Pattern Recognition 96 (2019) 106991

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Adaptation of person re-identification models for on-boarding new camera(s)

Rameswar Panda[a,1,*], Amran Bhuiyan[b], Vittorio Murino[c], Amit K. Roy-Chowdhury[a]

[a] Department of ECE, University of California, Riverside, USA
[b] LIVIA, École de Technologie Supérieure, Université du Québec, Montréal, Canada
[c] Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Italy

ARTICLE INFO

Article history:
Received 11 November 2018
Revised 16 July 2019
Accepted 31 July 2019
Available online 31 July 2019

Keywords:
Person re-identification
Camera network
Model adaptation
Limited supervision
Camera on-boarding,

ABSTRACT

Existing approaches for person re-identification have concentrated on either designing the best feature representation or learning optimal matching metrics in a static setting where the number of cameras are fixed in a network. Most approaches have neglected the dynamic and open world nature of the re-identification problem, where one or multiple new cameras may be temporarily on-boarded into an existing system to get additional information or added to expand an existing network. To address such a very practical problem, we propose a novel approach for adapting existing multi-camera re-identification frameworks with limited supervision. First, we formulate a domain perceptive re-identification method based on geodesic flow kernel that can effectively find the best source camera (already installed) to adapt with newly introduced target camera(s), without requiring a very expensive training phase. Second, we introduce a transitive inference algorithm for re-identification that can exploit the information from best source camera to improve the accuracy across other camera pairs in a network of multiple cameras. Third, we develop a target-aware sparse prototype selection strategy for finding an informative subset of source camera data for data-efficient learning in resource constrained environments. Our approach can greatly increase the flexibility and reduce the deployment cost of new cameras in many real-world dynamic camera networks. Extensive experiments demonstrate that our approach significantly outperforms state-of-the-art unsupervised alternatives whilst being extremely efficient to compute.

1. Introduction

Person re-identification (re-id), which addresses the problem of matching people across non-overlapping views in a multi-camera system, has drawn a great deal of attention in the last few years [1]. Much progress has been made in developing methods that seek either the best feature representations (e.g., [2,3]) or propose to learn optimal matching metrics (e.g., [4,5]). While they have obtained reasonable performance on commonly used benchmark datasets, we believe that these approaches have not yet considered a fundamental related problem: *Given a camera network where the inter-camera transformations/distance metrics have been learned in an intensive training phase, how can we on-board new camera(s) into the installed system with minimal additional effort?* This is an important problem to address in many realistic re-identification scenarios, where one or multiple new cameras may be temporarily inserted into an existing system to get additional information.

To illustrate such a problem, let us consider a scenario with $\mathcal{N}$ cameras for which we have learned the "optimal" pair-wise distance metrics, so providing high re-identification accuracy for all camera pairs. However, during a particular event, a new camera may be temporarily introduced to cover a certain related area that is not well-covered by the existing network of $\mathcal{N}$ cameras (see Fig. 1 for an example). Despite the dynamic and open nature of the world, almost all work in re-identification assume a *static* and *closed* world model of the re-id problem where the number of cameras are fixed in a network. Given a newly introduced camera, traditional re-id methods will try to relearn the inter-camera transformations/distance metrics using a costly training phase. This is impractical since labeling data in the new camera and then learning transformations with the others is time-consuming, and defeats the entire purpose of temporarily introducing the additional camera. Thus, there is a pressing need to develop *unsupervised* approaches for integrating new camera(s) into an existing re-identification framework with limited supervision.

* Corresponding author.
E-mail address: rpand002@ucr.edu (R. Panda).
[1] The author is currently a research scientist at IBM Research, Cambridge, MA 02142.

https://doi.org/10.1016/j.patcog.2019.106991
0031-3203/

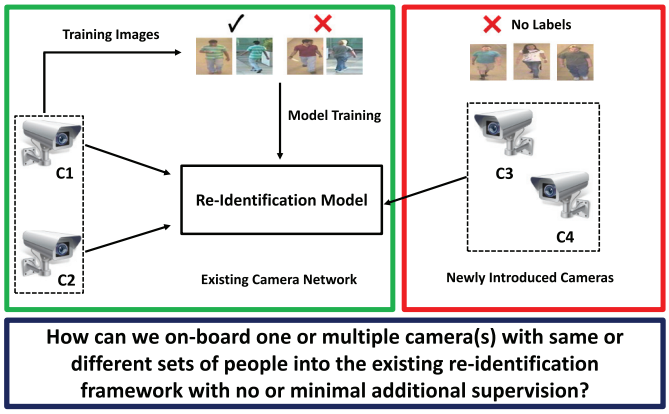


Fig. 1. Consider an existing network with two cameras c_1 and c_2 where we have learned a re-identification model using pair-wise training data from both of the cameras. During the operational phase, two new cameras c_3 and c_4 are introduced to cover a certain area that is not well covered by the existing 2 cameras. Most of the existing methods do not consider such dynamic nature of a re-id model. In contrast, we propose an unsupervised approach for on-boarding new camera(s) into the existing re-identification framework by exploring: *what is the best source camera(s) to pair with the new cameras and how can we exploit the best source camera(s) to improve the matching accuracy across the other existing cameras?*

Domain adaptation [6] has recently been successful in many vision problems such as object recognition [7,8] and activity classification [9] with multiple classes or domains. The main objective is to scale learned systems from a source domain to a target domain without requiring prohibitive amount of training data in the target domain. Considering newly introduced camera(s) as target domain, we pose an important question in this paper: *Can unsupervised domain adaptation be leveraged upon for on-boarding new camera(s) into person re-identification frameworks with limited supervision?*

Unlike object recognition [7], domain adaptation for person re-identification has additional challenges. A central issue in domain adaptation is *which source to transfer from*. When there is only one source of information available which is highly relevant to the task of interest, then domain adaptation is much simpler than in the more general and realistic case where there are multiple sources of information of greatly varying relevance. Re-identification in a dynamic network falls into the latter, more difficult case. Specifically, given multiple source cameras (already installed) and a target camera (newly introduced), *how can we select the best source camera to pair with the target camera?* The problem can be easily extended to multiple additional cameras being introduced.

Moreover, once the best source camera is identified, *how can we exploit this information to improve the re-identification accuracy of other camera pairs?* For instance, let us consider c_1 being the best source camera for the newly introduced camera c_3 in Fig. 1. Once the pair-wise distance metric between c_1 and c_3 is obtained, can we exploit this information to improve the re-identification accuracy across (c_2–c_3)? This is an especially important problem because it will allow us to now match data in the newly inserted target camera c_3 with all the previously installed cameras.

Given a network with thousands of cameras involving large number of images, finding the best source camera for a newly introduced camera can involve intensive computation of the pair-wise kernels over the whole set of images. Thus, it is important to automatically select an informative subset of the source data to pair with the target domain data. Specifically, *can we select an informative subset of source camera data that share similar characteristics as target camera data and use those for model adaptation in resource constrained environments?* This is crucial to increase the flexibility and decrease the deployment cost of newly introduced cameras in large-scale dynamic camera networks.

1.1. Overview of solution strategy

We first propose an unsupervised approach based on geodesic flow kernel [8,10] that can effectively find the best source camera to adapt with a target camera. Given camera pairs, each consisting of 1 (out of $\mathcal{N}$) source camera and a target camera, we first compute a kernel over the subspaces representing the data of both cameras and then use it to find the kernel distance across the source and target camera. Then, we rank the source cameras based on the average distance and choose the one with lowest distance as the best source camera to pair with the target camera. This is intuitive since a camera which is closest to the newly introduced camera will give the best re-identification performance on the target camera and hence, is more likely to adapt better than others. In other words, a source camera with lowest distance with respect to a target camera indicates that both of the sensors could be similar to each other and their features may be similarly distributed. Note that we learn the kernel with the labeled data from the source camera only.

We then introduce a transitive inference algorithm for person re-identification that can exploit information from best source camera to improve accuracy across other camera pairs. Reminding the previous example in Fig. 1 in which source camera c_1 best matches with target camera c_3, our proposed transitive algorithm establishes a path between camera pair ($c_2 - c_3$) by marginalization over the domain of possible appearances in best source camera c_1. Specifically, c_1 plays the role of a "connector" between c_2 and c_3. Experiments show that this approach consistently increases the overall re-identification accuracy in multiple networks by improving matching performance across camera pairs, while exploiting side information from best source camera.

Moreover, we also propose a source-target selective adaptation strategy that uses a subset of source camera data instead of all existing data to compute the kernels for finding the best source camera to pair with a target camera. Our key insight is that not all images in a source camera are equally effective in terms of adaptability and hence using an informative subset of images from the existing source cameras whose characteristics are similar to those of the target camera can well adapt the models in resource constrained environments. We develop a target-aware sparse prototype selection strategy using $\ell_{2,1}$-norm optimization to select a subset of source data that can efficiently describe the target set. Experiments demonstrate that our source-target selective learning strategy achieves the same performance as the full set while only using about 30% of images from the source cameras. Interestingly, our approach with prototype selection outperforms the compared methods that use all existing source data by a margin of about 8%-10% in rank-1 accuracy with only requiring about 10% of source camera data while introducing new cameras.

1.2. Contributions

We address a novel, and very practical problem in this paper—how to add one or more cameras temporarily to an existing network and exploit it for person re-identification, without also adding a very expensive training phase. Towards solving this problem, we make the following contributions: (i) an unsupervised re-identification approach based on geodesic flow kernel that can find the best source camera to adapt with newly introduced target camera(s) in a dynamic camera network; (ii) a transitive inference algorithm to exploit side information from the best source camera to improve the matching accuracy across other source-target camera pairs; (iii) a target-aware sparse prototype selection strategy using $\ell_{2,1}$-norm optimization to select an informative subset of source camera data for data-efficient learning in resource constrained environments; (iv) rigorous experiments validating the advantages

of our approach over existing alternatives on multiple benchmark datasets with variable number of cameras.

