

Joint Temporal-Spatial Information and Common Network Consistency Constraint for Person Re-identification

Zhaoxi Cheng, Hua Yang^(✉), and Lin Chen

Institute of Image Communication and Network Engineering,
Department of Electronic Engineering, Shanghai Jiao Tong University,
Shanghai 200240, China
{czzx536,hyang,SJChenLin}@sjtu.edu.cn

Abstract. Recent research of person Re-identification most focuses on exploring person appearance feature and distance measure between specific cameras pair. In this paper, a joint temporal-spatial information and common network consistency constraint framework is proposed to improve the re-identification performance in all pairwise cameras. First, a correction function is introduced for describing the influence factor of temporal-spatial information on similarity scores. Then the amended similarity score strategy is provided to tradeoff between the person appearance and temporal-spatial information. Finally the whole global optimization problem of the jointing temporal-spatial and common consistency constraints is solved by integer programming method. Using the multi-cameras RAiD dataset, the experiment results validate that the proposed framework obtains significantly better performance compared to the state of the art camera network person re-identification methods.

Keywords: Person re-identification · Temporal-spatial information
Network consistency

1 Introduction

Recently, person re-identification (re-ID) has become increasingly popular in the community. Its aim is seeking a person of interest in previous time and other cameras. There are many general challenges such as scale, altitude, illumination and view angle changes because the person images are obtained from different cameras deployed in non-overlapping scenes.

Facing these problems, most existing traditional person re-identification approaches focus on constructing distinctive visual features and learning an optimal metric. The feature-based methods measure the similarity of person in images with features [1–4]. These features mostly are color feature, shape feature, gradient feature, texture feature, or special features obtained by learning, such as BIF [5], covariance descriptors [6] and LOMO [7]. It is difficult to achieve better effect by using standard distance measurement, treating each of

the apparent features equally. To solve this problem, some researchers have proposed methodologies based on metric learning [8–10]. In a framework based on metric learning, a series of training data is used to learn a non-Euclidean measure so that the distance between the features of the correct match is the smallest and the distance between the mismatched pairs is greatest. Lately, a cross-view quadratic discriminant analysis (XQDA) method is proposed [7], which achieves good performance.

Both of the above two approach only depend on visual information and the effect is not very satisfactory. Further, deep learning method is proposed to learn better feature automatically, leading to some better results. CNN (Convolutional Neural Networks) has been widely utilized to extract spatial information of the single-shot pedestrians and achieved success [11–13], since it establish multidimensional model of pedestrians and can provide a better and robust feature representation. However in practical surveillance scenes, there are many cameras distributing in a large area of region, each of these cameras cover its own region which is non-overlapping with other camera's region. In the situation where a person is captured by difference camera, the time span may be very large, leading to a problem that the appearance change of a individual person may be very significant. At this time, it is difficult to provide sufficient identity discrimination using visual information purely. Among the images of persons, there are temporal-spatial information, and among the cameras, there are network consistent information. All the method mentioned above is mainly regard person re-identification problem as a retrieval task based on camera pair, with no consideration of camera network information. It remains a problem of great value to propose a re-identification framework considering the camera networks to improve the accuracy.

To address the problem, in 2014, Abir et al. first consider camera network for person re-identification [14], introducing the concept of consistent in network. First make a hypothesis: each pedestrian in each camera just appeared once, then define global similarity as the objective function and take network consistency as constraint. Finally the problem is transformed into integer linear programming problem. For each camera pairs in this camera network, the result has improve compared with traditional method. But there are also some limitation that its constraints is relatively strong, and it did not use the spatial-temporal information. Huang et.al. proposed a method of converting the spatial-temporal information and the similarity score obtained by the person re-identification of camera pair into probability [15]. They use probability to represent similarity and temporal-spatial information, then multiple the two probability and use the result as modified similarity. They obtain some improvement in person re-identification but the method doesn't consider the consistent problem in camera network.

