Consistent Re-identification in a Camera Network

Abir Das*, Anirban Chakraborty*, and Amit K. Roy-Chowdhury**

Dept. of Electrical Engineering, University of California, Riverside, CA 92521, USA

Abstract. Most existing person re-identification methods focus on finding similarities between persons between pairs of cameras (camera pairwise re-identification) without explicitly maintaining consistency of the results across the network. This may lead to infeasible associations when results from different camera pairs are combined. In this paper, we propose a network consistent re-identification (NCR) framework, which is formulated as an optimization problem that not only maintains consistency in re-identification results across the network, but also improves the camera pairwise re-identification performance between all the individual camera pairs. This can be solved as a binary integer programing problem, leading to a globally optimal solution. We also extend the proposed approach to the more general case where all persons may not be present in every camera. Using two benchmark datasets, we validate our approach and compare against state-of-the-art methods.

Keywords: Person re-identification, Network consistency.

1 Introduction

In many computer vision tasks it is often desirable to identify and monitor people as they move through a network of non-overlapping cameras. While many object tracking algorithms can achieve reasonable performance for a single camera, it is a more challenging problem for a network of cameras where issues such as changes of scale, illumination, viewing angle and pose start to arise. For non-overlapping cameras it is extremely challenging to associate the same persons at different cameras as no information is obtained from the "blind gaps" between them. This inter-camera person association problem is known as the person re-identification problem.

Person re-identification across non-overlapping fields-of-view (FOVs) is a well studied topic. Most widely used re-identification approaches focus on pairwise re-identification. Although the re-identification accuracy for each camera pair is high, it can be inconsistent if results from 3 or more cameras are considered. Matches of targets given independently by every pair of cameras might not conform to one another and in turn, can lead to inconsistent mappings. Thus, in

* The first two authors should be considered as joint first authors.
** Corresponding author.

D. Fleet et al. (Eds.): ECCV 2014, Part II, LNCS 8690, pp. 330–345, 2014.

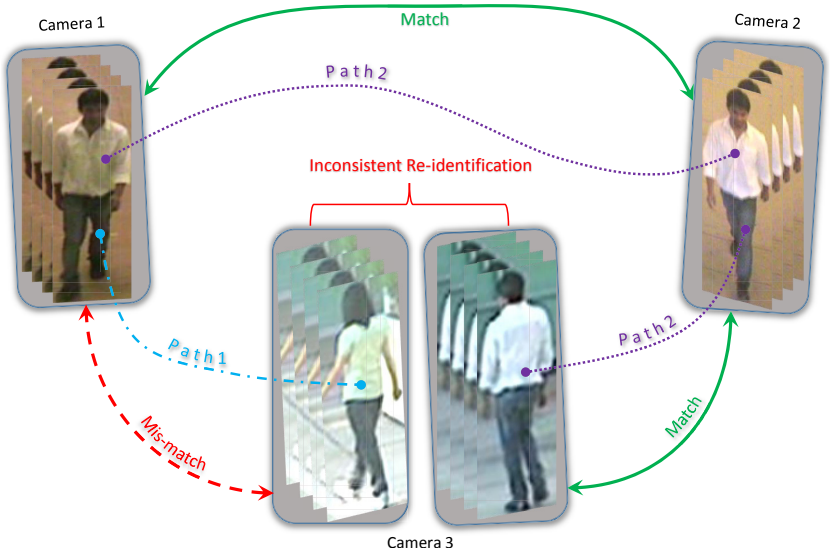

Fig. 1. Example of inconsistency in re-identification: Among the 3 possible re-identification results, 2 are correct. The match of the target in camera 1 to camera 3 can be found in two ways. The first one is the direct pairwise re-identification result between camera 1 and 3 (shown as 'Path 1'), and 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.

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. 1. Even though camera pairs 1-2 and 2-3 have correct re-identification of the target, the false match between the targets in camera pair 1-3 makes the overall re-identification across the triplet inconsistent. In this paper 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 motivation as well as 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. The pairwise re-identification similarity scores obtained using any feasible approach are the input to the proposed method. Unlike assigning a match for which the similarity score is maximum among a set of probable candidates, 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) which can be solved using standard branch and cut, branch and bound or dynamic search algorithms [20]. The application of the proposed formulation is not limited only to person re-identification, but can also be applied in solving other correspondence problems between multiple nodes/instances arising out of the same object at different instants, *e.g.*, object tracking, optical flow, feature correspondences etc.

