Learning Affine Hull Representations for Multi-Shot Person Re-Identification

Srikrishna Karanam, *Student Member, IEEE,* Ziyan Wu, *Member, IEEE,* Richard J. Radke, *Senior Member, IEEE*

***Abstract*—We consider the person re-identification problem, assuming the availability of a sequence of images for each person, commonly referred to as video-based or multi-shot re-identification. We approach this problem from the perspective of learning discriminative distance metric functions. While existing distance metric learning methods typically employ the average feature vector as the data exemplar, this discards the inherent structure of the data. To overcome this issue, we describe the image sequence data using affine hulls. We show that directly computing the distance between the closest points on these affine hulls as in existing recognition algorithms is not sufficiently discriminative in the context of person re-identification. To this end, we incorporate affine hull data modeling into the traditional distance metric learning framework, learning discriminative feature representations directly using affine hulls. We perform extensive experiments on several publicly available datasets to show that the proposed approach improves the performance of existing metric learning algorithms irrespective of the feature space employed to perform metric learning. Furthermore, we advance the state of the art on iLIDS-VID, PRID, and SAIVT, with absolute rank-1 performance improvements of 6.0%, 11.4%, and 6.0% respectively.**

***Index Terms*—Re-identification, camera network, video analytics.**

I. Introduction

RECOGNIZING the same person as s/he moves through a network of cameras with non-overlapping views, called re-identification or re-id, is a fundamental problem in video analytics, with crucial applications in security and surveillance. Consequently, the re-id problem has drawn increasing attention from the computer vision community. Much related research has focused on the single-shot version of the problem [1]–[9], wherein it is assumed that only one image per person per camera view is available. However, this is not the case in practical and real-world applications of re-id. For instance, once a person of interest is identified in a "probe" camera view, it is natural for that person to be tracked in a surveillance application to obtain an image sequence. Furthermore, all candidates observed in the target, or "gallery", camera view are typically also tracked. In such a "tag and track" application of re-id [10], [11], instead of a single image, we have a sequence or track of images for each person. Therefore, re-id in practice is a problem of matching image sequences rather than solitary images.

This problem formulation raises two critical questions: (1) how do we describe the available "multiple instance" data for each person? and (2) how do we exploit this data description to learn feature representations that are sufficiently discriminative to perform accurate re-id? Traditional re-id approaches are typically based on learning feature representations so that probe and gallery feature points corresponding to the same person are close in the learned feature space whereas those corresponding to different people are relatively far [1], [6], [12], [13]. A naïve application of this approach to our problem would be to either take the average of the available feature points as an exemplar or consider all the available feature points individually, in which case we would have to deal with millions of constraints and implementation infeasibility. Furthermore, and more crucially, such approaches fail to capture the inherent structure of the multiple instance data of each person.

Representing multiple instance data in the context of recognition has been a longstanding problem in machine learning and is typically studied as a multiple instance learning (MIL) problem [14]. While traditional MIL represents data as bags of feature points and recognizes a bag as positive if it contains at least one positive instance, we need a different interpretation in the context of re-id. In re-id, we have multiple feature points corresponding to a single person, all of which are positive instances. Subsequently, the representation of this data as an "image set" is more appropriate. Developing recognition algorithms based on image sets has been an active research area, with several approaches based on constructing affine or convex hulls of the data and considering the distance between the closest points on these hulls [15]–[17].

In this paper, we show that such a distance computation between affine hulls is not discriminative enough for re-id. Consequently, a natural question to address is: what are good representations of these affine hulls that make subsequent reasoning more accurate? Traditionally, distance metric learing approaches are adopted to learn discriminative representations of feature vectors. Can we learn such discriminative metric learning-based feature representations in the context of affine hulls? Traditional distance metric learning is typically based on solitary feature points and extending it to the case involving affine hulls of data is a non-trivial problem. In this regard, how do we directly learn distance metrics using affine hulls? It is natural to expect computational difficulties when dealing with affine hull representations of large amounts of data. So, is there

S. Karanam and Z. Wu are with Siemens Corporation, Corporate Technology, Princeton, NJ 08540 USA (e-mail: {srikrishna.karanam,ziyan.wu}@siemens.com).

R.J. Radke is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: rjradke@ecse.rpi.edu).

Corresponding author: S. Karanam.

a computationally efficient way of learning such discriminative feature representations?

In this paper, we propose an approach that addresses all questions raised above in a principled and intuitive manner. To be specific, our contributions are discussed below.

- **Data description:** We tackle the problem of describing the multiple instance data inherent in multi-shot re-id by constructing affine hulls. Such a mathematical representation provides for an intuitive description of the available data.
- **Learning discriminative affine hull representations:** We show that the traditionally used approach of computing the distance between affine hulls in algorithms based on image sets is not sufficiently discriminative in the context of re-id. We overcome this problem by learning discriminative feature representations learned directly from these affine hulls.
- **Improving metric learning:** We empirically demonstrate the efficacy of discriminative affine hull representations in the context of several existing metric learning algorithms. Specifically, we demonstrate the improvements achieved when compared to distance metrics learned using the traditional approach of taking the average feature point as the data exemplar.
- **Extensive empirical validation:** We perform extensive experiments on several state-of-the-art metric learning algorithms in the context of a wide variety of features using publicly available multi-shot re-id datasets to validate the impact of the proposed approach. We improve the state of the art, measured in terms of mean rank-1 performance, by 6.0%, 11.4%, and 6.0% on iLIDS-VID, PRID, and SAIVT datasets respectively.

II. Related work

Person Re-Identification. Much prior work in re-id can be categorized into two key themes: appearance modeling and metric learning. Since the early work of Gray and Tao [18] that divided an image into horizontal strips and extracted color and texture histograms, several appearance modeling schemes have been proposed. Bazzani *et al.* [19] designed a scheme in which local features describing the chromatic content, the spatial arrangement of color, and the recurrence of textured patterns were accumulated into a single descriptor. Ma *et al.* [20] used spatial, intensity, and gradient information at each pixel and encoded these local descriptors into Fisher vectors. Zhao *et al.* [5] proposed so-called dense features, wherein the image was regularly divided into numerous local patches and texture histograms and SIFT features [21] from each patch were concatenated into a single descriptor. Some recent methods include LOMO [22], where color and LBP features are constructed in conjunction with the retinex transform, and GOG [23], where local image patches are modeled in a hierarchical fashion using Gaussian distributions. End-to-end feature learning methods such as finetuning existing convolutional neural network models have also been explored [24], [25]. These learned features are typically used in conjunction with metric learning methods to perform re-id. A structured survey of these and several other related appearance modeling methods can be found in [26].