Motivated by above mentioned works, we propose a camera network person re-identification framework, which combine the visual information, temporal-spatial information and consistent information and lead to an improvement of the camera pairwise re-identification performance between all the cameras. Our

contribution is summarized by the following three aspects. (1) A novel framework of camera network person re-identification system with a unified symbol architecture is introduced. The framework we proposed joint the temporal-spatial information, network consistency and pairwise person re-identification method. (2) A correction function is introduced describing the influence factor of temporal-spatial information on similarity scores. And an amended similarity score strategy is provided to tradeoff between the person appearance and temporal-spatial information. Specially addition of the correction function and the original pairwise similarity score is proposed, with a coefficient controlling the ratio between original score and influence factor. (3) Jointing temporal-spatial information, common consistence constraints and pairwise person re-identification method, the whole global optimization problem is deviate which is solved by integer programming method, lead to the final result of camera network person re-identification problem.

2 Our Method

2.1 Overall Framework

In this section we describe the camera network person re-identification system. The proposed method can be represented using a framework diagram shown in Fig. 1. For each camera pair, original similarity score between every two person in different camera are computed by using existing methods of feature representation and metric learning. Temporal-spatial information is represented by correction function defined in (3). The consistency is represented as a constraint of an optimization problem which joint original score and temporal-spatial information. Finally we transform the whole problem into an integer programming problem.

2.2 Proposed Method

Symbol definition. First of all, we should define a set of symbol that unify all discussion below. Suppose there are M Cameras $\{C_1, C_2, \dots, C_M\}$ in a camera network. Each camera control its own region $\{region1, region2, regionM\}$ and has a corresponding scenes position coordinates $\{(x_1, y_1), (x_2, y_2), \dots, (x_M, y_M)\}$, which is the central location of its view region.

Then there are $C_M^2 = \frac{m(m-1)}{2}$ camera pairs theoretically, For each pair, there is a corresponding distance between the camera pair $d_{ij} = ((x_i - x_j)^2 + (y_i - y_j)^2)^{\frac{1}{2}}$ ($i < j$). The intuitive diagram of this network consist of M cameras is shown in Fig. 2(a).

For arbitrary camera C_i , there are n_i ($i = 1, 2, \dots, M$) person having ever appeared. We denote P_j^i as the j th person captured by the i th camera, then the n_i person captured by the i th camera is specifically $\{P_1^i, P_2^i, \dots, P_j^i, \dots, P_{n_i}^i\}$, each of which has a corresponding capture time. The capture time of person P_j^i is t_j^i ,

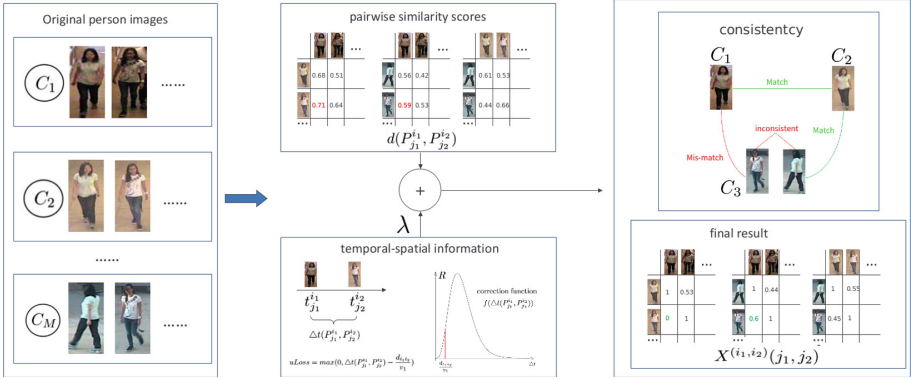


Fig. 1. Framework diagram: pairwise similarity scores is calculated by traditional method. Temporal-spatial information is modelled as correction function. Then pairwise similarity scores and correction function are added, with a coefficient λ to tradeoff the two addends. Finally the sum and network consistency are jointed, leading to the whole global optimization problem.