The proposed method is further generalized to a more challenging scenario in person re-identification when all persons are not present in all the cameras. This challenging scenario of dealing with a variable number of people has been addressed mainly in single cameras by methods relying on learning person specific discriminating signature [3,8]. For multi camera re-identification, a simple way has been to apply a threshold to the similarity score between persons in different 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) in the binary IP framework.

We compare the performance of our approach to state-of-the-art person re-identification methods using a publicly available benchmark dataset - WARD [16] having 3 cameras, and a new 4 camera dataset, RAiD (Reidentification Across indoor-outdoor Dataset) introduced by us. More details about the datasets are provided in sections 4.1 and 4.2.

2 Related Works and Our Contributions

In the last few years there has been increasing attention in the field of person re-identification across camera networks. The proposed approaches addressing the pairwise re-identification problem across non-overlapping cameras can be roughly divided into 3 categories, (i) discriminative signature based methods [2,3,15,16], (ii) metric learning based methods [4,1,23], and (iii) transformation learning based methods [11,18]. Multiple local features (color, shape and texture) are used in [3,15,16] to compute discriminative signatures for each person using multiple images. Similarity between person images is computed by measuring the distance between shape descriptors of color distributions projected in the log-chromaticity space [12] or by using an unsupervised salient feature learning framework in [24]. The authors in [9], propose a metric learning framework whereby a set of training data is used to learn an optimal non-Euclidean metric which minimizes the distance between features of pairs of true matches, while maximizing the same between pairs of wrong matches. Some of the recent works try to improve the re-identification performance by learning a relaxed Mahalanobis metric defined on pairs of true and wrong matches, by learning multiple metrics in a transfer learning set up [14] or by maintaining redundancy in colorspace using a local Fisher discriminant analysis based metric [17]. Works exploring transformation

of features between cameras tried to learn a brightness transfer function (BTF) between appearance features [18], a subspace of the computed BTFs [11], linear color variations model [10], or a Cumulative BTF [19] between cameras. Some of these works [10,11] learned space-time probabilities of moving targets between cameras which may be unreliable if camera FoVs are significantly non-overlapping. As the above methods do not take consistency into account, applying them to a camera network does not give consistent re-identification. Since the proposed method is built upon the pairwise similarity scores, any of the above methods can be the building block to generate the camera pairwise similarity between the targets.

There have been a few correspondence methods proposed in recent years in other aspects of computer vision, *e.g.*, point correspondence in multiple frames and multi target tracking that are relevant to the proposed method. In one of the early works [21], finding point correspondences in monocular image sequences is formulated as finding a graph cover and solved using a greedy method. A suboptimal greedy solution strategy was used in [22] to track multiple targets by finding a maximum cover path of a graph of detections where multiple features like color, position, direction and size determined the edge weights. In [6], the authors linked detections in a tracking scenario across frames by solving a constrained flow optimization. The resulting convex formulation of finding k-shortest node-disjoint paths guaranteed the global optima. However, this method does not actively use appearance features into the data association process which might lead to ID switches among different pairs of cameras resulting in inconsistency. An extension of the work using sparse appearance preserving tracklets was proposed in [5]. With known flow direction, a flow formulation of re-identification will be consistent. But in a re-identification problem with no temporal or spatial layout information, the flow directions are not natural and thus re-identification performance may widely vary with different choices of temporal or spatial flow.

Contributions of the paper: To summarize, the contributions of the proposed approach to the problem of person re-identification are the followings. Network consistent person re-identification problem is formulated as an optimization problem which not only maintains consistency across camera pairwise re-identification results, but also improves the re-identification performance across different camera pairs. To the best of our knowledge, this is the first time that consistency in re-identification across a network of cameras is explored, and an optimization strategy with consistency information from additional cameras is used to improve the otherwise standard camera pairwise re-identification. Due to the formulation of the optimization problem as a binary IP, it is guaranteed to reach the global optima as opposed to the greedy approaches applied in some of the correspondence methods. The method is not tuned to any camera pairwise similarity score generation approach. Any existing re-identification strategy giving independent camera pairwise similarity scores can be incorporated into the framework. The proposed method is also extensible to situations where every person is not present in every camera.

3 Network Consistent Re-identification Framework

In this section we describe the proposed approach in details. The Network Consistent Re-identification (NCR) method starts with the camera pairwise similarity scores between the targets. Section 4 gives a brief description of the process in which the pairwise similarity scores for each person is generated. First we describe the notation and define the terminologies associated to this problem that would be used throughout the rest of the section before delving deeper into the problem formulation.