2. Related work

Person re-identification has been studied from different perspectives (see [1] for a recent survey). Here, we focus on some representative methods closely related to our work.

Supervised Re-identification. Most existing person re-identification techniques are based on supervised learning. These methods either seek the best feature representation [2,3,11,12] or learn discriminant metrics/dictionaries [13–17] that yield an optimal matching score between two cameras or between a gallery and a probe image. Recently, deep learning methods have shown significant performance improvement on person re-id [18–24]. Combining feature representation and metric learning with an end-to-end deep neural networks is also a recent trend in re-identification [25–27]. Considering that a modest-sized camera network can easily have hundreds of cameras, these supervised re-id models will require huge amount of labeled data which are difficult to collect in real-world settings. In an effort to bypass tedious labeling of training data in supervised re-id models, there has been recent interest in using active learning for labeling examples in an interactive manner [28–31]. However, all these approaches consider a static camera network unlike the problem domain we consider.

Unsupervised Re-identification. Unsupervised learning models have received little attention in person re-identification because of their weak performance on benchmarking datasets compared to supervised methods. Representative methods along this direction use either hand-crafted appearance features [32,33] or saliency statistics [34] for matching persons without requiring huge amount of labeled data. Dictionary learning based methods have also been utilized in an unsupervised setting [35,36]. Recently, Generative Adversarial Networks (GAN) has also been used in semi-supervised settings [37,38]. Although being scalable in real-world settings, these approaches have not yet considered the dynamic nature of the re-identification problem, where new cameras can be introduced at any time to an existing network.

Open World Re-Identification. Open world recognition has been introduced in [39] as an attempt to move beyond the static setting to a dynamic and open setting where the number of training images/classes are not fixed in recognition. Recently there have been few works in person re-identification [40,41] by assuming that gallery and probe sets contain different identities of persons. Unlike such approaches, we consider another yet important aspect of open world person re-identification where the camera network is dynamic and the system has to incorporate a new camera with minimal additional effort.

Domain Adaptation. Domain adaptation [6], which aims to adapt a source domain to a target domain, has been successfully used in many areas of computer vision, e.g., object classification, and action recognition. Despite its applicability in classical vision tasks, domain adaptation for re-identification still remains as a challenging and under addressed problem. Recently, domain adaptation for re-id has begun to be considered [42–44]. However, these studies consider only improving the re-identification performance in a static camera network with fixed number of cameras. Furthermore, most of these approaches learn supervised models using labeled data from the target domain.

This paper has significant differences with our preliminary work in [45]. First, we develop a target-aware sparse prototype selection strategy for selecting a subset of source camera data to pair with a target camera while computing kernels (Section 3.4). This is especially an important problem as it will increase the flexibility and decrease the deployment cost of newly introduced cameras in many real world dynamic camera networks. Second, we extend our approach to more realistic scenarios where multiple cameras can be introduced to the network at the same time and show the effectiveness of our approach in a large-scale network of 16 cameras (Section 3.5). We also consider different identities of person appearing in the newly introduced camera as in many real world settings (Section 3.6). Third, we conduct comprehensive experiments to analyze the effect of feature representation and subspace dimension on the re-identification performance along with new experiments involving large number of images and cameras, different sets of people in target camera and model adaptation with prototype selection for resource-constrained environments (Section 4).

3. Proposed methodology

To on-board new camera(s) into an existing person re-identification framework, we first formulate an unsupervised approach based on geodesic flow kernel to find the best source camera (Section 3.2) and then propose a transitive inference algorithm to exploit information from the best source camera for improving matching accuracies across other source-target camera pairs (Section 3.3). Next, we describe the details on our target-aware sparse prototype selection strategy to select an informative subset of source camera data in Section 3.4.

3.1. Initial setup

Our proposed framework starts with an installed camera network where the discriminative distance metrics between each camera pairs is learned using a off-line intensive training phase. Let there be $\mathcal{N}$ cameras in a network and the number of possible camera pairs is $\binom{\mathcal{N}}{2}$. Let $\{(\mathbf{x}_i^{\mathcal{A}}, \mathbf{x}_i^{\mathcal{B}})\}_{i=1}^{m}$ be a set of training samples, where $\mathbf{x}_i^{\mathcal{A}} \in \mathbb{R}^D$ represents feature representation of a training sample from camera view $\mathcal{A}$ and $\mathbf{x}_i^{\mathcal{B}} \in \mathbb{R}^D$ represents feature representation of the same person in a different camera view $\mathcal{B}$.

Given the training data, we follow KISS metric learning (KISSME) [46] and compute the pairwise matrices such that distance between images of the same individual is less than distance between images of different individuals. The basic idea of KISSME is to learn the Mahalanobis distance by considering a log likelihood ratio test of two Gaussian distributions. The likelihood ratio test between dissimilar pairs and similar pairs can be written as

$$\mathcal{R}(\mathbf{x}_i^{\mathcal{A}}, \mathbf{x}_j^{\mathcal{B}}) = \log \frac{\frac{1}{\mathcal{C}_{\mathcal{D}}} \exp(-\frac{1}{2}\mathbf{x}_{ij}^T \Sigma_{\mathcal{D}}^{-1} \mathbf{x}_{ij})}{\frac{1}{\mathcal{C}_{\mathcal{S}}} \exp(-\frac{1}{2}\mathbf{x}_{ij}^T \Sigma_{\mathcal{S}}^{-1} \mathbf{x}_{ij})} \tag{1}$$

where $\mathbf{x}_{ij} = \mathbf{x}_i^{\mathcal{A}} - \mathbf{x}_j^{\mathcal{B}}$, $\mathcal{C}_{\mathcal{D}} = \sqrt{2\pi|\Sigma_{\mathcal{D}}|}$, $\mathcal{C}_{\mathcal{S}} = \sqrt{2\pi|\Sigma_{\mathcal{S}}|}$, $\Sigma_{\mathcal{D}}$ and $\Sigma_{\mathcal{S}}$ are covariance matrices of dissimilar and similar pairs respectively. With simple manipulations, (1) can be written as $\mathcal{R}(\mathbf{x}_i^{\mathcal{A}}, \mathbf{x}_j^{\mathcal{B}}) = \mathbf{x}_{ij}^T \mathbf{M} \mathbf{x}_{ij}$, where $\mathbf{M} = \Sigma_{\mathcal{S}}^{-1} - \Sigma_{\mathcal{D}}^{-1}$ is the Mahalanobis distance between covariances associated to a pair of cameras. We perform an Eigen-analysis to ensure $\mathbf{M} \in \mathbb{R}^{D \times D}$ is positive semi-definite [46].

Note that our approach is agnostic to the choice of metric learning algorithm used to learn the optimal metrics across camera pairs in an existing network. We adopt KISSME in this work since it is simple to compute and has shown to perform satisfactorily on the person re-identification problem.

3.2. Discovering the best source camera

Objective. Given an existing camera network where optimal camera pair-wise matching metrics are computed using the above training phase, our first objective is to select the best source camera which has the lowest kernel distance with respect to the newly

inserted camera. Towards this, we adopt an unsupervised strategy based on geodesic flow kernel [8,10] to compute the distances without requiring any labeled data from the new cameras.

Approach Details. Our approach consists of the following steps: (i) compute geodesic flow kernels between the new (target) camera and other existing cameras (source); (ii) use the kernels to determine the distance between them; (iii) rank the source cameras based on distance with respect to the target camera and choose the one with the lowest as best source camera.

Let $\{\mathcal{X}^s\}_{s=1}^{\mathcal{N}}$ be the $\mathcal{N}$ source cameras and $\mathcal{X}^{\mathcal{T}}$ be the newly introduced target camera. To compute the kernels in an unsupervised way, we extend a previous method [10] that adapts classifiers in the context of object recognition to the re-identification in a dynamic camera network. The main idea of our approach is to compute the low-dimensional subspaces representing data of two cameras (one source and one target) and then map them to two points on a Grassmanian. Intuitively, if these two points are close by on the Grassmanian, then the computed kernel would provide high matching performance on the target camera. In other words, both of the cameras could be similar to each other and their features may be similarly distributed over the corresponding subspaces. For simplicity, let us assume we are interested in computing the kernel matrix $\mathbf{K}^{\mathcal{ST}} \in \mathbb{R}^{D\times D}$ between the source camera $\mathcal{X}^{\mathcal{S}}$ and a newly introduced target camera $\mathcal{X}^{\mathcal{T}}$. Let $\tilde{\mathcal{X}}^{\mathcal{S}} \in \mathbb{R}^{D\times d}$ and $\tilde{\mathcal{X}}^{\mathcal{T}} \in \mathbb{R}^{D\times d}$ denote the d-dimensional subspaces, computed using Partial Least Squares (PLS) and Principal Component Analysis (PCA) on the source and target camera, respectively. Note that we can not use PLS on the target camera since it is a supervised dimension reduction technique and requires label information for computing the subspaces.