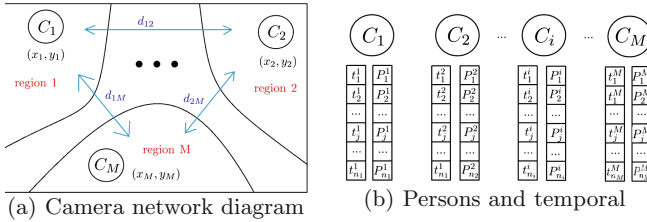


Fig. 2. Symbol architecture of the camera network re-identification framework

where $i = 1, 2, \dots, M, j = 1, 2, \dots, n_i$. These information can be described by a graph shown as Fig. 2(b).

For simply discussing, we first assume that the same N persons are present in each of the M cameras. Now for each camera pair, for example $(i_1, i_2), (i_1, i_2 = 1, 2, \dots, M)$. We can obtain the similarity score of each person pair in different camera by performing feature representation and metric learning. For person $P_{j_1}^{i_1}$ and $P_{j_2}^{i_2}$, its similarity score is denoted as $d(P_{j_1}^{i_1}, P_{j_2}^{i_2}), (j_1, j_2 = 1, 2, \dots, N)$. We can put all these similarity score of persons between camera (i_1, i_2) into an $N \times N$ matrix, called similarity matrix, denoted as $D^{(i_1, i_2)}$. The entry $D^{(i_1, i_2)}(a, b)$ represent $d(P_a^{i_1}, P_b^{i_2})$. For each camera pair we can obtain a similarity matrix, finally all the $\frac{M(M-1)}{2}$ similarity matrices can be obtained. We define a overall similarity matrix \mathbf{D} .

$$\mathbf{D} = [D^{(1,2)}, D^{(1,3)}, \dots, D^{(1,M)}, D^{(2,1)}, D^{(2,2)}, \dots, D^{(2,M)}, \dots, D^{(M-1,M)}] \quad (1)$$

Joint temporal-spatial constraint and consistency. This temporal-spatial constraint is Built on a hypothesis that the velocity of person should be in a reasonable range depend on practical occasions, specifically the walking velocity should have an upper bound v_1 under normal circumstances. Then if person A and person B are captured by two camera, their time interval should also in a range computed by distance of two camera and velocity bound. Specifically, consider person $P_{j_1}^{i_1}$ and $P_{j_2}^{i_2}$, they are captured by camera C_{i_1} and C_{i_2} . The distance of these two camera is $d_{i_1 i_2}$ and the time interval of these two person is defined as formula (2).

$$\Delta t(P_{j_1}^{i_1}, P_{j_2}^{i_2}) = t_{j_1}^{i_1} - t_{j_2}^{i_2} \tag{2}$$

Approximately, we can assume a condition that $\Delta t(P_{j_1}^{i_1}, P_{j_2}^{i_2})$ should be in the range of interval $(\frac{d_{i_1 i_2}}{v_1}, \infty)$. If the condition is not satisfied, these two person is not the same person very likely, then the similarity score $d(P_{j_1}^{i_1}, P_{j_2}^{i_2})$ is changed to zero, else similarity score $d(P_{j_1}^{i_1}, P_{j_2}^{i_2})$ is amended by adding a correction term $R(P_{j_1}^{i_1}, P_{j_2}^{i_2})$. The correction term is a function (Eq 3) of time interval $\Delta t(P_{j_1}^{i_1}, P_{j_2}^{i_2})$ defined in interval $(\frac{d_{i_1 i_2}}{v_1}, \infty)$ which we called correction function in this paper.

$$R(P_{j_1}^{i_1}, P_{j_2}^{i_2}) = f(\Delta t(P_{j_1}^{i_1}, P_{j_2}^{i_2})), \quad \Delta t(P_{j_1}^{i_1}, P_{j_2}^{i_2}) \in (\frac{d_{i_1 i_2}}{v_1}, \infty) \tag{3}$$

