A Dataset for Persistent Multi-Target Multi-Camera Tracking in RGB-D

A Dataset for Persistent Multi-Target Multi-Camera Tracking in RGB-D

Ryan Layne[1] Sion Hannuna[2] Massimo Camplani[2] Jake Hall[2]
Timothy M. Hospedales[3] Tao Xiang[1] Majid Mirmehdi[2] Dima Damen[2]
[1]Queen Mary University, [2]Bristol University, [3]Edinburgh University
r.d.c.layne@qmul.ac.uk

Abstract

Video surveillance systems are now widely deployed to improve our lives by enhancing safety, security, health monitoring and business intelligence. This has motivated extensive research into automated video analysis. Nevertheless, there is a gap between the focus of contemporary research, and the needs of end users of video surveillance systems. Many existing benchmarks and methodologies focus on narrowly defined problems in detection, tracking, re-identification or recognition. In contrast, end users face higher-level problems such as long-term monitoring of identities in order to build a picture of a person's activity across the course of a day, producing usage statistics of a particular area of space, and that these capabilities should be robust to challenges such as change of clothing. To achieve this effectively requires less widely studied capabilities such as spatio-temporal reasoning about people identities and locations within a space partially observed by multiple cameras over an extended time period. To bridge this gap between research and required capabilities, we propose a new dataset LIMA that encompasses the challenges of monitoring a typical home / office environment. LIMA contains 4.5 hours of RGB-D video from three cameras monitoring a four room house. To reflect the challenges of a realistic practical application, the dataset includes clothes changes and visitors to ensure the global reasoning is a realistic open-set problem. In addition to raw data, we provide identity annotation for benchmarking, and tracking results from a contemporary RGB-D tracker – thus allowing focus on the higher level monitoring problems.

1. Introduction

Video surveillance systems are now widely deployed to improve our lives by enhancing safety and security and to provide business intelligence [29] and healthcare monitoring/assist independent living [28]. Although the majority of surveillance systems involve outdoor spaces, there is an increasing demand for "smart" living and working spaces that are able to leverage human behavioural or identity information for enhanced functionality. These considerations have motivated extensive work in person detection [3], recognition [30] and search [31], re-identification [13] and tracking [4] – all challenging problems, particularly with arbitrary camera views, lighting, shadows, and cluttered backgrounds. While these are critical capabilities, they are somewhat lower level problems than those that concern the typical end users. In practice, the value provided by video surveillance systems comes from higher level problems such as longer-term monitoring of people and identities, in order to build a picture of a person's activities, or the usage pattern of a given physical space. The study of these higher-level tasks is addressed in multi-target multi-camera [26, 36] and identity-aware [33] tracking, but is relatively less well developed. This is in part because of lack of public datasets and benchmarks, which we aim to address here.

Research in person recognition, within-camera tracking and cross-camera re-identification provides improved modules for use by multi-target multi-camera (MTMC) systems. However, this research is often disconnected from the practical requirements of a realistic system. For example, re-identification benchmarks are often formalised as closed-world problems between two isolated cameras – while practical scenarios are always open-set [8, 21]. Similarly, tracking benchmarks are typically defined as relatively short-term problems, where there is limited change in a person's appearance, or long periods without an observation from *any* camera. In contrast to the practical case where, particularly in smart homes/offices, people may change clothes [14] or spend time out of views of all cameras while in rooms where privacy is expected. Existing work in multi-target multi-camera tracking exploits within-camera tracking, and multi-camera optimisation to find globally coherent estimate of person identities across space and time [26, 33]. This is a crucial strategy, but can struggle with these realistic challenges of appearance change and protracted disappearance preventing straightforward camera 'handoff' [26].

In this paper we introduce the LIMA (**L**ong term **I**dentity aware **M**ulti-target multi-camera**A** tracking) dataset that includes these challenges. The dataset is designed to promote research into better global optimisation strategies for joint MTMC tracking and better features and soft/hard biometrics for dealing with appearance change. By making this dataset publicly available, we aim to move research focus closer towards more practically relevant, multi-object object or target for consistency? or just MTMC multi-camera tracking problems. The LIMA dataset has the following important characteristics: (i) It is more reflective of real-world smart home/office conditions, including clothes change and an open-set of people to be monitored. (ii) It includes multi-camera identity annotation and benchmarking. (iii) It includes contemporary RGB-D video data that is typically not studied by previous work, but can be exploited to help address the greater challenge posed by such realistic scenarios. (iv) It includes person-tracking data. Since no tracker is perfect, identity reasoning must be robust to realistic tracking inaccuracies and errors. LIMA is compared to the most related prior datasets in Table 1.

2. Related Work

2.1. Underpinning Capabilities

Multi-target multi-camera tracking is often underpinned by key capabilities including (within-camera) tracking and (cross-camera) re-identification.

Tracking Within-camera tracking is now very well studied [4, 12]. Typical approaches [32] for fully automated tracking use the track-by-detection paradigm, stitching together detections in the form of moving regions highlighted by background subtraction, or by person detector [3] hits.

