REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
04-12-2014	FINAL	07/29/2009-09/30/2014

4. TITLE AND SUBTITLE

Wide-Area Cooperative Biometric Tagging, Tracking and Locating in a Multimodal Sensor Network

5a. CONTRACT NUMBER

N00014-09-C-0388

5b. GRANT NUMBER

N/A

5c. PROGRAM ELEMENT NUMBER

N/A

6. AUTHOR(S)

Amit Roy-Chowdhury, Bir Bhanu, Tim Faltemier

5d. PROJECT NUMBER

N/A

5e. TASK NUMBER

N/A

5f. WORK UNIT NUMBER

N/A

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of California, Riverside

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Public

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report highlights the fundamental contributions that this project has made with respect to the problem of multi-target tracking in a large camera network. Details of the algorithms and results on different datasets are provided. The camera network tracking code will be provided on a disk.
The following are the sections of this report.
1. Camera Network Tracking. This is the main algorithm for tracking in a camera network. The code corresponds to this algorithm. The code has also been provided to Progeny for use in other projects. Detailed experimental results are shown using this framework.
2. Multi-Target Tracking. This is a method to track multiple targets in single and

15. SUBJECT TERMS

camera network, tracking, person detection, reidentification

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Amit Roy-Chowdhury
U	U	U	UU	25	19b. TELEPHONE NUMBER *(Include area code)* 951-827-7886

Reset

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
04 DEC 2014		**29-07-2009 to 30-09-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Wide-Area Cooperative Biometric Tagging, Tracking and Locating in a Multimodal Sensor Network	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California, Riverside, Riverside,,CA,92521**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**26**	

Wide Area Scene Analysis (WASA) Final Report

Project Name and Contract Number: N00014-09-C-0388 - Wide-Area Cooperative Biometricrragging, Tracking and Locating in a Multimodal Sensor Network

PIs: Amit Roy-Chowdhury, Bir Bhanu (University of California, Riverside); Tim Faltemier (Progeny Systems Corp.)

December 4, 2014

This report highlights the fundamental contributions that this project has made with respect to the problem of multi-target tracking in a large camera network. Details of the algorithms and results on different datasets are provided. The camera network tracking code will be provided on a disk.

The following are the sections of this report.

1. Camera Network Tracking. This is the main algorithm for tracking in a camera network. The code corresponds to this algorithm. The code has also been provided to Progeny for use in other projects. Detailed experimental results are shown using this framework.

2. Multi-Target Tracking. This is a method to track multiple targets in single and multiple views. It is an essential step in the camera network tracking framework.

3. Person Detection. This section describes the approach to do person detection under varying illumination conditions, object pose, and partial occlusion. It is the first step towards the camera network tracking method.

4. Consistent Person Re-Identification. This section describes a method to obtain consistent results in re-identifying people across large variations of lighting, appearance and pose. The idea is that if a person is identified to be similar in two camera

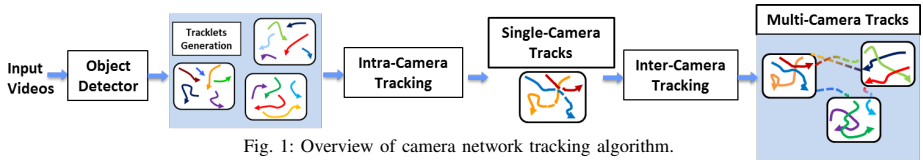


Fig. 1: Overview of camera network tracking algorithm.

pairs like (1,2) and (2,3), they should also be identified in the pair (2,3). This improves the overall re-identification results significantly.

I. Camera Network Tracking

In this section, we describe the camera network tracking algorithm considering the spatio-temporal relationships between tracklets. We convert the tracking problem to the tracklet association problem and find out the best subset of associations. An overview of the system is given in Fig. 1 where the details of each part of the system are given in the sections below. To test different challenges in the problem of camera network tracking, we recorded a new camera network tracking (CamNeT) dataset, which is more challenging than existing publicly available datasets. In CamNeT, the cameras comprise part of an actual surveillance system distributed along the corridors and open courtyard of a building. We evaluated the tracking algorithms [14] on this dataset.

A. CamNeT Dataset

In the CamNeT data collection procedure, several persons (8-25) in different subsets 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 person in each scenario varied from 25 to 50. There are four scenarios in CamNeT. We use scenario 1 as an example to show the properties of CamNeT. Four indoor cameras and four outdoor cameras were used on a sunny day. The indoor cameras covered most of the corridors. All the indoor cameras had front/back views of the persons. The persons who were not close to

Fig. 2: Entry and exit points for each camera for one camera setup.

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. Sometimes the view of one person could be blocked by another person who was walking together with him in side view. Approximately 20% to 30% of the open area is covered by active cameras. In Fig. 2, some representative entry and exit points for every camera are shown, where typical time gaps between camera views can be estimated.

B. Multi-Camera Tracking Algorithm With Social Grouping Model

Our tracking algorithm considers 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 generate detection responses for every person and then apply a tracker with particle filter to remove false positives and associate the remaining detections into tracklets for 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 single camera tracks (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. After intra-camera tracking is done,

different features of each SCT are generated to better distinguish two persons. We use appearance features, Histogram of Oriented Gradients (HOG) features, Pyramid Histogram of Oriented Gradients (PHOG) features and texture features to calculate feature distances. Besides, 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 find the linear brightness transfer function (BTF) in color space. Then, when people are in groups we can consider the inter-relationships 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 $\mathcal{X}_{i'}^{C'_n}$. 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

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

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

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

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.

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 Φ, 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 Φ, 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;

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 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 overlapping part of the two tracklets are very close to each other and these two tracklets can be viewed as belonging to the same group.