The function f can be in any form as long as it satisfy the following condition: f is first an increment function, then f reach the peak at a certain point, and then f become a decrease function, finally when $\Delta t(P_{j_1}^{i_1}, P_{j_2}^{i_2}) \rightarrow \infty, f \rightarrow 0$. We can consider it as a factor reflecting a kind of probability, such that $\Delta t(P_{j_1}^{i_1}, P_{j_2}^{i_2})$ is more likely at the middle of interval $(\frac{d_{i_1 i_2}}{v_1}, \infty)$ than at the two ends. In this paper we use Chi-square distribution probability density function as the correction function, as shown in Fig. 3(a), which satisfy the condition discussed above. Inspired by the velocity approximation of person mentioned above, we can define a function called upper loss($uLoss$) as formula (4).

$$uLoss = \max(0, \Delta t(P_{j_1}^{i_1}, P_{j_2}^{i_2}) - \frac{d_{i_1 i_2}}{v_1}) \tag{4}$$

Then the similarity $d(P_{j_1}^{i_1}, P_{j_2}^{i_2})$ is corrected to $d'(P_{j_1}^{i_1}, P_{j_2}^{i_2})$ according to the addition of original score and the correction term as shown in Eq. (5). The thought of the addition relation is similar as the rate distortion theory in information theory. Its main advantage is that we can use an factor to control the contribution ratio between original score and correction term I

$$d'(P_{j_1}^{i_1}, P_{j_2}^{i_2}) = (d(P_{j_1}^{i_1}, P_{j_2}^{i_2}) + \lambda R(P_{j_1}^{i_1}, P_{j_2}^{i_2}))I(uLoss \neq 0) \tag{5}$$

where $I()$ is indicative function, λ is a coefficient that control the proportion between similarity and correction term. So the similarity is added by a correction term, but if the time interval is beyond the pre-specified range, the similarity

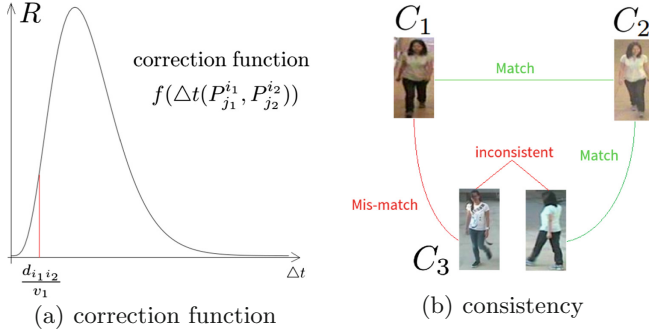


Fig. 3. Joint temporal-spatial information and consistency

is changed to zero. After traversing all the entry of $D^{(i_1, i_2)}$, we can get similarity matrix $D^{(i_1, i_2)}$ as median result. Repeating the above process we can get all the $\frac{M(M-1)}{2}$ median similarity matrices and the overall corrected similarity matrix D' .

$$D' = [D^{(1,2)}, D^{(1,3)}, \dots, D^{(1,M)}, D^{(2,1)}, D^{(2,2)}, \dots, D^{(2,M)}, \dots, D^{(M-1,M)}] \quad (6)$$

After getting the overall corrected similarity matrix D' , we use consistent constraint to further optimize the similarity matrix. The thought of this part is almost from the previous work [14]. Here we just give the train of thought shortly.

For each camera pair C_{i_1}, C_{i_2} , we define an assignment variable $x(P_{j_1}^{i_1}, P_{j_2}^{i_2})$ for all person pair $P_{j_1}^{i_1}$ and $P_{j_2}^{i_2}$, ($j_1, j_2 = 1, 2, \dots, N$).

$$x(P_{j_1}^{i_1}, P_{j_2}^{i_2}) = \begin{cases} 1 & \text{if } P_{j_1}^{i_1} \text{ and } P_{j_2}^{i_2} \text{ are the same person} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

We can put all these assignment variable of persons between camera (i_1, i_2) into an $N \times N$ matrix, called assignment matrix, denoted as $X^{(i_1, i_2)}$. Its entry $X^{(i_1, i_2)}(j_1, j_2)$ equal to 1 if $P_{j_1}^{i_1}$ and $P_{j_2}^{i_2}$ are the same person, equal to 0 otherwise.

