A Camera Network Tracking (CamNeT) Dataset and Performance Baseline

Shu Zhang[1], Elliot Staudt[1], Tim Faltemier[2], and Amit K. Roy-Chowdhury[1]

[1]Department of Electrical and Computer Engineering, University of California, Riverside
{szhang,estaudt,amitrc}@ee.ucr.edu
[2]Progeny Systems Corporation
{tfaltemier}@progeny.net

Abstract

In this paper, we propose a novel Non-Overlapping Camera Network Tracking Dataset (CamNeT) for evaluating multi-target tracking algorithms. The dataset is composed of five to eight cameras covering both indoor and outdoor scenes at a university. This dataset consists of six scenarios. Within each scenario are challenges relevant to lighting changes, complex topographies, crowded scenes, and changing grouping dynamics. Persons with predefined trajectories are combined with persons with random trajectories. Ground truth data for predefined trajectories is provided for each camera. Also, a baseline multi-target tracking system is presented. The tracking results using the baseline system are provided, which can be compared with future works. The work provides a comprehensive multi-camera dataset for performance evaluation in this challenging application domain, as well as an initial set of results.

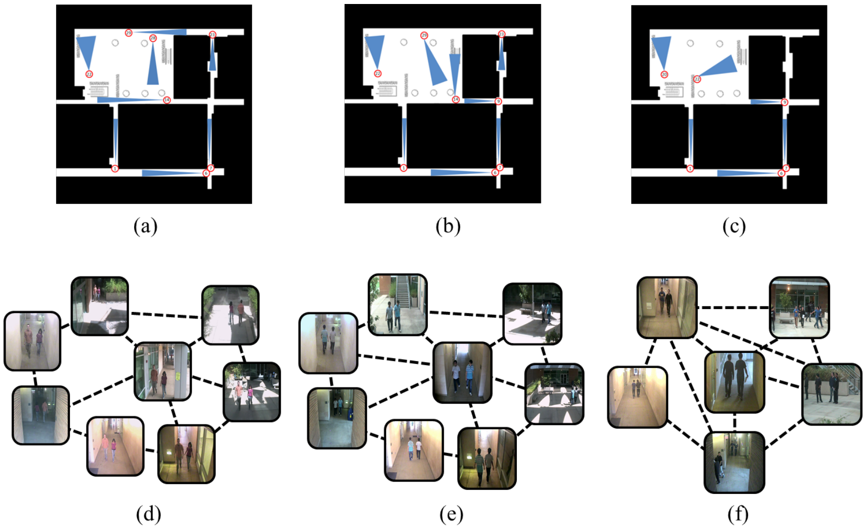
1. Introduction

The problem of multi-target tracking remains challenging, yet is a fundamental task for higher level automated video content analysis. Wide-area camera networks pose challenges that are unique to their application domain. These challenges include large blind areas between cameras, significant changes in the pose of targets, and differences in scene illumination between cameras. Moreover, single camera issues, like occlusion and appropriate feature selection, carry-over into the multi-camera domain and affect the overall performance. Though there are some existing multi-camera tracking works, they all use their own datasets lacking any standardization.

In this paper, we present a camera network tracking (CamNeT) dataset, specially designed for the problem of multi-target tracking. Differing from highly cited works [8, 9] where three cameras are used for testing tracking algorithms, five to eight cameras are used in this dataset. These cameras comprise part of an actual surveillance system distributed along the corridors and open courtyard of a building. Three different camera configurations are used in the proposed dataset. The layout of each configuration can be seen in Fig. 1 (a), (b) and (c). Lighting conditions vary from indoor scenes to outdoor scenes and cause appearance information to be more volatile than in other multi-camera datasets. The proposed dataset consists of six scenarios, one performed in the configuration in Fig. 1 (a), three performed in the configuration in Fig. 1 (b), and two performed in the configuration in Fig. 1 (c). Since temporal information is very important for tracking, a UTC time stamp is provided for every frame in each camera to compensate for the occurrence of frame loss.

To the best of our knowledge, there are no public multi-camera surveillance videos with non-overlapping views, especially for the purpose of tracking. Though multiple works report their tracking results with multi-camera non-overlapping views, none of them reported their results on a publicly available dataset. This makes a comparison between different tracking algorithms very difficult.