Let there be m cameras in a network. The number of possible camera pairs is $\binom{m}{2} = \frac{m(m-1)}{2}$. For simplicity we, first, assume, that the same n person are present in each of the cameras. In section 3.4 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' in this framework.

2. Similarity score matrix: Let $\mathbf{C}^{(p,q)}$ denote the similarity score matrix between camera p and camera q. Then $(i,j)^{th}$ cell in $\mathbf{C}^{(p,q)}$ denotes the similarity score between the persons $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$.

3. Assignment matrix: 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 associations between targets across cameras can be represented using 'assignment matrices', one for each camera pair. Each element $x_{i,j}^{p,q}$ of the assignment matrix $\mathbf{X}^{(p,q)}$ between the camera pair (p,q) is defined as follows:

$$x_{i,j}^{p,q} = \begin{cases} 1 & \text{if } \mathcal{P}_i^p \text{ and } \mathcal{P}_j^q \text{ are the same targets} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

As a result $\mathbf{X}^{(p,q)}$ is a permutation matrix, *i.e.,* only one element per row and per column is 1, all the others are 0. Mathematically, $\forall\, x_{i,j}^{p,q} \in \{0,1\}$

$$\sum_{j=1}^{n} x_{i,j}^{p,q} = 1 \ \forall i = 1 \text{ to } n \text{ and } \sum_{i=1}^{n} x_{i,j}^{p,q} = 1 \ \forall j = 1 \text{ to } n. \tag{2}$$

4. 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. It should be noted that there will be no edge between the nodes of the same camera, *i.e.,* two targets from the same camera. There are two attributes connected to each edge. They are the similarity score $c_{i,j}^{p,q}$ and the association value $x_{i,j}^{p,q}$.

5. 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 $\mathcal{P}_j^q$. The set of association values on all the edges between the nodes is denoted as $\mathcal{L}$, *i.e.,* $x_{i,j}^{p,q} \in \mathcal{L}$, $\forall i,j = [1,\cdots,n]$, $\forall p,q = [1,\cdots,m]$

and $p < q$. Finally, the set of all paths between any two nodes $\mathcal{P}_i^p$ and $\mathcal{P}_j^q$ is represented as $\mathcal{E}(\mathcal{P}_i^p, \mathcal{P}_j^q)$ and any path $e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q) \in \mathcal{E}(\mathcal{P}_i^p, \mathcal{P}_j^q)$.

3.1 Global Similarity of Association

For the camera pair (p, q), the sum of the similarity scores of association is given by $\sum_{i,j=1}^{n} c_{i,j}^{p,q} x_{i,j}^{p,q}$. Summing over all possible camera pairs the global similarity score can be written as

$$\mathbf{C} = \sum_{\substack{p,q=1 \\ p<q}}^{m} \sum_{i,j=1}^{n} c_{i,j}^{p,q} x_{i,j}^{p,q} \tag{3}$$

3.2 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 mathematically expressed by the set of equations (2). This is true for all possible pairs of cameras which can be expressed as,

$$\begin{aligned} \sum_{j=1}^{n} x_{i,j}^{p,q} &= 1 \ \forall i = 1 \text{ to } n \ \forall p, q = 1 \text{ to } m,\ p < q \\ \sum_{i=1}^{n} x_{i,j}^{p,q} &= 1 \ \forall j = 1 \text{ to } n \ \forall p, q = 1 \text{ to } m,\ p < q \end{aligned} \tag{4}$$

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

$$x_{i,j}^{p,q} \geq \left(\sum_{(\mathcal{P}_k^r, \mathcal{P}_l^s) \in e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)} x_{k,l}^{r,s} \right) - |e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)| + 1, \tag{5}$$

$\forall$ paths $e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q) \in \mathcal{E}(\mathcal{P}_i^p, \mathcal{P}_j^q)$, where $|e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)|$ denotes the cardinality of the path $|e^{(z)}(\mathcal{P}_i^p, \mathcal{P}_j^q)|$, *i.e.,* the number of edges in the path. The relationship holds true for all i and all j. For the case of a triplet of cameras the constraint in eqn. (5) simplifies to,

$$x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 2 + 1 = x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \tag{6}$$

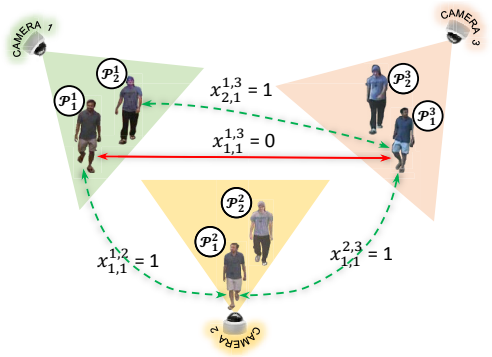

Fig. 2. An illustrative example showing that inconsistent re-identification is captured by the loop constraint given by eqn. (6) for a simple scenario involving 2 persons in 3 cameras.

