Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking

Ergys Ristani[1], Francesco Solera[2], Roger S. Zou[1],
Rita Cucchiara[2], and Carlo Tomasi[1]

[1] Computer Science Department, Duke University, Durham, USA
[2] Department of Engineering, University of Modena and Reggio Emilia, Modena, Italy

Abstract. To help accelerate progress in multi-target, multi-camera tracking systems, we present (i) a new pair of precision-recall measures of performance that treats errors of all types uniformly and emphasizes correct identification over sources of error; (ii) the largest fully-annotated and calibrated data set to date with more than 2 million frames of 1080p, 60fps video taken by 8 cameras observing more than 2,700 identities over 85 minutes; and (iii) a reference software system as a comparison baseline. We show that (i) our measures properly account for bottom-line identity match performance in the multi-camera setting; (ii) our data set poses realistic challenges to current trackers; and (iii) the performance of our system is comparable to the state of the art.

Keywords: Performance Evaluation, Multi Camera Tracking, Identity Management, Multi Camera Data Set, Large Scale Data Set

1 Introduction

Multi-Target, Multi-Camera (MTMC) tracking systems automatically track multiple people through a network of cameras. As MTMC methods solve larger and larger problems, it becomes increasingly important (i) to agree on straightforward performance measures that consistently report bottom-line tracker performance, both within and across cameras; (ii) to develop realistically large benchmark data sets for performance evaluation; and (iii) to compare system performance end-to-end. This paper contributes to these aspects.

Performance Measures. Multi-Target Tracking has been traditionally defined as continuously following multiple objects of interest. Because of this, existing performance measures such as CLEAR MOT report how often a tracker makes what types of incorrect decisions. We argue that some system users may instead be more interested in how well they can determine who is where at all times.

To see this distinction, consider the scenario abstractly depicted in Figure 1(a) and 1(c). Airport security is following suspect A spotted in the airport lobby. They need to choose between two trackers, 1(a) and 1(c). Both tag

* This material is based upon work supported by the National Science Foundation under grants CCF-1513816 and IIS-1543720 and by the Army Research Office under grant W911NF-16-1-0392.

the suspect as identity 1 and track him up to the security checkpoint. System 1(a) makes a single mistake at the checkpoint and henceforth tags the suspect as identity 2, so it loses the suspect at the checkpoint. After the checkpoint, system 1(c) repeatedly flips the tags for suspect A between 1 and 2, thereby giving police the correct location of the suspect several times also between the checkpoint and the gate, and for a greater overall fraction of the time. Even though system 1(a) incurs only one ID switch, airport security is likely to prefer system 1(c), which reports the suspect's position longer—multiple ID switches notwithstanding—and ultimately leads to his arrest at the gate.

We do not claim that one measure is better than the other, but rather that different measures serve different purposes. *Event-based* measures like CLEAR MOT help pinpoint the source of some errors, and are thereby informative for the designer of certain system components. In the interest of users in applications such as sports, security, or surveillance, where preserving identity is crucial, we propose two *identity-based* measures (ID precision and ID recall) that evaluate how well computed identities conform to true identities, while disregarding where or why mistakes occur. Our measures apply both within and across cameras.

Data Set. We make available a new data set that has more than 2 million frames and more than 2,700 identities. It consists of 8×85 minutes of 1080p video recorded at 60 frames per second from 8 static cameras deployed on the Duke University campus during periods between lectures, when pedestrian traffic is heavy. Calibration data determines homographies between images and the world ground plane. All trajectories were manually annotated by five people over a year, using an interface we developed to mark trajectory key points and associate identities across cameras. The resulting nearly 100,000 key points were automatically interpolated to single frames, so that every identity comes with single-frame bounding boxes and ground-plane world coordinates across all cameras in which it appears. To our knowledge this is the first dataset of its kind.

Reference System. We provide code for an MTMC tracker that extends a single-camera system that has shown good performance [1] to the multi-camera setting. We hope that the conceptual simplicity of our system will encourage plug-and-play experimentation when new individual components are proposed.

We show that our system does well on a recently published data set [2] when previously used measures are employed to compare our system to the state of the art. This comparison is only circumstantial because most existing results on MTMC tracking report performance using *ground-truth* person detections and *ground-truth* single-camera trajectories as inputs, rather than using the results from *actual* detectors and single-camera trackers. The literature typically justifies this limitation with the desire to measure only what a multi-camera tracker *adds* to a single-camera system. This justification is starting to wane as MTMC tracking systems approach realistically useful performance levels. Accordingly, we evaluate our system end-to-end, and also provide our own measures as a baseline for future research.

2 Related Work

We survey prior work on MTMC performance measures, data sets, and trackers.
Measures. We rephrase existing MTMC performance measures as follows.

- A *fragmentation* occurs in frame t if the tracker switches the identity of a trajectory in that frame, but the corresponding ground-truth identity does not change. The number of fragmentations at frame t is ϕ_t, and $\Phi = \sum_t \phi_t$.
- A *merge* is the reverse of a fragmentation: The tracker merges two different ground truth identities into one between frames t' and t. The number of merges at frame t is γ_t, and $\Gamma = \sum_t \gamma_t$.
- A *mismatch* is either a fragmentation or a merge. We define $\mu_t = \phi_t + \gamma_t$ and $M = \sum_t \mu_t$.

When relevant, each of these error counts is given a superscript w (for "within-camera") when the frames t' and t in question come from the same camera, and a superscript h (for "handover") otherwise.

The number of *false positives* fp_t is the number of times the tracker detects a target in frame t where there is none in the ground truth, the number of *false negatives* fn_t is the number of true targets missed by the tracker in frame t, and tp_t is the number of true positive detections at time t. The capitalized versions TP, FP, FN are the sums of tp_t, fp_t, and fn_t over all frames (and cameras, if more than one), and the superscripts w and h apply here as well if needed.

Precision and *recall* are the usual derived measures, $P = TP/(TP + FP)$ and $R = TP/(TP + FN)$.