The goal of distance metric learning is to learn a new feature space in which feature vectors of the same person stay close whereas the feature vectors of different people are far apart. This is typically mathematically represented in terms of pairwise constraints on the available feature vectors. Prosser *et al.* [1] used the pairwise constraints to formulate a ranking problem in the support vector machine (SVM) framework. Mignon and Jurie [12] used the pairwise similar and dissimilar constraints in a logistic loss minimization problem to learn the feature transformation matrix. Xiong *et al.* [6] recently proposed kernel versions of some of these popularly used metric learning algorithms. Martinel *et al.* [27] identified salient image regions to construct robust appearance descriptors, following which a Mahalanobis distance metric was learned to capture inter-camera variations. Liao *et al.* [22] employed quadratic discriminant analysis to formulate an eigendecomposition problem similar in spirit to the traditional Fisher discriminant analysis [28] and learned a cross-view discriminative subspace. In similar spirit to learning distance metrics, Martinel *et al.* [29] proposed the concept of warp functions that essentially capture all possible non-linear feature transformations of person images from one camera view to the other. A discussion of related metric learning methods can be found in [30]–[33] and the references therein. A systematic experimental evaluation of feature extraction and metric learning algorithms can be found in the paper by Karanam *et al.* [34].

Methods directly tackling the multi-shot re-id problem have also been proposed. Even in this case, there has been prior work along the lines of appearance modeling and metric learning, although a more accurate way of applying metric learning in this case would be to learn a ranking function given the multiple instance data for each person. Wang *et al.* [35] formulated the multi-shot re-id problem as an image sequence matching problem, taking the associated temporal aspect into account to rank video fragments using a compact descriptor based on quantized spatial and temporal gradients, called HOG3D [36]. Liu *et al.* [37] also approached the problem in the same spirit, extending the popular 2-dimensional Fisher vector representation to incorporate spatial and temporal information to design a 3-dimensional Fisher vector representation, called STFV3D. On the other hand, a few algorithms formulate the problem in terms of learning a ranking function. Li *et al.* [38] learned multiple personally-discriminative random forests to classify the multiple instance data corresponding to each person. Lisanti *et al.* [39] hypothesized that the feature vector of a particular image of a person in one camera view can be expressed as a sparse linear combination of the feature vectors of all the available images of the same person in some other camera view, thereby formulating a sparse recovery problem to rank gallery candidates. Li *et al.* [40] learned a discriminative feature space by iteratively learning a Fisher transformation matrix and hierarchically clustering image sequences. Karanam *et al.* [41], [42] formulated the multi-shot re-id problem as one of recovering block sparse coefficient vectors, demonstrating a generic ranking framework that can be used to improve the performance of existing metric learning methods.

Image set description and recognition. Most prior work in image set recognition models image sets as affine or convex hulls of the available data and subsequently determines the distance between the closest points on these hulls to perform recognition. Cevikalp and Triggs [15] used this idea to develop both linear and non-linear hull distance algorithms for face recognition. Subsequently, several other hull distance algorithms have been proposed. Hu *et al.* [16] hypothesized that the dissimilarity between image sets can be sparsely approximated from their respective image samples and formulated the so-called sparse approximated nearest points (SANP). Yang *et al.* [17] formulated a regularized linear hull distance algorithm to generate regularized nearest points (RNP) on hulls of data. Wu *et al.* [43] incorporated collaborative representation into the framework of RNP and re-formulated the objective to come up with a compute-efficient hull distance algorithm. Algorithms that directly learn distance metrics using hull data description schemes have also been proposed. Zhu *et al.* [44] proposed point-to-set and set-to-set learning methods, formulating the distance metric learning problem as a convex optimization problem in an SVM-like framework.

III. Proposed approach

In this section, we describe the proposed approach to learn discriminative affine hull representations. We first begin with a brief introduction to the distance metric learning problem as well as describing data using affine hulls.

A. The distance metric learning problem

The goal of distance metric learning is to learn a new feature space where the feature vectors belonging to the same people are close whereas the feature vectors belonging to different people are far apart. Formally, let $\mathbf{p}_1$ and $\mathbf{g}_1$ be two feature vectors of the person with index 1 in the probe and gallery cameras respectively. Let $\mathbf{g}_2$ be the feature vector of some other person in the gallery camera. The goal of distance metric learning can be mathematically formulated as learning a distance function, $d(\mathbf{x}, \mathbf{y})$, that takes in two feature vectors $\mathbf{x}$ and $\mathbf{y}$ as inputs, and satisfies the following ranking relationship:

$$d(\mathbf{p}_1, \mathbf{g}_1) < d(\mathbf{p}_1, \mathbf{g}_2) \tag{1}$$

Using such "pairwise constraints", most distance metric learning methods learn a feature space transformation matrix $\mathbf{T}$ that is used to compare feature vectors. In the transformed feature space, the distance function typically takes the following form:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{T}^\top \mathbf{x} - \mathbf{T}^\top \mathbf{y}\|_2 \tag{2}$$

With this background, we formally introduce the overall mathematical formulation that is generally employed to learn distance metrics for the re-id problem. Let $\mathcal{P} = \{\mathbf{p}_1, \dots, \mathbf{p}_N\}$ be feature vectors corresponding to N people in the probe camera. Similarly, let $\mathcal{G} = \{\mathbf{g}_1, \dots, \mathbf{g}_N\}$ be the N feature vectors corresponding to the same N people in the gallery camera. Any distance metric learning method typically first constructs pairs of "similar" and "dissimilar" feature vectors. Let

$$\mathcal{S} = \{(\mathbf{p}_i, \mathbf{g}_i), i = 1, \dots, N\} \tag{3}$$

denote the set of all the N pairs of similar feature vectors and

$$\mathcal{N} = \{(\mathbf{p}_i, \mathbf{g}_j) | i, j = 1, \dots, N, i \neq j\} \tag{4}$$

denote the set of possible pairs of feature vectors belonging to different people. A very general distance metric learning problem in the context of re-id can be posed as the following mathematical optimization problem:

$$\min_{\mathbf{T}} l(\mathcal{S}, \mathcal{N}) + \lambda \mathcal{R}(\mathcal{S}, \mathcal{N}) \tag{5}$$

where $l(\mathcal{S}, \mathcal{N})$ is the loss function, parameterized by $\mathbf{T}$, that the metric learning algorithm seeks to minimize and $\mathcal{R}(\mathcal{S}, \mathcal{N})$ is an optional regularization function to prevent the learned distance metric from overfitting.

B. Issues with distance metric learning in the multi-shot setting

In the multi-shot setting of the re-id problem, we have a set of feature vectors for each person. A natural extension of metric learning to this case would be to construct pairwise constraints from all the available feature vectors, which would run into millions! Apart from the obvious computational difficulties, such an approach disregards the underlying structure of the set of feature vectors, potentially resulting in a transformed feature space that gives sub-optimal re-id performance.

C. Describing data using affine hulls

The issues discussed above can be addressed by describing data as affine hulls. Given the feature vectors $\mathcal{P} = \{\mathbf{p}_1, \dots, \mathbf{p}_n\}$, $\mathbf{p}_i \in \mathbb{R}^r$ corresponding to n images of a certain person, the affine hull [45] of this data is the smallest affine subspace containing the data. Formally, if $\mu = \frac{1}{n}\sum_{i=1}^{n} \mathbf{p}_i$ is the mean vector of the data points and $\mathbf{U} \in \mathbb{R}^{r \times t}$ is an orthonormal basis describing the data, the affine hull of $\mathcal{P}$ can be written as the set $\mathbf{H}(\mathcal{P}) = \{\mathbf{x} = \mathbf{U}\mathbf{v} + \mu \mid \mathbf{v} \in \mathbb{R}^t\}$. An illustration of this concept is provided in Figure 1.