An example involving 3 cameras and 2 persons is illustrated with the help of Fig. 2. Say, the raw similarity score between pairs of targets across cameras 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, when these associations are combined together over the entire network, 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. (6). $x_{1,1}^{1,3} = 0$ but $x_{1,1}^{1,2} + x_{1,1}^{2,3} - 1 = 1$, thus violating the constraint.

For a generic scenario involving a large number of cameras where similarity scores between every pair of cameras may not be available, the loop constraint equations (*i.e.,* eqn. (5)) have to hold for every possible triplet, quartet, quintet (and so on) of cameras. On the other hand, if 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,

$$\begin{gathered} x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \\ \forall\, i,j = [1,\cdots n],\, \forall\, p,q,r = [1,\cdots m],\ \text{and}\ p < r < q \end{gathered} \tag{7}$$

3.3 Overall Optimization Problem

Thus, by combining the objective function in eqn. (3) with the constraints in eqn. (4) and eqn. (7) we pose the overall optimization problem as,

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

$$\begin{aligned}
&\text{subject to } \sum_{j=1}^{n} x_{i,j}^{p,q} = 1 \ \forall i = [1, \cdots, n] \ \forall p, q = [1, \cdots, m],\ p < q \\
&\sum_{i=1}^{n} x_{i,j}^{p,q} = 1 \ \forall j = [1, \cdots, n] \ \forall p, q = [1, \cdots, m],\ p < q \\
&x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \\
&\forall\, i, j = [1, \cdots n],\ \forall\, p, q, r = [1, \cdots m],\ \text{and } p < r < q \\
&x_{i,j}^{p,q} \in \{0, 1\}\ \forall i, j = [1, \cdots, n],\ \forall p, q = 1 \text{ to } m,\ p < q
\end{aligned} \tag{8}$$

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

3.4 Network Consistent Re-identification for Variable Number of Targets

As explained in the previous sub-section, the NCR problem can be solved by solving the binary IP formulated in eqn. (8). However, there may be situations when every person does not go through every camera. In such cases, the values of assignment variables in every row or column of the assignment matrix can all be 0. In other words, a person from any camera p can have *at most* one match from another camera q. As a result, the association constraints now change from equalities to inequalities as follows:

$$\begin{aligned}
&\sum_{j=1}^{n_q} x_{i,j}^{p,q} \leq 1 \ \forall i = [1, \cdots, n_p] \ \forall p, q = [1, \cdots, m],\ p < q \\
&\sum_{i=1}^{n_p} x_{i,j}^{p,q} \leq 1 \ \forall j = [1, \cdots, n_q] \ \forall p, q = 1 \text{ to } m,\ p < q,
\end{aligned} \tag{9}$$

where n_p snd n_q are the number of persons in camera p and q respectively. But with this generalization, it is easy to see that the objective function (ref. eqn. (8)) 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 is due to the fact that the current objective function assigns reward to both true positive (TP) associations (correctly matching a person present in both cameras) and false positive (FP) associations (wrongly associating a match to a person who is present in only one camera). Thus the optimal solution may contain many false positive associations. This situation can be avoided by incorporating a modification in the objective function as follows:

$$\sum_{\substack{p,q=1 \\ p<q}}^{m} \sum_{i,j=1}^{n_p, n_q} (c_{i,j}^{p,q} - k) x_{i,j}^{p,q}, \tag{10}$$

where 'k' is any value in the range of the similarity scores. This modification leverages upon the idea that, typically, similarity scores for most of the TP

matches in the data would be much larger than majority of the FP matches. 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. The choice of the parameter 'k' depends on the similarity scores generated by the chosen method, and thus can vary from one pairwise similarity score generating methods to another. 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. A simple method to choose 'k' could be running NCR for different values of 'k' over the training data and choosing the one giving the maximum accuracy on the cross validation data. So, for this more generalized case, the NCR problem can be formulated as follows,