For each camera pair we can obtain a assignment matrix, finally all the $\frac{M(M-1)}{2}$ assignment matrices can be obtained. We define a overall assignment matrix X .

$$X = [X^{(1,2)}, X^{(1,3)}, \dots, X^{(1,M)}, X^{(2,1)}, X^{(2,2)}, \dots, X^{(2,M)}, \dots, X^{(M-1,M)}] \quad (8)$$

For the camera pair (i_1, i_2) , the sum of the similarity scores of assignment is given by the equation below.

$$\sum_{j_1, j_2} d'(P_{j_1}^{i_1}, P_{j_2}^{i_2})x(P_{j_1}^{i_1}, P_{j_2}^{i_2})$$

Summing over all possible camera pairs the global similarity score can be written as

$$C = \sum_{i_1 < i_2} \sum_{j_1, j_2} d'(P_{j_1}^{i_1}, P_{j_2}^{i_2}) x(P_{j_1}^{i_1}, P_{j_2}^{i_2}) = D' X^T \tag{9}$$

Equation (9) is the objective function whom we want to maximize. There are two kinds of constraints: assignment constraint and loop constraint.

1. **Assignment constraint:** A person from any camera C_{i_1} can have only one match from another camera C_{i_2} . mathematically, $\forall x(P_{j_1}^{i_1}, P_{j_2}^{i_2}) \in \{0, 1\}$

$$\begin{cases} \sum_{j_1} x(P_{j_1}^{i_1}, P_{j_2}^{i_2}) = 1 & \forall j_2 = 1, 2, \dots, N \quad \forall i_1 < i_2 \in \{1, 2, \dots, M\} \\ \sum_{j_2} x(P_{j_1}^{i_1}, P_{j_2}^{i_2}) = 1 & \forall j_1 = 1, 2, \dots, N \quad \forall i_1 < i_2 \in \{1, 2, \dots, M\} \end{cases} \tag{10}$$

As a result only one element per row and per column is 1 in each assignment matrix $X^{(i_1, i_2)}$.

2. **Loop constraint:** Figure 3(b) illustrate the inconsistency phenomenon: For a triplet of cameras $\{C_1, C_2, C_3\}$, suppose $(P_{j_1}^1, P_{j_2}^2)$, $(P_{j_2}^2, P_{j_3}^3)$ and $(P_{j_1}^1, P_{j_4}^3)$ is matched by similarity score independently, when these assignments are combined together, it leads to an infeasible scenario: $P_{j_3}^3$ and $P_{j_4}^3$ are the same person.

This infeasibility can be corrected by formula (11), called loop constraint [14].

$$\begin{aligned} x(P_{j_1}^{i_1}, P_{j_2}^{i_2}) &\geq x(P_{j_2}^{i_2}, P_{j_3}^{i_3}) + x(P_{j_3}^{i_3}, P_{j_1}^{i_1}) - 1 \\ \forall j_1, j_2 &= 1, 2, \dots, N \quad i_1 < i_2 < i_3 \in \{1, 2, \dots, M\} \end{aligned} \tag{11}$$

The overall optimization problem is shown below. The constraints of this integer programming problem reflect the consistent information.