Given both of the subspaces, the closed loop solution to the geodesic flow kernel across two cameras is defined as

$$\mathbf{x}_i^{\mathcal{S}^T}\mathbf{K}^{\mathcal{ST}}\mathbf{x}_j^{\mathcal{T}} = \int_0^1 (\psi(\mathbf{y})^T\mathbf{x}_i^{\mathcal{S}})^T(\psi(\mathbf{y})\mathbf{x}_j^{\mathcal{T}})\ d\mathbf{y} \tag{2}$$

where $\mathbf{x}_i^{\mathcal{S}}$ and $\mathbf{x}_j^{\mathcal{T}}$ represent feature descriptor of ith and jth sample in source and target camera respectively. $\psi(\mathbf{y})$ is the geodesic flow parameterized by a continuous variable $\mathbf{y} \in [0, 1]$ and represents how to smoothly project a sample from the original D-dimensional feature space onto the corresponding low dimensional subspace. The geodesic flow $\psi(\mathbf{y})$ can be defined as [10],

$$\psi(\mathbf{y}) = \begin{cases} \tilde{\mathcal{X}}^{\mathcal{S}} & \text{if } \mathbf{y} = 0 \\ \tilde{\mathcal{X}}^{\mathcal{T}} & \text{if } \mathbf{y} = 1 \\ \tilde{\mathcal{X}}^{\mathcal{S}}\mathcal{U}_1\mathcal{V}_1(\mathbf{y}) - \tilde{\mathcal{X}}_o^{\mathcal{S}}\mathcal{U}_2\mathcal{V}_2(\mathbf{y}) & \text{otherwise} \end{cases} \tag{3}$$

where $\tilde{\mathcal{X}}_o^{\mathcal{S}} \in \mathbb{R}^{D\times(D-d)}$ is the orthogonal matrix to $\tilde{\mathcal{X}}^{\mathcal{S}}$ and $\mathcal{U}_1, \mathcal{V}_1, \mathcal{U}_2, \mathcal{V}_2$ are given by the following pairs of SVDs,

$$\mathcal{X}^{\mathcal{S}^T}\mathcal{X}^{\mathcal{T}} = \mathcal{U}_1\mathcal{V}_1\mathcal{P}^T,\ \ \mathcal{X}_o^{\mathcal{S}^T}\mathcal{X}^{\mathcal{T}} = -\mathcal{U}_2\mathcal{V}_2\mathcal{P}^T \tag{4}$$

With the above defined matrices, $\mathbf{K}^{\mathcal{ST}}$ can be computed as

$$\mathbf{K}^{\mathcal{ST}} = \begin{bmatrix} \tilde{\mathcal{X}}^{\mathcal{S}}\mathcal{U}_1 & \tilde{\mathcal{X}}_o^{\mathcal{S}}\mathcal{U}_2 \end{bmatrix} \mathcal{G} \begin{bmatrix} \mathcal{U}_1^T\mathcal{X}^{\mathcal{S}^T} \\ \mathcal{U}_2^T\mathcal{X}_o^{\mathcal{S}^T} \end{bmatrix} \tag{5}$$

where $\mathcal{G} = \begin{bmatrix} \text{diag}[1 + \frac{\sin(2\theta_i)}{2\theta_i}] & \text{diag}[\frac{(\cos(2\theta_i)-1)}{2\theta_i}] \\ \text{diag}[\frac{(\cos(2\theta_i)-1)}{2\theta_i}] & \text{diag}[1 - \frac{\sin(2\theta_i)}{2\theta_i}] \end{bmatrix}$ and $[\theta_i]_{i=1}^d$ represents the principal angles between source and target camera. Once we compute all pairwise geodesic flow kernels between a target camera and source cameras using (5), our next objective is to find the distance across all those pairs. A source camera which is closest to the new camera is more likely to adapt better than others. We follow [47] to compute distance between a target and source camera pair. Specifically, given a kernel matrix $\mathbf{K}^{\mathcal{ST}}$, the distance between data points of a source and target camera is defined as

$$\mathbf{D}^{\mathcal{ST}}(\mathbf{x}_i^{\mathcal{S}}, \mathbf{x}_j^{\mathcal{T}}) = \mathbf{x}_i^{\mathcal{S}^T}\mathbf{K}^{\mathcal{ST}}\mathbf{x}_i^{\mathcal{S}} + \mathbf{x}_j^{\mathcal{T}^T}\mathbf{K}^{\mathcal{ST}}\mathbf{x}_j^{\mathcal{T}} - 2\mathbf{x}_i^{\mathcal{S}^T}\mathbf{K}^{\mathcal{ST}}\mathbf{x}_j^{\mathcal{T}} \tag{6}$$

where $\mathbf{D}^{\mathcal{ST}} \in \mathbb{R}^{n_s\times n_t}$ represents the kernel distance matrix defined over a source and target camera. n_s and n_t represent the number of images in source and target camera respectively. We compute the average of $\mathbf{D}^{\mathcal{ST}}$ and consider it as the distance between two cameras. Finally, we chose the one that has the lowest distance a best source camera to pair with the newly introduced camera.

Remark 1. Note that we do not use any labeled data from the target camera to either compute the geodesic flow kernels in (5) or the kernel distance matrices in (6). Hence, our proposed approach can be applied integrate new cameras in a large-scale camera network with minimal additional effort.

Remark 2. We assume that the newly introduced camera will be close to at least one of the installed ones since we consider them to be operating in the same time window with same set of people appear in all camera views, as in most prior works except the work in [40]. However, our proposed adaptation approach is not limited to this constrained setting as we compute the view similarity in a completely unsupervised manner and hence can be easily applied in real-world settings where different sets of people appear in different camera views. To the best of our knowledge, this is first work which can be employed in fully open world re-identification systems considering both dynamic network and different identity of persons across cameras (see illustrative experiments in Section 4.7).

Remark 3. We also assume that person detections are available apriori before learning the re-identification models. However, in the dynamic environment addressed in this paper an important issue is the person detector for which the new camera could be even more challenging than for the re-id algorithm. Thus, it is critical to jointly adapt the person detectors and re-identification models for optimal performance in real world dynamic camera networks–we leave this as an interesting future work.

3.3. Transitive inference for re-identification

Objective. In the previous section we have presented an unsupervised approach for finding best source camera to pair with the target camera. Once the best source camera is identified, another question that remains in adapting models is: *can we exploit the best source camera information to improve the re-identification accuracy across other camera pairs?* Specifically, our objective is to exploit $\mathbf{K}^{\mathcal{S}^\star\mathcal{T}}$ and pair-wise optimal metrics learned in Section 3.1 to improve the matching accuracies of the target camera in a network.

Approach Details. Let $\{\mathbf{M}^{ij}\}_{i,j=1,i<j}^{\mathcal{N}}$ be the optimal pairwise metrics learned in a network of $\mathcal{N}$ cameras following Section 3.1 and $\mathcal{S}^\star$ be the best source camera for a newly introduced target camera $\mathcal{T}$ following Section 3.2.

Motivated by the effectiveness of Schur product (a.k.a. Hadamard product) for improving the matrix consistency and reliability in multi-criteria decision making [48], we develop a simple yet effective transitive algorithm for exploiting information from the best source camera. Our problem naturally fits to such decision making systems since our goal is to establish a path between two cameras via the best source camera. Given the best source camera $\mathcal{S}^\star$, we compute the kernel matrix between remaining source and target camera as follows,

$$\tilde{\mathbf{K}}^{\mathcal{ST}} = \mathbf{M}^{\mathcal{SS}^\star} \odot \mathbf{K}^{\mathcal{S}^\star\mathcal{T}},\ \ \forall[\mathcal{S}]_{i=1}^{\mathcal{N}},\ \ \mathcal{S} \neq \mathcal{S}^\star \tag{7}$$

where $\tilde{\mathbf{K}}^{\mathcal{ST}} \in \mathbb{R}^{D\times D}$ represents the updated kernel matrix between source camera $\mathcal{S}$ and target camera $\mathcal{T}$ by exploiting information from best source camera $\mathcal{S}^\star$. The operator $\odot$ denotes Schur product of two matrices. Eq. (7) establishes an indirect path between

camera pair $(\mathcal{S}, \mathcal{T})$ by marginalization over the domain of possible appearances in best source camera $\mathcal{S}^\star$. In other words, camera $\mathcal{S}^\star$ plays a role of connector between the target camera $\mathcal{T}$ and all other source cameras.

Summarizing, to incorporate new camera(s) in an existing network, we use the kernel matrix $\mathbf{K}^{\mathcal{S}^\star\mathcal{T}}$ in (5) to obtain the re-id accuracy across the new camera and best source camera, whereas we use the updated kernel matrices, computed using (7) to find the matching accuracy across the target camera and remaining source cameras in an existing network.

3.4. Learning kernels with prototype selection

Objective. For many applications with limited computation and communication resources, there is an imperative need of methods that could extract an informative subset from the source camera data for computing the kernels instead of all existing data. Thus, our main objective in this section is to develop a prototype selection strategy for finding a subset of source camera data that share similar characteristics as the target camera and then use those for discovering the best source camera in Section 3.2.

Approach Details. Motivated by sparse subset selection [49], we develop an efficient optimization framework to extract a sparse set of source camera images that are informative about the given source camera as well as informative about the target camera. We formulate the following objective function,

$$\min_{\mathcal{Z}^s \in \mathbb{R}^{n_s \times n_s}, \mathcal{Z}^\mathcal{T} \in \mathbb{R}^{n_t \times n_t}} \frac{1}{2}\left(\|\mathcal{X}^s - \mathcal{X}^s\mathcal{Z}^s\|_F^2 + \alpha\|\mathcal{X}^\mathcal{T} - \mathcal{X}^s\mathcal{Z}^\mathcal{T}\|_F^2\right) + \lambda\left(\|\mathcal{Z}^s\|_{2,1} + \|\mathcal{Z}^\mathcal{T}\|_{2,1}\right) \tag{8}$$

where $\alpha > 0$ balances the penalty between errors in the reconstruction of source camera data $\mathcal{X}^s \in \mathbb{R}^{D \times n_s}$ and errors in the reconstruction of target camera data $\mathcal{X}^\mathcal{T} \in \mathbb{R}^{n_s \times n_t}$. $||\mathcal{Z}^s||_{2,1} = \sum_{i=1}^{m} ||\mathcal{Z}_i^s||_2$ and $||\mathcal{Z}_i^s||_2$ is the ℓ_2-norm of the ith row of $\mathcal{Z}^s$. $\lambda > 0$ is a sparsity regularization parameter.