The overall inter-camera tracking system is encapsulated in the optimization of

an energy function shown in Fig. 1. 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. Suppose there are N cameras and the camera set is $\mathbf{C} = \{C_1, C_2, ..., C_N\}$. We use L to represent if two SCTs in different cameras can be associated or not, then

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

where $\mathcal{X}_i^{C_n}$ represents the i^{th} SCT in camera view C_n. We define the overall problem of multi-camera tracking as

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

where D is a distance function between two SCTs.

However, there are constraints which may reduce the number of possible associations, e.g., grouping behavior and prior knowledge of camera network topology. The prior knowledge of topology includes both spatial and temporal cues. If we use U to represent the location adjacency between C_n and C'_n, then

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

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

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

$$\begin{aligned} arg\min_{L} \sum_{i,i'} L(\mathcal{X}_i^{C_n}, \mathcal{X}_{i'}^{C'_n}) \cdot D(\mathcal{X}_i^{C_n}, \mathcal{X}_{i'}^{C'_n}) + \lambda_2 \cdot \sum_{i,i'} \Phi(\mathcal{X}_i^{C_n}, \mathcal{X}_{i'}^{C'_n}) \\ \text{s.t.} \quad P_{Tran}(C_n, C'_n) = c \end{aligned} \tag{6}$$

where P_{Tran} is the transition probability and c is a constant between 0 and 1.

C. Experimental Results

We test our tracking system on two subsets of CamNeT, which cover the two different scenarios. In our evaluation metrics, TL represents tracking length, XFrag represents crossing fragments, and XIDS represents crossing ID-switches. 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.

In our experiments on scenario 1, we generate 1456 tracklets and 322 SCTs for all the 8 cameras using our basic tracker. Table I 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.

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

TABLE I: Tracking results of scenario 1. 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.

Fig. 3 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

	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
TL	85.0%	78.9%	77.3%	70.0%	75.0%
XFrag	29	36	36	52	40
XIDS	23	26	32	44	34

TABLE II: Tracking results of scenarios 2 to 6.

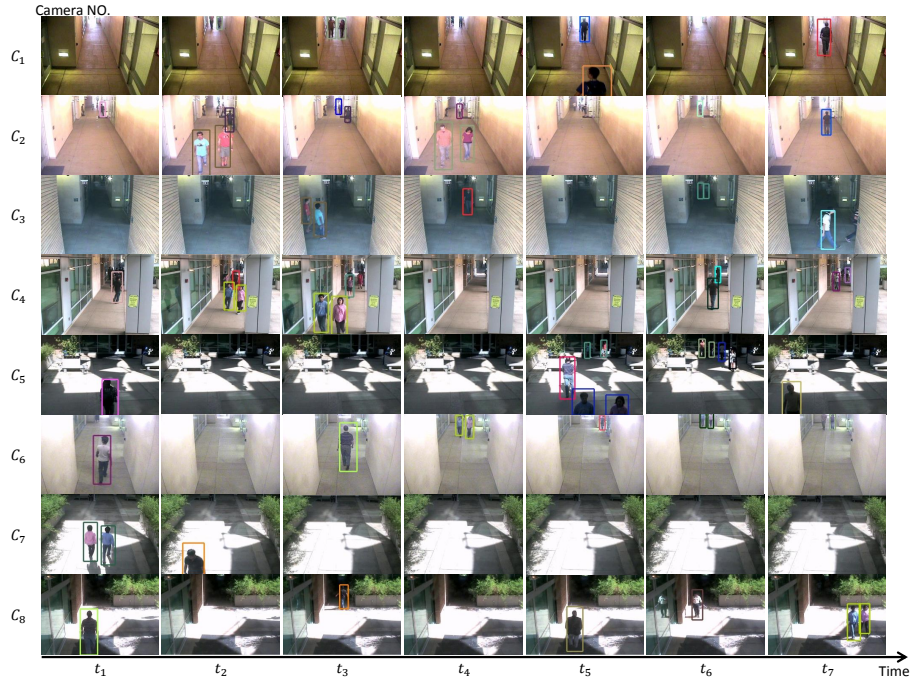


Fig. 3: Tracking result of scenario 1. Each row is the view from a different camera. Each column is a snapshot from all the cameras at a particular time instant. Bounding boxes of the same color from one time instant to the next represent re-associated targets. Bounding boxes of the same color within camera views represent a collection of people recognized as a group.

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

II. Multi-Target Tracking

We significantly improve tracking accuracy and efficiency for multi-target tracking in real-world surveillance cameras. In the following, we briefly introduce two proposed

methods: elementary grouping model [5] in single camera setup and reference set based appearance model [6] in multi-camera setup. Note that the elementary grouping model in this section is different from the social grouping model in Sec. I-B. In the elementary grouping model, we mainly consider the case of single camera setup. In a single camera view, all the group changes, e.g. a person leaving a group and joining another group, can easily be observed which makes building graphs of groups feasible. However, in the social grouping model in Sec. I-B, we mainly consider the case of multi-camera setup, where group changes in the blind areas may not be captured by the cameras. The grouping graph that is built in this section is therefore not suitable for the multi-camera scenario.

A. Tracking with Elementary Grouping Model

The widely used data association-based tracking methods are likely to fail under challenging conditions where appearance or motion of the target changes abruptly and drastically. Unlike most existing tracking approaches that use only low level information (e.g., time, appearance, and motion) to build the affinity model and consider each target as an independent agent, we learn online social grouping behavior to provide additional information for producing more robust tracklets affinities. An elementary grouping model is proposed to construct a grouping graph where each node represents a pair of tracklets that form an elementary group (a group of two targets) and each edge indicates the connected two nodes (elementary groups) have at least one target in common. The group trajectories of any two linked nodes are used to estimate the probability of the other target in each group being the same subject. The elementary grouping model is summarized in Figure 4. The size of a group may change dynamically as people join and leave the group, but a group of any size can be considered as a set of elementary groups. Therefore, focusing on finding elementary groups instead of the complete group makes our approach capable of modelling flexible group evolvement in the real world. Note that the social group