Given a probe image set $\mathcal{P}$ and a gallery image set $\mathcal{G}$, image set based recognition algorithms typically first construct the affine hulls of these two sets. To compute the extent of similarity/dissimilarity between $\mathcal{P}$ and $\mathcal{G}$, the general workflow is to determine the two points, one on each of the two affine hulls, that are closest to each other. Subsequently, the distance between these two points is used to represent the distance between the two sets $\mathcal{P}$ and $\mathcal{G}$. An illustration of this concept is provided in Figure 2.

Formalizing this notion, if $\mathbf{s}$ and $\mathbf{t}$ represent the two nearest points on the affine hulls $\mathbf{H}(\mathcal{P}) = \{\mathbf{x}_p = \mathbf{U}_p\mathbf{v}_p + \mu_p \mid \mathbf{v}_p \in \mathbb{R}^t\}$ and $\mathbf{H}(\mathcal{G}) = \{\mathbf{x}_g = \mathbf{U}_g\mathbf{v}_g + \mu_g \mid \mathbf{v}_g \in \mathbb{R}^t\}$, we solve the following optimization problem to find them:

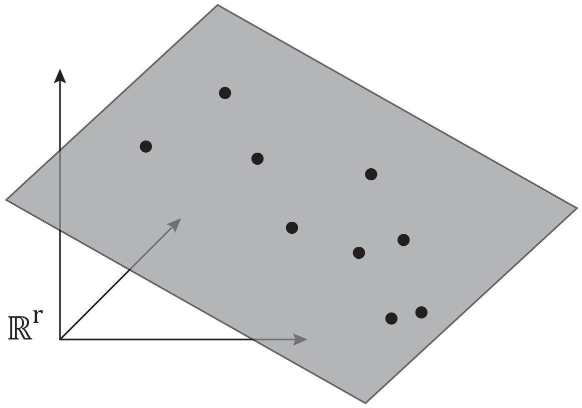


Fig. 1. The affine hull of data samples is a t-dimensional affine subspace in the r-dimensional space of feature vectors.

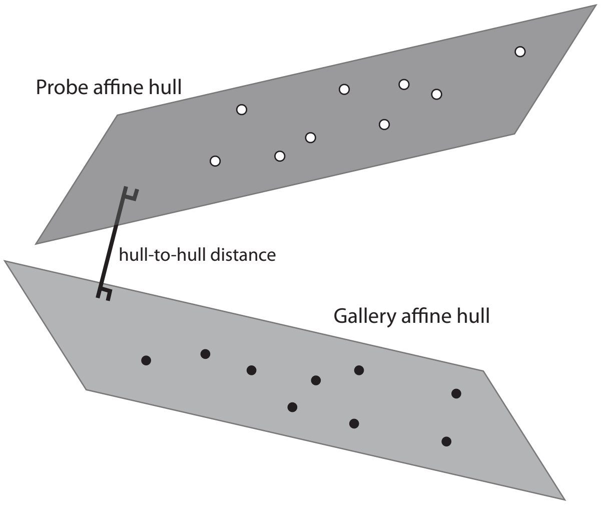


Fig. 2. The distance between two affine hulls is the length of the shortest line connecting one point on each subspace.

$$\min_{\mathbf{v}_p,\mathbf{v}_g} \quad \|\mathbf{x}_p - \mathbf{x}_g\|_2^2 \tag{6}$$

The closest points are then given by $\mathbf{s} = \mathbf{U}_p\mathbf{v}_p^* + \mu_p$ and $\mathbf{t} = \mathbf{U}_g\mathbf{v}_g^* + \mu_g$, where $\mathbf{v}_p^*$ and $\mathbf{v}_g^*$ are the optimal solutions to the minimization problem in Equation 6. The distance between $\mathcal{P}$ and $\mathcal{G}$ is then simply taken as $d = \|\mathbf{s} - \mathbf{t}\|$. Most hull distance algorithms differ in how they formulate the optimization problem of Equation 6. While AHISD [15] uses the same formulation as above, algorithms like SANP [16] and RNP [17] incorporate some kind of regularization into the problem formulation, typically based on the l_1 or l_2 norm to determine the closest points. Once the closest points are determined, computing the distance between the image sets reduces to the same Euclidean distance computation as above.

D. Learning Discriminative Affine Hull Representations: Improving Metric Learning Algorithms

Before describing the proposed approach, we lay out the notation used in the subsequent sections. We use $\mathcal{P}_i$ to denote the set of feature vectors corresponding to the images of the person with index i in the probe camera of the training set. Similarly, $\mathcal{G}_i$ denotes the set of feature vectors corresponding to the images of the same person in the gallery camera of the training set. Let N_p be the number of unique people in the probe and N_g be the number of unique people in the gallery. Let $(\mathbf{s}_i, \mathbf{t}_j)$ be the pair of closest points on the affine hulls of $\mathcal{P}_i$ and $\mathcal{G}_j$.

Our key insight is that directly computing the distance between the closest points on the affine hulls of $\mathcal{P}_i$ and $\mathcal{G}_i$ will not lead to accurate re-id results because this would be a unsupervised, suboptimal approach. To this end, we propose to learn distance metrics that result in discriminative representations of these affine hulls. Essentially, the formulation is in the same spirit as traditional metric learning algorithms that formulate pairwise constraints on the average feature points. However, the key idea is that we now formulate these constraints on pairs of closest points computed using affine hulls of the sets of image data available for each person. To make this more clear, let $\mathcal{P}_i$, $\mathcal{G}_i$, and $\mathcal{G}_j$ be three sets of feature vectors. Let $(\mathbf{s}_i, \mathbf{t}_i)$ and $(\mathbf{s}_i, \mathbf{t}_j)$ be the pairs of closest points on the affine hulls of $\{\mathcal{P}_i,\ \mathcal{G}_i\}$ and $\{\mathcal{P}_i,\ \mathcal{G}_j\}$ respectively, computed using some hull distance algorithm. Furthermore, for the sake of discussion here, let $\mathbf{a}_i^p$, $\mathbf{a}_i^g$, and $\mathbf{a}_j^g$ be the average feature points of the sets $\mathcal{P}_i$, $\mathcal{G}_i$, and $\mathcal{G}_j$ respectively.

Traditional metric learning algorithms formulate learning a new feature space with respect to constraints on the average feature points. Specifically, a new feature space is learned such that

$$d(\mathbf{a}_i^p, \mathbf{a}_i^g) < d(\mathbf{a}_i^p, \mathbf{a}_j^g) \tag{7}$$

where $d(\mathbf{x}, \mathbf{y})$ is the learned distance metric. In other words, the goal is to ensure the average feature points of the same person in the probe and gallery views are close whereas those of different people are relatively far. In contrast, we propose to enforce constraints on the pairs of closest points on the affine hulls of these sets. Specifically, our approach learns a new feature space such that

$$d(\mathbf{s}_i, \mathbf{t}_i) < d(\mathbf{s}_i, \mathbf{t}_j) \tag{8}$$

where $d(\mathbf{x}, \mathbf{y})$ is the learned distance metric. In addition to exploiting the underlying structure of the data, this approach is also relatively robust to noise. While the average feature point of a set is skewed by the presence of a few outliers, this is not the case for the pair of closest points determined from affine hulls. The idea, and its difference from the traditional approach, is illustrated in Figure 3.