$$\arg \min_{x(P_{j_1}^{i_1}, P_{j_2}^{i_2})} \sum_{i_1 < i_2 < M} \sum_{i, j=1}^N x(P_{j_1}^{i_1}, P_{j_2}^{i_2}) (d(P_{j_1}^{i_1}, P_{j_2}^{i_2}) + \lambda R(P_{j_1}^{i_1}, P_{j_2}^{i_2})) I(uLoss \neq 0) \tag{12}$$

$$\begin{aligned} s.t. \quad &\sum_{j_1} x(P_{j_1}^{i_1}, P_{j_2}^{i_2}) = 1 \quad \forall j_2 = 1, 2, \dots, N \quad \forall i_1 < i_2 \in \{1, 2, \dots, M\} \\ &\sum_{j_2} x(P_{j_1}^{i_1}, P_{j_2}^{i_2}) = 1 \quad \forall j_1 = 1, 2, \dots, N \quad \forall i_1 < i_2 \in \{1, 2, \dots, M\} \\ &x(P_{j_1}^{i_1}, P_{j_2}^{i_2}) \geq x(P_{j_2}^{i_2}, P_{j_3}^{i_3}) + x(P_{j_3}^{i_3}, P_{j_1}^{i_1}) - 1 \\ &\forall j_1, j_2 = 1, 2, \dots, N \quad i_1 < i_2 < i_3 \in \{1, 2, \dots, M\} \end{aligned} \tag{13}$$

The objective function (12) and constraints (13) form a binary integer program (BIP). Mathematically, this BIP problem has a mature solution approach to solve X [16].

After getting X , we can further modify similarity $d'(P_{j_1}^{i_1}, P_{j_2}^{i_2})$ to the final version $d''(P_{j_1}^{i_1}, P_{j_2}^{i_2})$ according to the Eq. (14). Finally, we obtain the final overall similarity matrix D'' .

$$d''(P_{j_1}^{i_1}, P_{j_2}^{i_2}) = \begin{cases} d'(P_{j_1}^{i_1}, P_{j_2}^{i_2}) & \text{if } x(P_{j_1}^{i_1}, P_{j_2}^{i_2}) = 0 \\ 1 & \text{if } x(P_{j_1}^{i_1}, P_{j_2}^{i_2}) = 1 \end{cases} \quad (14)$$

3 Experiment and Discussion

3.1 Experiment Setup

Data Set and Evaluation. Common dataset of state of the art person re-identification, such as VIPeR [17], CUHK01 [18], is obtained from two camera that is not satisfied our demand. In paper [14], whose issue is about camera network person re-identification, the author create the RAiD [14] dataset and use RAiD to perform experiment. The RAiD dataset is collected from four cameras covering large areas, camera 1 and 2 is indoor, camera 3 and 4 is outdoor, so it has large illumination variation. There are 43 person walking through this camera network, 41 of which appear in all the 4 cameras. We can just use these 41 subjects to unfold our experiment.

The final results are shown in terms of Cumulative Matching Characteristic (CMC) curves reflecting recognition rate.

Pairwise Similarity Score Generation. There are a lot of method to generate camera pairwise similarity score. In this paper we aim at discussing the role of the combination of temporal-spatial information and consistent constraint in camera network, so the selection is not very important, as long as stay constant through the whole experiment. In this paper we starts with extracting appearance features in the form of HSV color histogram from the images of the targets. Then we generate the similarity scores by using ICT [19], a recent work 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 and use RBF kernel SVM as classifier. For RAiD data set, we use 21 persons for training RBF kernel SVM while the rest 20 were used in testing. We take 10 images each person for both training and testing. The SVM parameter were selected using 4-fold cross validation. For each test we ran 5 independent trials and report the average results.

