P (1995) Social force model for pedestrian dynamics. Phys Rev E 51(5):4282–4286
- Henriques JF, Caseiro R, Batista J (2011) Globally optimal solution to multi-object tracking with merged measurements. In: Proc. IEEE Int. Conf. Comput. Vis., pp 2470–2477
- Hess R, Fern A (2009) Discriminatively trained particle filters for complex multi-object tracking. In: Proc. IEEE Int.



Table 27: Benchmark result comparison between offline and online methods on the Town Center data set

Method	MOTA $\uparrow$	MOTP $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	MT $\uparrow$	ML $\downarrow$	F1 $\uparrow$	IDS $\downarrow$	FM $\downarrow$
Offline	0.700 $\pm$ 0.052	0.727 $\pm$ 0.032	0.797 $\pm$ 0.102	0.723 $\pm$ 0.086	0.848 $\pm$ 0.014	0.055 $\pm$ 0.006	0.757 $\pm$ 0.089	45.000 $\pm$ 28.717	32.333 $\pm$ 9.292
Online	0.681 $\pm$ 0.046	0.695 $\pm$ 0.012	0.710 $\pm$ 0.002	0.641 $\pm$ 0.001	0.605 $\pm$ 0.059	0.076 $\pm$ 0.004	0.673 $\pm$ 0.001	262.333 $\pm$ 139.848	387.000 $\pm$ 93.338

Table 28: Benchmark result comparison between offline and online methods on the CAVIAR data set

Method	MOTA $\uparrow$	MOTP $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	MT $\uparrow$	ML $\downarrow$	F1 $\uparrow$	IDS $\downarrow$	FM $\downarrow$
Offline	0.872 $\pm$ 0.000	0.763 $\pm$ 0.000	0.963 $\pm$ 0.014	0.869 $\pm$ 0.048	0.860 $\pm$ 0.036	0.023 $\pm$ 0.020	0.926 $\pm$ 0.008	8.222 $\pm$ 4.024	20.222 $\pm$ 14.864
Online	0.729 $\pm$ 0.192	0.872 $\pm$ 0.000	0.904 $\pm$ 0.121	0.802 $\pm$ 0.064	0.780 $\pm$ 0.095	0.043 $\pm$ 0.027	0.859 $\pm$ 0.107	14.200 $\pm$ 3.271	31.800 $\pm$ 17.908

Table 29: Benchmark result along years on the PETS2009-S2L1 dataset

Year	MOTA $\uparrow$	MOTP $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	MT $\uparrow$	ML $\downarrow$	F1 $\uparrow$	IDS $\downarrow$	FM $\downarrow$
2009	0.806 $\pm$ 0.041	0.526 $\pm$ 0.010	0.755 $\pm$ 0.092	0.530 $\pm$ 0.000	-	-	0.622 $\pm$ 0.031	-	-
2010	0.810 $\pm$ 0.000	0.570 $\pm$ 0.000	0.850 $\pm$ 0.000	0.580 $\pm$ 0.000	-	-	0.690 $\pm$ 0.000	-	-
2011	0.811 $\pm$ 0.073	0.690 $\pm$ 0.085	0.956 $\pm$ 0.034	0.881 $\pm$ 0.047	0.730 $\pm$ 0.084	0.043 $\pm$ 0.036	0.916 $\pm$ 0.023	21.286 $\pm$ 19.465	32.000 $\pm$ 20.017
2012	0.877 $\pm$ 0.068	0.706 $\pm$ 0.076	0.966 $\pm$ 0.024	0.943 $\pm$ 0.033	0.890 $\pm$ 0.062	0.000 $\pm$ 0.000	0.953 $\pm$ 0.009	9.000 $\pm$ 9.274	8.333 $\pm$ 6.028
2013	0.932 $\pm$ 0.034	0.762 $\pm$ 0.045	0.986 $\pm$ 0.006	0.975 $\pm$ 0.021	0.937 $\pm$ 0.087	0.000 $\pm$ 0.000	0.980 $\pm$ 0.013	9.429 $\pm$ 5.769	11.200 $\pm$ 2.490
2014	0.921 $\pm$ 0.059	0.770 $\pm$ 0.054	0.986 $\pm$ 0.003	0.948 $\pm$ 0.027	0.963 $\pm$ 0.037	0.009 $\pm$ 0.019	0.967 $\pm$ 0.015	6.600 $\pm$ 3.209	9.000 $\pm$ 5.292
2015	0.930 $\pm$ 0.037	0.774 $\pm$ 0.202	0.986 $\pm$ 0.000	0.970 $\pm$ 0.000	0.950 $\pm$ 0.000	0.000 $\pm$ 0.000	0.978 $\pm$ 0.000	1.500 $\pm$ 2.121	4.000 $\pm$ 0.000

- Conf. Comput. Vis. Pattern Recognit., pp 240–247
- Hofmann M, Haag M, Rigoll G (2013a) Unified hierarchical multi-object tracking using global data association. In: Proc. IEEE Int. Conf. Advanced Video Signal-Based Surveillance, pp 22–28
- Hofmann M, Wolf D, Rigoll G (2013b) Hypergraphs for joint multi-view reconstruction and multi-object tracking. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 3650–3657
- Hu M, Ali S, Shah M (2008) Detecting global motion patterns in complex videos. In: Proc. IEEE Int. Conf. Pattern. Recognit., pp 1–5
- Hu W, Tan T, Wang L, Maybank S (2004) A survey on visual surveillance of object motion and behaviors. IEEE Trans Syst Man Cybern Part C-Appl Rev 34(3):334–352