To put the proposed approach in a more formal framework, as before, let $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_N\}$ be N sets of feature vectors corresponding to N people in the probe camera and $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N\}$ be the corresponding gallery feature sets. We form pairs of positive and negative feature sets and compute the closest pair of points from the corresponding affine hulls, thus generating sets of similar feature vectors

$$\mathcal{S} = \{(\mathbf{s}_i^{sim}, \mathbf{t}_i^{sim}), i = 1, \ldots, N\} \tag{9}$$

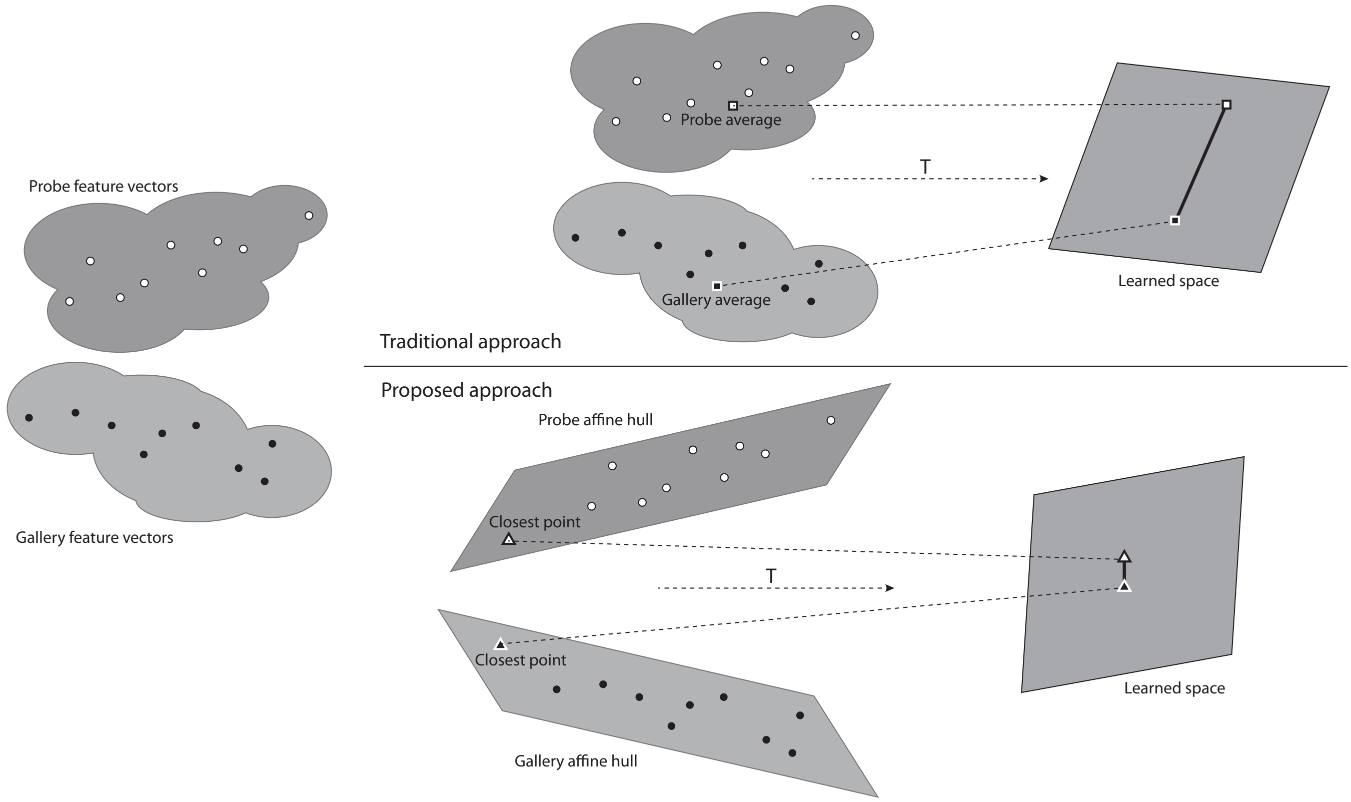


Fig. 3. An illustration of the main idea presented in this paper. In the traditional approach (top half of figure), the distance (in a learned transformation space) between the average feature vectors of multi-shot data (squares) is used to characterize the similarity between the probe and gallery sets. In the proposed approach (bottom half of figure), the learned space is based on the pairs of points defining the affine hull distance (triangles), which better characterizes the set-to-set similarity.

and dissimilar feature vectors

$$\mathcal{N} = \{(\mathbf{s}_i^{dis}, \mathbf{t}_j^{dis}) | i, j = 1, \ldots, N, i \neq j\}. \tag{10}$$

Note that each pair $(\mathbf{s}_i, \mathbf{t}_j)$ in both similar and dissimilar sets is formed from the corresponding feature sets $(\mathcal{P}_i, \mathcal{G}_j)$. Once we have constructed the training data from pairs of closest points, the metric learning problem to learn the distance metric can then be formulated as before:

$$\min_{\mathbf{T}} l(\mathcal{S}, \mathcal{N}) + \lambda \mathcal{R}(\mathcal{S}, \mathcal{N}) \tag{11}$$

While this is a generic problem formulation, we give some algorithm-specific details as to how to incorporate the proposed approach into the metric learning framework. In PCCA-like algorithms that learn Mahalanobis distance metrics, the process is straightforward. Specifically, as described in the original paper, PCCA learns the matrix $\mathbf{T}$ by minimizing an objective function based on logistic loss. Specifically, the loss function used is

$$l(\mathcal{S}, \mathcal{N}) = \sum_{i=1}^{n} l_\beta(t_i(d^2(\mathbf{x}, \mathbf{y}) - 1)) \tag{12}$$

where n is the number of training samples, $t_i = 1$ if the i^{th} training sample $(\mathbf{x}_i, \mathbf{y}_i)$ is a positive example and $t_i = -1$ if the i^{th} training sample is a negative pair, $l_\beta(x) = \frac{1}{\beta}\log(1+\beta x)$ is the generalized logistic function, and $d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^\top \mathbf{T}(\mathbf{x} - \mathbf{y})$ is the distance function. Once we have the pairs of similar and dissimilar feature vectors $\mathcal{S}$ and $\mathcal{N}$ as constructed above, we can easily use this data to minimize the objective function in Equation 12.

In FDA-like metric learning algorithms [11], [28], the distance metric is learned by solving generalized eigenvalue problems involving data scatter matrices. Specifically, the projection matrix $\mathbf{T}$ is learned as

$$\mathbf{T} = \underset{\mathbf{T}}{argmax} \text{ trace}\{(\mathbf{T}^\top \mathbf{A}_w \mathbf{T})^{-1} \mathbf{T}^\top \mathbf{A}_b \mathbf{T}\}$$

Again, the proposed approach can be easily incorporated into this framework by constructing the within-class ($\mathbf{A}_w$) and between-class ($\mathbf{A}_b$) matrices using pairs of closest points on affine hulls obtained as described above. In this case, the loss function used to learn the projection matrix is

$$l(\mathcal{S}, \mathcal{N}) = \text{ trace}\{(\mathbf{T}^\top \mathbf{A}_w \mathbf{T})^{-1} \mathbf{T}^\top \mathbf{A}_b \mathbf{T}\}$$

which is maximized to determine $\mathbf{T}$.