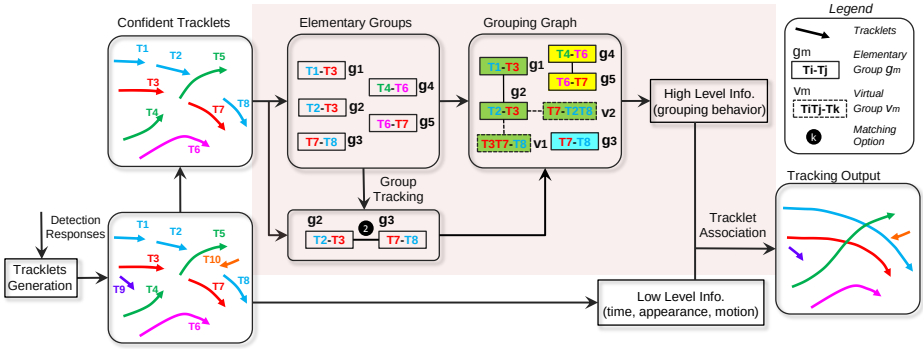


Fig. 4: Block diagram of our tracking system. Tracklets with the same color contain the same target. Best viewed in color.

Method	MT	ML	Frag	IDS	Time
Baseline Model 1	74.7%	6.7%	11	12	1.5s
Baseline Model 2	78.7%	6.7%	10	8	4.2s
GBM Model [12]	89.3%	2.7%	7	5	50s
Our Model	90.7%	2.7%	6	5	4.6s

TABLE III: Comparison of tracking results on CAVIAR dataset. The number of trajectories in ground-truth (GT) is 75.

in this report refers to a number of individuals with correlated movements and does not indicate a group of people who know each other.

We evaluate our approach on two widely used public single-camera pedestrian tracking datasets: the CAVIAR dataset [1] and the TownCentre dataset [4]. The following metrics are used for performance comparison: the number of trajectories in ground-truth (GT), the ratio of mostly tracked trajectories (MT), the ratio of mostly lost trajectories (ML), the number of fragments (Frag) and ID switches (IDS). We compare our approach with the basic affinity model (Baseline Model 1), elementary grouping model without group tracking (Baseline Model 2) and the Grouping Behavior model (GBM). Results in Table III and Table IV suggest that our approach provides better performance and is much more efficient computationally compared with state-of-the-art method.

B. Tracking with Reference Set Based Appearance Model

Tracking multiple targets in non-overlapping cameras is challenging since the observations of the same targets are often separated by time and space. There might be

Method	MT	ML	Frag	IDS	Time
Baseline Model 1	76.8%	7.7%	37	60	350s
Baseline Model 2	78.6%	6.8%	34	46	457s
GBM Model [12]	83.2%	5.9%	28	39	4861s
Our Model	85.5%	5.9%	26	36	465s

TABLE IV: Comparison of tracking results on TownCentre dataset. The number of trajectories in ground-truth (GT) is 220.

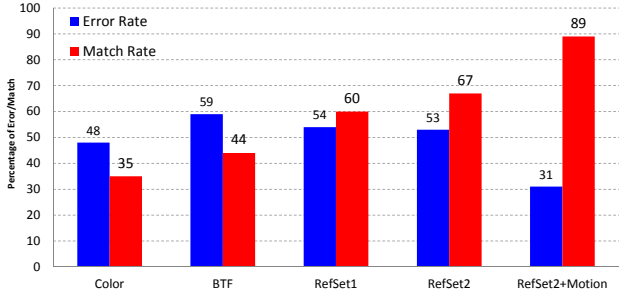


Fig. 5: Comparison of the proposed method and other baseline models on MultiCam dataset.

significant appearance change of a target across camera views caused by variations in illumination conditions, poses, and camera imaging characteristics. Consequently, the same target may appear very different in two cameras. Therefore, associating tracks in different camera views directly based on their appearance similarity is difficult and prone to error. In most previous methods the appearance similarity is computed either using color histograms or based on pre-trained Brightness Transfer Function (BTF) that maps color between cameras. However, BTF is not suitable for a camera network that has a large *within* camera illumination change. To address this problem, we propose a novel reference set based appearance model to improve multi-target tracking in a network of non-overlapping cameras. Contrary to previous work, a reference set is constructed for a pair of cameras, containing subjects appearing in both camera views. For track association, instead of directly comparing the appearance of two targets in different camera views, they are compared indirectly via the reference set. Besides global color histograms, texture and shape features are extracted at different locations of a target, and AdaBoost is used to learn the discriminative power of each feature.

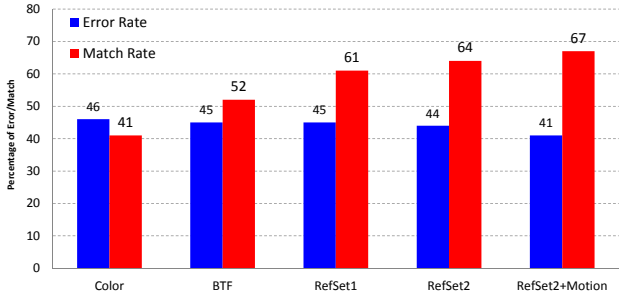


Fig. 6: Comparison of the proposed method and other baseline models on VideoWeb dataset.