- Hu W, Li X, Luo W, Zhang X, Maybank S, Zhang Z (2012) Single and multiple object tracking using log-euclidean riemannian subspace and block-division appearance model. IEEE Trans Pattern Anal Mach Intel 34(12):2420–2440
- Huang C, Wu B, Nevatia R (2008) Robust object tracking by hierarchical association of detection responses. In: Proc. Eur. Conf. Comput. Vis., pp 788–801
- Ishiguro K, Yamada T, Ueda N (2008) Simultaneous clustering and tracking unknown number of objects. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1–8
- Izadinia H, Saleemi I, Li W, Shah M (2012) (mp)2t: Multiple people multiple parts tracker. In: Proc. Eur. Conf. Comput. Vis., pp 100–114
- Jiang H, Fels S, Little JJ (2007) A linear programming approach for multiple object tracking. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1–8
- Jin Y, Mokhtarian F (2007) Variational particle filter for multi-object tracking. In: Proc. IEEE Int. Conf. Comput. Vis., pp 1–8
- Kalal Z, Mikolajczyk K, Matas J (2012) Tracking-learning-detection. IEEE Trans Pattern Anal Mach Intel 34(7):1409–1422
- Kasturi R, Goldgof D, Soundararajan P, Manohar V, Garofolo J, Bowers J, Boonstra M, Korzhova V, Zhang J (2009) Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. IEEE Trans Pattern Anal Mach Intel 31(2):319–336
- KC AK, De Vleeschouwer C (2013) Discriminative label propagation for multi-object tracking with sporadic appearance features. In: Proc. IEEE Int. Conf. Comput. Vis., pp 2000–2007
- Keni B, Rainer S (2008) Evaluating multiple object tracking performance: the clear mot metrics. EURASIP J Image Video Process
- Khan Z, Balch T, Dellaert F (2004) An mcmc-based particle filter for tracking multiple interacting targets. In: Proc. Eur. Conf. Comput. Vis., pp 279–290
- Khan Z, Balch T, Dellaert F (2005) Mcmc-based particle filtering for tracking a variable number of interacting targets. IEEE Trans Pattern Anal Mach Intel 27(11):1805–1819
- Khan Z, Balch T, Dellaert F (2006) Mcmc data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements. IEEE Trans Pattern Anal Mach Intel 28(12):1960–1972
- Kim IS, Choi HS, Yi KM, Choi JY, Kong SG (2010) Intelligent visual surveillance-a survey. Int J Control Autom Syst 8(5):926–939
- Koller D, Weber J, Malik J (1994) Robust multiple car tracking with occlusion reasoning. In: Proc. Eur. Conf. Comput. Vis., pp 189–196
- Kratz L, Nishino K (2010) Tracking with local spatio-temporal motion patterns in extremely crowded scenes. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 693–700
- Kratz L, Nishino K (2012) Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes. IEEE Trans Pattern Anal Mach Intel 34(5):987–1002
- Kuo CH, Nevatia R (2011) How does person identity recognition help multi-person tracking? In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1217–1224
- Kuo CH, Huang C, Nevatia R (2010) Multi-target tracking by on-line learned discriminative appearance models. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 685–692
- Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 2169–2178
- Leal-Taixé L, Pons-Moll G, Rosenhahn B (2011) Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In: Proc. IEEE Int. Conf. Comput. Vis. Workshops, pp 120–127

- Leal-Taixé L, Pons-Moll G, Rosenhahn B (2012) Branch-and-price global optimization for multi-view multi-target tracking. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, IEEE, pp 1987–1994
- Leal-Taixé L, Milan A, Reid I, Roth S, Schindler K (2015) Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:150401942*
- Leibe B, Schindler K, Van Gool L (2007) Coupled detection and trajectory estimation for multi-object tracking. In: *Proc. IEEE Int. Conf. Comput. Vis.*, pp 1–8