Temporal-spatial Information Generation. In order to verify the role of temporal-spatial information, dataset should provide relative labels for each person, specifically time when the corresponding person is captured by a camera, and the location of each camera. Without losing rationality, we can take some hypothesis for these temporal-spatial information. Firstly, because every person appears in all these 4 cameras, for each person we can generate a random permutation of $\{1, 2, 3, 4\}$ representing the path of the corresponding person. For example, suppose the generating permutation is $\{2, 4, 1, 3\}$, the meaning is that the person first captured in the camera network by camera C_2 when he is walking in this camera network, then was captured by the sequence of C_4 , and then

C_1 , finally C_3 . Secondly, we give each person a random velocity in the range of (0.8 m/s, 1.2 m/s) by experience. Thirdly, four camera is given a random coordinate (x, y) in the range that $x \in (-500 \text{ m}, 500 \text{ m})$, $y \in (-500 \text{ m}, 500 \text{ m})$, and the distance between each camera pair can be calculated. Finally, For each person we assume a time, denoted as t_0 , when the person was first captured in this camera network. t_0 is random variable and follow normal distribution, actually all sample is in the range (100 s, 300 s). Given the generated information mentioned above, we can calculate the precision time when each person was captured by each camera. Approximately, we get the temporal-spatial information of each person in RAiD dataset.

3.2 Experiment Results and Analysis

We compare the performance for camera pairs 1-2, 1-3, 1-4, 2-3, 2-4 and 3-4 respectively, as shown in Figs.4(a)–(f). Here we denoted temporal-spatial as “T-S”. The parameters of experiment are v_1 , λ and degrees of freedom n . Our choice is that $v_1 = 2.4 \text{ m/s}$, $\lambda = 10$ and $n = 700$. There are two cases, one is for camera pair 1-2 and for camera pair 3-4 where appearance variation is not so much, the other is for camera pair 1-3, 1-4, 2-3, 2-4 where there is significant lighting variation. As can be seen, the proposed method greatly improves CMC compared with conventional methods (ICT), and the method of network consistent re-identification (NCR) for both the case. The improvements are obvious after introducing spatial-temporal and consistent information. For the indoor camera pair 1-2 and the outdoor camera pair 3-4, where lighting variation is not significant, the rank1, rank2 and rank5 performance is shown in Table 1.

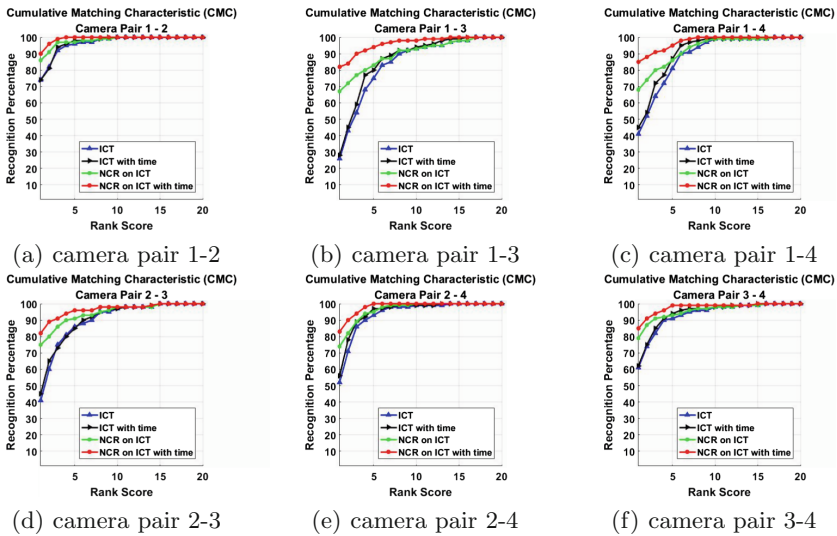


Fig. 4. CMC curves for each camera pairs

Table 1. Comparison with state-of-the-art NCR methods on RAiD datasets (%).

camera pairs	1-2			1-3			1-4			2-3			2-4			3-4		
CMC Rank R	R=1	R=2	R=5	R=1	R=2	R=5	R=1	R=2	R=5	R=1	R=2	R=5	R=1	R=2	R=5	R=1	R=2	R=5
ICT [19]	74	82	96	26	43	75	41	52	81	41	60	86	52	71	93	61	74	91
ICT with T-S	74	81	98	28	45	80	45	54	87	45	65	85	56	78	97	62	75	94
NCR [14]	86	91	97	67	72	83	68	74	86	75	80	91	74	82	95	79	87	93
NCR with T-S	90	96	100	82	84	94	85	88	95	82	89	96	83	90	100	85	91	99

As can be seen, the proposed method applied on similarity scores generated by ICT (NCR on ICT with T-S) achieve 90% and 85% rank 1 performance respectively.