Single-camera, multi-object tracking performance is typically measured by the Multiple Object Tracking Accuracy (MOTA):

$$\text{MOTA} = 1 - \frac{FN + FP + \Phi}{T} \tag{1}$$

and related scores (MOTP, MT, ML, FRG) [3–5]. MOTA penalizes detection errors ($FN + FP$) and fragmentations (Φ) normalized by the total number T of true detections. If extended to the multi-camera case, MOTA and its companions under-report across-camera errors, because a trajectory that covers n_f frames from n_c cameras has only about n_c across-camera detection links between consecutive frames and about $n_f - n_c$ within camera ones, and $n_c \ll n_f$. To address this limitation handover errors [6] and multi-camera object tracking accuracy (MCTA) [2, 7] measures were introduced, which we describe next.

Handover errors focus only on errors across cameras, and distinguish between fragmentations Φ^h and merges Γ^h. Fragmentations and merges are divided further into crossing (Φ^h_X and Γ^h_X) and returning (Φ^h_R and Γ^h_R) errors. These more detailed handover error scores help understand different types of tracker failures, and within-camera errors are quantified separately by standard measures.

MCTA condenses all aspects of system performance into one measure:

$$\text{MCTA} = \underbrace{\frac{2PR}{P+R}}_{F_1} \underbrace{\left(1 - \frac{M^w}{T^w}\right)}_{\text{within camera}} \underbrace{\left(1 - \frac{M^h}{T^h}\right)}_{\text{handover}} . \tag{2}$$

Dataset	IDs	Duration	Cams	Actors	Overlap	Blind Spots	Calib.	Resolution	FPS	Scene	Year
Laboratory [8]	3	2.5 min	4	Yes	Yes	No	Yes	320x240	25	Indoor	2008
Campus [8]	4	5.5 min	3	Yes	Yes	No	Yes	320x240	25	Outdoor	2008
Terrace [8]	7	3.5 min	4	Yes	Yes	No	Yes	320x240	25	Outdoor	2008
Passageway [9]	4	20 min	4	Yes	Yes	No	Yes	320x240	25	Mixed	2011
Issia Soccer [11]	25	2 min	6	No	Yes	No	Yes	1920x1080	25	Outdoor	2009
Apidis Basket. [12]	12	1 min	7	No	Yes	No	Yes	1600x1200	22	Indoor	2008
PETS2009 [10]	30	1 min	8	Yes	Yes	No	Yes	768x576	7	Outdoor	2009
NLPR MCT 1 [2]	235	20 min	3	No	No	Yes	No	320x240	20	Mixed	2015
NLPR MCT 2 [2]	255	20 min	3	No	No	Yes	No	320x240	20	Mixed	2015
NLPR MCT 3 [2]	14	4 min	4	Yes	Yes	Yes	No	320x240	25	Indoor	2015
NLPR MCT 4 [2]	49	25min	5	Yes	Yes	Yes	No	320x240	25	Mixed	2015
Dana36 [14]	24	N/A	36	Yes	Yes	Yes	No	2048x1536	N/A	Mixed	2012
USC Campus [6]	146	25 min	3	No	No	Yes	No	852x480	30	Outdoor	2010
CamNeT [13]	50	30 min	8	Yes	Yes	Yes	No	640x480	25	Mixed	2015
DukeMTMC (ours)	2834	85 min	8	No	Yes	Yes	Yes	1920x1080	60	Outdoor	2016

Table 1: Summary of existing data sets for MTMC tracking. Ours is in the last row.

This measure multiplies the F_1 detection score (harmonic mean of precision and recall) by a term that penalizes within-camera identity mismatches (M^w) normalized by true within-camera detections (T^w) and a term that penalizes wrong identity handover mismatches (M^h) normalized by the total number of handovers. Consistent with our notation, T^h is the number of true detections (true positives TP^h plus false negatives FN^h) that occur when consecutive frames come from different cameras.

Comparing to MOTA, MCTA multiplies within-camera and handover mismatches rather than adding them. In addition, false positives and false negatives, accounted for in precision and recall, are also factored into MCTA through a product. This separation brings the measure into the range $[0, 1]$ rather than $[-\infty, 1]$ as for MOTA. However, the reasons for using a product rather than some other form of combination are unclear. In particular, each error in any of the three terms is penalized inconsistently, in that its cost is multiplied by the (variable) product of the other two terms.

Data Sets. Existing multi-camera data sets allow only for limited evaluation of MTMC systems. Some have fully overlapping views and are restricted to short time intervals and controlled conditions [8–10]. Some sports scenarios provide quality video with many cameras [11, 12], but their environments are severely constrained and there are no blind spots between cameras. Data sets with disjoint views come either with low resolution video [2, 6, 13], a small number of cameras placed along a straight path [2, 6], or scripted scenarios [2, 8–10, 13, 14]. Most importantly, all existing data sets only have a small number of identities. Table 1 summarizes the parameters of existing data sets. Ours is shown in the last row. It contains more identities than all previous data sets *combined*, and was recorded over the longest time period at the highest temporal resolution (60 fps).

Systems. MTMC trackers rely on pedestrian detection [15] and tracking [16] or assume single-camera trajectories to be given [6, 13, 17–26]. *Spatial relations between cameras* are either explicitly mapped in 3D [13, 19], learned by tracking known identities [25, 27, 28], or obtained by comparing entry/exit rates across pairs of cameras [6, 18, 26]. Pre-processing methods may fuse data from partially overlapping views [29], while some systems rely on completely overlapping and unobstructed views [9, 17, 30–32]. People *entry and exit points* may be explicitly modeled on the ground [6, 18, 19, 26] or image plane [24, 27]. *Travel time* is also modeled, either parametrically [13, 27] or not [6, 19, 24–26].

Appearance is captured by color [6, 13, 18–21, 23–25, 29, 27] and texture descriptors [6, 13, 18, 20, 22, 29]. Lighting variations are addressed through color normalization [18], exemplar based approaches [20], or brightness transfer functions learned with [23, 25] or without supervision [13, 19, 24, 29]. Discriminative power is improved by *saliency* information [33, 34] or *learning* features specific to body parts [6, 18, 20–23, 27], either in the image [35–37] or back-projected onto an articulated [38, 39] or monolithic [40] 3D body model.
All MTMC trackers employ *optimization* to maximize the coherence of observations for predicted identities. They first summarize spatial, temporal, and appearance information into a graph of *weights* w_{ij} that express the affinity of node observations i and j, and then partition the nodes into identities either greedily through bipartite matching or, more generally, by finding either paths or cliques with maximal internal weights. Some contributions are as follows:

	Single-Camera	Cross-Camera	Both
Bipartite	[41–43]	[6, 18, 20, 22]	—
Path	[9, 44–46]	[25, 27]	[2, 29]
Clique	[1, 47–55]	[23]	Ours

Table 2: Optimization techniques employed by MTMC systems.