E. Re-Identification using learned representations

Given a probe person represented by the corresponding feature set $\mathcal{P}$ and N feature sets $\mathcal{G}_i$, $i = 1, \ldots, N$ of the people in the gallery, we employ the following procedure to determine the identity of the probe person:

- Construct affine hulls for each pair $\{\mathcal{P}, \mathcal{G}_i\}$ and find the pairs of closest points $(\mathbf{s}^p, \mathbf{s}_i^g)$, $i = 1, \ldots, N$ on these hulls using a hull distance algorithm.
- Determine the representations $\hat{\mathbf{s}}^p$ and $\hat{\mathbf{s}}_i^g$ of $\mathbf{s}^p$ and $\mathbf{s}_i^g$ respectively in the new feature space learned with a metric learning algorithm.
- Assign the identity of the gallery representation $\hat{\mathbf{s}}_i^g$ with the least Euclidean distance to $\hat{\mathbf{s}}^p$ as the identity of the probe person.

IV. Experiments and Results

A. Datasets

We empirically validate the proposed algorithm on the following publicly available multi-shot re-id datasets:

- iLIDS-VID: The iLIDS-VID [35] dataset contains image sequences of 300 people as seen from two cameras with non-overlapping views in an indoor airport environment. The length of the image sequences varies from 23 frames to 192 frames, with an average length of 73 frames.
- PRID 2011: The PRID 2011 [46] dataset contains image sequences of 385 people seen from one camera, and 749 people seen from an adjacent camera, both situated in an outdoor environment with non-overlapping views. To ensure evaluation consistency with [35], we only consider sequences corresponding to 178 people seen in both views. The average sequence length for these people is 100 frames.
- SAIVT-SoftBio: The SAIVT-SoftBio [47] dataset contains image sequences of 150 people passing through 8 cameras in an airport surveillance network. To ensure evaluation consistency with [47], we only consider two camera pairs: cameras 3 and 8 (which we call SAIVT-38) with 99 people and cameras 5 and 8 (which we call SAIVT-58) with 103 people.

B. Training/Testing protocol and parameters

For each dataset, we generate 10 sets of training and testing data splits. For iLIDS-VID and PRID 2011, we use the same equal-sized training and testing splits as provided in [35]. For SAIVT-38, we use images of 31 people as the training set and the rest as the testing set. For SAIVT-58, we use images of 33 people as the training set and the rest as the testing set.

We use the training data to learn a distance metric, as described below, and then compute the re-id performance for each split in the transformed feature space and report the average performance across all the splits. While most results described in this section are generated using the RNP [17] algorithm with parameters $\lambda_1 = 10$ and $\lambda_2 = 1$ to determine the closest points on affine hulls, we also analyze the performance of the proposed approach using an alternative hull distance algorithm, AHISD [15], to further demonstrate the general applicability of the proposed approach.

C. Features and metric learning algorithms and data normalization

Our approach is feature- and metric-agnostic and can be applied to any combination of existing feature extraction and metric learning algorithms. To this end, we consider a wide variety of algorithms that have become popular in the re-id community over the last four years. Specifically, we consider 8 existing feature extraction and 10 existing metric learning algorithms, noted in Table I, representing methods published through CVPR 2016. In addition, we also consider the L_2 metric, where we do not perform any feature space transformation and work in the original feature space. A detailed description of these algorithms can be found in the noted references as well as the recent systematic study by Karanam *et al.* [34]. For computational efficiency, we reduce the dimensionality of each feature space to 100 dimensions using the principal components analysis (PCA) algorithm. Subsequently, we employ the data normalization technique suggested in [48]. Specifically, we divide each component of the feature vector by its largest value across the training set, following which each feature vector is l_2-normalized.

D. Evaluation framework

As noted in the previous section, our proposed approach, which we name discriminative representations of affine hulls (DRAH), is independent of the feature space we work in and can be used to improve the performance of existing metric learning methods. To this end, we adopt the following evaluation framework: given a certain feature space, we first consider the average feature vector of the available multi-shot data and learn a transformed feature space, which we then use to compute the re-id performance. We name this approach AVER in all subsequent discussion. We then use our approach, DRAH, to model data as affine hulls, find closest points on them, and learn a transformed feature space using these closest points, which we then use to compute the re-id performance.

E. Results and discussion

We first evaluate the impact of modeling data as affine hulls in the absence of any metric learning. To this end, in AVER, we use the L_2 distance to rank gallery candidates. In DRAH, we construct affine hulls in the originally computed feature spaces, following which we use the L_2 distance between these pairs of closest points to rank gallery candidates. The rank-1 results of this experiment are shown in Table II.

Here, we also present some qualitative results comparing the performance of AVER and DRAH. To this end, we use GOG as the feature and the L_2 distance as the ranking algorithm to rank gallery candidates. In Figure 4, we show one example from each of the four test datasets, directly comparing the ranking performance between AVER and DRAH.

We further demonstrate the efficacy of our proposed approach in the context of several metric learning algorithms. The rank-1 results of this experiment are shown in Tables III through VI.

As can be noted from the results in these cases, modeling data as affine hulls is an effective strategy in dealing with the multi-shot aspect of the data, with DRAH generally giving better performance than AVER regardless of the feature space. We see much clearer trends in the case involving metric learning when compared to that without metric learning, with

TABLE I
EVALUATED (A) FEATURE EXTRACTION AND (B) METRIC LEARNING METHODS.

Feature	Year
ELF [18]	ECCV 08
LDFV [20]	ECCVW 12
AlexNet [49]	NIPS 12
gBiCov [50]	BMVC 12
SDC [5]	CVPR 13
HistLBP [6]	ECCV 14
LOMO [22]	CVPR 15
GOG [23]	CVPR 16

(a)

Metric	Year
L_2	–
FDA [28]	AE 1936
MFA [51]	PAMI 07
ITML [30]	ICML 07
LMNN [31]	JMLR 08
PCCA [12]	CVPR 12
KISSME [13]	CVPR 12
LFDA [52]	CVPR 13
kMFA [6]	ECCV 14
kLFDA [6]	ECCV 14
XQDA [22]	ICCV 15

(b)