For the camera pair 1-3, 1-4, 2-3 and 2-4, all of which is indoor-outdoor pair where lighting variation is significant relatively, the rank1, rank2 and rank5 performance is shown in Table 1. The proposed method applied on similarity scores generated by ICT (NCR on ICT with T-S) achieve 82%, 85%, 82% and 83% rank 1 performance respectively.

It can further be seen that when we apply only the T-S information to ICT, normally there is a slightly improvement of the rank 1 performance compared to their original rank 1 performance. This phenomenon reflect the effect of temporal-spatial information in the auxiliary aspect. Additionally, it can 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 relatively. For example, for camera pair 1-3, the rank 1 performance boosts up to 82% on application of NCR with T-S to ICT compared to their original rank 1 performance of 26% respectively. And with the help of T-S information, the rank 1 performance bring about a 15% increment compared with the NCR rank 1 performance of 67%. So it is meaningful to introduce the combination of temporal-spatial information and consistent information in the issue of person re-identification in camera network.

4 Conclusion

This paper puts forward a new framework that add temporal-spatial information in the system of camera network person re-identification. The proposed method boosts camera pairwise re-identification performance. The future directions of our research is the occasion where the hypotheses that same N person are present in each of the M cameras is removed. In additional, we will try to apply our approach to bigger networks.

Acknowledgments. This work was supported in Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 15DZ1207403, 17DZ1205602, NSF61771303).

References

1. Martinel, N., Micheloni, C.: Re-identify people in wide area camera network. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 31–36. IEEE (2012)
2. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. *Comput. Vis. Image Underst.* **117**(2), 130–144 (2013)
3. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person re-identification: what features are important? In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 391–401. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33863-2_39
4. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: BMVC, vol. 1, p. 6 (2011)
5. Ma, B., Su, Y., Jurie, F.: BiCov: a novel image representation for person re-identification and face verification. In: British Machine Vision Conference, 11 p. (2012)
6. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Boosted human re-identification using Riemannian manifolds. *Image Vis. Comput.* **30**(6), 443–452 (2012)
7. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)
8. Yang, L., Jin, R.: Distance metric learning: a comprehensive survey. *Mich. State Univ.* **2**, 78 (2006)
9. Alavi, A., Yang, Y., Harandi, M., Sanderson, C.: Multi-shot person re-identification via relational stein divergence. In: 2013 IEEE International Conference on Image Processing, pp. 3542–3546. IEEE (2013)
10. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2360–2367. IEEE (2010)
11. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1249–1258 (2016)
12. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015)
13. Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W., Li, S.Z.: Embedding deep metric for person re-identification: a study against large variations. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 732–748. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_44
14. Das, A., Chakraborty, A., Roy-Chowdhury, A.K.: Consistent re-identification in a camera network. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 330–345. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10605-2_22
15. Huang, W., Hu, R., Liang, C., Yu, Y., Wang, Z., Zhong, X., Zhang, C.: Camera network based person re-identification by leveraging spatial-temporal constraint and multiple cameras relations. In: Tian, Q., Sebe, N., Qi, G.-J., Huet, B., Hong, R., Liu, X. (eds.) MMM 2016. LNCS, vol. 9516, pp. 174–186. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-27671-7_15

16. Schrijver, A.: Theory of Linear and Integer Programming. Wiley, Hoboken (1998)
17. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol. 3 (2007)
18. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 31–44. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37331-2_3
19. Avraham, T., Gurvich, I., Lindenbaum, M., Markovitch, S.: Learning implicit transfer for person re-identification. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 381–390. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33863-2_38