- Leibe B, Schindler K, Cornelis N, Van Gool L (2008) Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Trans Pattern Anal Mach Intel* 30(10):1683–1698
- Li K, Miller ED, Chen M, Kanade T, Weiss LE, Campbell PG (2008) Cell population tracking and lineage construction with spatiotemporal context. *Med Image Anal* 12(5):546–566
- Li X, Hu W, Shen C, Zhang Z, Dick A, Hengel AVD (2013) A survey of appearance models in visual object tracking. *ACM Trans Intell Syst Technol* 4(4):58
- Li Y, Huang C, Nevatia R (2009) Learning to associate: Hybridboosted multi-target tracker for crowded scene. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 2953–2960
- Liu Y, Li H, Chen YQ (2012) Automatic tracking of a large number of moving targets in 3d. In: *Proc. Eur. Conf. Comput. Vis.*, pp 730–742
- Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 60(2):91–110
- Lu WL, Ting JA, Little JJ, Murphy KP (2013) Learning to track and identify players from broadcast sports videos. *IEEE Trans Pattern Anal Mach Intel* 35(7):1704–1716
- Luo W, Kim TK (2013) Generic object crowd tracking by multi-task learning. In: *Proc. Brit. Mach. Vis. Conf.*, pp 73.1–73.13
- Luo W, Kim TK, Stenger B, Zhao X, Cipolla R (2014) Bi-label propagation for generic multiple object tracking. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 1290–1297
- Luo W, Stenger B, Zhao X, Kim TK (2015) Automatic topic discovery for multi-object tracking. In: *Proc. AAAI Conf. Artif. Intell.*, pp 3820–3826
- McLaughlin N, Martinez Del Rincon J, Miller P (2013) On-line multiperson tracking with occlusion reasoning and unsupervised track motion model. In: *Proc. IEEE Int. Workshop Perform. Eva. Track. Surveillance*, pp 37–42
- Meijering E, Dzyubachyk O, Smal I, van Cappellen WA (2009) Tracking in cell and developmental biology. *Semin Cell Dev Biol* 20(8):894–902
- Milan A, Schindler K, Roth S (2013) Detection- and trajectory-level exclusion in multiple object tracking. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 3682–3689
- Milan A, Roth S, Schindler K (2014) Continuous energy minimization for multitarget tracking. *IEEE Trans Pattern Anal Mach Intel* 36(1):58–72
- Mitzel D, Leibe B (2011) Real-time multi-person tracking with detector assisted structure propagation. In: *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, pp 974–981
- Mitzel D, Horbert E, Ess A, Leibe B (2010) Multi-person tracking with sparse detection and continuous segmentation. In: *Proc. Eur. Conf. Comput. Vis.*, pp 397–410
- Mordohai P, Medioni G (2010) Dimensionality estimation, manifold learning and function approximation using tensor voting. *J Mach Learn Res* 11:411–450
- Nillius P, Sullivan J, Carlsson S (2006) Multi-target tracking-linking identities using bayesian network inference. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 2187–2194
- Okuma K, Taleghani A, De Freitas N, Little JJ, Lowe DG (2004) A boosted particle filter: Multitarget detection and tracking. In: *Proc. Eur. Conf. Comput. Vis.*, pp 28–39
- Park Y, Lepetit V, Woo W (2008) Multiple 3d object tracking for augmented reality. In: *Proc. IEEE/ACM Int. Symp. Mix. Augment. Real.*, pp 117–120
- Pellegrini S, Ess A, Schindler K, Van Gool L (2009) You'll never walk alone: Modeling social behavior for multi-target tracking. In: *Proc. IEEE Int. Conf. Comput. Vis.*, pp 261–268
- Pellegrini S, Ess A, Van Gool L (2010) Improving data association by joint modeling of pedestrian trajectories and groupings. In: *Proc. Eur. Conf. Comput. Vis.*, pp 452–465