The objective function is intuitive: minimization of (8) favors selecting a sparse set of prototypes that simultaneously reconstructs the source camera data $\mathcal{X}^s$ via $\mathcal{Z}^s$, as well as the target camera data $\mathcal{X}^\mathcal{T}$ via $\mathcal{Z}^\mathcal{T}$, with high accuracy. Specifically, rows in $\mathcal{Z}^s$ provide information on relative importance of each image in describing the source camera $\mathcal{X}^s$, while rows in $\mathcal{X}^\mathcal{T}$ give information on relative importance of each image in $\mathcal{X}^s$ in describing target camera $\mathcal{X}^\mathcal{T}$. Given the two sparse coefficient matrices, our next goal is to select a unified set of images from source camera that share similar characteristics with target camera. To achieve this, we propose to minimize the following objective function:

$$\min_{\mathcal{Z}^s, \mathcal{Z}^\mathcal{T}} \frac{1}{2}\left(\|\mathcal{X}^s - \mathcal{X}^s\mathcal{Z}^s\|_F^2 + \alpha\|\mathcal{X}^\mathcal{T} - \mathcal{X}^s\mathcal{Z}^\mathcal{T}\|_F^2\right) + \lambda\left(\|\mathcal{Z}^s\|_{2,1} + \|\mathcal{Z}^\mathcal{T}\|_{2,1}\right) + \beta||\mathcal{Z}_c||_{2,1} \; s.t. \; \mathcal{Z}_c = [\mathcal{Z}^s|\mathcal{Z}^\mathcal{T}] \tag{9}$$

where $\ell_{2,1}$-norm on the consensus matrix $\mathcal{Z}_c \in n_s \times (n_s + n_t)$ enables $\mathcal{Z}^s$ and $\mathcal{Z}^\mathcal{T}$ to have the similar sparse patterns and share the common components. In each round of the optimization, the updated sparse coefficient matrices in the former rounds can be used to regularize the current optimization criterion. Thus, it can uncover the shared knowledge of $\mathcal{Z}^s$ and $\mathcal{Z}^\mathcal{T}$ by suppressing irrelevant images that are less effective in terms of adaptability to the newly introduced camera.

Optimization. Since problem (9) is non-smooth involving multiple $\ell_{2,1}$-norms, it is difficult to optimize directly. Motivated by the effectiveness of Half-quadratic optimization [50], we devise an iterative algorithm to solve (9) by minimizing its augmented function alternatively as shown in Algorithm 1. More details on the optimization are included in the supplementary material.

Algorithm 1 Algorithm for Solving Problem (9).

Input: Feature matrices $\mathcal{X}^s$ and $\mathcal{X}^\mathcal{T}$;Parameters α, λ, β, set $t = 0$
Initialize $\mathcal{Z}^s$ and $\mathcal{Z}^\mathcal{T}$ randomly, set $\mathcal{Z}_c = [\mathcal{Z}^s|\mathcal{Z}^\mathcal{T}]$
Output: Optimal sparse coefficient matrix $\mathcal{Z}_c$.
while *not converged* **do**
1. Compute P^t, Q^t and R^t as:
$P_{ii} = \frac{1}{2\sqrt{||\mathcal{Z}_i^s||_2^2 + \epsilon}}$, $Q_{ii} = \frac{1}{2\sqrt{||\mathcal{Z}_i^\mathcal{T}||_2^2 + \epsilon}}$,
$R_{ii} = \frac{1}{2\sqrt{||\mathcal{Z}_{ci}||_2^2 + \epsilon}}$
2. Compute $\mathcal{Z}^{s^{t+1}}$ and $\mathcal{Z}^{\mathcal{T}^{t+1}}$ as:
$\mathcal{Z}^s = (\mathcal{X}^{s^T}\mathcal{X}^s + 2\lambda P + 2\beta R)^{-1}\mathcal{X}^{s^T}\mathcal{X}^s$
$\mathcal{Z}^\mathcal{T} = (\alpha\mathcal{X}^{s^T}\mathcal{X}^s + 2\lambda Q + 2\beta R)^{-1}\alpha\mathcal{X}^{s^T}\mathcal{X}^\mathcal{T}$
3. Compute $\mathcal{Z}_c^{t+1}$ as: $\mathcal{Z}_c^{t+1} = [\mathcal{Z}^{s^{t+1}} \mid \mathcal{Z}^{\mathcal{T}^{t+1}}]$;
4. $t = t + 1$;
end while

Once the problem (9) is solved, we first sort the source camera images by decreasing importance according to the ℓ_2 norms of the rows of $\mathcal{Z}_c$. To summarize, we first learn the pair-wise kernels across all the unlabeled target camera data and selected prototypes from the source camera to discover the best camera as in Section 3.2. Second, we adopt the same transitive inference algorithm mentioned in Section 3.3 to exploit the information from the best source camera to improve the person re-identification accuracy across remaining source-target camera pairs.

3.5. Extension to multiple newly introduced cameras

Our approach is not limited to a single camera and can be easily extended to even more realistic scenarios where multiple cameras are introduced to an existing network at the same time. Given multiple newly introduced cameras, one can follow two different strategies to adapt re-identification models in dynamic camera networks. Specifically, one can easily find a common best source camera based on lowest average distance to pair with all the new cameras or multiple best source cameras, one for each target camera, in an unsupervised way similar to the above approach (see experiments in Section 4.3).

3.6. Extension to semi-supervised adaptation

Although our framework is designed for unsupervised adaptation of re-identification models, it can be easily extended if labeled data from the newly introduced camera become available. Specifically, the label information from target camera can be encoded while computing subspaces. That is, instead of using PCA for estimating the subspaces, we can use Partial Least Squares (PLS) to compute the discriminative subspaces on the target data by exploiting the labeled information. PLS has shown to be effective in finding discriminative subspaces by projecting labeled data into a common subspace [51]. This essentially leads to semi-supervised adaptation in a camera network (see experiments in Section 4.6).

4. Experiments

In this section, we evaluate the performance of our approach by performing several experiments on multiple benchmark datasets.

4.1. Datasets and settings

Datasets. We conduct experiments on five different benchmark datasets to verify the effectiveness of our framework,

namely WARD [52], RAiD [53], SAIVT-SoftBio [54], Shinpuhkan2014 [55] and Market-1501 [56]. The number of cameras in WARD, RAiD and SAIVT-SoftBio are 3, 4, and 8 respectively. Shinpuhkan2014 dataset with 16 cameras is one of the largest publicly available dataset in terms of number of cameras, while the Market-1501 dataset is one of the largest dataset in terms of number of images containing 32,668 images across 6 cameras. Since Market-1501 dataset is not designed for camera pair-wise re-identification, we pre-process it according to our experimental setting and choose 605 persons who are present across all cameras. More details on the datasets are available in the supplementary material.

Feature Extraction and Matching. The feature extraction stage consists of extracting Local Maximal Occurrence (LOMO) feature [57] for person representation. The descriptor has 26,960 dimensions. We apply principal component analysis to reduce the dimensionality to 100 in all our experiments, as in [46]. Without low-dimensional feature, it is computationally infeasible to inverse covariance matrices as discussed in [46]. We use kernel distance [47] (Eq. (6)) to compute both distance between cameras and matching scores.

Performance Measures. We show results using Cumulative Matching Characteristic (CMC) curves and normalized Area Under Curve (nAUC) values, as is common practice in re-identification literature. CMC curve is a plot of recognition performance versus ranking score and represents the expectation of finding correct match in the top k matches. nAUC gives an overall score of how well a re-id method performs irrespective of the dataset size.

Experimental Settings. All the images for each dataset are normalized to 128×64 for being consistent with the evaluations carried out by state-of-the-art methods [3,33,53]. Following the literature [46,53,57], the train and test set are kept disjoint by picking half of the available data for training set and rest of the half for testing. We repeated each task 10 times by randomly picking 5 images from each identity both for train and test time. The subspace dimension for all the possible combinations are kept 50.

Compared Methods. We compare our approach with both unsupervised and supervised alternatives as follows.

(a) Unsupervised Methods. We compare our approach with several unsupervised alternatives which fall into two categories: (i) hand-crafted feature-based methods including `CPS` [33] and `SDALF` [3], (ii) two domain adaptation based methods (`Best-GFK` and `Direct-GFK`) based on geodesic flow kernel [10]. For `Best-GFK` baseline, we compute the re-id performance of a camera pair by applying the kernel matrix, $\mathbf{K}^{\mathcal{S}^\star\mathcal{T}}$ computed between best source and target camera [10], whereas in `Direct-GFK` baseline, we use the kernel matrix computed directly across source and target camera using (5). The purpose of comparing with `Best-GFK` is to show that the kernel matrix computed across the best source and target camera does not produce optimal re-id performance in computing matching performance across other source cameras and the target camera. On the other hand, the purpose of comparing with `Direct-GFK` baseline is to explicitly show the effectiveness of our transitive algorithm in improving re-id performance in a dynamic camera network.

We use publicly available codes for `CPS` and `SDALF` and tested on our experimented datasets. We use the same features as the proposed one and kept the parameters same as mentioned in the published works. We also implement both `Best-GFK` and `Direct-GFK` baselines under the same experimental settings to have a fair comparison with our proposed method.