There do exist some datasets with overlapping camera views. The Videoweb Activities dataset [5] is a dataset that has been widely used. It has multiple activities among more than 10 cameras. However, it does not contain a non-overlapping view scenario. Similarly, the Multiple Cameras Fall dataset [1] has 8 cameras monitoring one meeting room with overlapping views. In this case, the purpose of the dataset is totally different from the one we are proposing. MuHAVi [12] uses 8 cameras with overlapping views to collect 17 action classes which are performed by 14 actors. All these datasets are not specifically designed for tracking purposes; instead they are more suitable for activity analysis. The PETS 2009 dataset [15] is one of the


(a) (b) (c)

(d) (e) (f)

Figure 1. Camera configurations and an example of every camera view. (a)(d) are from scenario 1, (b)(e) are from scenario 2, and (c)(f) are from scenario 6. (a) (b) and (c) are camera configurations for scenario 1, scenarios 2-4 and scenarios 5-6 respectively. Note that the camera in the middle of (f) is only used in scenario 6 while scenario 5 leaves a larger blind area between cameras. The black regions indicate that there is no path through these regions, while the white regions represent available paths. Each camera view is represented by a blue triangle. The camera numbers are listed using a red circle. (c) and (d) are examples, where two persons are shown in 8 different indoor and outdoor cameras, highlighting the challenges of working with such networks. These two persons have widely differing appearances in different camera views. The dotted lines represent possible path connections between two camera views.

most popular multi-view datasets for tracking. 8 cameras with overlapping views are used to monitor persons' walking behaviors. There are only three scenes in the tracking subset of the dataset, where each scene only lasts for around 40 seconds. Other datasets are designed for the problem of object re-identification. The Dana36 dataset [11] contains more than 23,000 images depicting 15 persons and 9 vehicles. Both overlapping and non-overlapping scenarios are provided in this dataset. However, as stated in the paper, this dataset is not suitable for tracking because of the specifics of data acquisition (multiple passes). The 3DPeS dataset [2] is another collection designed for the problem of person re-identification. Both of these two datasets lack temporal information, because of which they cannot be used for tracking. There are some multi-camera datasets that are more suited to the CamNeT use-case. The GRID dataset [10] contains 250 pedestrian image pairs taken from 8 disjoint camera views. However, such a multi-camera dataset does not fit into the problem of multi-target tracking since no full video is provided. Instead, only person re-identification can be performed.

Some research papers report multi-target tracking results [3, 4, 7, 8, 9, 13]. None of these papers use the same dataset to evaluate their algorithms. That lack of consistency exposes a clear need for a dataset that can serve as a suitable platform for each of these and future tracking algorithms to be tested for a non-overlapping use-cases. In addition, a common dataset would need to provide a collection of challenges that lie at the frontier of robust tracking system capabilities.

This dataset is more challenging than other non-overlapping multi-camera datasets used in the literature because

1). The number of cameras in the tracking literature is usually between 2 and 5, while we use 5 to 8.

2). Every pair of cameras has more than one path from one to the other as shown in Fig. 1 (d)-(f).

3). Our dataset has both indoor and outdoor scenarios. The lighting conditions and features of each target are significantly different in each camera.

4). The number of targets in each camera can vary from 1 to 10 per frame, often making tracking difficult. In scenario 1 to 4, there are around 10 persons in every scenario that walk a predefined path. A minimum of 20 additional people walk uncontrolled, adding to scene clutter. In scenario 5 to 6, there are around 25 persons in every scenario with significant occlusions. More activities are introduced in scenario 5 to 6, i.e., people talking, group merging, group splitting, long-term occlusion, and etc.

Our contributions are as follows.

(1) This is the first public multi-camera dataset with non-overlapping views which is specially designed for multi-target tracking. Cameras are synchronized across all camera views, and global time information is provided to detect

frame loss.

(2) There are 6 scenarios in which every scenario lasts at least five minutes with 5 to 8 cameras. The videos are rich with person activities. This is different from the dataset used in existing work [7], in which the dataset used is sparse with respect to person activities.

(3) The CamNeT dataset provides single person and group walking behavior across different cameras under both indoor and outdoor scenarios. In each scenario, the paths of around 10 - 25 people are predefined while several unknown persons move through the scene and make multi-target tracking extremely hard.

(4) The detailed annotations for subjects walking predefined paths are provided.

(5) We provide detailed preliminary results with a baseline tracking algorithm.

2. Camera Network Tracking

2.1. Database collection

In the CamNeT data collection procedure, several persons (8-25) in different subsets were asked to follow specific paths in the camera network. These persons either walked alone or in a group. In some cases, subjects would split from one group and join another group. In addition, multiple unknown persons trafficked the data collection areas. The total number of persons in each scenario varied from 25 to 50.