In this paper, we extend a previous clique method [1] to formulate within- and across-camera tracking in a unified framework, similarly to previous MTMC flow methods [2, 29]. In contrast with [23], we handle identities reappearing in the same camera and differently from [8, 9] we handle co-occuring observations in overlapping views naturally, with no need for separate data fusion methods.

3 Performance Measures

Current event-based MTMC tracking performance measures count mismatches between ground truth and system output through *changes* of identity over time. The next two Sections show that this can be problematic both within and across cameras. The Section thereafter introduces our proposed measures.

3.1 Within-Camera Issues

With event-based measures, a truly-unique trajectory that switches between two computed identities over n frames can incur penalties that are anywhere between 1, when there is exactly one switch, and $n-1$, in the extreme case of one identity switch per frame. This can yield inconsistencies if correct identities are crucial. For example, in all cases in Figure 1, the tracker covers a true identity `A` with computed identities `1` and `2`. Current measures would make cases (b) and (c) equally bad, and (a) much better than the other two.

And yet the key mistake made by the tracker is to see two identities where there is one. To quantify the extent of the mistake, we need to decide which of the two computed identities we should match with `A` for the purpose of performance evaluation. Once that choice is made, every frame in which `A` is assigned to the wrong computed identity is a frame in which the tracker is in error.

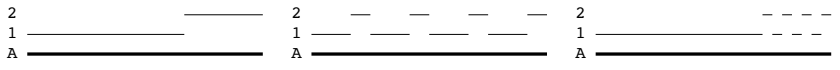


Fig. 1: Where there is one true identity A (thick line, with time in the horizontal direction), a tracker may mistakenly compute identities 1 and 2 (thin lines) broken into two fragments (a) or into eight (b, c). Identity 1 covers 67% of the true identity's trajectory in (a) and (b), and 83% of it in (c). Current measures charge one fragmentation error to (a) and 7 to each of (b) and (c). Our proposed measure charges 33% of the length of A to each of (a) and (b), and 17% to (c).

Since the evaluator—and not the tracker—makes this choice, we suggest that it should favor the tracker to the extent possible. If this is done for each tracker under evaluation, the choice is fair. In all cases in Figure 1, the most favorable choice is to tie A to 1, because this choice explains the largest fraction of A.

Once this choice is made, we measure the number of frames over which the tracker is wrong—in the example, the number of frames of A that are not matched to 1. In Figure 1, this measure makes (a) and (b) equally good, and (c) better than either. This penalty is consistent because it reflects precisely what the choice made above maximizes, namely, the number of frames over which the tracker is correct about who is where. In (a) and (b), the tracker matches ground truth 67% of the time, and in (c) it matches it 83% of the time.

Figure 1 is about fragmentation errors. It can be reinterpreted in terms of merge errors by exchanging the role of thick and thin lines. In this new interpretation, choosing the longest ground-truth trajectory as the correct match for a given computed trajectory explains as much of the tracker's output as possible, rather than as much of the ground truth. In both directions, our *truth-to-result matching* criterion is to let ground truth and tracker output explain as much of each other's data as possible, in a way that will be made quantitative later on.

3.2 Handover Issues

Event-based measures often evaluate handover errors separately from within-camera errors: Whether a mismatch is within-camera or handover depends on the identities associated to the very last frame in which a trajectory is seen in one camera, and on the very first frame in which it is seen in the next—a rather brittle proposition. In contrast, our measure counts the number of incorrectly matched frames, regardless of other considerations: If only one frame is wrong, the penalty is small. For instance, in the cases shown in Figure 2, current measures either charge a handover penalty when the handover is essentially correct (a) or fail to charge a handover penalty when the handover is essentially incorrect (b). Our measure charges a one-frame penalty in case (a) and a penalty nearly equal to the trajectory length in camera II in case (b), as appropriate. These cases are not just theoretical. In Section 6, we show that 74% of the 5,549 handovers computed by our tracker in our data set show similar phenomena.

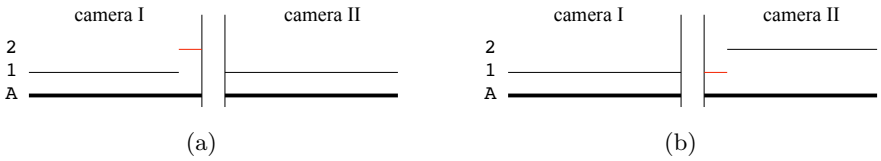


Fig. 2: (a) Ground-truth trajectory `A` is handed over correctly between cameras, because it is given the same computed identity `1` throughout, except that a short fragment in camera I is mistakenly given identity `2` (red). This counts as a handover error with existing measures. (b) `A` is handed over incorrectly, but a short fragment in camera II mistakenly given identity `1` (red) makes existing measures *not* count it as a handover error. Existing measures would charge a within-camera fragmentation and an across-camera fragmentation to (a) and one within-camera fragmentation to (b), even if assignment (a) is much better than (b) in terms of target identification.

These issues are exacerbated in measures, such as MCTA, that combine measures of within-camera mismatches and handover mismatches into a single value by a product (Eq. 2). If one of the anomalies discussed above changes a within-camera error into a handover error or *vice versa*, the corresponding contribution to the performance measure can change drastically, because the penalty moves from one term of the product to another: If the product has the form wh ("within" and "handover"), then a unit contribution to w has value h in the product, and changing that contribution from w to h changes its value to w.

3.3 The Truth-To-Result Match

To address these issues, we propose to measure performance not by *how often* mismatches occur, but by *how long* the tracker correctly identifies targets. To this end, ground-truth identities are first matched to computed ones. More specifically, a bipartite match associates one ground-truth trajectory to exactly one computed trajectory by minimizing the number of mismatched frames over all the available data—true and computed. Standard measures such as precision, recall, and F_1-score are built on top of this truth-to-result match. These scores then measure the number of mismatched or unmatched detection-frames, regardless of where the discrepancies start or end or which cameras are involved.