(b) Supervised Methods. We compare with several supervised alternatives which fall into two categories: (i) feature transformation based methods including `FT` [11], `ICT` [58], `WACN` [52], (ii) metric learning based methods including `KISSME` [46], `LDML` [59], `XQDA` [57] and `MLAPG` [15]. Our model can operate with any initial network setup and hence we show our results with both KISSME and Logistic Discriminant-based Metric Learning (LDML) [59], denoted as `Ours-K` and `Ours-L`, respectively. Note that we could not compare with recent deep learning based methods as they are mostly specific to a static setting and also their pairwise camera results are not available on the experimented datasets. We did not re-implement such methods in our dynamic setting as it is very difficult to exactly emulate all the implementation details.

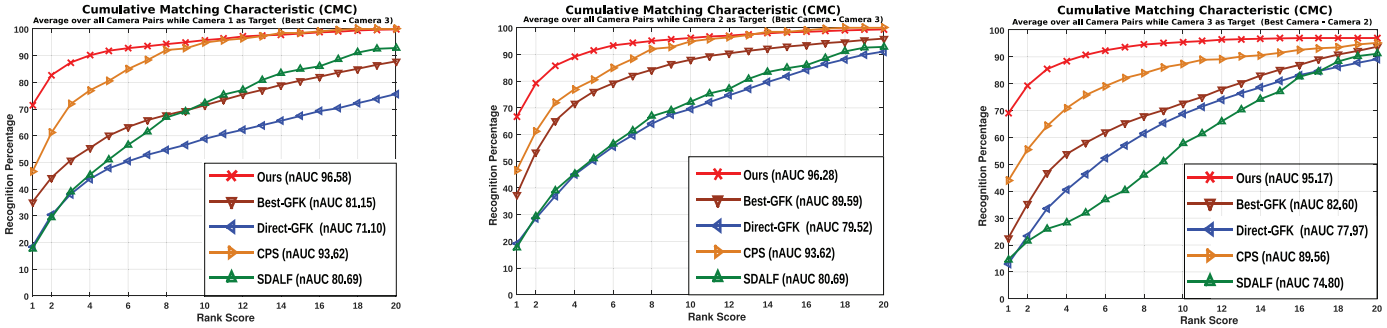
To report existing feature transformation based methods results, we use prior published performances from [53]. For metric learning based methods, we use publicly available codes and test on our experimented datasets.

4.2. Re-identification by introducing a new camera

Goal. The main goal of this experiment is to analyze (a) the performance of our unsupervised approach while finding the best source camera to pair with the target camera (Section 3.2) and (b) performance of our transitive inference approach for exploiting the information from best source camera to improve the re-identification accuracy of other camera pairs? (Section 3.3)

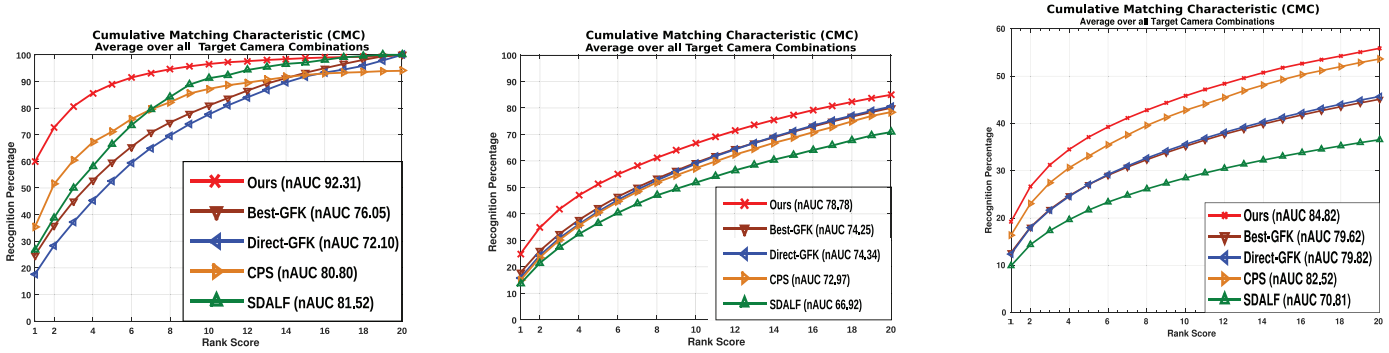
Implementation Details. We considered one camera as newly introduced target camera and all the other as source cameras. We considered all the possible combinations for conducting experiments. We first pick which source camera matches best with the target one, and then use the proposed transitive algorithm to compute the re-id performance across remaining camera pairs.

Results. Fig. 2 show the results for all possible combinations on the 3 camera WARD dataset, whereas Fig. 3 shows the average performance over all possible combinations by inserting one



(a) Camera 1 as Target (b) Camera 2 as Target (c) Camera 3 as Target

Fig. 2. CMC curves for WARD dataset with 3 cameras. Plots (a, b, c) show the performance of different methods while introducing camera 1, 2 and 3 respectively to a dynamic network. Please see the text in Section 4.2 for the analysis of the results.



(a) RAiD (b) SAIVT-SoftBio (c) Market-1501

Fig. 3. CMC curves averaged over all target camera combinations, introduced one at a time. (a) Results on RAiD dataset with 4 cameras (b) Results on SAVIT-SoftBio dataset with 8 cameras, and (c) Results on Market-1501 dataset with 6 cameras.

(a) WARD (b) SAIVT-SoftBio

Fig. 4. Effectiveness of our transitive algorithm in person re-identification on (a) WARD and (b) SAIVT-SoftBio datasets. Top row: Our matching result using the transitive algorithm. Middle row: matching the same person using `Best-GFK`. Bottom row: matching the same person using `Direct-GFK`. Visual comparison of top 10 matches shows that `Ours` perform best in matching persons across camera pairs by exploiting information from the best source camera. More qualitative results are included in the supplementary material. Best viewed in color.

camera on RAiD, SAIVT-SoftBio and Market-1501 datasets respectively. The following observations can be made from the figures: (i) the proposed framework for re-identification consistently outperforms all compared unsupervised methods on all datasets by a considerable margin, including the Market-1501 dataset with significantly large number of images and person identities. (ii) among the alternatives, CPS is the most competitive. However, the gap is still significant due to the two introduced components working in concert: discovering the best source camera and exploiting its information for re-identification. The rank-1 performance improvements over CPS are 23.44%, 24.50%, 9.98% and 2.85% on WARD, RAiD, SAIVT-SoftBio and Market-1501 datasets respectively. (iii) `Best-GFK` works better than `Direct-GFK` in most cases, suggesting that kernel computed across the best source camera and target camera can be applied to find the matching accuracy across other camera pairs. (iv) Finally, the performance gap between our method and `Best-GFK` (maximum improvement of 17% in nAUC on RAiD) shows the effectiveness of our transitive algorithm in exploiting information from the best source camera while computing re-identification accuracies across different source-target camera pairs (see Fig. 4 for some qualitative examples).

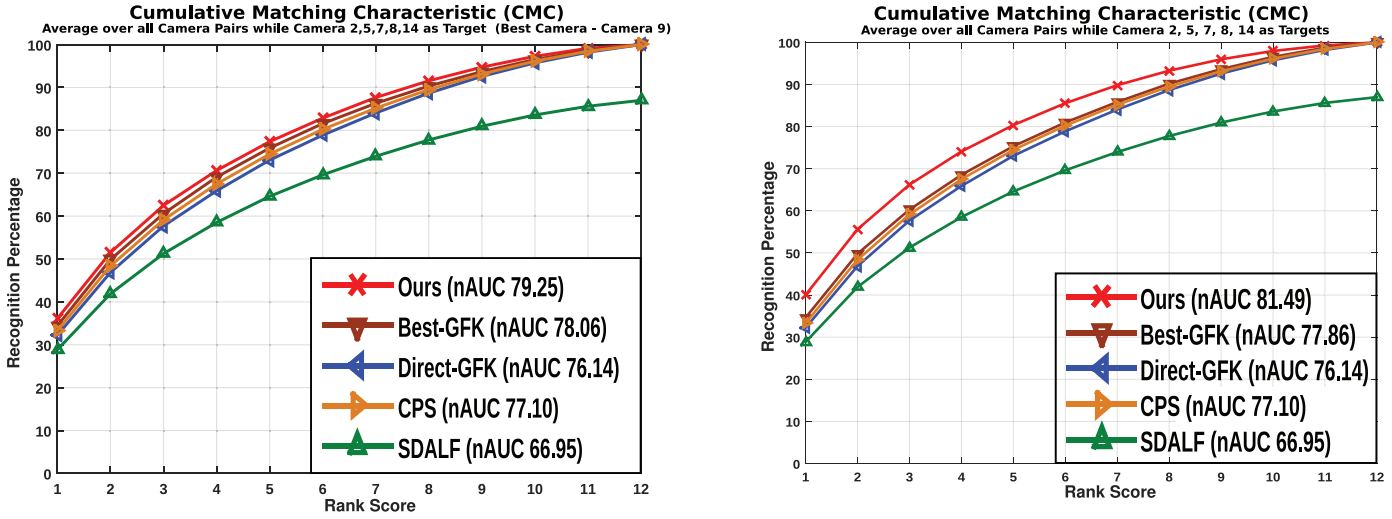
We also compare our approach with a CNN-based deep learning method (ResNet-50 [60] classifier) on SAIVT-SoftBio dataset. We train the network in identification setting and fine-tune from the ImageNet pre-trained model using only source camera images (without any labeled images from the target camera). Once the model is finetuned, we evaluate re-identification using the learned feature representations. Our approach performs significantly better than the ResNet-50 baseline (Rank-1: 24.92% vs 21.67%) which once again suggests that our approach is more effective by exploiting information from best source camera via a transitive inference. We believe the low performance of ResNet-50 baseline is due to lack of enough labeled data as well as lack of learning feature transferability across source and target cameras.