In scenario 1, four indoor cameras and four outdoor cameras were used on a sunny day. The indoor cameras covered most of the corridors as shown in Fig. 1. All the indoor cameras had front/back views of the persons. Thus the persons who were not close to the camera were small within the camera frame. In the outdoor scenarios, there were strong shadows on the ground. Four cameras covered a small part of the courtyard. Different from the indoor camera views, which had one-to-one path connections, the courtyard is large and a person could have different walking choices from one camera view to another. The outdoor cameras had both front/back views and side views of each person. It is noted that sometimes the view of one person could be blocked by another person who was walking together with him/her. In scenarios 2-4, 5 indoor cameras and 3 outdoor cameras were used. We changed some of the camera configurations so that different scenarios could be explored. In scenarios 5 and 6, around 25 persons walked along different paths. We varied the number of cameras; 5 cameras were used in scenario 5 and 6 were used in scenario 6. There are more areas which are not covered by the cameras. Rich person activities are considered in these two scenarios. Persons can walk together, stay together while talking to each other, merge to/split from a group within or outside a camera view, etc. The large amount of unknown behaviors in the blind areas between cameras, the large number of persons, and the heavily cluttered scenes make the provided tracking problem extremely challenging. In each setup approximately 20% to 30% of the open area is covered by active cameras.

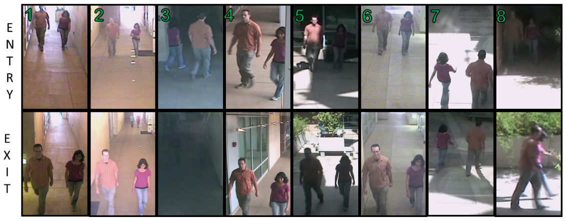


Figure 2. Entry and exit points for each camera for one setup.

Each scenario lasted from 5 to 7 minutes. Though the frame rate for every scenario was 25 frames per second, problems with network communication caused frame loss in one or more cameras. Network communication issues and slightly different start times for video recording between cameras resulted in the problem that every video has different lengths. To solve this problem, our dataset includes global time information. Each frame of every video has a corresponding UTC timestamp. This means that temporal correspondences between cameras can be relied upon, which is required for tracking in multiple cameras. The selected frames can be found in Fig. 3.

2.2. Dataset Characteristics

Compared to existing datasets, CamNeT represents significant challenges. One of the main challenges is varying lighting conditions. Fig. 1 shows the appearance variations under different camera views. Lighting was also subject to change within camera views. The courtyard contained areas of shadow and bright sunlit illumination. Furthermore, persons whose paths were predefined entered each camera's field of view at least twice for scenarios 1-4. However, the direction they faced was not necessarily fixed for each camera. Such wide variations in appearance makes appearance-only tracking methods fail.

The dotted line in Fig. 1 shows possible paths from one camera to another. The camera network represents a complex topology where there exists more than one path between cameras. Therefore, the spatial information between tracklets is relevant, but not necessarily predictive. With this information, typical time gaps between camera views can be estimated, but not solely relied upon for movement prediction. Some representative entry and exit points for every camera are shown in Fig. 2.

Grouping patterns are also variable for this dataset. In scenarios 1-3, persons with planned trajectories stayed in one group or walked alone. However, in scenario 4 there were instances of a person leaving one group and joining another. This is further muddled by the presence of

Table 1. Comparison between different datasets. OV represents overlapping views and NOV denotes non-overlapping views.

	# of camera	OV/NOV	Max # of persons per camera	Highest Resolution	Pers Height (pixels)	Indoor or Outdoor	Max # of persons per scenario
VideoWeb	4-8	OV	8	640 x 480	50-350	outdoor	12
Dana36	36	both	3	2048 x 1536	200-600	both	15
3DPeS	2-8	NOV	≤ 5	704 x 576	50-100	outdoor	unknown
PETS09	4-8	OV	8	768 x 576	80-100	outdoor	30
CamNeT	5-8	NOV	10	640 x 480	50-350	both	39

crowds. In scenarios 5-6, group merge and split events could happen in the blind area between cameras. In each scenario of our dataset, more than 20 persons pass through the scene. In some instances up to 10 persons appear at the same time in one camera view. Since grouping information can change, these scenarios represent the most challenging tracking problems.

To better explain the characteristics of CamNeT, Table 1 is provided to compare our dataset to other camera network datasets. Note that the first three datasets in Table 1 do not suit the purpose of tracking because of the non-availability of temporal information or time synchronization across cameras. The fourth dataset is used for tracking; however it is not designed for non-overlapping views. The proposed dataset is much more suitable for tracking in an non-overlapping camera network.