To compute the optimal truth-to-result match, we construct a bipartite graph $G = (V_T, V_C, E)$ as follows. Vertex set V_T has one "regular" node τ for each true trajectory and one "false positive" node f_γ^+ for each computed trajectory γ. Vertex set V_C has one "regular" node γ for each computed trajectory and one "false negative" node f_τ^-, for each true trajectory τ. Two regular nodes are connected with an edge $e \in E$ if their trajectories overlap in time. Every regular true node τ is also connected to its corresponding f_τ^-, and every regular computed node γ is also connected to its corresponding f_γ^+.

The cost on an edge $(\tau, \gamma) \in E$ tallies the number of false negative and false positive frames that would be incurred if that match were chosen. Specifically, let $\tau(t)$ be the sequence of detections for true trajectory τ, one detection for each

frame t in the set $\mathcal{T}_\tau$ over which τ extends, and define $\gamma(t)$ for $t \in \mathcal{T}_\gamma$ similarly for computed trajectories. The two simultaneous detections $\tau(t)$ and $\gamma(t)$ are a *miss* if they do not overlap in space, and we write

$$m(\tau, \gamma, t, \Delta) = 1 \; . \tag{3}$$

More specifically, when both τ and γ are regular nodes, spatial overlap between two detections can be measured either in the image plane or on the reference ground plane in the world. In the first case, we declare a miss when the area of the intersection of the two detection boxes is less than Δ (with $0 < \Delta < 1$) times the area of the union of the two boxes. On the ground plane, we declare a miss when the positions of the two detections are more than $\Delta = 1$ meter apart. If there is no miss, we write $m(\tau, \gamma, t, \Delta) = 0$. When either τ or γ is an irregular node (f_τ^- or f_γ^+), any detections in the other trajectory are misses. When both τ and γ are irregular, m is undefined. We define costs in terms of binary misses, rather than, say, Euclidean distances, so that a miss between regular positions has the same cost as a miss between a regular position and an irregular one. Matching two irregular trajectories incurs zero cost because they are empty.

With this definition, the cost on edge $(\tau, \gamma) \in E$ is defined as follows:

$$c(\tau, \gamma, \Delta) = \underbrace{\sum_{t \in \mathcal{T}_\tau} m(\tau, \gamma, t, \Delta)}_{\text{False Negatives}} + \underbrace{\sum_{t \in \mathcal{T}_\gamma} m(\tau, \gamma, t, \Delta)}_{\text{False Positives}} \; . \tag{4}$$

A minimum-cost solution to this bipartite matching problem determines a one-to-one matching that minimizes the cumulative false positive and false negative errors, and the overall cost is the number of mis-assigned detections for all types of errors. Every (τ, γ) match is a True Positive ID ($IDTP$). Every (f_γ^+, γ) match is a False Positive ID ($IDFP$). Every (τ, f_τ^-) match is a False Negative ID ($IDFN$). Every (f_γ^+, f_τ^-) match is a True Negative ID ($IDTN$).

The matches (τ, γ) in $IDTP$ imply a *truth-to-result* match, in that they reveal which computed identity matches which ground-truth identity. In general not every trajectory is matched. The sets

$$MT = \{\tau \mid (\tau, \gamma) \in IDTP\} \quad \text{and} \quad MC = \{\gamma \mid (\tau, \gamma) \in IDTP\} \tag{5}$$

contain the *matched ground-truth trajectories* and *matched computed trajectories*, respectively. The pairs in $IDTP$ can be viewed as a bijection between MT and MC. In other words, the bipartite match implies functions $\gamma = \gamma_m(\tau)$ from MT to MC and $\tau = \tau_m(\gamma)$ from MC to MT.

3.4 Identification Precision, Identification Recall, and F_1 Score

We use the $IDFN$, $IDFP$, $IDTP$ counts to compute identification precision (IDP), identification recall (IDR), and the corresponding F_1 score IDF_1. More specifically,

$$IDFN = \sum_{\tau \in AT} \sum_{t \in \mathcal{T}_\tau} m(\tau, \gamma_m(\tau), t, \Delta) \tag{6}$$

$$IDFP = \sum_{\gamma \in AC} \sum_{t \in \mathcal{T}_\gamma} m(\tau_m(\gamma), \gamma, t, \Delta) \tag{7}$$

$$IDTP = \sum_{\tau \in AT} \text{len}(\tau) - IDFN \;=\; \sum_{\gamma \in AC} \text{len}(\gamma) - IDFP \tag{8}$$

where AT and AC are all true and computed identities in MT and MC.

$$IDP = \frac{IDTP}{IDTP + IDFP} \tag{9}$$

$$IDR = \frac{IDTP}{IDTP + IDFN} \tag{10}$$

$$IDF_1 = \frac{2\,IDTP}{2\,IDTP + IDFP + IDFN} \tag{11}$$

Identification precision (recall) is the fraction of computed (ground truth) detections that are correctly identified. IDF_1 is the ratio of correctly identified detections over the average number of ground-truth and computed detections. ID precision and ID recall shed light on tracking trade-offs, while the IDF_1 score allows ranking all trackers on a single scale that balances identification precision and recall through their harmonic mean.

Our performance evaluation approach based on the truth-to-result match addresses all the weaknesses mentioned earlier in a simple and uniform way, and enjoys the following desirable properties: (1) *Bijectivity:* A correct match (with no fragmentation or merge) between true identities and computed identities is one-to-one. (2) *Optimality:* The truth-to-result matching is the most favorable to the tracker. (3) *Consistency:* Errors of any type are penalized in the same currency, namely, the number of misassigned or unassigned frames. Our approach also handles overlapping and disjoint fields of view in exactly the same way—a feature absent in all previous measures.