Re-identification Re-ID [13] is another extensively studied area which focuses on matching people in different camera views. It is related to the intra/within-camera tracking by detection in that it seeks to match images of person detections. However, it is more difficult because the view angle, lighting and pose typically change more dramatically. Re-ID may be used by MTMC systems to associate within-camera tracks in the process of performing multi-camera tracking. Typical approaches focus on obtaining view-invariant appearance features [11], and/or learning matching models specific to a given pair of camera views to be matched [20]. Relatively less studied directions include enhancing re-identification using soft-biometrics like attributes [18], height [25, 1], shape [1, 15], or movement style [14]. Such techniques are likely to be increasingly important when addressing realistic longer-duration home/office data where identity should be estimated correctly despite that people are likely to change clothes.

2.2. Multi-Target Multi-Camera Tracking

MTMC tracking is a key higher level capability of great importance to end-users. Systems may address within-camera tracking and cross-camera matching in an end-to-end manner [33], or more commonly rely on combining the output of separate modules for within-camera tracking and re-identification [9, 17, 6]. In both cases, one of the key challenges to address in order to achieve good performance, is to jointly infer the identities of all persons in the network, in order to maximise the overall coherence of the estimate [9, 17, 6, 33, 26, 8]. For example, under the priors that: people move in a spatio-temporally smooth way across the camera network, one person cannot be in two distinct places at the same time, people spend the expected amount of time in blind-spots, and people's transitions across camera reflect the learned [6] or given [36] camera connectivity matrix. Many methods make some simplifying assumptions limiting their generality [26]. In the most general case, with no simplifying assumptions, a MTMC method should handle situations where: identities can re-appear in the same camera (there is no one-way transition across the camera network), some cameras overlap and see an identity at the same time, while other cameras are disjoint and separated by a large blind-spot, and that there is an a-priori unknown and varying number of people observed by the whole network [26].

The global optimisation method used in many MTMC methods is predicated on the distribution of subject's appearance lying on a suitable manifold [34, 33]. Learning cross-camera appearance models can alleviate the impact of differing camera views on the manifold assumption, but in practical home/office scenarios where people are likely to change clothes, the discontinuity in appearance can seriously violate this assumption. How to deal with this systematically in a MTMC scenario is an open question.

2.3. Datasets

Some related datasets in the area of MTMC tracking are summarised in Table 1. Many suffer from either being small.

One particular challenge which is not addressed by existing methods and datasets is that of appearance (clothes) change. Existing methods are typically targeted at relatively short-term tracking, so subjects' observed appearance is only indirectly affected by the usual camera viewpoints and lighting, etc. However, in home / office settings, and over longer monitoring periods, people are likely to change clothes, which will hamper existing methods which assume subject's actual appearance does not change. This problem could be alleviated by biometrics such as face recognition [33], height/shape [1, 25] or movement [14] biometrics, assuming such cues can be appropriately integrated to constrain the global MTMC optimisation. To this end

Dataset	Depth	App Change	Video	#Cam	#People	Open-World	Rooms	Raw Video Avail	Tracking
LIMA	✓	✓	04:30	3	22	✓	3+1	✓	✓
SoftBio [5]	X	X	05:30	8	152	X	1	✓	X
Care Short [34]	X	X	00:06	15	13	X	3	X	✓
Care Long [33]	X	X	07:45	15	49	X	3	X	✓
Duke [26]	X	X	01:25	8	2834	✓	?	✓	✓
RAID [9]	X	X	?	4	43	✓	?	X	✓
CamNet [36]	X	X	00:30	8	50	✓	6	✓	✓
USC [17]	X	X	00:25	3	146	✓	3	X	✓

Table 1. Comparison of Datasets. Open-World: Whether the same set of people are tracked throughout, or people enter/leave during the recording. Rooms: Number of distinct rooms, or open space. Tracking: Whether within-camera tracking is provided. '?' indicates unspecified.

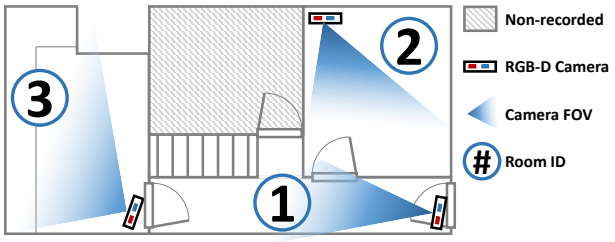

Figure 1. LIMA Dataset: Illustrative Schematic of the Building

our dataset also includes depth data. RGB-D cameras are now increasingly pervasive, however the use of this sensing modality to disambiguate MTMC problems is currently unexplored. Using such depth data provides numerous opportunities including: deriving shape/height cues [1, 25, 2] for better robustness to appearance change; investigating depth rather than RGB based appearance features; depth-based 3D positioning in order to better estimate people's proximity to specific entry/exit zones within a camera [17, 6, 24].