The resolution for each frame is 640 by 480 pixels. This is nowhere near the best resolution available, however it is not uncommon. The purpose of this dataset is to provide a challenging group of videos that require advanced tracking algorithms to correctly track across cameras. The resolution and consequent size of tracked objects, being 50 to 350 pixels in height, is seen as following the spirit of the challenge. As well, the appearance information for a person can vary greatly within a camera view falling off dramatically at the edges. The lack of detail combined with the other factors present in this dataset requires advanced context models to fill in the gaps where direct observations will fail.

2.3. Annotations

To better test and compare results with this dataset, the annotations of the ground truth of the persons whose walking paths were predefined are provided. The ground truth of a person is expressed by the camera number, the frame number, the person's upper left corner image coordinate (horizontal and vertical coordinates) and the size of the target (width and height). We save every person's ground truth in a text file with the name as the ID of this person. The exact UTC time can be obtained by looking for the timestamps for every frame in every camera. Only when a viable full body appears in the scene do we label the ground truth of this person.

We show a thorough experimental evaluation of the system. We also show how the overall performance decreases when some aspects of the algorithm are removed.

3. Baseline Algorithm

To evaluate the effectiveness of each proposed algorithm, we provide a baseline algorithm considering the spatio-temporal relationships between tracklets. Input to the multi-target tracking system was the collection of recorded videos for a particular time period. We used the detector [6] to generate detection responses for every person and then a basic tracker with particle filter to remove false positives and associate the remaining detections into tracklets for every camera. The problem of how to associate these sets of tracklets and find out the best subset of associations was then broached. Our camera network tracking framework was invoked where a camera to camera feature transformation scheme and the proposed social group model (SGM) across cameras are used. An overview of the system is given in Fig. 4 where the details of each part of the system are given in the sections below.

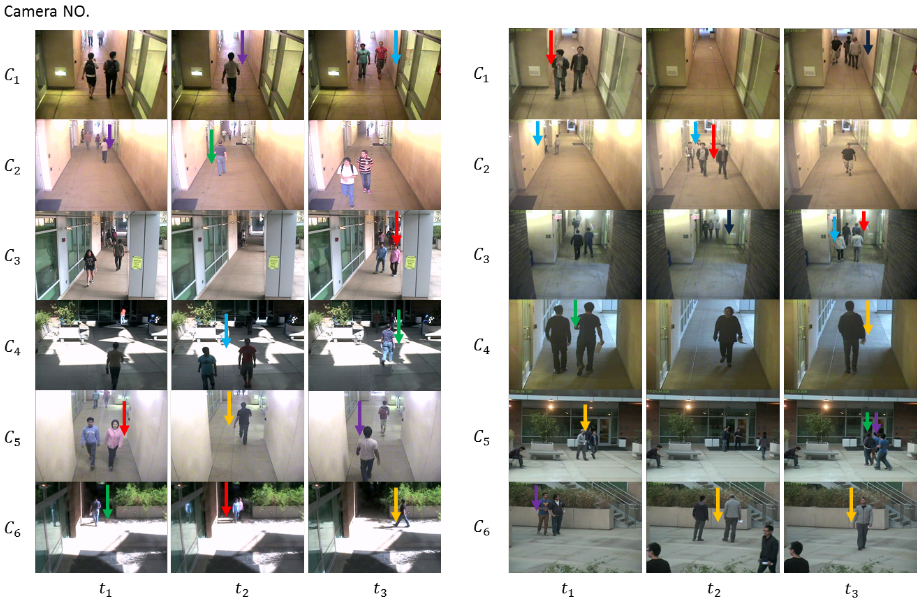
3.1. Inter- and Intra-Camera Tracking

The tracklets generated from a basic tracker in every camera are assumed to be a set of short, reliable tracks. To reduce the high dimension of associations, the first task in a multi-camera tracking framework is to create long, robust single camera tracks (SCTs) for each camera. To realize this goal, features of each tracklet were generated first, and the Bhattacharyya distance was used to calculate the appearance-distance between each feature. We used both appearance and motion information to group tracklets into SCTs for every person in every camera.

The input of the inter-camera tracking system is the output of the intra-camera tracking system, which are a set of long, robust SCTs. Each SCT represents a target in a single camera and the goal of the inter-camera system is to associate all SCTs in a high dimensional space.