We apply our reference set based appearance model with weighted features (RefSet2) on the test set, and introduce three baseline models for comparison: (1) using Bhattacharyya distance of holistic color histograms directly to measure the appearance similarity (Color); (2) generating appearance model based on the BTF model in [10] (BTF); (3) our proposed reference set based appearance model with only holistic color histograms as appearance feature (RefSet1). Two metrics are used for evaluation: $ErrorRate = \frac{Error}{N_{result}}$, $MatchRate = \frac{Match}{N_{GT}}$, where $Error$ and $Match$ are the number of incorrectly and correctly associated track pairs in the result, N_{result} and N_{GT} are the number of track associations in the result and the ground-truth, respectively. The tracking results on MultiCam dataset (captured on the UCR campus) and VideoWeb dataset [8] demonstrate that when using the proposed method, we achieve the highest match rate and the lowest error rate compared to all the baseline models. When a motion model that measures the walking direction of a target is integrated into the tracking system (RefSet2+Motion), the error rates are reduced and more track pairs are correctly associated. The comparison between RefSet1 and RefSet2 demonstrates that by using features of various types and extracted at different locations we can get more information than using global color histograms only, as they capture the appearance information that is overlooked by color histograms. Note that the ViedoWeb dataset is originally designed for complex real-world activity recognition, participants in this dataset have more non-linear motion and heavy interactions than

that in the MutliCam dataset. Therefore, the overall tracking performance on this dataset is not as good as that on the MultiCam dataset. Also, non-linear motion and interactions among individuals make it difficult to predict accurate motion direction of a target. Thus, after integrating motion model with RefSet2, the improvement on both error rate and match rate is small.

III. Person Detection

We significantly increased both the speed and accuracy of our object detectors and specifically our face/body detection through a variety of improvements. This in turn provides enhanced capabilities for tracking of non-cooperative targets in challenging environments. In the following, we briefly summarize the algorithmic improvements.

A. Crosstalk Cascades

The primary algorithmic improvement that we focused on was the implementation of a crosstalk cascade [9] within the detection framework. In a traditional scanning approach, evaluation of windows proceeds independently. This is suboptimal as detector responses at nearby locations tend to be highly correlated. Instead, a crosstalk cascade exploits these correlations by allowing adjacent detectors to communicate, thus coupling evaluation of nearby windows. This allows the detector to rapidly discard regions that are unlikely to contain the object of interest, and focus on aggressive evaluation of promising image regions (Fig. 7). In addition, this technology can also be leveraged to speed up multiple unrelated detectors, for instance, detectors for different types of objects that are nonetheless correlated at some level, e.g., in terms of their similarity of distinguishing image features.

B. Algorithmic Optimization

In addition to the integration of a crosstalk cascade, we also identified some other inefficiencies in the code that could be exploited to further speed up the execution,

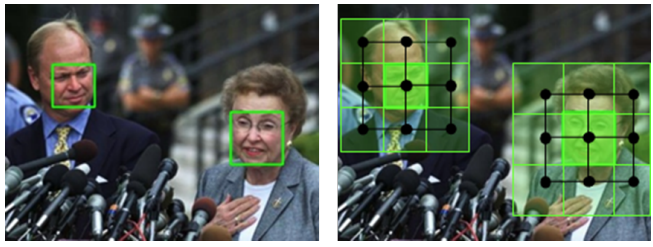


Fig. 7: Left: The face detections obtained via cross-talk cascades, an innovative type of classifier that leverages information from adjacent scanning windows in order to rapidly discard unlikely regions and produce robust detections in promising regions. Right: Illustration of this process, with promising regions highlighted in green.

as well as several other minor changes to improve detection performance. An empirical comparison of our face detection algorithm against other recently published approaches demonstrates we achieve state-of-the-art detection accuracy (Fig. 8) but at extremely high throughput.

In fact, our results found that the use of a crosstalk cascade in our detection framework combined with the other improvements results in an order-of-magnitude speedup in the overall throughput of the detector. Recent experiments with the optimized detector demonstrate that it is capable of processing at speeds exceeding 100 frames per second (fps) on 640x480 video or 30+ fps (faster than real-time) on 1080x720 full-HD video.

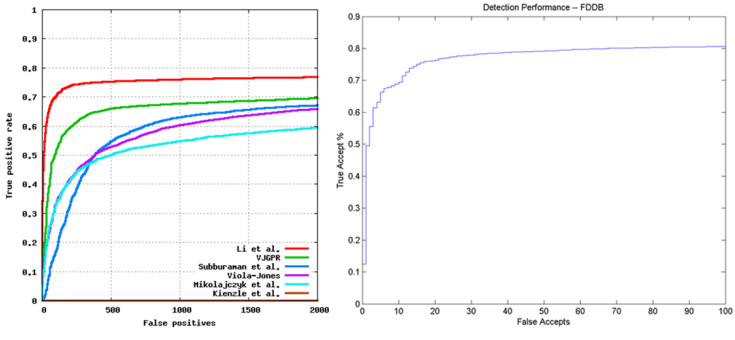


Fig. 8: Left: Face detection performance of a number of recent algorithms on the well-known FDDB dataset (source: http://vis-www.cs.umass.edu/fddb/results.html). Right: Performance of our new face detector, which is competitive with the current state-of-the-art at higher throughput.

Fig. 9: Examples of fiducial points identified by our new landmark model across a wide range of poses including pitch, yaw, and roll.

C. Dense Landmarks

Finally, we have developed an improved landmark model which finds robust fiducial points on the face across a wide range of poses and environmental conditions (Fig. 9). The landmarks are somewhat denser than those obtained by our previous approach and enable improved pose-estimation and face tracking through changes in pose.

Features extracted at each of these landmarks are also used in our algorithm, a feature-aided tracking and object recognition framework that is robust to occlusions and objects disappearing from the field of view. We are still evaluating the relative value of the different landmarks but have already seen improvements in tracking performance, especially in extreme poses (greater than 45 degrees).

IV. Person Re-identification

We addressed person re-identification in a camera network by exploiting the requirement of consistency of re-identification results. The proposed method not only boosts camera pairwise re-identification performance but also can handle a largely unaddressed problem of matching variable number of persons across cameras. Even if the re-identification accuracy for each camera pair is high, it can be inconsistent if results from 3 or more cameras are considered. Thus, in person re-identification across a camera network, multiple paths of correspondences may exist between targets from any two cameras, but ultimately all these paths must point to the same correspondence maps for each target in each camera. An example scenario is shown in Fig. 10.