$$\begin{aligned}
&\underset{\substack{x_{i,j}^{p,q} \\ i=[1,\cdots,n_p] \\ j=[1,\cdots,n_q] \\ p,q=[1,\cdots,m]}}{\operatorname{argmax}} \left(\sum_{\substack{p,q=1 \\ p<q}}^{m} \sum_{i,j=1}^{n_p,n_q} (c_{i,j}^{p,q} - k) x_{i,j}^{p,q} \right) \\
&\text{subject to } \sum_{j=1}^{n_q} x_{i,j}^{p,q} = 1 \ \forall i = [1,\cdots,n_p] \ \forall p,q = [1,\cdots,m],\ p < q \\
&\sum_{i=1}^{n_p} x_{i,j}^{p,q} = 1 \ \forall j = [1,\cdots,n_q] \ \forall p,q = [1,\cdots,m],\ p < q \\
&x_{i,j}^{p,q} \geq x_{i,k}^{p,r} + x_{k,j}^{r,q} - 1 \\
&\forall\, i = [1,\cdots,n_p],\ j = [1,\cdots,n_q],\ \forall\, p,q,r = [1,\cdots m],\ \text{and } p < r < q \\
&x_{i,j}^{p,q} \in \{0,1\} \ \forall i = [1,\cdots,n_p],\ j = [1,\cdots,n_q],\ \forall p,q = [1,\cdots,m],\ p < q
\end{aligned} \tag{11}$$

A rigorous proof showing that the problem in eqn. (8) is a special case of the more generalized problem described in this section can be found in the supplementary.[1]

4 Experiments

Datasets and Performance Measures: To validate our approach, we performed experiments on two benchmark datasets - WARD [16] and one new dataset RAiD introduced in this work. Though state-of-the-art methods for person re-identification *e.g.*, [8,3,13] evaluate their performances using other datasets too (*e.g.*, ETHZ, CAVIAR4REID, CUHK) these do not fit our purposes since these are either two camera datasets or several sequences of different 2 camera datasets.

[1] Supplementary materials are available at `www.ee.ucr.edu/~amitrc/publications.php`

WARD is a 3 camera dataset and RAiD is a 4 camera dataset. Results are shown in terms of recognition rate as Cumulative Matching Characteristic (CMC) curves and normalized Area Under Curve (nAUC) values (provided in the supplementary), as is the common practice in the literature. The CMC curve is a plot of the recognition percentage versus the ranking score and represents the expectation of finding the correct match inside top t matches. nAUC gives an overall score of how well a re-identification method performs irrespective of the dataset size. In the case where every person is not present in all cameras, we show the accuracy as total number of true positives (true matches) and true negatives (true non-matches) divided by the total number of unique people present. All the results used for comparison were either taken from the corresponding works or by running codes which are publicly available or obtained from the authors on datasets for which reported results could not be obtained. We did not re-implement other methods as it is very difficult to exactly emulate all the implementation details.

Pairwise Similarity Score Generation: The camera pairwise similarity score generation starts with extracting appearance features in the form of HSV color histogram from the images of the targets. Before computing these features, the foreground is segmented out to extract the silhouette. Three salient regions (head, torso and legs) are extracted from the silhouette as proposed in [3]. The head region S^H is discarded, since it often consists of a few and less informative pixels. We additionally divide both body and torso into two horizontal sub-regions based on the intuition that people can wear shorts or long pants, and short or long sleeves tops.

Given the extracted features, we generate the similarity scores by learning the way features get transformed between cameras in a similar approach as [18,11]. Instead of using feature correlation matrix or the feature histogram values directly, we capture the feature transformation by warping the feature space in a nonlinear fashion inspired by the principle of Dynamic Time Warping (DTW). The feature bin number axis is warped to reduce the mismatch between feature values of two feature histograms from two cameras. Considering two non-overlapping cameras, a pair of images of the same target is a feasible pair, while a pair of images between two different targets is an infeasible pair. Given the feasible and infeasible transformation functions from the training examples, a Random Forest (RF) [7] classifier is trained on these two sets. The camera pairwise similarity score between targets are obtained from the probability given by the trained classifier of a test transformation function as belonging to either the set of feasible or infeasible transformation functions. In addition to the feature transformation based method, similarity scores are also generated using the publicly available code of a recent work - ICT [2] where pairwise re-identification was posed as a classification problem in the feature space formed of concatenated features of persons viewed in two different cameras.