- Perera AA, Srinivas C, Hoogs A, Brooksby G, Hu W (2006) Multi-object tracking through simultaneous long occlusions and split-merge conditions. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 666–673
- Pérez P, Hue C, Vermaak J, Gangnet M (2002) Color-based probabilistic tracking. In: *Proc. Eur. Conf. Comput. Vis.*, pp 661–675
- Pirsiavash H, Ramanan D, Fowlkes CC (2011) Globally-optimal greedy algorithms for tracking a variable number of objects. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 1201–1208
- Poiesi F, Mazzon R, Cavallaro A (2013) Multi-target tracking on confidence maps: An application to people tracking. *Comput Vis Image Underst* 117(10):1257–1272
- Porikli F, Tuzel O, Meer P (2006) Covariance tracking using model update based on lie algebra. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 728–735
- Possegger H, Mauthner T, Roth PM, Bischof H (2014) Occlusion geodesics for online multi-object tracking. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 1306–1313
- Qin Z, Shelton CR (2012) Improving multi-target tracking via social grouping. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 1972–1978
- Reid DB (1979) An algorithm for tracking multiple targets. *IEEE Trans Autom Control* 24(6):843–854
- Reilly V, Idrees H, Shah M (2010) Detection and tracking of large number of targets in wide area surveillance. In: *Proc. Eur. Conf. Comput. Vis.*, pp 186–199
- Rodriguez M, Ali S, Kanade T (2009) Tracking in unstructured crowded scenes. In: *Proc. IEEE Int. Conf. Comput. Vis.*, pp 1389–1396
- Rodriguez M, Sivic J, Laptev I, Audibert JY (2011) Data-driven crowd analysis in videos. In: *Proc. IEEE Int. Conf. Comput. Vis.*, pp 1235–1242
- Ryoo MS, Aggarwal JK (2008) Observe-and-explain: A new approach for multiple hypotheses tracking of humans and objects. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 1–8
- Scovanner P, Tappen MF (2009) Learning pedestrian dynamics from the real world. In: *Proc. IEEE Int. Conf. Comput. Vis.*, pp 381–388
- Segal AV, Reid I (2013) Latent data association: Bayesian model selection for multi-target tracking. In: *Proc. IEEE Int. Conf. Comput. Vis.*, pp 2904–2911
- Shafique K, Lee MW, Haering N (2008) A rank constrained continuous formulation of multi-frame multi-target tracking problem. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 1–8

- Shi J, Tomasi C (1994) Good features to track. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 593–600
- Shi X, Ling H, Xing J, Hu W (2013) Multi-target tracking by rank-1 tensor approximation. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 2387–2394
- Shi X, Ling H, Hu W, Yuan C, Xing J (2014) Multi-target tracking with motion context in tensor power iteration. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 3518–3525
- Shu G, Dehghan A, Oreifej O, Hand E, Shah M (2012) Part-based multiple-person tracking with partial occlusion handling. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1815–1821
- Shu G, Dehghan A, Shah M (2013) Improving an object detector and extracting regions using superpixels. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 3721–3727
- Song B, Jeng TY, Staudt E, Roy-Chowdhury AK (2010) A stochastic graph evolution framework for robust multi-target tracking. In: Proc. Eur. Conf. Comput. Vis., pp 605–619
- Spampinato C, Chen-Burger YH, Nadarajan G, Fisher RB (2008) Detecting, tracking and counting fish in low quality unconstrained underwater videos. Proc Int Conf Comput Vis Theory Appl 2008:514–519
- Spampinato C, Palazzo S, Giordano D, Kvasidis I, Lin FP, Lin YT (2012) Covariance based fish tracking in real-life underwater environment. In: Proc. Int. Conf. Comput. Vis. Theory Appl., pp 409–414