3.5 Additional Comparative Remarks

Measures of Handover Difficulty. Handover errors in current measures are meant to account for the additional difficulty of tracking individuals across cameras, compared to tracking them within a single camera's field of view. If a system designer were interested in this aspect of performance, a similar measure could be based on the difference between the total number of errors for the multi-camera solution and the sum of the numbers of single-camera errors:

$$E_M - E_S \quad \text{where} \quad E_M = IDFP_M + IDFN_M \quad \text{and} \quad E_S = IDFP_S + IDFN_S\,. \tag{12}$$

The two errors can be computed by computing the truth-to-result mapping twice: Once for all the data and once for each camera separately (and then adding the single-camera errors together). The difference above is nonnegative, because the

multi-camera solution must account for the additional constraint of consistency across cameras. Similarly, simple manipulation shows that ID precision, ID recall, and IDF_1 score are sorted the other way:

$$IDP_S - IDP_M \geq 0 \quad , \quad IDR_S - IDR_M \geq 0 \quad , \quad F_{1S} - F_{1M} \geq 0$$

and these differences measure how well the overall system can associate across cameras, given within-camera associations.

Comparison with CLEAR MOT. The first step in performance evaluation matches true and computed identities. In CLEAR MOT the event-based matching defines the best mapping sequentially at each frame. It minimizes Euclidean distances (within a threshold Δ) between unmatched detections (true and computed) while matched detections from frame $t-1$ that are still within Δ in t are preserved. Although the per-frame identity mapping is 1-to-1, the mapping for the entire sequence is generally many-to-many.

In our identity-based measures, we define the best mapping as the one which minimizes the total number of mismatched frames between true and computed IDs for the entire sequence. Similar to CLEAR MOT, a match at each frame is enforced by a threshold Δ. In contrast, our reasoning is not frame-by-frame and results in an ID-to-ID mapping that is 1-to-1 for the entire sequence.

The second step evaluates the goodness of the match through a scoring function. This is usually done by aggregating mistakes. MOTA aggregates FP, FN and Φ while we aggregate IDFP and IDFN counts. The notion of fragmentation is not present in our evaluation because the mapping is strictly 1-to-1. In other words our evaluation only checks whether every detection of an identity is explained or not, consistently with our definition of tracking. Also, our aggregated mistakes are binary mismatch counts instead of, say, Euclidean distances. This is because we want all errors to be penalized in the same currency. If we were to combine the binary IDFP and IDFN counts with Euclidean distances instead of IDTP, the unit of error would be ambiguous: We won't be able to tell whether the tracker under evaluation is good at explaining identities longer or following their trajectories closer.

Comparison with Identity-Aware Tracking. Performance scores similar to ours were recently introduced for this specific task [56]. The problem is defined as computing trajectories for a known set of true identities from a database. This implies that the truth-to-result match is determined during tracking and not evaluation. Instead, our evaluation applies to the more general MTMC setting where the tracker is agnostic to the true identities.

4 Data Set

Another contribution of this work is a new, manually annotated, calibrated, multi-camera data set recorded outdoors on the Duke University campus with 8 synchronized cameras (Fig. 3)*. We recorded 6,791 trajectories for *2,834 different*

* `http://vision.cs.duke.edu/DukeMTMC`

identities (distinct persons) over *1 hour and 25 minutes for each camera*, for a total of more than 10 video hours and more than 2 million frames. There are on average 2.5 single-camera trajectories per identity, and up to 7 in some cases.

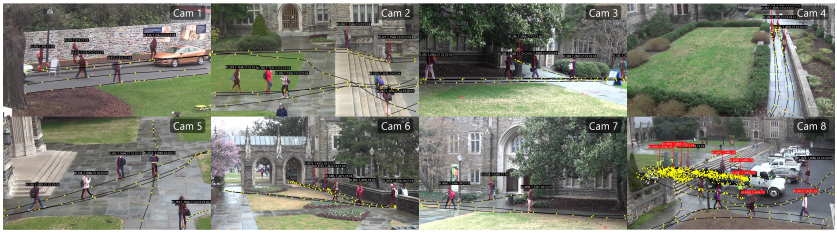


Fig. 3: Images and annotations of our DukeMTMC data set for frame 30890.

The cumulative trajectory time is more than 30 hours. Individual camera density varies from 0 to 54 people per frame, depending on the camera. There are 4,159 hand-overs and up to 50 people traverse blind spots at the same time. More than 1,800 self-occlusion events happen (with 50% or more overlap), lasting 60 frames on average. Our videos are recorded at 1080p resolution and 60 fps to capture spatial and temporal detail. Two camera pairs (2-8 and 3-5) have small overlapping areas, through which about 100 people transit, while the other cameras are disjoint. Full annotations are provided in the form of trajectories of each person's foot contact point with the ground. Image bounding boxes are also available and have been semi-automatically generated. The first 5 minutes of video from all the cameras are set aside for validation or training, and the remaining 80 minutes per camera are for testing.

Unlike many multi-camera data sets, ours is not scripted and cameras have a wider field of view. Unlike single-camera benchmarks where a tracker is tested on very short videos of different challenging scenarios, our data set is recorded in a fixed environment, and the main challenge is persistent tracking under occlusions and blind spots.

People often carry bags, backpacks, umbrellas, or bicycles. Some people stop for long periods of time in blind spots and the environment rarely constrains their paths. So transition times through blind spots are often but not always informative. 891 people walk in front of only one camera—a challenge for trackers that are prone to false-positive matches across cameras.

Working with this data set requires efficient trackers because of the amount of data to process. To illustrate, it took 6 days on a single computer to generate all the foreground masks with a standard algorithm [57] and 7 days to generate all detections on a cluster of 192 cores using the DPM detector [58]. Computing appearance features for all cameras on a single machine took half a day; computing all tracklets, trajectories, and identities together also took half a day with the proposed system (Sec. 5). People detections and foreground masks are released along with the videos.

Limitations. Our data set covers a single outdoor scene from fixed cameras. Soft lighting from overcast weather could make tracking easier. Views are mostly disjoint, which disadvantages methods that exploit data from overlapping views.

5 Reference System

We provide a reference MTMC tracker that extends to multiple cameras a system that was previously proposed for single camera multi-target tracking [1]. Our system takes target detections from any detection system, aggregates them into tracklets that are short enough to rely on a simple motion model, then aggregates tracklets into single camera trajectories, and finally connects these into multi-camera trajectories which we call *identities*.