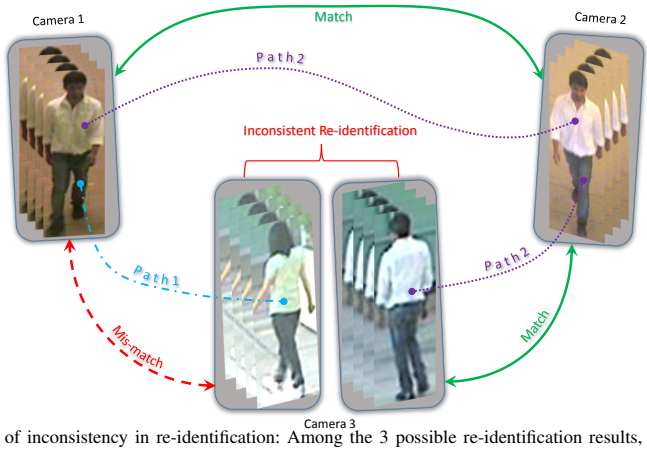


Fig. 10: Example of inconsistency in re-identification: Among the 3 possible re-identification results, 2 are correct. The match of the target in camera 1 to camera 3 can be found in two ways. The first one is the direct pairwise re-identification result between camera 1 and 3 (shown as 'Path 1'), and the second one is the indirect re-identification result in camera 3 given via the matched person in camera 2 (shown as 'Path 2'). The two outcomes do not match and thus the re-identification of the target across 3 cameras is not consistent.

We propose a novel re-identification scheme across multiple cameras by incorporating the consistency requirement. We show that the consistency requirement not only makes the interpretation of re-identification more meaningful, but also makes the pairwise re-identification accuracy high. Since consistency across the camera network is the building block of the proposed method, we term this as the 'Network Consistent Re-identification' (NCR) strategy.

To achieve a consistent and optimal re-identification, we pose the problem of re-identification as an optimization problem that minimizes the global cost of associating pairs of targets on the entire camera network constrained by a set of consistency criteria. Our formulation picks the assignments for which the total similarity of all matches is the maximum, as well as the constraint that there is no inconsistency in the assignment among any pair of cameras given any other intermediate camera. The resulting optimization problem is translated into a binary integer program (IP).

The proposed method is also generalized to a more challenging scenario in person re-identification when all persons are not present in all the cameras. With our formulation we show that we can address this largely unaddressed challenge of multicamera

person re-identification by employing a reward for true negatives (no association for an isolated person in one camera). We compare the performance of our approach to state-of-the-art person re-identification methods using a publicly available benchmark dataset - WARD [11] (3 cameras), and a new 4 camera dataset, RAiD [7] introduced by us.

A. Network Consistent Re-identification Framework

The Network Consistent Re-identification (NCR) method starts with the camera pairwise similarity scores between the targets. The camera pairwise similarity score is generated by learning the way features get transformed between cameras. The notation and terminologies associated to this framework are described next.

Let there be m cameras in a network. For simplicity we, first, assume, that the same n person are present in each of the cameras. In section IV-B1 we will extend the formulation for a variable number of targets.

1. Node: The i^{th} person in camera p is denoted as $\mathcal{P}_i^p$ and is called a 'node'. The similarity score between two nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ is denoted as $c_{i,j}^{p,q}$.

2. Assignment variable: We need to know the association between the persons $\mathcal{P}_i^p$ and $\mathcal{P}_j^q, \forall i,j = \{1, \cdots n\}$ and $\forall p,q = \{1, \cdots m\}$. The association between two nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ is expressed by the variable $x_{i,j}^{p,q}$. $x_{i,j}^{p,q}$ is a binary variable which takes the value 1 if $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ are the same targets or 0 otherwise.

3. Edge: An 'edge' between two nodes $\mathcal{P}_i^p$, and $\mathcal{P}_j^q$ from two different cameras is a probable association between the i^{th} person in camera p and the j^{th} person in camera q. There are two attributes connected to each edge. They are the similarity score $c_{i,j}^{p,q}$ and the association value $x_{i,j}^{p,q}$.

4. Path: A 'path' between two nodes $(\mathcal{P}_i^p, \mathcal{P}_j^q)$ is a set of edges that connect $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ without traveling through a camera twice. A path between $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ can be represented as the set of edges $e(\mathcal{P}_i^p, \mathcal{P}_j^q) = \{(\mathcal{P}_i^p, \mathcal{P}_a^r), (\mathcal{P}_a^r, \mathcal{P}_b^s), \cdots (\mathcal{P}_c^t, \mathcal{P}_j^q)\}$, where $\{\mathcal{P}_a^r, \mathcal{P}_b^s, \cdots \mathcal{P}_c^t\}$ are the set of intermediate nodes on the path between $\mathcal{P}_i^p$ and

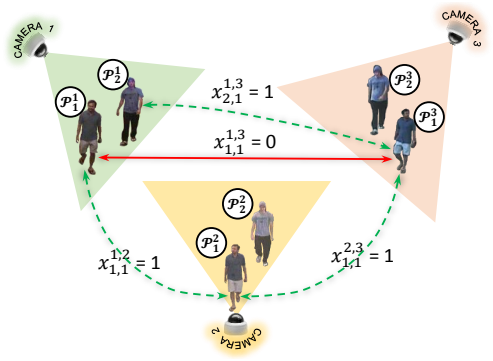


Fig. 11: An illustrative example showing that inconsistent re-identification is captured by the loop constraint given by eqn. (10) for a simple scenario involving 2 persons in 3 cameras.