3.1.1 Feature Generation

After intra-camera tracking is done, different features of each SCT are generated to better distinguish two persons. We use appearance features in HSV space, HOG features,



(a) (b)

Figure 3. Selected frames of selected cameras from the proposed dataset. (a) and (b) are two scenarios. The horizontal axis represents the time and the vertical axis shows the camera numbers in different scenarios. The time is not synchronized in this presentation because we want to show as many tracks as possible. In (a), the same group or individual is represented by the same color of arrow. For instance, the group consisting of the person in pink and the person in blue shows up in camera 5, 6 and 3 respectively. The features and sizes of them are highly distinguished in these three cameras, especially in C_6 at time t_2. In (b), the scenario is even more challenging than (a). The group denoted by the red arrow in camera 1 and the group denoted by light blue at t_1 merge to a large group in C_2 at t_2. A similar scenario can be found with the green and purple arrow. The three-person group in C_3 is denoted by the dark blue arrow at t_2. However, only two of them can be found in C_1 at t_3. The group with the yellow arrow at t_1 and t_2 splits to two individuals at t_3.

PHOG features and texture features to calculate feature distances.

3.1.2 Feature Transformation across Cameras

In our camera setup, there are both indoor and outdoor scenarios with very different lighting conditions. Therefore, the appearance of the same person might vary widely across cameras. So normalized appearance features are important for reducing the effect of lighting variance. We use the method in [7] to find the linear brightness transfer function (BTF) in color space.

3.1.3 Social Grouping Model

We observe that people often walk with others. Therefore, when people are in groups we can consider the interrelationships between them rather than tracking each person separately. We exploit both the spatial and temporal information between neighboring targets to build a social grouping model (SGM) in one camera. If we are confident for at least one person's association, this increases our confidence for associations made for other people in the same group.

If $\mathcal{X}$ represents a SCT, we calculate the motion similarity between two pairs of SCTs in two cameras C_n and C'_n, which is represented by $\mathcal{X}_i^{C_n}$ and $\mathcal{X}_{i'}^{C'_n}$. We adopt the definition of a group in [16], where a moving group is a collection of people who move at similar speeds and in similar directions. A group is created when two or more people walk together for enough time within a distance threshold. At a given time t, let τ be defined as

$$\tau = \min\{w(\mathcal{X}_i^{C_n}), h(\mathcal{X}_i^{C_n}), w(\mathcal{X}_j^{C_n}), h(\mathcal{X}_j^{C_n})\} \qquad (1)$$

where $w(\mathcal{X}_i^{C_n})$ and $h(\mathcal{X}_i^{C_n})$ are the width and height of the bounding box of SCT i in at time t. If the the distance between two SCTs $d(\mathcal{X}_i^{(T)}, \mathcal{X}_{i'}^{(T')})$ satisfies the following condition

$$d(\mathcal{X}_i^{(T)}, \mathcal{X}_{i'}^{(T')}) = ||\mathcal{X}_i^{C_n} - \mathcal{X}_j^{C_n}|| < \alpha \cdot \tau \qquad (2)$$

with α be a control parameter and (T) be a time window T, we can say that the tracklet $\mathcal{X}_i^{C_n}$ and $\mathcal{X}_j^{C_n}$ are in the same group in camera C_n if the condition holds for 80% of time. We will still find a grouping function Φ which represents if two SCTs belong to the same group under two different camera views. The overall algorithm of SGM across cameras is given in Algorithm 1.

In Algorithm 1, θ is a controlled threshold. Φ is a grouping cue matrix, where $\Phi_{i,j} = 0$ means tracklets i and j are