3. LIMA Dataset Description

Scenarios Our dataset is aligned with a home/office monitoring scenario, where a relatively small number (compared to outdoor crowd scene datasets) of people should be monitored over a longer period of time. The cameras were set up to observe the kitchen, main hallway, and living area on the ground floor of a typical residential home, as illustrated in Figure 1. There are several other rooms (bedroom and bathroom) in the home not recorded for privacy reasons. This creates the realistic challenge that people in those rooms may be invisible to the camera network for extended periods of time.

We captured data in 13 "sessions" across three days. In each session we simulated as closely as possible the real world activities of cohabiting families or friends. Within each session a series of normal activities (eating, callisthenics, napping, reading, etc.) were procedurally generated for each individual who then performed these tasks in sequence but otherwise in an unconstrained manner. Natural interaction between participants was encouraged – we instructed participants to perform actions in real-time as much as possible except for their duration, due to recording storage and processing constraints. Each session lasted between 15-20 minutes and contains 2-4 people, with each person performing an average of 8 actions per session. To explore the open-world session, 'residents' occasionally left, and 'guests' occasionally visited. In addition, to explore the realistic challenge of varying appearance in such an environment, participants also changed their clothing at a random point in their schedule. This was typically done in the realistic way of going into the bedroom to change. This meant that the actual cloth changing event is out of the view of the camera network, providing a particular challenge for long-term tracking. The data[1] is illustrated in Figure 2. Table 2 provides a breakdown of the data statistics and annotation content available.

Technical Description The data was captured with three commercially available ASUS Xtion PRO cameras[2], at 640x480 pixel resolution and $\approx$ 25 FPS in both RGB and depth channels. The full dataset consists of 4.5 hours of RGB-D video from three synchronised cameras. The data is provided both in raw format suitable for applying end-to-end systems with RGB/RGB-D tracking algorithms; as well as in preprocessed format with within-camera tracks from a commercial RGB-D tracker provided. In this case tracklets, and extracted bounding boxes are provided.

Pre-processing and Annotation We use a commercially available off-the-shelf software, OpenNI's NITE[3] to detect people and generate tracklets. The tracklets are then exhaustively annotated manually at the person detection level, with each bounding box labelled with a ground-truth identity label. Due to the limitations of the commercial tracker, a portion of the tracklets cover multiple person identities. This happens for the most part due to tracker confusion be-

[1]Download data from: http://www.irc-sphere.ac.uk/work-package-2/ReID

[2]https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/

[3]http://openni.ru/files/nite/

Figure 2. LIMA Dataset: Example RGB frames (top) and depth sensor measurements (bottom, darker is nearer) from cameras 1, 2, and 3 (left-to-right)

tween people, or people and background shapes. To simulate future improvements to tracker reliability, we remove detections from a tracklet that are not labelled with the majority identity label. For building a reliable tracklet appearance model, we remove detections that indicate a prone position, such as lying down, (detection aspect ratio $r < 0.75$) and lastly subsample tracklet's detections when they are more than 400 prior to feature extraction.

Evaluation We aim to support two types of MTMC applications: (i) MTMC tracking, in which all people are to be tracked, without any prior knowledge of the occupants [26], and (ii) Identity-aware tracking [33], in which there is an enrolment database. Both need to correctly stitch tracks across cameras. In the MTMC case guests/visitors are to be tracked equally with residents. In the identity-aware case, it needs to assign the long term tracks the correct identity from the enrolled database, so visitors provide potential distractors to recognition.

To support the identity-aware scenario, each session also contains a brief enrolment sequence, where the home residents for that sequence appear in isolation, so an appearance and/or biometric model can be built to support their recognition in future identity-aware tracking.

MTMC is harder in that there is no prior knowledge about the number of people and their appearance. Identity-aware tracking is easier due to prior knowledge about the number and appearance of the enrolled residents, but harder in that long-term tracks must not only correctly associate the same person's within-camera tracks over time, but those long-term tracks should be assigned the right ID from the enrolment database, and differentiated from guests. In other words, there are two ways to generate identity-aware errors, but only one way to generate MTMC errors.

Day 1	**Sessions**	**Frames**	**BBs**	**Ts**	**IDs**
Cam1	4	144258	37223	179	13 (10)
Cam2		133096	118610	126	
Cam3		128464	119571	87	
Total		**405818 (133086)**	**275404**	**392**	
Day 2	**Sessions**	**Frames**	**BBs**	**Ts**	**IDs**
Cam1	4	88189	20996	140	12 (9)
Cam2		95475	53527	124	
Cam3		84169	40516	66	
Total		**267833 (83760)**	**115039**	**330**	
Day 2	**Sessions**	**Frames**	**BBs**	**Ts**	**IDs**
Cam1	5	112877	33260	234	16 (15)
Cam2		114018	65605	164	
Cam3		107598	50399	98	
Total		**334493 (104544)**	**149264**	**496**	**41 (22)**

Table 2. Specifications of LIMA dataset. We provide pre-processed tracklets (Ts) and person detections (Bounding Boxes, BBs) for 13 sessions. The number of total people across all sessions can be found in column IDs, with the number of unique individuals present that day in parentheses. The number of annotated frames provided is also listed in parentheses.