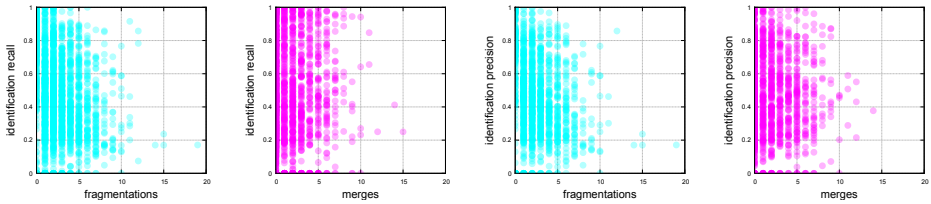
In each of these layers, a graph $\mathcal{G} = (V, E)$ has observations (detections, tracklets, or trajectories) for nodes in V, and edges in E connect any pairs of nodes i, j for which *correlations* w_{ij} are provided. These are real values in $[-1, 1]$ that measure evidence for or against i and j having the same identity. Values of $\pm\infty$ are also allowed to represent hard evidence. A Binary Integer Program (BIP) solves the *correlation clustering* problem [59] on $\mathcal{G}$: Partition V so as to maximize the sum of the correlations w_{ij} assigned to edges that connect co-identical observations and the penalties $-w_{ij}$ assigned to edges that straddle identities. Sets of the resulting partition are taken to be the desired aggregates.

Solving this BIP is NP-hard and the problem is also hard to approximate [60], hence the need for our multi-layered solution to keep the problems small. To account for unbounded observation times, solutions are found at all levels over a sliding temporal window, with solutions from previous overlapping windows incorporated into the proper BIP as "extended observations". For additional efficiency, observations in all layers are grouped heuristically into a number of subgroups with roughly consistent appearance and space-time locations.

Our implementation includes default algorithms for the computation of appearance descriptors and correlations in all layers. For appearance, we use the methods from the previous paper [1] in the first layers and simple striped color histograms [61] for the last layer. Correlations are computed from both appearance features and simple temporal reasoning.

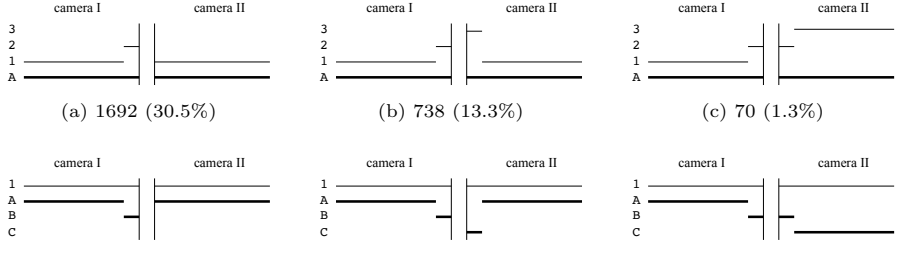
6 Experiments

This Section shows that (i) traditional event based measures are not good proxies for a tracker's ID precision or ID recall, defined in Section 3; (ii) handover errors, as customarily defined, cause frequent problems in practice; and (iii) the performance of our reference system, when evaluated with existing measures, is comparable to that of other recent MTMC trackers. We also give detailed performance numbers for our system on our data under a variety of performance measures, including ours, to establish a baseline for future comparisons.



(a) (b) (c) (d)

Fig. 4: Scatter plots of ground-truth trajectory ID recall (a, b) and ID precision (c, d) versus the number of trajectory fragmentations (a, c) and merges (b, d). Correlation coefficients are -0.24, -0.05, -0.38 and -0.41. This confirms that event- and identity-based measures quantify different aspects of tracker performance.



(a) 1692 (30.5%) (b) 738 (13.3%) (c) 70 (1.3%)

(d) 1065 (19.2%) (e) 496 (8.9%) (f) 58 (1.0%)

Fig. 5: [See Figure 2 for the interpretation of these diagrams.] In about 74% (4,119 out of 5,549) of the handovers output by our reference system on our data set, a short trajectory close to the handover causes a marked discrepancy between event-based, traditional performance measures and our identity-based measures. A handover fragmentation error (a, b) or merge error (d, e) is declared where the handover is essentially correct. A handover fragmentation error (c) or merge error (f) is not declared where the handover is essentially incorrect. Each caption shows the number of occurrences and the percentage of the total number of computed handovers.

ID Recall, ID Precision and Mismatches. Figure 4 shows that fragmentations and merges correlate poorly with ID recall and ID precision, confirming that event- and identity-based measures quantify different aspects of performance.

Truth-to-Result Mapping. Section 3 and Figure 2 describe situations in which traditional, event-based performance measures handle handover errors differently from ours. Figure 5 shows that these discrepancies are frequent in our results.

Traditional System Performance Analysis. Table 3 (top) compares our reference method to existing ones on the NLPR MCT data sets [2] and evaluates performance using the existing MCTA measure. The results are obtained under the commonly used experimental setup where all systems start with the same input of ground-truth single-camera trajectories. On average, our baseline system ranks second out of six by using our simple default appearance features. The highest ranked method [18] uses features based on discriminative learning.

Systems	**NLPR 1**	**NLPR 2**	**NLPR 3**	**NLPR 4**	**Avg. Rank**
USC [18]	**0.9152**	**0.9132**	0.5163	0.7052	2.25
Ours	0.7967	0.7336	0.6543	**0.7616**	2.5
GE [2]	0.8353	0.7034	**0.7417**	0.3845	2.75
hfutdspmct [7]	0.7425	0.6544	0.7368	0.3945	3.5
CRIPAC-MCT [62]	0.6617	0.5907	0.7105	0.5703	4
Adb-Team [7]	0.3204	0.3456	0.1382	0.1563	6

	CLEAR MOT Measures									Our Measures		
Cam	**FP↓**	**FN↓**	**IDS↓**	**FRG↓**	**MOTA↑**	**MOTP↑**	**GT**	**MT↑**	**ML↓**	**IDP↑**	**IDR↑**	**IDF$_1$↑**
1	9.70	52.90	178	366	37.36	67.57	1175	105	128	79.17	44.97	57.36
2	21.48	29.19	866	1929	49.17	61.70	1106	416	50	69.11	63.78	66.34
3	7.04	39.39	134	336	53.50	63.57	501	229	42	81.46	55.11	65.74
4	10.61	33.42	107	403	55.92	66.51	390	128	21	79.23	61.16	69.03
5	3.48	23.38	162	292	73.09	70.52	644	396	33	84.86	67.97	75.48
6	38.62	48.21	1426	3370	12.94	48.62	1043	207	91	48.35	43.71	45.91
7	8.28	29.57	296	675	62.03	60.73	678	373	53	85.23	67.08	75.07
8	1.29	61.69	270	365	36.98	69.07	1254	369	236	90.54	35.86	51.37
1-8									Upper bound	72.25	50.96	59.77
1-8									Baseline	52.35	36.46	42.98

Table 3: *Top Table:* MCTA score comparison on the existing NLPR data sets, starting from ground truth single camera trajectories. The last column contains the average dataset ranks. *Bottom Table:* Single-camera (white background) and multi-camera (grey background) results on our DukeMTMC data set. For each separate camera we report both standard multi-target tracking measures as well as our new measures.