$\mathcal{P}^q_j$. The set of association values on all the edges between the nodes is denoted as $\mathcal{L}$, *i.e.,* $x^{p,q}_{i,j} \in \mathcal{L}$, $\forall i, j = [1, \cdots, n]$, $\forall p, q = [1, \cdots, m]$ and $p < q$. Finally, the set of all paths between any two nodes $\mathcal{P}^p_i$ and $\mathcal{P}^q_j$ is represented as $\mathcal{E}(\mathcal{P}^p_i, \mathcal{P}^q_j)$ and any path $e^{(z)}(\mathcal{P}^p_i, \mathcal{P}^q_j) \in \mathcal{E}(\mathcal{P}^p_i, \mathcal{P}^q_j)$.

1) Global Similarity of Association: The global similarity score of association can be obtained by summing the similarity scores over all camera pairs as,

$$\mathbf{C} = \sum_{\substack{p,q=1 \\ p<q}}^{m} \sum_{i,j=1}^{n} c^{p,q}_{i,j} x^{p,q}_{i,j} \tag{7}$$

2) Set of Constraints: The set of constraints are as follows.

1. Association constraint: A person from any camera p can have only one match from another camera q. This is true for all possible pairs of cameras which can be expressed as,

$$\sum_{j=1}^{n} x^{p,q}_{i,j} = 1 \ \forall i = 1 \text{ to } n \ \forall p, q = 1 \text{ to } m,\ p < q, \ \sum_{i=1}^{n} x^{p,q}_{i,j} = 1 \ \forall j = 1 \text{ to } n \ \forall p, q = 1 \text{ to } m,\ p < q \tag{8}$$

2. Loop constraint: This constraint comes from the consistency requirement. Given two nodes $\mathcal{P}^p_i$ and $\mathcal{P}^q_j$, it can be noted that for consistency, a logical 'AND' relationship between the association value $x^{p,q}_{i,j}$ and the set of association values $\{x^{p,r}_{i,a}, x^{r,s}_{a,b}, \cdots x^{t,q}_{c,j}\}$ of a possible path between the nodes has to be maintained. The association value between the two nodes $\mathcal{P}^p_i$ and $\mathcal{P}^q_j$ has to be 1 if the association values corresponding

to all the edges of any possible path between these two nodes are 1. Keeping the binary nature of the association variables and the association constraint in mind the relationship can be compactly expressed as,

$$x_{i,j}^{p,q} \geq \left(\sum_{(\mathcal{P}_k^r, \mathcal{P}_l^s) \in e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)} x_{k,l}^{r,s} \right) - |e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)| + 1, \tag{9}$$

$\forall$ paths $e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q) \in \mathcal{E}(\mathcal{P}_i^p, \mathcal{P}_j^q)$. The relationship holds true for all i and all j. For the case of a triplet of cameras the constraint in eqn. (9) simplifies to,

$$x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 2 + 1 = x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \tag{10}$$

An example involving 3 cameras and 2 persons is shown in Fig. 11. Say, the raw similarity score suggests associations between $(\mathcal{P}_1^1, \mathcal{P}_1^2), (\mathcal{P}_1^2, \mathcal{P}_1^3)$ and $(\mathcal{P}_2^1, \mathcal{P}_1^3)$ independently. However, it leads to an infeasible scenario - $\mathcal{P}_1^1$ and $\mathcal{P}_2^1$ are the same person. This infeasibility is also correctly captured through the constraint in eqn. (10). $x_{1,1}^{1,3} = 0$ but $x_{1,1}^{1,2} + x_{1,1}^{2,3} - 1 = 1$, thus violating the constraint.

For a generic scenario where the similarity scores between all persons for every possible pair of cameras are available, the loop constraints on quartets and higher order loops are not necessary. If loop constraint is satisfied for every triplet of cameras then it automatically ensures consistency for every possible combination of cameras taking 3 or more of them. So the loop constraint for the network of cameras become,

$$x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \ \forall\, i, j = [1, \cdots n], \forall\, p, q, r = [1, \cdots m], \text{ and } p < r < q \tag{11}$$

B. Overall Optimization Problem

Thus, by combining the objective function in eqn. (7) with the constraints in eqn. (8) and eqn. (11) we pose the overall optimization problem as,

$$\underset{\substack{x_{i,j}^{p,q} \\ i,j=[1,\cdots,n] \\ p,q=[1,\cdots,m]}}{\operatorname{argmax}} \left(\sum_{\substack{p,q=1 \\ p<q}}^{m} \sum_{i,j=1}^{n} c_{i,j}^{p,q} x_{i,j}^{p,q} \right)$$

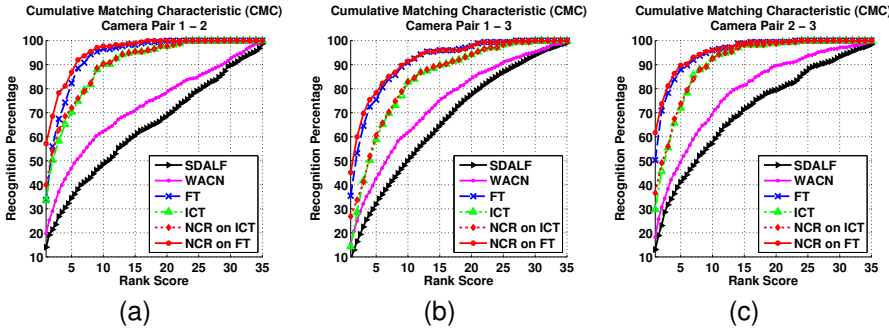


Fig. 12: CMC curves for the WARD dataset. Results and comparisons in (a), (b) and (c) are shown for the camera pairs 1-2, 1-3, and 2-3 respectively.

$$\begin{aligned}
&\text{subject to } \sum_{j=1}^{n} x_{i,j}^{p,q} = 1 \ \forall i = [1, \cdots, n] \ \forall p, q = [1, \cdots, m], \ p < q \\
&\sum_{i=1}^{n} x_{i,j}^{p,q} = 1 \ \forall j = [1, \cdots, n] \ \forall p, q = [1, \cdots, m], \ p < q \\
&x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \\
&\forall \, i, j = [1, \cdots n], \ \forall \, p, q, r = [1, \cdots m], \ \text{and } p < r < q \\
&x_{i,j}^{p,q} \in \{0, 1\} \ \forall i, j = [1, \cdots, n], \ \forall p, q = 1 \text{ to } m, \ p < q
\end{aligned} \tag{12}$$

The above optimization problem for optimal and consistent re-identification is a binary integer program.