4. Methodology

4.1. Benchmark Protocols

Multi-Target Multi-Camera Tracking An MTMC system is assumed to label all N tracklets from within-camera tracking as one of K global identities (K is effectively a person count within a given video sequence), where typically $K \ll N$. For evaluation, the K estimated global identities are assigned to K' ground truth identities by exhaustive search. Based upon this assignment, standard measures such as precision, recall and F_1-score are calculated between the K' true labels in the dataset and estimated labels of all the N local tracks [26]. The evaluation considers

both micro and macro averages over these quantities. Macro average is more influenced by people who appear often and micro average treats each person equally.

Identity-Aware Tracking For identity-aware tracking, we report both averaged absolute accuracy as well as the same retrieval oriented precision/recall/F_1 measure metrics.

4.2. Baseline Model

Overall Framework To provide a flexible and unconstrained baseline for MTMC, we adapt the constrained clustering algorithm E^2CP [23]. For N input tracklets $\{x_i\}_1^N$, an $N \times N$ affinity matrix A is generated, where A_{ij} is the similarity of tracklets x_i and x_j based on appearance or soft-biometric cues. Furthermore, a sparse constraint matrix $C \in \{+1, 0, -1\}^{N \times N}$ is defined. Must-link constraints $C_{i,j} = +1$ mean that x_i and x_j are known to be the same identity, e.g., due to simultaneous visibility at an overlapped viewpoint, or a hard-biometric match. Cannot-link constraints $C_{i,j} = -1$ mean that x_i and x_j are known to be different identity, e.g., due to simultaneous visibility from disjoint cameras. Finally, $C_{i,j} = 0$ means there are no known constraints on identity. We then apply the method in [23] to perform spectral clustering on the affinity matrix A subject to the constraints in C.

Affinity Matrix Generation To populate the inter-tracklet affinity matrix, directly comparing tracklet appearance is ineffective due to cross-camera appearance change. We therefore apply the KISS metric learner [16]. The affinity between tracklets x_i and x_j is given by the exponentiated Mahalanobis distance between their d-dimensional appearance features.

$$\log(A_{i,j}) = -\gamma(x_i - x_j)M(x_i - x_j)^T \qquad (1)$$

KISS uses an identity-annotated training set, and trains Mahalanobis matrix M to maximise the ratio $\frac{A_{i,j \in \mathcal{S}}(M)}{A_{j,k \in \mathcal{D}}(M)}$ between the affinity of a set of matched people $\mathcal{S}$ and a set of mismatched people $\mathcal{D}$. In this way, M is trained to promote features, and pairwise feature interactions (it is a full covariance matrix) that make matching people have high affinity (and mismatching people have low affinity). Given a training set that includes matching and mismatching pairs of people in multiple views, the metric is trained to ensure affinities hold as expected across views.

KISS is trained using 2,701 identity pairs from PETA [10] and CUHK Campus [19] (CUHK02,P1) as auxiliary sources of data. Given the need to compute the $d \times d$ full covariance matrix M, and apply it online at runtime, we reduce all features using PCA to $d = 500$.

Constraint Matrix Generation For the main experiment, we populate the constraint matrix through two simple heuristics: **H1:** If two tracks are visible at the same time in the same camera, they must be different people and require a cannot-link constraint. **H2:** If two tracks are visible at the same time in completely disjoint cameras, they must be different people and require a cannot-link constraint.

From these two heuristics we were able to generate 1,166 negative constraints across the whole dataset. We do not have an easy way to generate positive constraints since there is minimal inter-camera overlap in LIMA dataset. However, we also explore the impact of additional positive constraints that may be obtainable for example from hard-biometric matches such as face [33], and additional negative constraints, that could be obtained, for example through soft-biometric mismatches such as different genders.

Within vs Across Camera Matching It is important to note that while a key challenge for Identity-Aware/MTMC is cross-camera track association, the within-camera tracker does not always track users for their entire time spent in one camera view, i.e., within-camera tracking may result in fragmented tracklets. Thus, the MTMC system needs to associate tracklets *both within and across camera*.

Estimating the Number of Residents In our constrained clustering approach to MTMC, the number of clusters K corresponds to the number of unique identities established by the overall tracking framework. In most scenarios, the total number of entities to track in a given video segment is not prior knowledge, so we need to estimate K.

In order to estimate K, we employ two standard methods: (i) "silhouette" [27] $s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$ and (ii) "Calinski and Harabasz" [7] $CH = (\frac{SSB}{N-1})(\frac{SSW}{N-k})^{-1}$ indices. The silhouette score s_i for the ith point in a cluster reflects relative similarity between its member cluster a_i and the other clusters b_i. The CH index measures relative dispersion within a cluster (SSW) and between other clusters (SSB) as measured by sum-of-squares. We estimate K by searching $K \in \{2 \dots 8\}$ and taking the K that maximises the average of the two metrics.