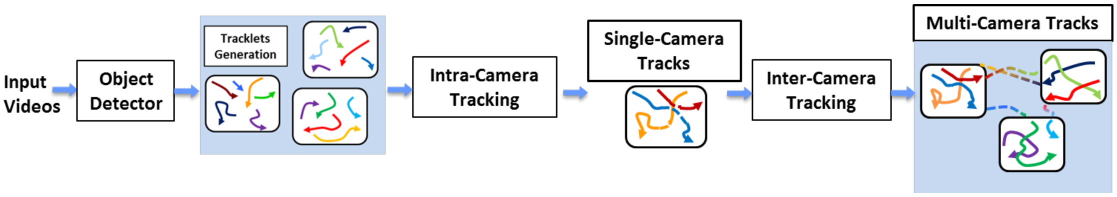


Figure 4. Overview of baseline camera network tracking algorithm.

```
Algorithm 1 Overview of Social Grouping Model
Input:
    -SCTs from the intra-camera tracking scheme (Assuming $p$ SCTs in $C_n$ and $q$ SCTs in $C'_n$);
    -A zero initialized grouping matrix $\Phi$, the size of which is $(p+q)\times(p+q)$;
Build a matrix $G_1$ which is $p\times p$ and another matrix $G_2$ which is $q\times q$. These two matrices are to label if two SCTs are close to each other for enough time or not;
Find pairs of SCTs from the same camera which satisfy Equ. (2) in 80% of the time windows $(T)$ and $(T')$ individually in the corresponding position of $G_1$ and $G_2$;
for $i$ from 1 to $p$ do
    for $i'$ from 1 to $q$ do
        if $d(\mathcal{X}_i^{(T)}, \mathcal{X}_{i'}^{(T')}) < \theta$ then
            check if there is at least one $j$ and one $j'$ which make $G_1(i,j)=1$ and $G_2(i',j')=1$;
            if YES then
                if $E_v(j,j')=1$ and $E_p(j,j')<\delta_p$ then
                    $\Phi(\mathcal{X}_i^{(T)}, \mathcal{X}_{i'}^{(T')}) = -1$;
                    $\Phi(\mathcal{X}_j^{(T)}, \mathcal{X}_{j'}^{(T')}) = -1$;
                end if
            end if
        end if
    end for
end for
Output:
    The grouping matrix $\Phi$, where $\Phi(i,i')=-1$ means the two SCTs in different time windows belong to a same group and $\Phi(i,i')=0$ means otherwise;
```

not in the same group in the given two time windows (T) and (T'), while $\Phi_{i,j}=-1$ means they are. Note that $\Phi_{i,j}$ does not represent two tracklets in the same time window; instead it represents two tracklets in different time windows. d represents the feature distance between two tracklets. In this algorithm, if an element in the matrix G_1 or G_2 equals to 1, this means that the overlapped part of the two tracklets are very close to each other and these two tracklets can be viewed as belonging to the same group.

3.2. Tracking Algorithm in a Non-Overlapping Camera Network

The overall camera network tracking system is encapsulated in the optimization of an energy function shown in Fig. 4. The goal of the energy function is to combine different features of SCTs, which are generated by the intra-camera tracking module, and then compare each SCT in order to find a one-to-one mapping between each SCT. This one-to-one mapping is then used to generate the final track for the wide area. Suppose there are N cameras and the camera set is $\mathbf{C}=\{C_1, C_2, ..., C_N\}$. If we use L to represent if two SCTs in different cameras can be associated or not, then

$$L(\mathcal{X}_i^{C_n}, \mathcal{X}_{i'}^{C'_n}) = \begin{cases} 1, & \text{if } \mathcal{X}_i^{C_n} \rightarrow \mathcal{X}_{i'}^{C'_n}, \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $\mathcal{X}_i^{C_n}$ represents the i^{th} SCT in camera view C_n and "$\rightarrow$" denotes that the two tracklets can be associated. We define the overall problem of multi-camera tracking as

$$arg\min_{L} \sum_{i,i'} L(\mathcal{X}_i^{C_n}, \mathcal{X}_{i'}^{C'_n}) \cdot D(\mathcal{X}_i^{C_n}, \mathcal{X}_{i'}^{C'_n}) \quad (4)$$

where D is a distance function between two SCTs.

However, there are a couple of constraints which may reduce the number of possible associations. For example, grouping behavior is an important cue we observe when people are walking together. Also, similar to [3], prior knowledge of camera network topology is another important cue for intra-camera tracklet association. The prior knowledge of topology includes both spatial and temporal cues. For the spatial cues, we can know if it is possible for a person walk from one camera to another. Temporal constraint can tell us how much time it typically takes for a person to walk from one camera to another. Assuming we detected every person in every camera, if we use U to represent the location adjacency between C_n and C'_n, then

$$U(C_n, C'_n) = \begin{cases} 1, & \text{if } C_n \rightarrow C'_n, \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where $C_n \rightarrow C'_n$ means these two cameras have location adjacency.

If the mean transition time from C_n to C'_n is $\bar{t}$ and the standard deviation for each transition time is $\sigma(t)$, the temporal transition probability V is a Gaussian function

$$V(C_n, C'_n) = G(\bar{t}, \sigma(t)) \tag{6}$$

The overall transition probability between two cameras is:

$$P_{Tran}(C_n, C'_n) = U(C_n, C'_n) \cdot V(C_n, C'_n) \tag{7}$$

Adding both group constraints and the topology constrains to the overall energy function for a inter-camera system, it becomes to

$$\begin{aligned} arg \min_{L} \sum_{i,i'} L(\mathcal{X}_i^{C_n}, \mathcal{X}_{i'}^{C'_n}) \cdot D(\mathcal{X}_i^{C_n}, \mathcal{X}_{i'}^{C'_n}) + \\ \lambda_2 \cdot \sum_{i,i'} \Phi(\mathcal{X}_i^{C_n}, \mathcal{X}_{i'}^{C'_n}) \\ \text{s.t.} \quad P_{Tran}(C_n, C'_n) = c \end{aligned} \tag{8}$$

where c is a constant between 0 and 1.