- Stalder S, Grabner H, Van Gool L (2010) Cascaded confidence filtering for improved tracking-by-detection. In: Proc. Eur. Conf. Comput. Vis., pp 369–382
- Sugimura D, Kitani KM, Okabe T, Sato Y, Sugimoto A (2009) Using individuality to track individuals: clustering individual trajectories in crowds using local appearance and frequency trait. In: Proc. IEEE Int. Conf. Comput. Vis., pp 1467–1474
- Sun Z, Bebis G, Miller R (2006) On-road vehicle detection: A review. IEEE Trans Pattern Anal Mach Intel 28(5):694–711
- Takala V, Pietikainen M (2007) Multi-object tracking using color, texture and motion. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.
- Tang S, Andriluka M, Milan A, Schindler K, Roth S, Schiele B (2013) Learning people detectors for tracking in crowded scenes. In: Proc. IEEE Int. Conf. Comput. Vis., pp 1049–1056
- Tang S, Andriluka M, Schiele B (2014) Detection and tracking of occluded people. Int J Comput Vis 110(1):58–69
- Tang S, Andres B, Andriluka M, Schiele B (2015) Subgraph decomposition for multi-target tracking. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 5033–5041
- Tomasi C, Kanade T (1991) Detection and tracking of point features. Tech. Rep. CMU-CS-91-132, School of Computer Science, Carnegie Mellon Univ.
- Tuzel O, Porikli F, Meer P (2006) Region covariance: A fast descriptor for detection and classification. In: Proc. Eur. Conf. Comput. Vis., pp 589–600
- Wang X (2013) Intelligent multi-camera video surveillance: A review. Pattern Recognit Lett 34(1):3–19
- Wen L, Li W, Yan J, Lei Z, Yi D, Li SZ (2014) Multiple target tracking based on undirected hierarchical relation hypergraph. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1282–1289
- Wolf JK, Viterbi AM, Dixon GS (1989) Finding the best set of  $k$  paths through a trellis with application to multitarget tracking. IEEE Trans Aerosp Electron Syst 25(2):287–296
- Wu B, Nevatia R (2006) Tracking of multiple, partially occluded humans based on static body part detection. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 951–958
- Wu B, Nevatia R (2007) Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. Int J Comput Vis 75(2):247–266
- Wu Y, Lim J, Yang MH (2013a) Online object tracking: A benchmark. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 2411–2418
- Wu Z, Kunz TH, Betke M (2011) Efficient track linking methods for track graphs using network-flow and set-cover techniques. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1185–1192
- Wu Z, Thangali A, Sclaroff S, Betke M (2012) Coupling detection and data association for multiple object tracking. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1948–1955
- Wu Z, Zhang J, Betke M (2013b) Online motion agreement tracking. In: Proc. Brit. Mach. Vis. Conf., pp 63.1–63.10
- Xing J, Ai H, Lao S (2009) Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1200–1207
- Xing J, Ai H, Liu L, Lao S (2011) Multiple player tracking in sports video: a dual-mode two-way bayesian inference approach with progressive observation modeling. IEEE Tran Image Process 20(6):1652–1667
- Yamaguchi K, Berg AC, Ortiz LE, Berg TL (2011) Who are you with and where are you going? In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1345–1352
- Yan X, Wu X, Kakadiaris IA, Shah SK (2012) To track or to detect? an ensemble framework for optimal selection. In: Proc. Eur. Conf. Comput. Vis., pp 594–607
- Yan X, Kakadiaris IA, Shah SK (2014) What do i see? modeling human visual perception for multi-person tracking. In: Proc. Eur. Conf. Comput. Vis., pp 314–329