Features To encode each tracklet's appearance, we extract bounding boxes and then extract LOMO [20] or KCCA [22] features from box's RGB channel, and HOG features for depth channel. After some heuristics to remove partially occluded detections near to transitions, we perform within-tracklet subsampling, and then averaging across frames to obtain a single RGB and D feature vector for each tracklet.

Identity-Aware For the identity-aware setting, we train multi-class SVM person recognisers with linear kernels, using the features from the enrolment portion of each session as input, and identities from enrolment as labels. We then compare two identity-aware tracking methods:
M1: Classify each tracklet by multi-class SVM and assign the most likely identity.
M2: Classify each tracklet by multi-class SVM. Use the resulting high confidence ($p > 0.7$) (mis)matches to generate additional constraints, and re-cluster the data with the pro-

posed constrained clustering. Then assign identities to all tracks in each cluster by majority vote among the identities assigned by SVM to the tracklets within the cluster. The idea here is that the application of the constrained clustering with a significantly increased number of constraints now is very likely to generate clusters reflecting the enrolled residents. Meanwhile, the predictions on those tracklets with low-confidence SVM predictions is made in a globally coherent way via their assignment to a visually similar and logically compatible cluster.

Discussion Our overall approach to MTMC is simpler than some contemporary alternatives [33, 26] but it has a number of advantages: (i) being flexible with no required assumptions on what movements are allowed, while any such available priors can be encoded as constraints, (ii) addressing both within-and across-camera track stitching, (iii) being easy to implement and fast to run, based on publicly available optimisation software [23, 16], and (iv) can be applied with minimal change to both vanilla and identity-aware variants of the MTMC problem.

Parameter Settings The E^2CP clustering algorithm has several parameters to tune, γ, the Gaussian scaling parameter for the exponentiated Mahalanobis distance in Equation 1, λ, the constraint propagation term for [23]'s model, and finally the number of k-nearest neighbours from which to build the initial graph, $\bar{k}$. For identity-aware settings, we use linear SVMs with a single, slack parameter c. Finally, we also learn scalar parameter α for weighting the fusion between RGB and HOG depth features , i.e., $A = (1-\alpha)A^{RGB} + \alpha A^{HOG}$. For each of the 13 sessions, we adopt a leave-one-out strategy to learn γ, α, $\bar{k}$ and c, and follow the authors of [23] in setting $\lambda = 0.8$ for our main experiments; however, we set $\lambda = 0.2$ for identity-aware experiments, since these generate denser constraints and it is beneficial to limit propagation to prevent saturation.

5. Benchmark Evaluations

We evaluate multi-target multi-camera tracking and identification in each of the 13 sessions within our dataset, and present results as averages over the per-session results.

5.1. Multi-Target Multi-Camera

In the first experiment, we evaluate MTMC performance. One key challenge in MTMC is that the total number of people is unknown. To explore this, we compare the results in the challenging case where the number of residents (K', Section 4.1) is unknown and easier case where we assume prior knowledge of the number of residents (K' known), although not enrolment.

To explore aspects of our methodology and dataset, we further compare the conditions: **±C:** With/without the constraints applied to global optimisation. **±D:** With/without including depth information in the tracklet representation.

The results are shown in Table 3. We make the following observations: (i) Comparing the top and bottom sections, we see that prior knowledge of the number of people to track makes the problem significantly easier. (ii) In general, ability to exploit constraints tends to improve performance over the unconstrained setting (+C vs -C). (iii) The impact of constrained clustering is greater in the more challenging case where the number of residents is not known. (iv) Generally, the inclusion of depth data improves performance compared to vanilla RGB tracking (+D vs -D), but not dramatically so. This is attributed to the relative simplicity of our depth feature (HOG) compared to highly refined RGB-based features. It also depends on the RGB features used: the KCCA feature seems to be more complementary to the HOG depth features. Better results may be obtained in future by developing better depth features, and/or exploiting depth to generate biometric rather than appearance cues. Overall, the results confirms that our baseline is credible, but they are not good enough to provide a useful practical system. This thus indicates that the dataset is a useful benchmark for future research.

5.2. Identity-Aware Tracking

In the second experiment, we evaluate identity-aware MTMC performance. To explore the significance of the clothes appearance change aspect of our dataset, we consider two conditions. In the first more realistic condition, people are enrolled once and and then during the testing sequence they change appearance in an a-priori unknown way. In the second condition, the enrolment sequence contains residents wearing each set of clothes that they might later change to in the testing sequence.