Experimental Setup: To be consistent with the evaluations carried out by state-of-the-art methods, images were normalized to 128×64. The H, S and V color histograms extracted from the body parts were quantized using 10 bins

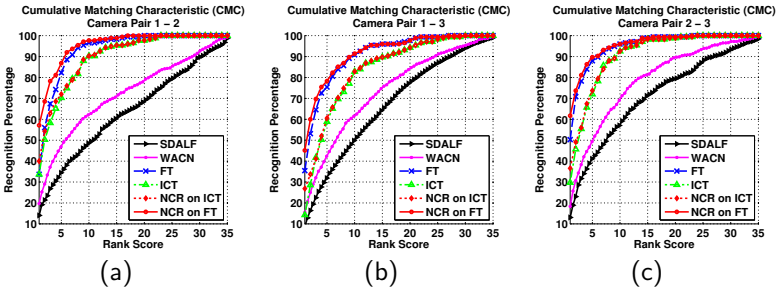

Fig. 3. CMC curves for the WARD dataset. Results and comparisons in (a), (b) and (c) are shown for the camera pairs 1-2, 1-3, and 2-3 respectively..

each. Image pairs of the same or different person(s) in different cameras were randomly picked to compute the feasible and infeasible transformation functions respectively. All the experiments are conducted using a multi-shot strategy where 10 images per person is taken for both training and testing The RF parameters such as the number of trees, the number of features to consider when looking for the best split, etc. were selected using 4-fold cross validation. For each test we ran 5 independent trials and report the average results.

4.1 WARD Dataset

The WARD dataset [16] has 4786 images of 70 different people acquired in a real surveillance scenario in three non-overlapping cameras. This dataset has a huge illumination variation apart from resolution and pose changes. The cameras here are denoted as camera $1, 2$ and 3. Fig. 3(a), (b) and (c) compare the performance for camera pairs $1-2, 1-3$, and $2-3$ respectively. The 70 people in this dataset are equally divided into training and test sets of 35 persons each. The proposed approach is compared with the methods SDALF [3], ICT [2] and WACN [16]. 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. The runner up in camera pair 1-2 is 'NCR on ICT' with rank 1 recognition percentage of 40. The runner up for the rest of the camera pairs is 'FT' with corresponding numbers for camera pairs 1-3 and 2-3 being 35.43 and 50.29 respectively Fig. 4 shows two example scenarios from this dataset where inconsistent re-identifications are corrected on application of NCR algorithm.

4.2 RAiD Dataset

Unlike the publicly available person re-identification datasets, Re-identification Across indoor-outdoor Dataset (RAiD) is collected so that a large number of

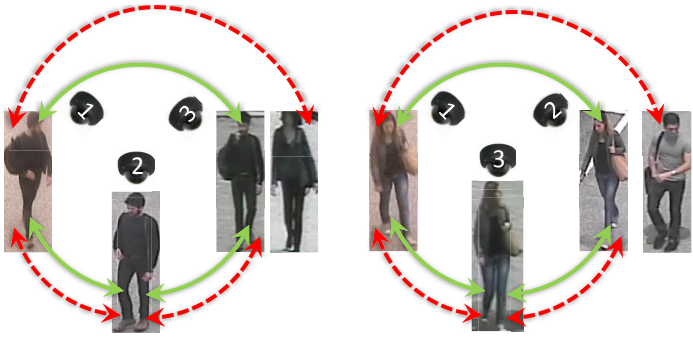

Fig. 4. Two examples of correction of inconsistent re-identification from WARD dataset. The red dashed lines denote re-identifications performed on 3 camera pairs independently by FT method. The green solid lines show the re-identification results on application of NCR on FT. The NCR algorithm exploits the consistency requirement and makes the resultant re-identification across 3 cameras correct.

people are seen in multiple cameras in a wide area camera network. This new dataset also has large illumination variation as this was collected using both indoor (camera 1 and 2) and outdoor cameras (camera 3 and 4). 43 subjects were asked to walk through these 4 cameras resulting in 6920 annotated images. 41 of the total 43 persons appear in all the cameras. We took these 41 persons to validate the proposed approach. The dataset is publicly available to download in `http://www.ee.ucr.edu/~amitrc/datasets.php`

The proposed approach is compared with the same methods as for the WARD dataset. 21 persons were used for training while the rest 20 were used in training. Figs. 5(a) - (f) compare the performance for camera pairs 1-2, 1-3, 1-4, 2-3, 2-4 and 3-4 respectively. We see that 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 (for the rest 4 camera pairs). Expectedly, for camera pairs 1-2 and 3-4 the performance of the proposed method is the best. For the indoor camera pair 1-2 the proposed method applied on similarity scores generated by feature transformation (NCR on FT) and on the similarity scores by ICT (NCR on 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. For the rest of the cases where there is significant illumination variation the proposed method is superior to all the rest.