1) Network Consistent Re-identification for Variable Number of Targets: There may be situations when every person does not go through every camera. In such cases, a person from any camera p can have *at most* one match from another camera q. The association constraints now change to:

$$\sum_{j=1}^{n_q} x_{i,j}^{p,q} \leq 1 \ \forall i = [1, \cdots, n_p] \ \forall p, q = [1, \cdots, m], \ p < q, \ \sum_{i=1}^{n_p} x_{i,j}^{p,q} \leq 1 \ \forall j = [1, \cdots, n_q] \ \forall p, q = 1 \text{ to } m, \ p < q, \tag{13}$$

But with this generalization, the objective function (ref. eqn. (12)) is no longer valid. Even though the provision of 'no match' is now available, the optimal solution will try to get as many association as possible across the network. This situation can be avoided by incorporating a modification in the objective function as follows:

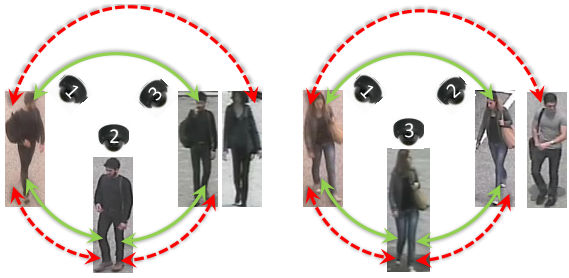


Fig. 13: Two examples of correction of inconsistent re-identification from WARD dataset. The red lines denote re-identifications performed on 3 camera pairs independently by FT method. The green lines show the re-identification results on application of NCR on FT. The NCR algorithm makes the resultant re-identification across 3 cameras correct.

$$\sum_{\substack{p,q=1\\p<q}}^{m} \sum_{i,j=1}^{n_p,n_q} (c_{i,j}^{p,q} - k) x_{i,j}^{p,q}, \tag{14}$$

where 'k' is any value in the range of the similarity scores. In the new cost function, instead of rewarding all positive associations we give reward to most of the TPs, but impose penalties on the FPs. As the rewards for all TP matches are discounted by the same amount 'k' and as there is penalty for FP associations, the new cost function gives us optimal results for both 'match' and 'no-match' cases. Ideally, the distributions of similarity scores of the TPs and FPs are non-overlapping and 'k' can be any real number from the region separating these two distributions. However, for practical scenarios where TP and FP scores overlap, an optimal 'k' can be learned from training data. So, for this more generalized case, the NCR problem can be formulated as follows,

$$\begin{aligned}
&\underset{\substack{x_{i,j}^{p,q}\\ i=[1,\cdots,n_p]\\ j=[1,\cdots,n_q]\\ p,q=[1,\cdots,m]}}{\operatorname{argmax}} \left(\sum_{\substack{p,q=1\\p<q}}^{m} \sum_{i,j=1}^{n_p,n_q} (c_{i,j}^{p,q} - k) x_{i,j}^{p,q} \right) \\
&\text{subject to } \sum_{j=1}^{n_q} x_{i,j}^{p,q} = 1 \ \forall i = [1,\cdots,n_p] \ \forall p,q = [1,\cdots,m],\ p<q \\
&\sum_{i=1}^{n_p} x_{i,j}^{p,q} = 1 \ \forall j = [1,\cdots,n_q] \ \forall p,q = [1,\cdots,m],\ p<q \\
&x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \\
&\forall\, i = [1,\cdots,n_p],\ j = [1,\cdots,n_q],\ \forall p,q,r = [1,\cdots m],\ \text{and } p<r<q \\
&x_{i,j}^{p,q} \in \{0,1\} \ \forall i = [1,\cdots,n_p],\ j = [1,\cdots,n_q],\ \forall p,q = [1,\cdots,m],\ p<q
\end{aligned} \tag{15}$$

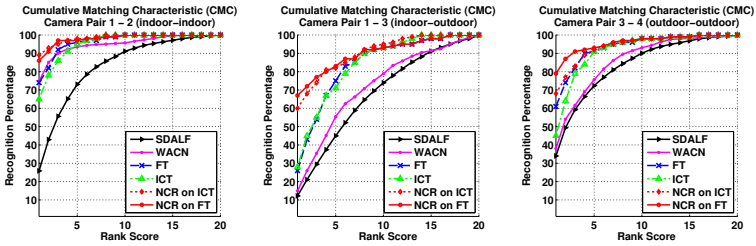


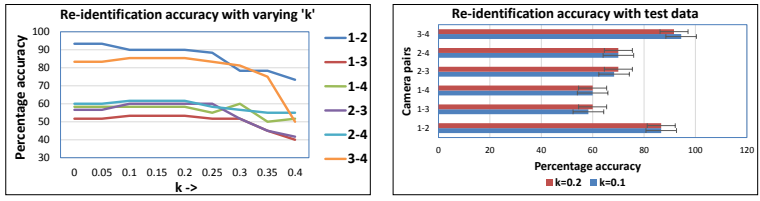
Fig. 14: CMC curves for RAiD dataset. In (a), (b), (c) comparisons are shown for the camera pairs 1-2 (both indoor), 1-3 (indoor-outdoor), and 3-4 (both outdoor) respectively.

C. Experiments

Datasets and Performance Measures: We performed experiments on two benchmark datasets - WARD [11] and one new dataset RAiD. The proposed approach is compared with the methods SDALF [3], ICT [2] and WACN [11] in terms of Cumulative Matching Characteristic (CMC) curves.