The results are shown in Table 4, where we compare the two enrolment conditions and two proposed tracking approaches. From the results we can first see that the multiple enrolment condition results are significantly higher than the more realistic single enrolment condition. This highlights the under-studied challenge of the appearance of appearance change in tracking. Given that existing methods are tuned on existing datasets and benchmarks, it suggests their performance may drop dramatically in practice when users inevitably change their appearance. Thus, our dataset provides an important opportunity to facilitate the development of methods that are more robust to these challenges. Secondly, we can see from the results that the enrolment does typically boost performance compared to all the MTMC conditions in Table 3. Finally, we see that our hybrid constrained+clustering recognition approach (M2) to identity-aware tracking does improve performance significantly compared to the naive approach (M1) of just recognising people based on their enrolment images. This is particularly so in the more realistic single-enrolment case. To

Known Resident Population K'					
KCCA	Micro Prec	Macro Prec	Micro Rec	Macro Rec	Micro F1
+C+D	0.771	0.787	0.771	0.790	0.739
-C+D	0.762	0.781	0.762	0.779	0.736
+C-D	0.756	0.776	0.756	0.776	0.730
-C-D	0.754	0.769	0.754	0.768	0.712
LOMO	Micro Prec	Macro Prec	Micro Rec	Macro Rec	Micro F1
+C+D	0.774	0.777	0.774	0.791	0.712
-C+D	0.754	0.758	0.754	0.771	0.691
+C-D	0.772	0.779	0.772	0.792	0.715
-C-D	0.752	0.768	0.752	0.778	0.714
Unknown Resident Population K'					
KCCA	Micro Prec	Macro Prec	Micro Rec	Macro Rec	Micro F1
+C+D	0.659	0.666	0.633	0.680	0.556
-C+D	0.651	0.653	0.626	0.647	0.517
+C-D	0.663	0.675	0.638	0.678	0.555
-C-D	0.630	0.620	0.608	0.620	0.477
LOMO	Micro Prec	Macro Prec	Micro Rec	Macro Rec	Micro F1
+C+D	0.658	0.672	0.632	0.697	0.545
-C+D	0.636	0.634	0.613	0.638	0.473
+C-D	0.684	0.685	0.653	0.701	0.537
-C-D	0.653	0.648	0.627	0.654	0.497

Table 3. Multi-target multi-camera tracking results, aggregating over sessions in our dataset. Evaluation by micro/macro precision and recall, and F1-score. $\pm C$ indicates with and without use of constraints in the joint inference. $\pm D$ indicates with and without depth-based appearance feature.

Multiple Enrolment						
KCCA	Acc	Micro Prec	Macro Prec	Micro Rec	Macro Rec	Micro F1
M1	0.785	0.837	0.844	0.837	0.852	0.730
M2	0.857	0.826	0.840	0.826	0.834	0.800
LOMO	Acc	Micro Prec	Macro Prec	Micro Rec	Macro Rec	Micro F1
M1	0.864	0.899	0.896	0.899	0.910	0.787
M2	0.922	0.891	0.912	0.891	0.909	0.883
Single Enrolment						
KCCA	Acc	Micro Prec	Macro Prec	Micro Rec	Macro Rec	Micro F1
M1	0.556	0.678	0.681	0.678	0.7111	0.346
M2	0.791	0.753	0.765	0.753	0.768	0.720
LOMO	Acc	Micro Prec	Macro Prec	Micro Rec	Macro Rec	Micro F1
M1	0.634	0.740	0.736	0.740	0.759	0.471
M2	0.830	0.793	0.803	0.793	0.813	0.767

Table 4. Identity-aware multi-target multi-camera tracking results, aggregating over sessions in our dataset. Evaluation by accuracy, micro/macro precision and recall, F1-score. M1: Track by recognition, M2: Track by recognition-enhanced constrained clustering.

Session	IDs	Acc	Micro Prec	Macro Prec	Micro Rec	Macro Rec	Micro F1
S01	3	0.922	0.838	0.852	0.838	0.848	0.829
S02	4	0.704	0.782	0.793	0.782	0.780	0.739
S03	3	0.762	0.767	0.736	0.767	0.746	0.628
S04	3	0.844	0.807	0.754	0.807	0.784	0.644
S05	3	0.984	0.988	0.985	0.988	0.987	0.986
S06	3	0.983	0.987	0.989	0.987	0.986	0.987
S07	3	1.000	1.000	1.000	1.000	1.000	1.000
S08	3	0.963	0.986	0.989	0.986	0.984	0.986
S09	3	0.742	0.768	0.793	0.768	0.795	0.759
S10	3	0.985	0.989	0.996	0.989	0.984	0.987
S11	4	0.882	0.765	0.758	0.765	0.763	0.713
S12	3	0.844	0.887	0.901	0.887	0.913	0.886
S13	3	0.982	0.986	0.988	0.986	0.983	0.985

Table 5. Identity-aware tracking results by session. Tracking model M2, and multiple-enrolment condition as in Table 4

illustrate the variability in the results across sessions, Table 5 presents the results for identity-aware tracking (M2, Multiple enrolment) broken down by session.

5.3. Further Analysis

Qualitative Results We present two qualitative examples illustrating how constraints help long-term tracking in Figure 3. Here, some residents have changed their clothes (first and second columns). The shift in appearance means that some of their tracklets (first and second columns) would then have been associated to incorrect clusters (third column), that are closer matches in feature space. However, in this case the presence of constraints in the clustering preclude a naive and incorrect association, and as a result they associate correctly (fourth column).