In all the camera pairs, the top two performances come from the NCR method applied on two different camera pairwise similarity score generating methods. It can further be seen that for camera pairs with large illumination variation (*i.e.* 1-3, 1-4, 2-3 and 2-4) the performance improvement is significantly large. For camera pair 1-3, 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

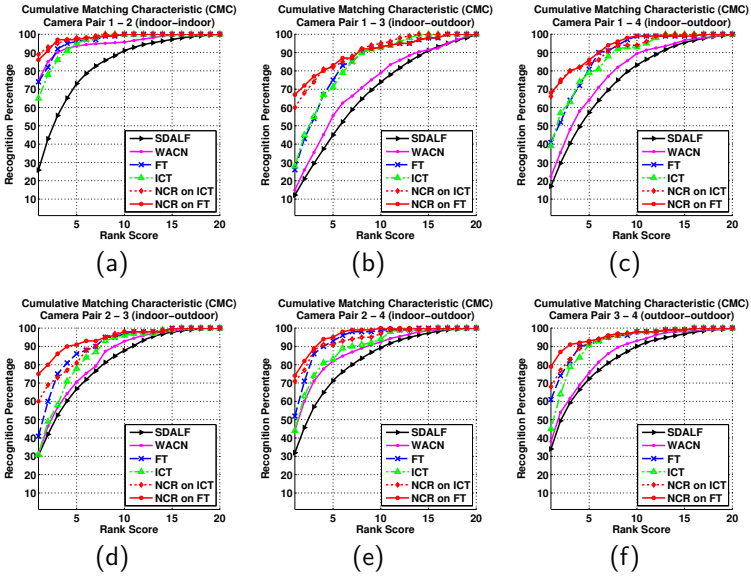

Fig. 5. CMC curves for RAiD dataset. In (a), (b), (c), (d), (e), (f) comparisons are shown for the camera pairs 1-2 (both indoor), 1-3 (indoor-outdoor), 1-4 (indoor-outdoor), 2-3 (indoor-outdoor), 2-4 (indoor-outdoor) and 3-4 (both outdoor) respectively.

the overall performance with the best absolute accuracy achieved for camera pairs consisiting of only indoor or only outdoor cameras. On the other hand, the relative improvement is significantly large in case of large illumination variation between the two cameras.

4.3 Re-identification with Variable Number of Persons

Next we evaluate the performance of the proposed method for the generalized setting when all the people may not be present in all cameras. For this purpose we chose two cameras (namely 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. For this experiment the accuracy of the proposed method is shown with similarity scores as obtained by learning the feature transformation between the camera pairs. The accuracy is calculated by taking both true positive and true negative matches into account and it is expressed as $\frac{(\#\text{ true positive}+\#\text{ true negative})}{\#\text{ of unique people in the testset}}$.

Since the existing methods do not report re-identification results on variable number of persons nor is the code available which we can modify easily to incorporate such a scenario, we can not provide a comparison of performance here. However we show the performance of the proposed method for different values of 'k'. The value of 'k' is learnt using 2 random partitions of the training data in the same scenario (*i.e.*, removing 40% of the people from camera 3 and 4).

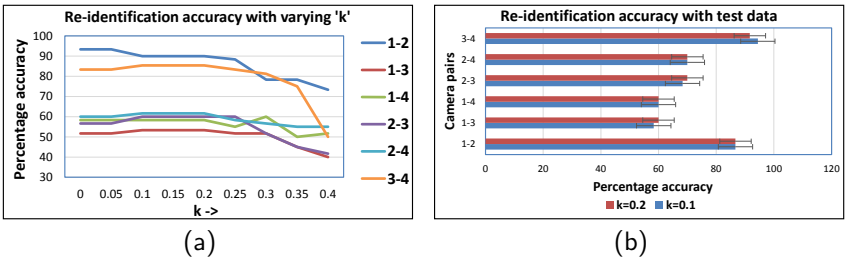

Fig. 6. performance of the NCR algorithm after removing 40% of the people from both camera 3 and 4. In (a) re-identification accuracy on the training data is shown for every camera pair by varying the parameter k after removing 40% of the training persons. (b) shows the re-identification accuracy on the test data for the chosen values of $k = 0.1$ and 0.2 when 40% of the test people were not present.