WARD Dataset The WARD dataset [11] consists of 70 different people acquired in a real surveillance scenario in 3 non-overlapping cameras (camera $1, 2$ and 3). Fig. 12(a), (b) and (c) compare the performance for camera pairs $1-2, 1-3$, and $2-3$ respectively. The legends 'NCR on FT' and 'NCR on ICT' imply that the NCR algorithm is applied on similarity scores generated by learning the feature transformation and by ICT respectively. For all 3 camera pairs the proposed method outperforms the rest. The difference is most clear in the rank 1 performance. For all the camera pairs 'NCR on FT' shows the best rank 1 performance of recognition percentages as high as 57.14, 45.15 and 61.71 for camera pairs 1-2, 1-3 and 2-3 respectively. Fig. 13 shows two example scenarios where inconsistent re-identifications are corrected.

RAiD Dataset Re-identification Across indoor-outdoor Dataset (RAiD) is collected so that a large number of people are seen in a wide area camera network. This new dataset has large illumination variation as it uses both indoor (camera 1 and 2) and outdoor cameras (camera 3 and 4). The dataset is publicly available to download in



(a) (b)

Fig. 15: performance of the NCR algorithm after removing 40% of the people from both camera 3 and 4. In (a) re-identification accuracy on the training data is shown for every camera pair by varying the parameter k after removing 40% of the training persons. (b) shows the re-identification accuracy on the test data for the chosen values of $k = 0.1$ and 0.2 when 40% of the test people were not present.

http://www.ee.ucr.edu/~amitrc/datasets.php. 21 persons were used for training while the rest 20 were used in training. Due to space constraints we report the results on 3 representative of the total 6 pairs of the cameras. Figs. 14(a) - (c) compare the performance for camera pairs 1-2, 1-3, and 3-4 respectively. The proposed method performs better than all the rest for both the cases when there is not much appearance variation (for camera pair 1-2 where both cameras are indoor and for camera pair 3-4 where both cameras are outdoor) and when there is significant lighting variation (camera pair 1-3). For the indoor camera pair 1-2 the proposed method applied on similarity scores generated by feature transformation and on the similarity scores by ICT achieve 86% and 89% rank 1 performance respectively. For the outdoor camera pair 3-4 the same two methods achieve 79% and 68% rank 1 performance respectively. It can further be seen that for camera pairs with large illumination variation (*i.e.* 1-3) the performance improvement is significantly large. For this camera pair, the rank 1 performance shoots up to 67% and 60% on application of NCR algorithm to FT and ICT compared to their original rank 1 performance of 26% and 28% respectively. Clearly, imposing consistency improves the overall performance. The relative improvement is significantly large in case of large illumination variation.

Re-identification with Variable Number of Persons Next we go for the generalized setting when all the people may not be present in all cameras. We chose two cameras (camera 3 and 4) and removed 8 (40% out of the test set containing 20 people)

randomly chosen people keeping all the persons intact in cameras 1 and 2. The accuracy is calculated by taking both true positive and true negative matches into account($\frac{(\#\text{ true positive}+\#\text{ true negative})}{\#\text{ of unique people in the testset}}$). The average accuracy for varying 'k' for all the 6 cameras are shown in Fig. 15(a). As shown, the accuracy remains more or less constant till $k = 0.25$. After that, the accuracy for camera pairs having the same people (namely camera pairs 1-2 and 3-4) falls rapidly, but for the rest of the cameras where the number of people are variable remains significantly constant. This is because the reward for 'no match' increases with the value of 'k' and for camera pair 1-2 and 3-4 there is no 'no match' case. So, any value of 'k' in the range $(0 - 0.25)$ is a reasonable choice. The accuracy of all the 6 cameras for $k = 0.1$ and 0.2 is shown in Fig. 15(b).

V. Conclusions

The main output of this project is the development of a multi-camera tracking framework and software that takes the raw video as input, detects the moving targets, computed spatio-temporal associations between them and finally obtains the multi-camera tracks. It addresses a number of challenging problems in computer vision - detection, tracking and multi-camera association. A software package has been developed and provided to our collaborators at Progeny. It is also being provided to ONR along with datasets on which it has been tested.

References

[1] Caviar dataset. http://homepages.inf.ed.ac.uk/rbf/caviardata1/

[2] Avraham, T., Gurvich, I., Lindenbaum, M., Markovitch, S.: Learning implicit transfer for person re-identification. In: European Conference on Computer Vision Workshop. pp. 381–390 (2012)

[3] Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. Computer Vision and Image Understanding 117(2), 130–144 (Nov 2013)

[4] Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2011)

[5] Chen, X., Qin, Z., An, L., Bhanu, B.: An online learned elementary grouping model for multi-target tracking. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2014)

[6] Chen, X., An, L., Bhanu, B.: Multi-target tracking in non-overlapping cameras using a reference set. IEEE Sensors Journal (2014) (accept, subject to mandatory corrections)

[7] Das, A., Chakraborty, A., Roy-Chowdhury, A.K.: Consistent re-identification in a camera network. In: European Conference on Computer Vision (2014)

[8] Denina, G., Bhanu, B., Nguyen, H., Ding, C., Kamal, A., Ravishankar, C., Roy-Chowdhury, A., Ivers, A., Varda, B.: Videoweb dataset for multi-camera activities and non-verbal communication. In: Distributed Video Sensor Networks. Springer London (2010)

[9] Dollár, P., Appel, R., Kienzle, W.: Crosstalk cascades for frame-rate pedestrian detection. In: European Conference on Computer Vision (2012)

[10] 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), 146–162 (Feb 2008)

[11] Martinel, N., Micheloni, C.: Re-identify people in wide area camera network. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop. Providence, RI (2012)

[12] Qin, Z., Shelton, C.R.: Improving multi-target tracking via social grouping. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2012)

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