Effectiveness of Constraints In the current experiments we were unable to obtain a significant number of constraints to fully explore the potential benefit of our flexible approach to MTMC. We therefore perform a synthetic analysis to explore the potential impacts of the positive and negative constraints. In practice, these additional constraints could be obtained through further research on biometrics (e.g., face) that can generate positive constraints [33], and soft biometrics (e.g., gender) that can generate negative constraints [35]. Figure 4 shows a graph of MTMC tracking performance against the number of synthesised constraints of both types. From this we can see that increasing the number of either type of constraints leads to a much better solution. Constraints are particularly helpful in our dataset, where they help to disambiguate identities across clothes change. It suggests that future work on biometrics, *etc*. will greatly improve tracking performance in our problem.

Sources of Error The main sources of error are due to: (i) Upstream errors in the underlying within-camera tracking module including the tracker loitering on person-shaped background regions, and classic failure under occlusion. (ii) Clothes change. This was expected to be a challenge, given the appearance based matching. Although some instances can be corrected with our constrained clustering, this remains an open problem to address through better constraint generation and biometric cue development. This could also be ameliorated by exploiting the fact that people normally change clothes in a particular room (bedroom, bathroom). So although out of sight, logic could be developed to look for clothes changes when people enter/exit typical changing rooms. (iii) People sitting and lying down – These are natural activities in home or office environments but challenge systems and particularly the latest generation of feature representations (E.g., KCCA and LOMO) designed specifically for monitoring upright walking pedestrians.

6. Discussion

We introduced a new dataset to promote and support research into realistic multi-target multi-camera tracking scenarios. Evaluation of a global optimisation algorithm on this dataset demonstrates that it is feasible, but better fea-

Figure 3. Qualitative illustration of success cases for constraints. People change from their starting clothing set to their second clothing set. Without negative constraints, the model incorrectly classifies another person due to having a similar appearance (red border). By including constraints, this can be mitigated (green border).

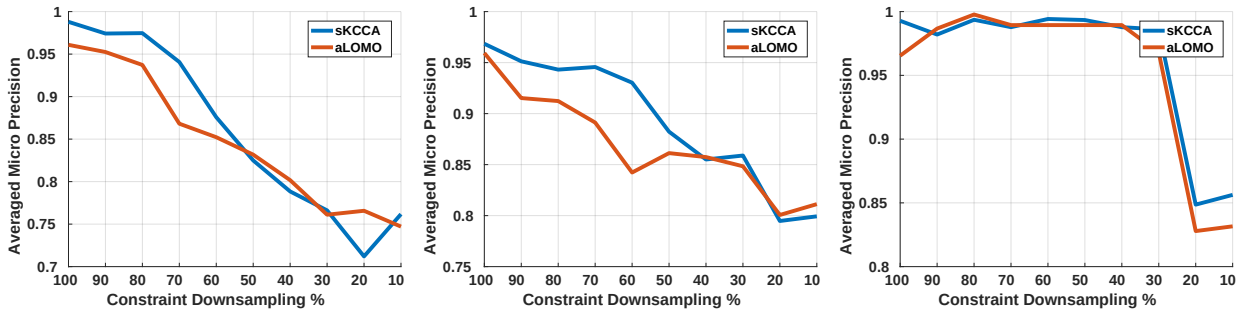

Figure 4. MTMC performance under varying density of synthesised constraints: from exhaustively synthesised to sparsely populated. Left: both positive and negative constraints. Middle: positive only. Right: negative constraints only.

tures, within-camera trackers and MTMC optimisers need to be developed to produce a system useful in realistic MTMC scenarios. Our dataset and annotation support a variety of future directions of investigation:

Depth Appearance Features In our experiments, HOG-based encoding of depth contributed only weakly to performance. However, research in depth appearance descriptors is in its infancy compared to decades of RGB descriptor research. Future depth descriptor research will benefit long-term tracking via better exploiting this modality.

Appearance and Pose Robustness The main challenges to overcome are the inclusion of clothes changes and people being in lying/sitting poses. These are under-studied but realistic challenges, and ours is the only MTMC dataset we know of to include clothes change in particular. Future research to develop methods robust to these covariates may use: soft-biometrics such as gait/movement style [14], which in our case can be fused with the appearance-driven affinity; biometric filters such as gender that can be used to add negative constraints [35], or hard-biometrics such as face. Cues such as face may not be visible in every track, so a strategy is needed to exploit them in a missing-data way [33]. In our baseline model this is straightforward by including them as constraints only where available. How to integrate all these cues given variable reliability and observability provides a challenge to support study of data fusion.

Depth-Biometrics Our depth-video also provides the opportunity for various novel soft-biometrics such as height and body-shape [2, 25, 1, 15] that can be further developed to increase robustness to clothes change. However, further work is needed as existing studies typically assume that people are walking pedestrians, and may not be robust to our home/office setting where people are often sitting/lying.