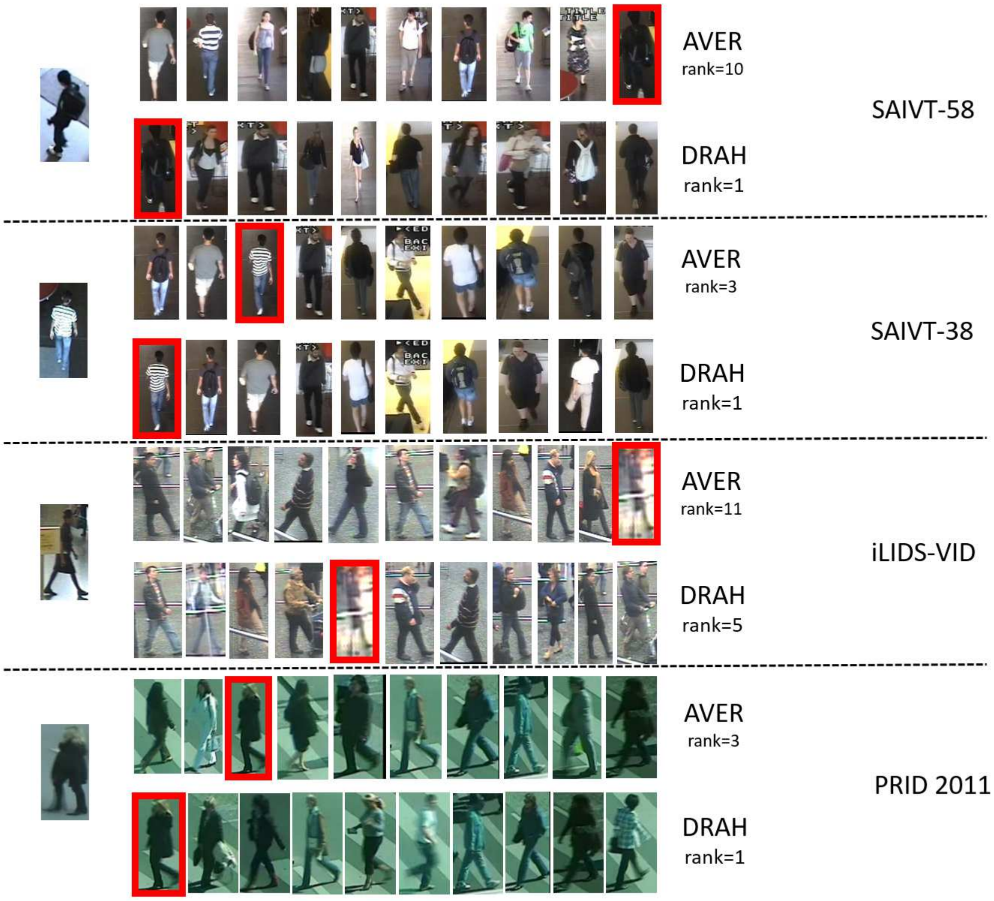


Fig. 4. Qualitative examples from each of the four test datasets to illustrate the impact of the proposed approach. For each dataset, we show two ranked gallery lists for a certain probe candidate. One list corresponds to using the traditional feature averaging scheme whereas the other list corresponds to the proposed approach. In each case, we see that the person of interest is ranked higher in the list corresponding to the proposed approach, DRAH, when compared to the traditional approach, AVER.

TABLE II
AVER vs. DRAH in the absence of any metric learning.

Metric	SAIVT-58		SAIVT-38		iLIDS-VID		PRID	
Rank	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH
ELF	11.1	**24.7**	19.3	**34.8**	10.5	**14.4**	16.3	**23.5**
LDFV	10.1	**14.3**	29.3	**37.6**	8.3	**9.2**	16.3	**20.0**
AlexNet	27.4	**38.6**	67.9	**73.5**	16.9	**22.8**	35.2	**42.3**
gBiCov	18.7	**31.3**	39.7	**52.9**	7.8	**12.5**	44.3	**47.8**
SDC	**10.4**	10.3	29.4	**36.0**	9.3	**11.9**	21.2	**27.1**
histLBP	5.0	**7.1**	23.2	**25.6**	7.2	**8.9**	17.4	**22.7**
LOMO	29.9	**45.7**	47.5	**65.6**	17.7	**26.9**	**50.7**	48.1
GOG	44.7	**53.9**	76.0	**82.8**	29.6	**35.1**	60.8	**64.0**

TABLE III
AVER vs. DRAH with metric learning: results on the SAIVT-58 dataset.

Metric	FDA		MFA		ITML		LMNN		PCCA		KISSME		LFDA		kMFA		kLFDA		XQDA	
Rank	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH
ELF	21.1	**41.4**	13.7	**20.1**	27.6	**36.7**	17.7	**33.7**	28.7	**31.9**	22.9	**43.3**	22.1	**43.3**	23.3	**40.9**	29.1	**51.0**	28.1	**33.1**
LDFV	26.7	**34.0**	13.4	**13.4**	28.7	**35.7**	19	**28.7**	29.9	**33.0**	25.7	**35.1**	26.4	**34.9**	20.1	**35.6**	28.6	**49.0**	33	**34.6**
AlexNet	29.6	**46.7**	25.9	**29.4**	33.9	**43.9**	28.9	**41.9**	29.9	**37.6**	29.9	**45.4**	30.7	**45.7**	19	**31.3**	31.6	**63.9**	34.4	**38.4**
gBiCov	19.4	**32.0**	20	**28.9**	25.4	**35.3**	24.1	**30.0**	22.9	**27.7**	19.3	**32.1**	19.4	**32.0**	27.6	**49.0**	26	**55.7**	24.6	**25.7**
SDC	18.7	**22.3**	10.0	**11.1**	23.7	**22.6**	14.4	**19.3**	21.7	**25.7**	18.0	**22.1**	18.9	**22.6**	18.3	**32.6**	25.7	**41.6**	**23.0**	22.3
histLBP	14.4	**20.0**	**6.3**	4.6	18.4	**23.9**	8.6	**15**	18.9	**20.4**	15.7	**21.4**	14.7	**20.0**	12.6	**26.6**	19.9	**33.0**	19.3	**21.6**
LOMO	45.1	**60.7**	29.1	**34.3**	45.3	**55.6**	34.9	**45.9**	47.9	**52.4**	44.0	**59.7**	43.7	**60.6**	24.4	**46.4**	46.3	**59.0**	50	**56.6**
GOG	53.4	**66.7**	43.6	**45.7**	57	**66.3**	49.3	**53.6**	52.7	**59.4**	52.3	**66.1**	54.3	**65.7**	26	**53.6**	51.9	**61.6**	56.7	**62.9**

TABLE IV
AVER vs. DRAH with metric learning: results on the SAIVT-38 dataset.

Metric	FDA		MFA		ITML		LMNN		PCCA		KISSME		LFDA		kMFA		kLFDA		XQDA	
Rank	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH
ELF	33.2	**50.0**	18.2	**29.0**	38.7	**49.9**	26.2	**36.9**	37.4	**43.8**	33.5	**50.1**	34.9	**50.3**	22.9	**39.7**	39.6	**59.3**	38.4	**42.1**
LDFV	53.8	**58.4**	31.5	**31.8**	53.4	**56.5**	40.7	**41.3**	49.0	**46.3**	53.5	**57.8**	56.2	**59.4**	**36.3**	33.7	52.4	**67.4**	**54.4**	48.5
AlexNet	74.4	**82.4**	58.8	**64.6**	75.6	**80.3**	59.6	**65.0**	62.9	**70.0**	74.6	**82.6**	74.0	**82.1**	42.1	**72.5**	68.2	**87.8**	**69.0**	67.6
gBiCov	48.1	**59.9**	40.3	**48.2**	50.4	**61.9**	44.3	**51.0**	42.9	**49.1**	47.9	**59.7**	48.1	**59.9**	42.9	**63.1**	47.6	**76.3**	**46.8**	44.9
SDC	45.6	**54.1**	29.9	**30.9**	51.5	**52.2**	36.3	**40.1**	47.6	**50.0**	46.6	**55.7**	44.4	**54.3**	33.5	**47.4**	50.6	**68.5**	**51.0**	46.6
histLBP	36.2	**50.4**	**20.9**	19.3	40.4	**45.7**	26.6	**32.6**	36.0	**41.9**	35.3	**51.8**	36.5	**51.2**	25.6	**42.4**	42.5	**62.6**	**42.4**	41.8
LOMO	70.6	**80.9**	41.9	**54.4**	68.8	**73.7**	52.6	**58.1**	67.9	**71.0**	68.1	**78.5**	68.7	**78.7**	39.7	**61.9**	68.5	**82.2**	**68.5**	67.4
GOG	85.4	**91.0**	70.6	**75.9**	85.4	**89.7**	72.8	**76.2**	81.9	**85.0**	85.4	**90.9**	84.3	**90.6**	62.4	**83.7**	86.0	**92.1**	**86.0**	82.8