The average accuracy over these two random partitions for varying 'k' for all the 6 cameras are shown in Fig. 6(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 due to the fact that 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, for these two camera pairs, the optimization problem (in eqn. (11)) reaches the global maxima at the cost of assigning 0 label to some of the true associations (for which the similarity scores are on the lower side). So, any value of 'k' in the range $(0-0.25)$ will be a reasonable choice. The accuracy of all the 6 cameras for $k = 0.1$ and 0.2 is shown in Fig. 6(b), where it can be seen that the performance is significantly high and does not vary much with different values of 'k.'

5 Conclusions

In this work we addressed the problem of person re-identification in a camera network by exploiting the requirement of consistency of re-identification results. A novel binary integer programing formulation of the problem is provided. 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. The future directions of our research will be not only to apply our approach to bigger networks with large numbers of cameras, and cope with wider space-time horizons but to apply also to other areas, (*e.g.*, social network analysis, medical imaging to name a few) where consistency is the key to robustness.

Acknowledgements. This work was partially supported by NSF grants IIS-1316934 and CNS-1330110. We acknowledge the authors of [16] for providing the code of WACN. We would also like to thank Andrew Yu, a current UCR undergraduate student, for helping in the annotation of the RAiD dataset.

References

1. Alavi, A., Yang, Y., Harandi, M., Sanderson, C.: Multi-shot person re-identification via relational stein divergence. In: IEEE International Conference on Image Processing (2013)
2. Avraham, T., Gurvich, I., Lindenbaum, M., Markovitch, S.: Learning implicit transfer for person re-identification. In: European Conference on Computer Vision, Workshops and Demonstrations. 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. Bellet, A., Habrard, A., Sebban, M.: A survey on metric learning for feature vectors and structured data. ArXiv e-prints (2013)
5. Ben Shitrit, H., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: IEEE International Conference on Computer Vision. pp. 137–144 (2011)
6. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(9), 1806–1819 (2011)
7. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
8. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: British Machine Vision Conference (2011)
9. Dikmen, M., Akbas, E., Huang, T.S., Ahuja, N.: Pedestrian recognition with a learned metric. In: Asian conference on Computer vision. pp. 501–512 (2010)
10. Gilbert, A., Bowden, R.: Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In: European Conference Computer Vision (2006)
11. 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)
12. Kviatkovsky, I., Adam, A., Rivlin, E.: Color invariants for person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(7), 1622–1634 (2013)
13. Li, W., Wang, X.: Locally aligned feature transforms across views. In: International Conference on Computer Vision and Pattern Recognition (2013)
14. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Asian Conference on Computer Vision. pp. 31–44 (2012)
15. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person re-identification : What features are important ? In: European Conference on Computer Vision, Workshops and Demonstrations. pp. 391–401. Springer Berlin Heidelberg, Florence, Italy (2012)
16. Martinel, N., Micheloni, C.: Re-identify people in wide area camera network. In: International Conference on Computer Vision and Pattern Recognition Workshops. pp. 31–36. IEEE, Providence, RI (Jun 2012)
17. Pedagadi, S., Orwell, J., Velastin, S.: Local fisher discriminant analysis for pedestrian re-identification. In: International Conference on Computer Vision and Pattern Recognition. pp. 3318–3325 (2013)
18. Porikli, F., Hill, M.: Inter-camera color calibration using cross-correlation model function. In: IEEE International Conference on Image Processing (ICIP). pp. 133–136 (2003)

19. Prosser, B., Gong, S., Xiang, T.: Multi-camera matching using bi-directional cumulative brightness transfer functions. In: British Machine Vision Conference (Sep 2008)
20. Schrijver, A.: Theory of linear and integer programming. John Wiley and Sons (1998)
21. Shafique, K., Shah, M.: A noniterative greedy algorithm for multiframe point correspondence. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(1), 51–65 (2005)
22. Taj, M., Maggio, E., Cavallaro, A.: Multi-feature graph-based object tracking. In: Multimodal Technologies for Perception of Humans, vol. 4122, pp. 190–199. Springer Berlin Heidelberg (2007)
23. Yang, L., Jin, R.: Distance metric learning : A comprehensive survey. Tech. rep., Michigan State University (2006)
24. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: International Conference on Computer Vision and Pattern Recognition (2013)