Online vs Batch Processing Our current evaluation is based on identity inference produced by batch processing an entire session. In practical scenarios users may be interested in incrementally produced tracking results. Our dataset and annotations support such benchmarks for future evaluation.

Acknowledgement This work was performed under the SPHERE IRC project funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/K031910/1.

References

[1] A. Albiol, J. Oliver, and J. M. Mossi. Who is who at different cameras: people re-identification using depth cameras. *IET Computer Vision*, 6(5), 2012.

[2] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino. Re-identification with RGB-D Sensors. In *ECCV, Workshop on Re-identification*, 2012.

[3] R. Benenson, M. Omran, J. Hosang, and B. Schiele. *Ten Years of Pedestrian Detection, What Have We Learned?* Springer, 2015.

[4] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, (1), 2008.

[5] A. Bialkowski, S. Denman, and P. Lucey. A database for person re-identification in multi-camera surveillance networks. In *Digital Image Computing: Techniques and Applications*, 2012.

[6] Y. Cai and G. Medioni. Exploring context information for inter-camera multiple target tracking. In *Workshop on the Applications of Computer Vision*, 2014.

[7] T. Calinski and J. Harabasz. A Dendrite Method for Cluster Analysis. *Communications in Statistics*, 3(1), 1974.

[8] B. Cancella and T. M. Hospedales. Open-World Person Re-Identification by Multi-Label Assignment Inference. In *BMVC*, 2014.

[9] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury. Consistent Re-identification in a Camera Network. In *ECCV*, 2014.

[10] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian Attribute Recognition At Far Distance. In *ACM International Conference on Multimedia*, 2014.

[11] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[12] J. Ferryman and A. Shahrokni. An overview of the pets 2009 challenge. *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2009.

[13] S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors. *Person Re-identification*. Springer, 2014.

[14] M. Gou, X. Zhang, A. Rates-Borras, S. Asghari-Esfeden, M. Sznaier, and O. Camps. Person re-identification in appearance impaired scenarios. In *BMVC*, 2016.

[15] D. Hernandez, M. Castrillon, and J. Lorenzo. People counting with re-identification using depth cameras. In *Imaging for Crime Detection (ICDP)*, SIANI, Spain, 2011.

[16] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. *CVPR*, 2012.

[17] C. Kuo, B. Huang, and R. Nevatia. Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In *ECCV*, 2010.

[18] R. Layne, T. M. Hospedales, and S. Gong. Person Re-identification by Attributes. In *BMVC*, 2012.

[19] W. Li and X. Wang. Locally Aligned Feature Transforms across Views. In *CVPR*, June 2013.

[20] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person Re-identification by Local Maximal Occurrence Representation and Metric Learning. In *CVPR*, 2015.

[21] S. Liao, Z. Mo, J. Zhu, Y. Hu, and S. Z. Li. Open-set person re-identification. *arXiv:1408.0872*, 2014.

[22] G. Lisanti, I. Masi, and A. Del Bimbo. Matching People across Camera Views using Kernel Canonical Correlation Analysis. In *IEEE International Conference on Distributed Smart Cameras*, 2014.

[23] Z. Lu and H. H. Ip. Constrained spectral clustering via exhaustive and efficient constraint propagation. In *ECCV*, 2010.

[24] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *CVPR*, 2004.

[25] M. Munaro, A. Fossati, A. Basso, E. Menegatti, and L. V. Gool. One-Shot Person Re-identification with a Consumer Depth Camera. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-identification*. Springer, London, 2014.

[26] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016.

[27] P. J. Rousseeuw. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics*, 20, 1987.

[28] F. Sadri. Ambient intelligence. *ACM Computing Surveys*, 43(4), Oct. 2011.

[29] C. Shan, F. Porikli, T. Xiang, and S. Gong, editors. *Video Analytics for Business Intelligence*. Springer, 2012.

[30] P. Tome, J. Fierrez, R. Vera-rodriguez, and M. S. Nixon. Soft Biometrics and Their Application in Person Recognition at a Distance. *IEEE Transactions on Information Forensics and Security*, 9(3), 2014.

[31] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *WACV*, Snowbird, Utah, 2009.

[32] B. Yang and R. Nevatia. An online learned crf model for multi-target tracking. In *CVPR*, 2012.

[33] S.-I. Yu, D. Meng, W. Zuo, and A. G. Hauptmann. The solution path algorithm for identity-aware multi-object tracking. In *CVPR*, 2016.

[34] S. I. Yu, Y. Yang, and A. Hauptmann. Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization. In *CVPR*, 2013.

[35] H. Zhang, J. R. Beveridge, B. A. Draper, and P. J. Phillips. On the effectiveness of soft biometrics for increasing face verification rates. *Computer Vision and Image Understanding*, 137(0):50 – 62, 2015.

[36] S. Zhang, E. Staudt, T. Faltemier, and A. K. Roy-Chowdhury. A camera network tracking (camnet) dataset and performance baseline. In *IEEE Winter Conference on Applications of Computer Vision*, 2015.