As mentioned above, to solve Equ. (8), D and Φ have to be computed. D is computed by a predefined distance function where Bhattacharyya distance is used in this work and Φ can be computed as determined in Sec. 3.1.3.

4. Preliminary Experimental Results

Our evaluation metrics in camera network tracking are based on ([14]).

1). Tracking length (TL): Percentage of completed trajectory which was correctly tracked.

2). Crossing fragments (XFrag): The number times that there is a link between two tracks within a specified tolerance, but missing in the tracking results.

3). Crossing ID-switches (XIDS): The total number of times that there is no link between two tracks in two cameras within a specified tolerance of the ground truth trajectories, but one or more links exist in the tracking results.

In our experiments, we assume that if the tracking results are within 0.5 meters of the ground truth, we consider the association between two tracklets is correct; otherwise it is wrong. We test our tracking system on two subsets of CamNeT, which cover the two different scenarios. The step-by-step results of scenario 1 are listed, where different combinations of models from the baseline algorithms are tested. The final tracking results of scenarios 2-6 are also provided.

In our experiments on scenario 1, we generate 1456 tracklets and 322 SCTs for all the 8 cameras using our basic tracker. Table 2 shows the tracking results of scenario 1. In order to demonstrate the significance of each model in our algorithm, we compare our results with the state-of-art method in [13]. We also consider the SGM in the implementation for fair comparison. The results show that when SGM is applied, the numbers of XIDS and XFrag reduce. Moreover, both temporal (i.e. the walking time from one camera to another) and spatial constraints (i.e. if a walking path exists between two camera views) are applied when we implement our algorithm. We take out one or both of these two constraints and show the importance of the effect of the topology.

Table 2. Tracking results of scenario 1, where "t-constraints" denotes the temporal constraints, "s-constraints" denotes the spatial constraints and 'st-constrains" represents the spatio-temporal constraints. The first row shows the results obtained using the method in [13]. The rest of the rows show results for different variants of the proposed method. The several constraints with/without which the proposed method is run are described in the first column.

	TL	**XFrag**	**XIDS**
Method in [13]	82.8%	24	23
Without SGM	84.1%	27	20
Without t-constraints	72.2%	21	75
Without s-constraints	56.6%	22	102
Without st-constraints	43.9%	18	156
With SGM and st-constraints	84.3%	27	15

Fig. 5 shows the tracking results over the data collection period. Each row represents the data collected for a particular camera, while each column represents the data collected at a specific time. The boxed individuals in each scene represent people being tracked. For groups of people determined to be walking together, the same color box is used to represent the pair. From one time instant to another, box color remains constant for the same people when correct associations are made within and between cameras.

The inter camera tracking results of scenario 2 to 6 are listed in Table 3. We use spatio-temporal constraints when reporting our results.

Table 3. Tracking results of scenarios 2 to 6 (S2-S6).

	S2	S3	S4	S5	S6
TL	85.0%	78.9%	77.3%	70.0%	75.0%
XFrag	29	36	36	52	40
XIDS	23	26	32	44	34

5. Conclusions

In this paper, we provide a new non-overlapping multi-camera dataset (CamNeT) for tracking. This dataset has 5 to 8 non-overlapping cameras, which cover around 20% to 30% of the open area. Due to the lighting conditions variations and crowded scenarios, this dataset is very challenging and can be seen as a standard dataset to work with. We also present a baseline camera network tracking system. We show preliminary results on our datasets which can be compared against any other methods.

References

[1] E. Auvinet, C. Rougier, J. Meunier, A. St-Amaud, and J. Rousseau. Multiple cameras fall dataset. Technical report 1350, DIRO - Université de Montréal, 2010.