4.3. Introducing multiple cameras

Goal. The aim of this experiment is to validate the effectiveness of our approach while introducing multiple cameras at the same time into an existing network. We investigate two different scenarios such as (a) one common best source camera for all target cameras and (b) multiple best source cameras, one for each target camera in a dynamic network.

Implementation Details. We conduct this experiment on Shinpuhkan2014 dataset [55] with of 16 cameras. We randomly chose 2, 3 and 5 cameras as the target cameras and treat the remaining cameras as the source cameras. For scenario (a), we pick the common best source camera based on the average distance and for scenario (b), we use multiple best source cameras, one for each target camera in the transitive inference.

Results. Fig. 5 show results of different methods in two different scenarios while randomly introducing 5 cameras on Shinpuhkan2014 dataset. Following observations can be made: (i) similar to the results in Section 4.2, our approach outperforms all compared methods in both scenarios. This indicates that the proposed method is very effective and can be applied to large-scale dynamic



(a) One common best source camera (b) Multiple best source cameras

Fig. 5. CMC curves for Shinpuhkan2014 dataset while introducing 5 cameras at the same time (Camera 2, 5, 7, 8, 14 as Targets). (a) Performance of different methods with **one common best source camera** for all the target cameras and (b) Performance with **multiple best source cameras**, one for each target camera while computing re-id performance across a network. Please see supplementary material for the results on 2 and 3 target cameras.

Table 1
Model adaptation with prototype selection. Numbers show rank-1 recognition scores in % averaged over all possible combinations of target cameras, introduced one at a time.

Methods	WARD	RAiD
SDALF	16.66	26.80
CPS	45.70	35.35
Direct-GFK	16.87	17.63
Best-GFK	32.72	24.74
Ours-Proto-10%	*54.88*	*45.61*
Ours-Proto-20%	*60.72*	*53.67*
Ours-Proto-30%	*68.65*	*58.92*
Ours	**68.99**	**59.84**

camera networks where multiple cameras can be introduced at the same time. (ii) The proposed adaptation approach works better with multiple best source cameras compared to a common best source camera used for transitive inference (about 5% improvement – see Fig. 5(b)). This is expected since multiple best source cameras can better exploit information from different best source cameras. Results with the integration of 2 and 3 cameras at the same time are included in the supplementary.

4.4. Learning kernels with prototype selection

Goal. The main objective of this experiment is to analyze the performance of our target-aware sparse prototype selection strategy by using the selected prototypes from source camera while learning the geodesic flow kernels (Section 3.4).

Implementation Details. The regularization parameters λ and β in (9) are taken as λ_0/γ where $\gamma = 50$ and λ_0 is analytically computed from the data [49]. α is empirically set to 0.5 and kept fixed for all results. We compare our approach with four variants of our method where 10%, 20%, and 30% of source camera images are selected as prototypes for estimating the pair-wise kernels.

Results. Table 1 shows the results on both WARD and RAiD datasets. We have the following observations: (i) our approach (Ours-Proto-30%) achieves the similar performance (difference of only less than 1%) as the full set with only 30% of source camera prototypes. This can greatly reduce the deployment cost of new cameras in many large-scale camera networks involving significantly large number of images. (ii) our approach with only 10% of selected prototypes (Ours-Proto-10%) significantly outperforms all compared methods that use all existing source data on both datasets. The rank-1 performance improvements over CPS are 9.18% and 10.26% on WARD and RAiD datasets respectively.

We also investigate the effectiveness of our target-aware sparse prototype selection strategy by comparing with randomly selecting 20% of prototypes, and found that the later produces inferior results with rank-1 accuracy of 27.54% and 19.82% on WARD and RAiD datasets respectively. We believe this is because our prototype selection strategy efficiently exploits the information of target camera (see Eq. (9)) to select an informative subset of source camera data which share similar characteristics as target camera.

4.5. Comparison with supervised re-identification

Goal. The main objective of this experiment is to compare the performance of our approach with supervised alternatives while on-boarding new cameras.

Implementation Details. Given a newly introduced camera, we use the metric learning based methods to relearn the pair-wise distance metrics using the same train/test split, as mentioned in Section 4.1. We show the average performance over all possible combinations by introducing one camera at a time.

Results. We have the following key findings from Table 2: (i) both variants of our unsupervised approach (Ours-K and Ours-L) ouperforms all the feature transformation based approaches on both datasets by a big margin. (ii) on WARD dataset with 3 cameras, our approach is very competitive on both settings: Ours-K outperforms KISSME and LDML whereas Ours-L overcomes MLAPG. This result suggests that our approach is more effective in matching persons across a newly introduced camera and existing source cameras by exploiting information from best source camera via a transitive inference. (iii) on the RAiD dataset with 4 cameras, the performance gap between our method and metric-learning based methods begins to appear. This is expected as with a large network involving a higher number of camera pairs, an unsupervised approach can not compete with a supervised one, especially, when the latter one is using an intensive training phase.

Table 2
Comparison with supervised methods. Numbers show rank-1 recognition scores in % averaged over all possible combinations of target cameras, introduced one at a time.

Methods	WARD	RAiD	Reference
FT	49.33	39.81	TPAMI2015 [11]
ICT	42.51	25.31	ECCV2012 [58]
WACN	37.53	17.71	CVPRW2012 [52]
KISSME	66.95	55.68	CVPR2012 [46]
LDML	58.66	61.52	ICCV2009 [59]
XQDA	**77.20**	**77.81**	TPAMI2015 [57]
MLAPG	72.26	77.68	ICCV2015 [15]
Ours-K	*68.99*	*59.84*	Proposed
Ours-L	*73.77*	*61.87*	Proposed

However, we would like to point out once more that in practice collecting labeled samples from a newly inserted camera is very difficult and unrealistic in actual scenarios.

4.6. Extension to semi-supervised adaptation

Goal. The objective of this experiment is to analyze the performance of our proposed approach by incorporating the labeled data from the target camera.

Implementation Details. We compare the proposed unsupervised approach with four variants of our method where 10%, 25%, 50% and 100% of the labeled data from target camera are used for estimating kernel matrix respectively. We follow same experimental strategy except that we use PLS instead of PCA to compute the discriminative subspaces in target camerain.

Results. We have the following key findings from Fig. 6: (i) As expected, the semi-supervised baseline `Ours-Semi-100%`, works best since it uses all the labeled data from target domain to compute the kernel matrix for finding the best source camera. (ii) Our method remains competitive to `Ours-Semi-100%` on both datasets (Rank-1 accuracy: 60.04% vs 59.84% on RAiD and 26.41% vs 24.92% on SAIVT-SoftBio dataset). However, note that collecting labeled samples from the target camera is very difficult in practice. (iii) Interestingly, the performance gap between our unsupervised method and other three semi-supervised baselines (`Ours-Semi-50%`, `Ours-Semi-25%`, and `Ours-Semi-10%`) are moderate on RAiD (Fig. 6-a), but on SAIVT-SoftBio, the gap is significant (Fig. 6-b). We believe this is probably due to the lack of enough labeled data in the target camera to give a reliable estimate of PLS subspaces.

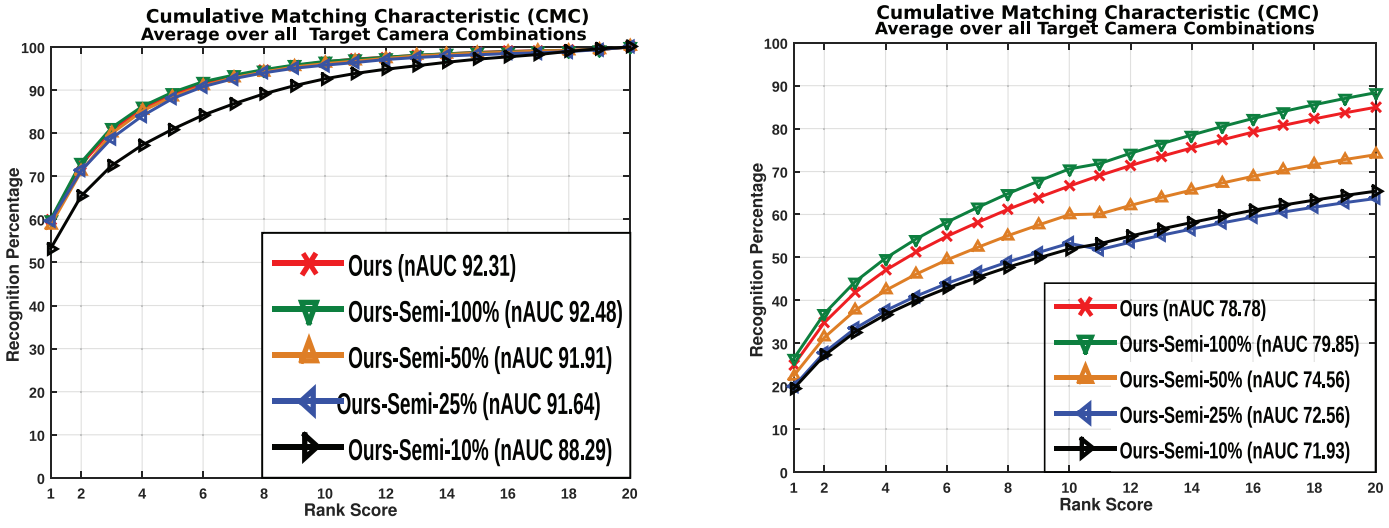
4.7. Analysis with different sets of people in the new camera

Goal. The goal of this experiment is to analyze the performance of our approach with different identities of people appearing in the target camera as in a real world setting. Note that the train and test set are still kept disjoint as in standard re-id settings.