System Performance Details. Table 3 (bottom) shows both traditional and new measures of performance, both single-camera and multi-camera, for our reference system when run on our data set. This table is meant as a baseline against which new methods may be compared.

From the table we see that our IDF_1 score and MOTA do not agree on how they rank the sequence difficulty of cameras 2 and 3. This is primarily because they measure different aspects of the tracker. Also, they are different in the relative value differences. For example, camera 6 appears much more difficult than 7 based on MOTA, but the difference is not as dramatic when results are inspected visually or when IDF_1 differences are considered.

7 Conclusion

We define new measures of MTMC tracking performance that emphasize correct identities over sources of error. We introduce the largest annotated and calibrated data set to date for the comparison of MTMC trackers. We provide a reference tracker that performs comparably to the state of the art by standard measures, and we establish a baseline of performance measures, both traditional and new, for future comparisons. We hope in this way to contribute to accelerating advances in this important and exciting field.

References

1. Ristani, E., Tomasi, C.: Tracking multiple people online and in real time. In: ACCV-12th Asian Conference on Computer Vision, Springer (2014)
2. Cao, L., Chen, W., Chen, X., Zheng, S., Huang, K.: An equalised global graphical model-based approach for multi-camera object tracking. ArXiv:11502.03532 [cs] (February 2015)
3. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP Journal on Image and Video Processing (246309) (2008) 1–10
4. Wu, B., Nevatia, R.: Tracking of multiple, partially occluded humans based on static body part detection. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 1., IEEE (2006) 951–958
5. Milan, A., Schindler, K., Roth, S.: Challenges of ground truth evaluation of multi-target tracking. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on, IEEE (2013) 735–742
6. Kuo, C.H., Huang, C., Nevatia, R.: Inter-camera Association of Multi-target Tracks by On-Line Learned Appearance Affinity Models. In Daniilidis, K., Maragos, P., Paragios, N., eds.: Computer Vision ECCV 2010. Number 6311 in Lecture Notes in Computer Science. Springer Berlin Heidelberg (2010) 383–396
7. Multi-Camera Object Tracking Challenge: ECCV workshop on visual surveillance and re-identification. `http://mct.idealtest.org` (2014)
8. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multi-Camera People Tracking with a Probabilistic Occupancy Map. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(2) (February 2008) 267–282
9. Berclaz, J., Fleuret, F., Türetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence (2011)
10. Ferryman, J., Shahrokni, A.: An overview of the PETS 2009 challenge. (2009)
11. D'Orazio, T., Leo, M., Mosca, N., Spagnolo, P., Mazzeo, P.L.: A semi-automatic system for ground truth generation of soccer video sequences. In: Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on, IEEE (2009) 559–564
12. De Vleeschouwer, C., Chen, F., Delannay, D., Parisot, C., Chaudy, C., Martrou, E., Cavallaro, A., et al.: Distributed video acquisition and annotation for sport-event summarization. In: NEM summit 2008:: Towards Future Media Internet. (2008)
13. Zhang, S., Staudt, E., Faltemier, T., Roy-Chowdhury, A.: A Camera Network Tracking (CamNeT) Dataset and Performance Baseline. In: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV). (January 2015) 365–372
14. Per, J., Kenk, V.S., Mandeljc, R., Kristan, M., Kovačič, S.: Dana36: A multi-camera image dataset for object identification in surveillance scenarios. In: Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on, IEEE (2012) 64–69
15. Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? In: ECCV 2014 Workshops. Volume 8926. (2015) 613–627
16. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: Motchallenge 2015: Towards a benchmark for multi-target tracking. arXiv:1504.01942 [cs] (April 2015) arXiv: 1504.01942.