DRAH giving better performance when compared to AVER in most feature-metric combinations. There seem to be a few exceptions, which also help complement the structure of the image data in the datasets used in this work. In particular, we observe these exceptions in a few combinations for the SAIVT-38 and PRID datasets, partly due to the fact that, relative to the other datasets, the variations across the available images for each person are not as pronounced, as noted in Karanam *et al.* [34]. This results in certain features producing closely clustered feature sets for which modeling data as affine hulls does not result in the desired benefit. Furthermore, the figures in the metric learning case, shown in Tables III through VI, are generally much higher when compared to those in Table II, suggest that modeling data as affine hulls and directly computing the distance between the closest points on these affine hulls is alone not sufficient to get good performance because this would be a purely unsupervised, suboptimal approach. Learning discriminative representations of these affine hulls

TABLE V
AVER vs. DRAH with metric learning: results on the iLIDS-VID dataset.

Metric	FDA		MFA		ITML		LMNN		PCCA		KISSME		LFDA		kMFA		kLFDA		XQDA	
Rank	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH
ELF	16.7	**25.5**	11.1	**20.2**	20.2	**22.4**	23.7	**32.1**	18.2	**25.3**	11.9	**16.3**	15.9	**25.9**	16.5	**60.1**	13.1	**53.9**	23.2	**26.1**
LDFV	21.3	**26.7**	13.4	**20.3**	20.1	**24.5**	24.5	**30.5**	25.5	**29.2**	16.7	**18.3**	21.5	**26.2**	20.1	**58.3**	18.7	**54.4**	28.1	**29.4**
AlexNet	21.5	**27.4**	16.0	**26.1**	19.5	**23.0**	29.9	**38.5**	22.4	**25.3**	12.3	**14.6**	20.8	**27.7**	25.9	**60.9**	18.7	**43.2**	**27.8**	27.4
gBiCov	8.1	**13.7**	7.2	**14.1**	12.3	**18.9**	16.5	**25.8**	14.7	**20.9**	3.7	**6.6**	8.1	**13.7**	11.5	**62.5**	8.2	**60.4**	13.1	**13.7**
SDC	16.3	**22.1**	12.1	**17.2**	15.0	**16.8**	21.5	**26.3**	18.2	**21.2**	12.5	**14.1**	18.4	**22.5**	16.5	**56.3**	15.8	**53.5**	21.6	**21.9**
histLBP	20.4	**26.9**	13.1	**18.0**	18.7	**21.5**	25.5	**27.3**	24.1	**26.3**	16.4	**17.3**	20.4	**25.9**	18.3	**54.2**	18.5	**49.3**	26	**28.6**
LOMO	36.7	**42.6**	26.0	**34.0**	30.5	**36.0**	44.1	**50.7**	32.9	**39.7**	25.3	**26.1**	36.8	**43.2**	37.9	**61.7**	32.1	**45.5**	42.3	**43.2**
GOG	40.6	**49.0**	30.4	**42.3**	41.0	**45.2**	47.1	**56.6**	37.5	**45.8**	29.2	**33.8**	40.4	**48.7**	43.5	**53.7**	37.0	**64.0**	45.7	**49.2**

TABLE VI
AVER vs. DRAH with metric learning: results on the PRID dataset.

Metric	FDA		MFA		ITML		LMNN		PCCA		KISSME		LFDA		kMFA		kLFDA		XQDA	
Rank	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH	AVER	DRAH
ELF	19.3	**20.1**	16.5	**24.8**	25.8	**33.8**	31.5	**40.7**	28.5	**31.6**	21.1	**25.3**	19.2	**22.4**	28.1	**58.9**	25.3	**53.7**	31.6	**31.9**
LDFV	32.2	**32.8**	24.0	**34.4**	**34.6**	32.0	**40.2**	42.4	38.9	**45.6**	33.8	**15.6**	**31.9**	31.5	42.8	**65.3**	38.5	**60.7**	**45.1**	43.8
AlexNet	27.2	**30.6**	31.2	**39.9**	36.7	**46.2**	44.2	**49.0**	30.2	**36.9**	30.4	**33.3**	28.3	**31.9**	41.6	**76.7**	37.5	**72.6**	40.2	**42.6**
gBiCov	22.0	**29.4**	34.9	**42.8**	29	**46.9**	51.3	**58.4**	36.5	**47.1**	24.4	**34.2**	22.0	**29.4**	48.3	**83.1**	39.9	**80.3**	38.9	**39.0**
SDC	23.9	**24.9**	19.1	**25.4**	30.6	**34.7**	35.3	**39.9**	33.5	**37.3**	27.3	**29.4**	23.8	**26.1**	31.5	**61.1**	29.1	**55.7**	**36.2**	35.6
histLBP	21.9	**22.5**	20.2	**25.1**	24.9	**30.4**	28.9	**38.2**	30.4	**32.0**	23.3	**24.9**	19.9	**21.2**	30.3	**60.3**	29.4	**58.7**	32.1	**33.4**
LOMO	**58.1**	57.8	48.0	**54.7**	61.6	**62.6**	68.5	**71.3**	64.3	**70.3**	61.3	**64.3**	57.4	**59.7**	65.1	**75.5**	67.3	**87.1**	**70.7**	**70.7**
GOG	61.6	**62.8**	60.1	**67.9**	67.2	**71.0**	71.5	**75.3**	61.1	**68.1**	67.1	**69.2**	61.2	**61.9**	70.6	**79.8**	70.8	**88.7**	**75.5**	74.0

using metric learning algorithms, a key theme of the approach proposed in this paper, is critical.

Finally, we summarize the results in Tables II through VI using two CMC curves for each dataset, shown in Figure 5. To this end, for each dataset, we compute the average performance across all the evaluated feature-metric combinations. We note that DRAH gives consistently better performance when compared to AVER, resulting in a normalized area under the CMC curve improvement of 3.5, 1.7, 5.8, and 1.7 and an average rank-1 improvement of 9.9%, 8.1%, 10.2%, and 8.3% on the SAIVT-58, SAIVT-38, iLIDS-VID, and PRID datasets respectively. The relatively higher performance improvement observed in the case of iLIDS-VID and SAIVT-58 is in line with the way images in these datasets are captured. As noted in Karanam *et al.* [34], iLIDS-VID and SAIVT-58 suffer from a higher degree of viewpoint and illumination variations when compared to the other datasets. Therefore, as noted earlier, the proposed approach is an effective strategy to deal with such multi-shot feature sets.