Implementation Details. We consider two scenarios as follows. *Scenario 1 with 0% overlap*: first 15 persons in source camera and next 20 persons in target camera for training on WARD dataset while we use first 13 persons in source camera and next 10 persons in target camera for training on RAiD dataset. *Scenario 2 with 50% overlap*: partial overlap of persons exists across source and target cameras, i.e., all the persons appearing in the source camera are present in the target camera but there exists some persons that only appear in target camera and not in source cameras. We consider first 13 persons in source camera and all 23 persons in target camera for training in this setting.

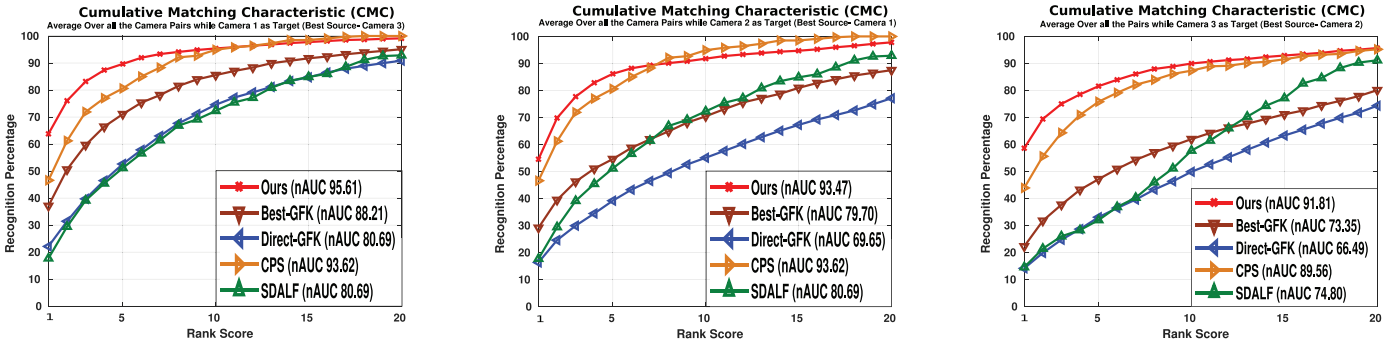
Results. Fig. 7 shows the re-id performance on WARD dataset with completely disjoint sets of people in the target camera. Following are the key observations from Fig. 7: (i) the proposed framework consistently outperforms all compared methods by a significant margin even though completely new persons appear in the target camera. (ii) similar to previous results with 100% overlap of persons across source and target cameras (see Fig. 2), `CPS` is still the most competitive. However, our approach outperforms `CPS` by a margin about 20% in rank-1 accuracy on WARD dataset. (iii) finally, the large performance gap between our method, `Direct-GFK` and `Best-GFK` ($\sim$30% in rank-1 accuracy) once again shows the effectiveness of our transitive algorithm in real-world scenarios where completely new person identities appear in the newly introduced camera.

Table 3 shows the performance of our approach with different percentage of overlap in person identities across source and target camera on RAiD dataset. As expected, the performance increases with increase in the percentage of overlap and achieves the maximum rank-1 accuracy of 59.84% when the same set of people



(a) RAiD (b) SAIVT-SoftBio

Fig. 6. Semi-supervised adaptation with labeled data. Plots (a,b) show CMC curves averaged over all target camera combinations on RAiD and SAIVT-SoftBio respectively.



(a) Camera 1 as Target (b) Camera 2 as Target (c) Camera 3 as Target

Fig. 7. Re-identification performance on WARD dataset with different sets of people in the target camera *(Scenario 1: 0% Overlap)*. Plots (a, b, c) show the performance of different methods while introducing camera 1, 2 and 3 respectively to a network.

Table 3
Performance comparison with different % of overlap in person identities across source and target camera. Numbers show rank-1 recognition scores in % averaged over all possible combinations of target cameras, introduced one at a time.

Datasets	0% Overlap	50% Overlap	100% Overlap
RAiD	50.83	56.81	59.84

appear in all camera views as in standard person re-identification setting. This is because kernel matrices are the best measure of similarity when there is complete overlap across two data distributions. Our approach outperforms all compared methods at 0% overlap on both WARD and RAiD datasets showing it's effectiveness in fully open world re-identification systems with both dynamic network and completely different sets of persons appearing in the newly introduced camera(s).

4.8. Additional results in the supplementary material

We include the following experiments and results in our supplementary material. (a) We perform experiment to verify the effectiveness of our approach by replacing KISSME [46] with LDML metric learning [59] as the initial set up and observe that our approach outperforms all compared methods in both WARD and RAiD datasets suggesting that the proposed adaptation technique works significantly well irrespective of the metric learning method used in the existing network. (b) We verify the effectiveness of our approach by changing the feature representation from LOMO feature with Weighted Histograms of Overlapping Stripes (WHOS) feature representation [57]. Our approach outperforms all compared methods which suggests that the proposed adaptation technique works significantly well irrespective of the feature used to represent persons in a camera network. Moreover, the significant improvement over `Best-GFK` (∼10%) shows that the proposed transitive algorithm is very effective in exploiting information from the best source camera irrespective of the feature representation. (c) We also analyze the performance of our method by changing the dimension of subspace used to compute the geodesic flow kernels and observe that dimensionality of the subspace has a little effect on the performance suggesting that our method is robust to the change in dimensionality of the subspace used to compute the geodesic kernels across target and source cameras.

Moreover, due to space constraint, we only report average CMC curves for most experiments in our main paper and leave the full CMC curves including more qualitative matching results in the supplementary material.

5. Conclusions and future works

In this paper, we presented an efficient yet scalable framework to adapt person re-identification models in a dynamic network, where one or multiple new cameras may be temporarily inserted into an existing system to get additional information. We developed an unsupervised approach based on geodesic flow kernel to find the best source camera to pair with newly introduced camera(s), without requiring a very expensive training phase. We then introduced a simple yet effective transitive inference algorithm that can exploit information from best source camera to improve the accuracy across other camera pairs. Moreover, we develop a source-target selective adaptation strategy that uses a subset of source data instead of all existing data to compute the kernels in resource constrained environments. Extensive experiments on several benchmark datasets well demonstrate the efficacy of our method over state-of-the-art methods.

In our current work, we explained how it is possible to onboard new camera(s) to an existing network with no additional supervision for the new cameras. However, transfer learning across networks is still a largely under-addressed problem with many challenges. Given multiple existing source networks and a newly installed target network with limited labeled data, we first need to find the relevance/similarity of each source network, or parts thereof, in terms of amount of knowledge that it can transfer to a target network. Developing efficient statistical measures for finding relevance in a multi-camera network with significant changes in viewing angle, lighting, and occlusion can be a very interesting future work. Furthermore, labeled data from source networks are often a subject of legal, technical and contractual constraints between data owners and customers. Thus, existing transfer learning approaches may not be directly applicable in such scenarios where the source data is absent. However, compared to the source data, the well-trained source model(s) are usually freely accessible in many applications and contain equivalent source knowledge as well. Leveraging person re-identification models in absence of source data via knowledge distillation [61], can be another interesting direction for future research.

Acknowledgment

This work was partially supported by NSF grant 1544969 and ONR grant N00014-19-1-2264.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2019.106991.