- Yang B, Nevatia R (2012a) Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1918–1925
- Yang B, Nevatia R (2012b) An online learned crf model for multi-target tracking. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 2034–2041
- Yang B, Nevatia R (2012c) Online learned discriminative part-based appearance models for multi-human tracking. In: Proc. Eur. Conf. Comput. Vis., pp 484–498
- Yang B, Huang C, Nevatia R (2011) Learning affinities and dependencies for multi-target tracking using a CRF model. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp 1233–1240
- Yang C, Duraiswami R, Davis L (2005) Fast multiple object tracking via a hierarchical particle filter. In: Proc. IEEE Int. Conf. Comput. Vis., pp 212–219
- Yang J, Vela PA, Shi Z, Teizer J (2009a) Probabilistic multiple people tracking through complex situations. In: Proc. IEEE Int. Workshop Perform. Eva. Track. Surveillance, pp xxx–xxx
- Yang M, Yu T, Wu Y (2007) Game-theoretic multiple target tracking. In: Proc. IEEE Int. Conf. Comput. Vis., pp 1–8
- Yang M, Lv F, Xu W, Gong Y (2009b) Detection driven adaptive multi-cue integration for multiple human tracking. In: Proc. IEEE Int. Conf. Comput. Vis., pp 1554–1561

- Yilmaz A, Javed O, Shah M (2006) Object tracking: A survey. *ACM Comput Surv* 38(4):13
- Yu Q, Medioni G, Cohen I (2007) Multiple target tracking using spatio-temporal markov chain monte carlo data association. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 1–8
- Yu T, Wu Y, Krahnstoever NO, Tu PH (2008) Distributed data association and filtering for multiple target tracking. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 1–8
- Zamir AR, Dehghan A, Shah M (2012) Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In: *Proc. Eur. Conf. Comput. Vis.*, pp 343–356
- Zhan B, Monekosso DN, Remagnino P, Velastin SA, Xu LQ (2008) Crowd analysis: a survey. *Mach Vis Appl* 19(5):345–357
- Zhang J, Presti LL, Sclaroff S (2012) Online multi-person tracking by tracker hierarchy. In: *Proc. IEEE Int. Conf. Advanced Video Signal-Based Surveillance*, pp 379–385
- Zhang L, van der Maaten L (2013) Structure preserving object tracking. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 1838–1845
- Zhang L, van der Maaten L (2014) Preserving structure in model-free tracking. *IEEE Trans Pattern Anal Mach Intel* 36(4):756–769
- Zhang L, Wu B, Nevatia R (2007) Detection and tracking of multiple humans with extensive pose articulation. In: *Proc. IEEE Int. Conf. Comput. Vis.*, pp 1–8
- Zhang L, Li Y, Nevatia R (2008) Global data association for multi-object tracking using network flows. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 1–8
- Zhang S, Wang J, Wang Z, Gong Y, Liu Y (2015) Multi-target tracking by learning local-to-global trajectory models. *Pattern Recognit* 48(2):580–590
- Zhao T, Nevatia R (2004) Tracking multiple humans in crowded environment. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 406–413
- Zhao X, Gong D, Medioni G (2012) Tracking using motion patterns for very crowded scenes. In: *Proc. Eur. Conf. Comput. Vis.*, pp 315–328
- Zhou B, Wang X, Tang X (2011) Random field topic model for semantic region analysis in crowded scenes from tracklets. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 3441–3448
- Zhou B, Tang X, Wang X (2012a) Coherent filtering: detecting coherent motions from crowd clutters. In: *Proc. Eur. Conf. Comput. Vis.*, pp 857–871
- Zhou B, Wang X, Tang X (2012b) Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In: *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp 2871–2878