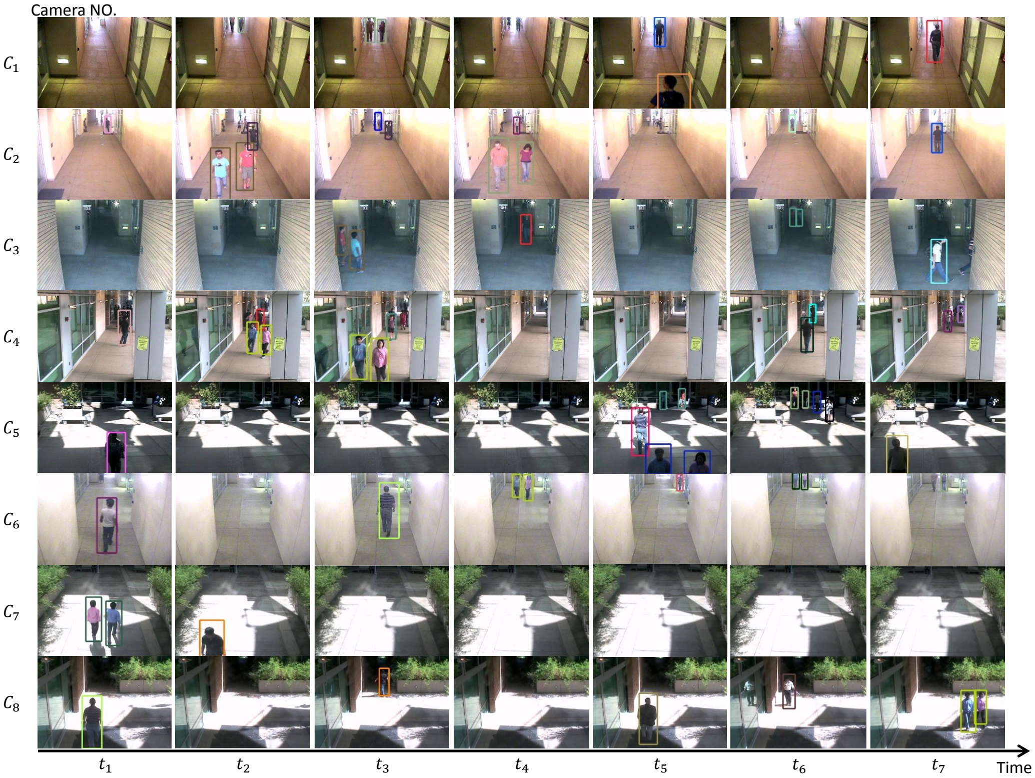


Figure 5. Tracking result of scenario 1. Each row is the view from a different camera. Each column is a snapshot from all the cameras at a particular time instant. Bounding boxes of the same color from one time instant to the next represent re-associated targets. Bounding boxes of the same color within camera views represent a collection of people recognized as a group.

[2] D. Baltieri, R. Vezzani, and R. Cucchiara. 3DPeS: 3D people dataset for surveillance and forensics. In *ACM Workshop on Multimedia Access to 3D Human Objects*, 2011.

[3] K. Chen, C. Lai, Y. Hung, and C. Chen. An adaptive learning method for target tracking across multiple cameras. In *CVPR*, 2008.

[4] X. Chen, K. Huang, and T. Tan. Direction-based stochastic matching for pedestrian recognition in non-overlapping cameras. In *ICIP*, 2011.

[5] G. Denina, B. Bhanu, H. Nguyen, C. Ding, A. Kamal, C. Ravishankar, A. K. Roy-Chowdhury, A. Ivers, and B. Varda. Videoweb dataset for multi-camera activities and non-verbal communication. Distributed Video Sensor Networks, Springer, 2011.

[6] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based model. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2007.

[7] A. Gilbert and R. Bowden. Tracking objects across cameras by incrementally learning inter-camera color calibration and patterns of activity. In *ECCV*, 2006.

[8] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, 2008.

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

[10] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *CVPR*, 2009.

[11] J. Per, V. S. Kenk, R. Mandeljc, M. Kristan, and S. Kovacic. Dana36: A multi-camera dataset for object identification in surveillance scenarios. In *IEEE Conf. on Advanced Video and Signal-Based Surveillance*, 2012.

[12] S. Singh, S. A. Velastin, and H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *2nd Workshop on Activity Monitoring by Multi-camera Surveillance Systems*, 2010.

[13] B. Song and A. K. Roy-Chowdhury. Robust tracking in a camera network: A multi-objective optimization framework. *IEEE journal of Selected Topics in Signal Processing*, 2(4):582–596, 2008.

[14] ICPR 2012 Contest. People tracking in wide baseline camera networks. `http://www.wide-baseline-camera-network-contest.org/?page_id=50`.

[15] PETS 2009. PETS 2009 benchmark data. `http://www.pets2009.net/`.

[16] B. Yang and R. Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *CVPR*, 2012.