References

[1] L. Zheng, Y. Yang, A. G. Hauptmann, Person re-identification: past, present and future, arXiv:1610.02984 (2016).
[2] Z. Wu, Y. Li, R.J. Radke, Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features, IEEE Trans. Pattern Anal. Mach. Intell. 37 (5) (2014) 1095–1108.
[3] L. Bazzani, M. Cristani, V. Murino, Symmetry-driven accumulation of local features for human characterization and re-identification, Comput. Vis. Image Understanding 117 (2) (2013) 130–144.
[4] S. Paisitkriangkrai, C. Shen, A. Van Den Hengel, Learning to rank in person re-identification with metric ensembles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1846–1855.
[5] H.-X. Yu, A. Wu, W.-S. Zheng, Cross-view asymmetric metric learning for unsupervised person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 994–1002.
[6] V.M. Patel, R. Gopalan, R. Li, R. Chellappa, Visual domain adaptation: a survey of recent advances, IEEE Signal Process. Mag. 32 (3) (2015) 53–69.
[7] K. Saenko, B. Kulis, M. Fritz, T. Darrell, Adapting visual category models to new domains, in: European Conference on Computer Vision, 2010, pp. 213–226.
[8] R. Gopalan, R. Li, R. Chellappa, Domain adaptation for object recognition: an unsupervised approach, in: 2011 International Conference on Computer Vision, 2011, pp. 999–1006.
[9] Z. Ma, Y. Yang, F. Nie, N. Sebe, S. Yan, A.G. Hauptmann, Harnessing lab knowledge for real-world action recognition, Int. J. Comput. Vis. 109 (1–2) (2014) 60–73.
[10] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2066–2073.
[11] N. Martinel, A. Das, C. Micheloni, A.K. Roy-Chowdhury, Re-identification in the function space of feature warps, IEEE Trans. Pattern Anal. Mach. Intell. 37 (8) (2014) 1656–1669.
[12] D. Li, Y. Gong, D. Cheng, W. Shi, X. Tao, X. Chang, Consistency-preserving deep hashing for fast person re-identification, Pattern Recognit. 94 (2019) 207–217.
[13] M. Koestinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.
[14] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2197–2206.
[15] S. Liao, S.Z. Li, Efficient PSD constrained asymmetric metric learning for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3685–3693.
[16] S. Karanam, Y. Li, R.J. Radke, Person re-identification with discriminatively trained viewpoint invariant dictionaries, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4516–4524.
[17] M. Cao, C. Chen, X. Hu, S. Peng, Towards fast and kernelized orthogonal discriminant analysis on person re-identification, Pattern Recognit. 94 (2019) 218–229.
[18] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, in: 2014 22nd International Conference on Pattern Recognition, 2014, pp. 34–39.
[19] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1335–1344.
[20] S. Zhou, J. Wang, D. Meng, X. Xin, Y. Li, Y. Gong, N. Zheng, Deep self-paced learning for person re-identification, Pattern Recognit. 76 (2018) 739–751.
[21] Z. Zhou, Y. Huang, W. Wang, L. Wang, T. Tan, See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4747–4756.
[22] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Y. Yang, Improving person re-identification by attribute and identity learning, Pattern Recognit. 95 (2019) 151–161.
[23] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, W. Gao, Attention driven person re-identification, Pattern Recognit. 86 (2019) 143–155.
[24] J. Meng, A. Wu, W.-S. Zheng, Deep asymmetric video-based person re-identification, Pattern Recognit. 93 (2019) 430–441.
[25] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.
[26] W. Li, R. Zhao, T. Xiao, X. Wang, DeepReID: deep filter pairing neural network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.
[27] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, C. Zhang, Alignedreid++: dynamically matching local information for person re-identification, Pattern Recognit. 94 (2019) 53–61.
[28] A. Das, R. Panda, A. Roy-Chowdhury, Active image pair selection for continuous person re-identification, in: 2015 IEEE International Conference on Image Processing, 2015, pp. 4263–4267.
[29] N. Martinel, A. Das, C. Micheloni, A.K. Roy-Chowdhury, Temporal model adaptation for person re-identification, in: European Conference on Computer Vision, 2016, pp. 858–877.
[30] H. Wang, S. Gong, X. Zhu, T. Xiang, Human-in-the-loop person re-identification, in: European Conference on Computer Vision, 2016, pp. 405–422.
[31] X. Xin, J. Wang, R. Xie, S. Zhou, W. Huang, N. Zheng, Semi-supervised person re-identification using multi-view clustering, Pattern Recognit. 88 (2019) 285–297.
[32] C. Liu, S. Gong, C.C. Loy, On-the-fly feature importance mining for person re-identification, Pattern Recognit. 47 (4) (2014) 1602–1615.
[33] D.S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification., in: Bmvc, vol. 1, 2011, p. 6.
[34] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3586–3593.
[35] E. Kodirov, T. Xiang, Z. Fu, S. Gong, Person re-identification by unsupervised l1 graph learning, in: European Conference on Computer Vision, 2016, pp. 178–195.
[36] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, J. Bu, Semi-supervised coupled dictionary learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3550–3557.
[37] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro (2017) 3754–3762.
[38] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer GAN to bridge domain gap for person re-identification (2018) 79–88.
[39] A. Bendale, T. Boult, Towards open world recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1893–1902.
[40] W.-S. Zheng, S. Gong, T. Xiang, Towards open-world person re-identification by one-shot group-based verification, IEEE Trans. Pattern Anal. Mach. Intell. 38 (3) (2015) 591–606.
[41] X. Zhu, B. Wu, D. Huang, W.-S. Zheng, Fast open-world person re-identification, IEEE Trans. Image Process. 27 (5) (2017) 2286–2300.
[42] R. Layne, T.M. Hospedales, S. Gong, Domain transfer for person re-identification, in: Proceedings of the 4th ACM/IEEE International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream, 2013, pp. 25–32.
[43] X. Wang, W.-S. Zheng, X. Li, J. Zhang, Cross-scenario transfer person reidentification, IEEE Trans. Circuits Syst. Video Technol. 26 (8) (2015) 1447–1460.
[44] A.J. Ma, J. Li, P.C. Yuen, P. Li, Cross-domain person reidentification using domain adaptation ranking SVMs, IEEE Trans. Image Process. 24 (5) (2015) 1599–1613.
[45] R. Panda, A. Bhuiyan, V. Murino, A.K. Roy-Chowdhury, Unsupervised adaptive re-identification in open world dynamic camera networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7054–7063.
[46] M. Koestinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.
[47] J.M. Phillips, S. Venkatasubramanian, A gentle introduction to the kernel distance, arXiv:1103.1625 (2011).
[48] G. Kou, D. Ergu, J. Shang, Enhancing data consistency in decision matrix: adapting hadamard model to mitigate judgment contradiction, Eur. J. Oper. Res. 236 (1) (2014) 261–271.
[49] E. Elhamifar, G. Sapiro, R. Vidal, See all by looking at a few: Sparse modeling for finding representative objects, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1600–1607.
[50] R. He, T. Tan, L. Wang, W.-S. Zheng, L21 regularized correntropy for robust feature selection, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2504–2511.
[51] W.R. Schwartz, A. Kembhavi, D. Harwood, L.S. Davis, Human detection using partial least squares analysis, in: 2009 IEEE 12th International Conference on Computer Vision (ICCV), 2009, pp. 24–31.
[52] N. Martinel, C. Micheloni, Re-identify people in wide area camera network, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 31–36.
[53] A. Das, A. Chakraborty, A.K. Roy-Chowdhury, Consistent re-identification in a camera network, in: European Conference on Computer Vision, 2014, pp. 330–345.
[54] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, P. Lucey, A database for person re-identification in multi-camera surveillance networks, in: 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA), 2012, pp. 1–8.
[55] Y. Kawanishi, Y. Wu, M. Mukunoki, M. Minoh, Shinpuhkan2014: a multi-camera pedestrian dataset for tracking people across multiple cameras, 20th Korea-Japan Joint Workshop on Frontiers of Computer Vision, vol. 5, 2014.
[56] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.
[57] G. Lisanti, I. Masi, A.D. Bagdanov, A. Del Bimbo, Person re-identification by iterative re-weighted sparse ranking, IEEE Trans. Pattern Anal. Mach. Intell. 37 (8) (2014) 1629–1642.
[58] T. Avraham, I. Gurvich, M. Lindenbaum, S. Markovitch, Learning implicit transfer for person re-identification, in: European Conference on Computer Vision, 2012, pp. 381–390.
[59] M. Guillaumin, J. Verbeek, C. Schmid, Is that you? Metric learning approaches for face identification, in: 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 498–505.
[60] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[61] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv:1503.02531 (2015).

Rameswar Panda is currently a Research Scientist at IBM Research AI, MIT-IBM Watson AI Lab, Cambridge, USA. Prior to joining IBM, he obtained his Ph.D in Electrical and Computer Engineering from University of California, Riverside in 2018. His primary research interests span the areas of computer vision, machine learning and multimedia. In particular, his current focus is on developing semi, weakly, unsupervised algorithms for solving different vision problems. His work has been published in top-tier conferences such as CVPR, ICCV, ECCV, MM as well as high impact journals such as TIP and TMM.

Amran Bhuiyan received the Bachelor degree in applied physics, electronic & communication engineering from University of Dhaka, Bangladesh in 2009, the M.Sc. degree in Computer Engineering and Information Technology from the Lucian Blaga University of Sibiu,Romania under the Erasmus Mundus external window in 2011 and the Ph.D. degree in Pattern Analysis and Computer Vision from the Istituto Italiano di Tecnologia, Genova, Italy. He is currently a Postdoctoral Researcher with LIVIA, ole de Technologie Suprieure, Universit du Qubec, Montral, Canada. His main research interests include computer vision, machine learning, person re-identification and video surveillance.

Vittorio Murino received the Laurea degree in electronic engineering and the Ph.D. degree in electronic engineering and computer science from the University of Genova, Genoa, Italy, in 1989 and 1993, respectively. He is a Full Professor with the University of Verona, Verona, Italy, and the Director of Pattern Analysis and Computer Vision (PAVIS) Department, Istituto Italiano di Tecnologia, Genoa, Italy. From 1995 to 1998, he was an Assistant Professor with the Department of Mathematics and Computer Science, University of Udine, Udine, Italy. Since 1998, he has been with the University of Verona. He was the Chairman of the Department of Computer Science from 2001 to 2007, where he was the Coordinator of the Ph.D. program in computer science from 1999 to 2003. He is responsible for several national and European projects, and an Evaluator of EU project proposals related to several frameworks and programs. He is currently with the Istituto Italiano di Tecnologia, leading PAVIS Department involved in computer vision, machine learning, and image analysis activities. He has co-authored more than 400 papers published in refereed journals and international conferences. His current research interests include computer vision, pattern recognition, and machine learning, more specifically, statistical and probabilistic techniques for image and video processing, with applications on (human) behavior analysis and related applications such as video surveillance, biomedical imaging, and bioinformatics. He has been an IAPR Fellow since 2006. He is a member of the technical committees of most significant computer vision and pattern recognition conferences and a Guest Co-Editor of special issues in relevant scientific journals. He is also an Editorial Board Member of Computer Vision and Image Understanding, Machine Vision and Applications and Pattern Analysis and Applications.

Amit K. Roy-Chowdhury received the Bachelors degree in electrical engineering from Jadavpur University, Calcutta, India, the Masters degree in systems science and automation from the Indian Institute of Science, Bangalore, India, and the Ph.D. degree in electrical engineering from the University of Maryland, College Park. He is a Professor of Electrical Engineering at University of California, Riverside. His research interests include image processing and analysis, computer vision, and video communications and statistical methods for signal analysis. His current research projects include intelligent camera networks, wide-area scene analysis, motion analysis in video, activity recognition and search, video-based biometrics (face and gait), biological video analysis, and distributed video compression. He is coauthor of “The Acquisition and Analysis of Videos over Wide Areas” He is the editor of the book “Distributed Video Sensor Networks”. He has been on the organizing and program committees of multiple conferences and serves on the editorial boards of a number of journal. He is a Fellow of the IEEE and IAPR.