17. Bredereck, M., Jiang, X., Korner, M., Denzler, J.: Data association for multi-object Tracking-by-Detection in multi-camera networks. In: 2012 Sixth International Conference on Distributed Smart Cameras (ICDSC). (October 2012) 1–6
18. Cai, Y., Medioni, G.: Exploring context information for inter-camera multiple target tracking. In: 2014 IEEE Winter Conference on Applications of Computer Vision (WACV). (March 2014) 761–768
19. Chen, K.W., Lai, C.C., Lee, P.J., Chen, C.S., Hung, Y.P.: Adaptive Learning for Target Tracking and True Linking Discovering Across Multiple Non-Overlapping Cameras. IEEE Transactions on Multimedia **13**(4) (August 2011) 625–638
20. Chen, X., An, L., Bhanu, B.: Multitarget Tracking in Nonoverlapping Cameras Using a Reference Set. IEEE Sensors Journal **15**(5) (May 2015) 2692–2704
21. Chen, X., Huang, K., Tan, T.: Direction-based stochastic matching for pedestrian recognition in non-overlapping cameras. In: 2011 18th IEEE International Conference on Image Processing (ICIP). (September 2011) 2065–2068
22. Daliyot, S., Netanyahu, N.S.: A Framework for Inter-camera Association of Multi-target Trajectories by Invariant Target Models. In Park, J.I., Kim, J., eds.: Computer Vision - ACCV 2012 Workshops. Number 7729 in Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 372–386
23. Das, A., Chakraborty, A., Roy-Chowdhury, A.K.: Consistent re-identification in a camera network. In: Computer Vision–ECCV 2014. Springer (2014) 330–345
24. Gilbert, A., Bowden, R.: Tracking Objects Across Cameras by Incrementally Learning Inter-camera Colour Calibration and Patterns of Activity. In Leonardis, A., Bischof, H., Pinz, A., eds.: Computer Vision ECCV 2006. Number 3952 in Lecture Notes in Computer Science. Springer Berlin Heidelberg (2006) 125–136
25. Javed, O., Shafique, K., Rasheed, Z., Shah, M.: Modeling inter-camera spacetime and appearance relationships for tracking across non-overlapping views. Computer Vision and Image Understanding **109**(2) (February 2008) 146–162
26. Makris, D., Ellis, T., Black, J.: Bridging the gaps between cameras. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. Volume 2. (June 2004)
27. Jiuqing, W., Li, L.: Distributed optimization for global data association in non-overlapping camera networks. In: 2013 Seventh International Conference on Distributed Smart Cameras (ICDSC). (October 2013) 1–7
28. Calderara, S., Cucchiara, R., Prati, A.: Bayesian-competitive consistent labeling for people surveillance. Pattern Analysis and Machine Intelligence, IEEE Transactions on **30**(2) (Feb 2008) 354–360
29. Zhang, S., Zhu, Y., Roy-Chowdhury, A.: Tracking multiple interacting targets in a camera network. Computer Vision and Image Understanding **134** (May 2015) 64–73
30. Ayazoglu, M., Li, B., Dicle, C., Sznaier, M., Camps, O.: Dynamic subspace-based coordinated multicamera tracking. In: 2011 IEEE International Conference on Computer Vision (ICCV). (November 2011) 2462–2469
31. Kamal, A., Farrell, J., Roy-Chowdhury, A.: Information Consensus for Distributed Multi-target Tracking. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2013) 2403–2410
32. Hamid, R., Kumar, R., Grundmann, M., Kim, K., Essa, I., Hodgins, J.: Player localization using multiple static cameras for sports visualization. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2010) 731–738

33. Martinel, N., Micheloni, C., Foresti, G.L.: Saliency weighted features for person re-identification. In: Computer Vision-ECCV 2014 Workshops. Springer International Publishing (2014) 191–208
34. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013)
35. Bedagkar-Gala, A., Shah, S.: Multiple person re-identification using part based spatio-temporal color appearance model. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. (Nov 2011) 1721–1728
36. Bedagkar-Gala, A., Shah, S.K.: Part-based spatio-temporal model for multi-person re-identification. Pattern Recognition Letters **33**(14) (2012) 1908 – 1915 Novel Pattern Recognition-Based Methods for Re-identification in Biometric Context.
37. Cheng, D., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Proceedings of the British Machine Vision Conference, BMVA Press (2011) 68.1–68.11 http://dx.doi.org/10.5244/C.25.68.
38. Baltieri, D., Vezzani, R., Cucchiara, R.: Learning articulated body models for people re-identification. In: Proceedings of the 21st ACM International Conference on Multimedia. MM '13, New York, NY, USA, ACM (2013) 557–560
39. Cheng, D., Cristani, M.: Person re-identification by articulated appearance matching. In Gong, S., Cristani, M., Yan, S., Loy, C.C., eds.: Person Re-Identification. Advances in Computer Vision and Pattern Recognition. Springer London (2014) 139–160
40. Baltieri, D., Vezzani, R., Cucchiara, R.: Mapping appearance descriptors on 3d body models for people re-identification. International Journal of Computer Vision **111**(3) (2015) 345–364
41. Brendel, W., Amer, M., Todorovic, S.: Multiobject tracking as maximum weight independent set. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1273–1280
42. Shu, G., Dehghan, A., Oreifej, O., Hand, E., Shah, M.: Part-based multiple-person tracking with partial occlusion handling. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 1815–1821
43. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. International Journal of Computer Vision **75**(2) (2007) 247–266
44. Izadinia, H., Saleemi, I., Li, W., Shah, M.: Mp2t: Multiple people multiple parts tracker. In Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., eds.: ECCV (6). Volume 7577 of Lecture Notes in Computer Science., Springer (2012) 100–114
45. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1201–1208
46. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
47. Butt, A.A., Collins, R.T.: Multiple target tracking using frame triplets. In: Computer Vision–ACCV 2012. Springer (2013) 163–176
48. Chari, V., Lacoste-Julien, S., Laptev, I., Sivic, J.: On pairwise costs for network flow multi-object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 5537–5545

49. Collins, R.T.: Multitarget data association with higher-order motion models. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 1744–1751
50. Dehghan, A., Assari, S.M., Shah, M.: Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In: CVPR. Volume 1. (2015) 2
51. Kumar, R., Charpiat, G., Thonnat, M.: Multiple object tracking by efficient graph partitioning. In: Computer Vision–ACCV 2014. Springer (2014) 445–460
52. Shafique, K., Shah, M.: A noniterative greedy algorithm for multiframe point correspondence. Pattern Analysis and Machine Intelligence, IEEE Transactions on **27**(1) (2005) 51–65
53. Tang, S., Andres, B., Andriluka, M., Schiele, B.: Subgraph decomposition for multi-target tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 5033–5041
54. Wen, L., Li, W., Yan, J., Lei, Z., Yi, D., Li, S.Z.: Multiple target tracking based on undirected hierarchical relation hypergraph. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE (2014) 1282–1289
55. Zamir, A., Dehghan, A., Shah, M.: Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In: Proceedings of the European Conference on Computer Vision (ECCV). (2012)
56. Yu, S.I., Meng, D., Zuo, W., Hauptmann, A.: The solution path algorithm for identity-aware multi-object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3871–3879
57. Yao, J., Odobez, J.M.: Multi-layer background subtraction based on color and texture. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE (2007) 1–8
58. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(9) (Sept 2010) 1627–1645
59. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. In: Foundations of Computer Science. (2002)
60. Tan, J.: A note on the inapproximability of correlation clustering. (2008)
61. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person re-identification: What features are important? In: Computer Vision–ECCV 2012. Workshops and Demonstrations, Springer (2012) 391–401
62. Chen, W., Cao, L., Chen, X., Huang, K.: A novel solution for multi-camera object tracking. In: Image Processing (ICIP), 2014 IEEE International Conference on, IEEE (2014) 2329–2333