While the results discussed above were generated using RNP as the hull distance algorithm, our proposed approach is equally applicable to other hull distance algorithms. To substantiate this point, we repeated all the above experiments using an alternative hull distance algorithm, AHISD [15], using the same evaluation protocols and combinations of features/metric learning algorithms. The average CMC curves obtained are shown in Figure 6, where we observe an average rank-1 improvement of 5.5%, 1.6%, 4.8%, and 1.2% on the SAIVT-58, SAIVT-38, iLIDS-VID, and PRID datasets respectively with DRAH over AVER.

F. Comparison with the state of the art

Finally, we compare the performance of the proposed approach with the state of the art in multi-shot re-id. As of ECCV 2016, the best reported rank-1 results on the iLIDS-VID and PRID datasets are 58.0% and 77.3%, achieved using a recurrent neural network (RNN) [25] and a convolutional neural network (CNN) [24] respectively. Our approach, with GOG [23] as the feature and kLFDA [6] as the metric learning algorithm, achieves a rank-1 performance of 64.0% and 88.7%, representing a substantial improvement of 6.0% and 11.4% respectively. On the SAIVT-38 and SAIVT-58 datasets, our approach, again with GOG as the feature space, results in a rank-1 performance improvement of 6.0% and 9.7% respectively over the best performing combination of GOG and XQDA [22] and GOG and ITML [30] respectively. These results are summarized in Tables VII and VIII.

These results suggest that using the pair of closest points instead of the average points is an effective strategy to deal with the multi-shot aspect of multi-shot person re-id. While average points are sensitive to noise and outliers, pairs of closest points better characterize set-to-set similarity of feature sets. Furthermore, as noted above and in Section IV-E, we achieve better performance by learning representations of these points using discriminative distance metrics and not solely relying on Euclidean distance. While this is expected with the

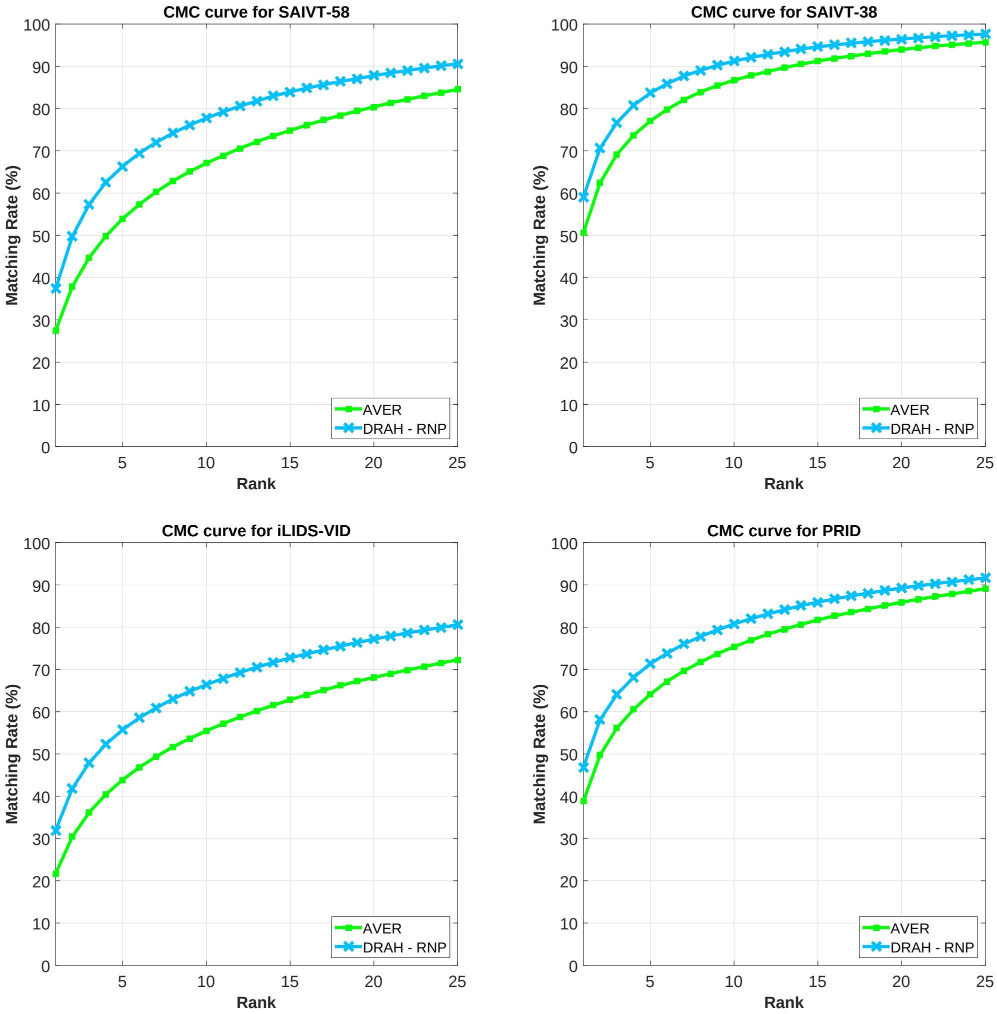


Fig. 5. Evaluating the impact of the DRAH over AVER across all the feature-metric algorithm combinations with the RNP hull distance algorithm.

added supervision in the metric learning case, these results also suggest that pairs of closest points provide for better training data for the metric learning algorithms, resulting in DRAH giving improved performance over AVER.

V. Conclusions and Future Work

We advocated for modeling the multi-shot data in multi-shot person re-id using affine hulls, and demonstrated that such a data modeling scheme can improve the performance of existing metric learning methods that use the average feature vector as the data exemplar. Furthermore, we demonstrated substantial improvements over the existing state of the art on three popular multi-shot re-id datasets.

A promising future research direction in the context of the proposed method would be to integrate multi-shot data modeling and ranking with metric learning. While most existing methods treat these two topics separately, developing a unified metric learning and multi-shot ranking framework that exploits the several aspects of multi-shot data can potentially lead to further performance gains. For instance, borrowing ideas from research in spatio-temporal feature learning [53], [54] would be a natural next step in developing such unified algorithms. Additionally, such multi-shot ranking and learning techniques can be integrated with representative sample selection schemes [55], [56], developing algorithms that can be used to select most discriminative fragments from the available multi-shot data for ranking gallery candidates.

We conclude with a discussion on scenarios where the proposed approach might fail and potential solutions to tackle the problem. In scenarios that enforce a dress code on people, appearance features computed using any of the feature extraction algorithms discussed here will result in a feature space where most gallery candidates will look alike. In scenarios that involve high crowd density, the person images captured by cameras will be affected by occlusions and background distractions. In this case, the feature space of the candidates will be noisy. In such scenarios, the proposed approach as well as the traditional approach will not give satisfactory results. To mitigate these problems, a possible solution would be to adopt a multi-modal approach to describe the appearance of person images. For instance, in the scenario involving a dress code, person faces can be detected using a face detector [57] and this information can be fused into the existing appearance

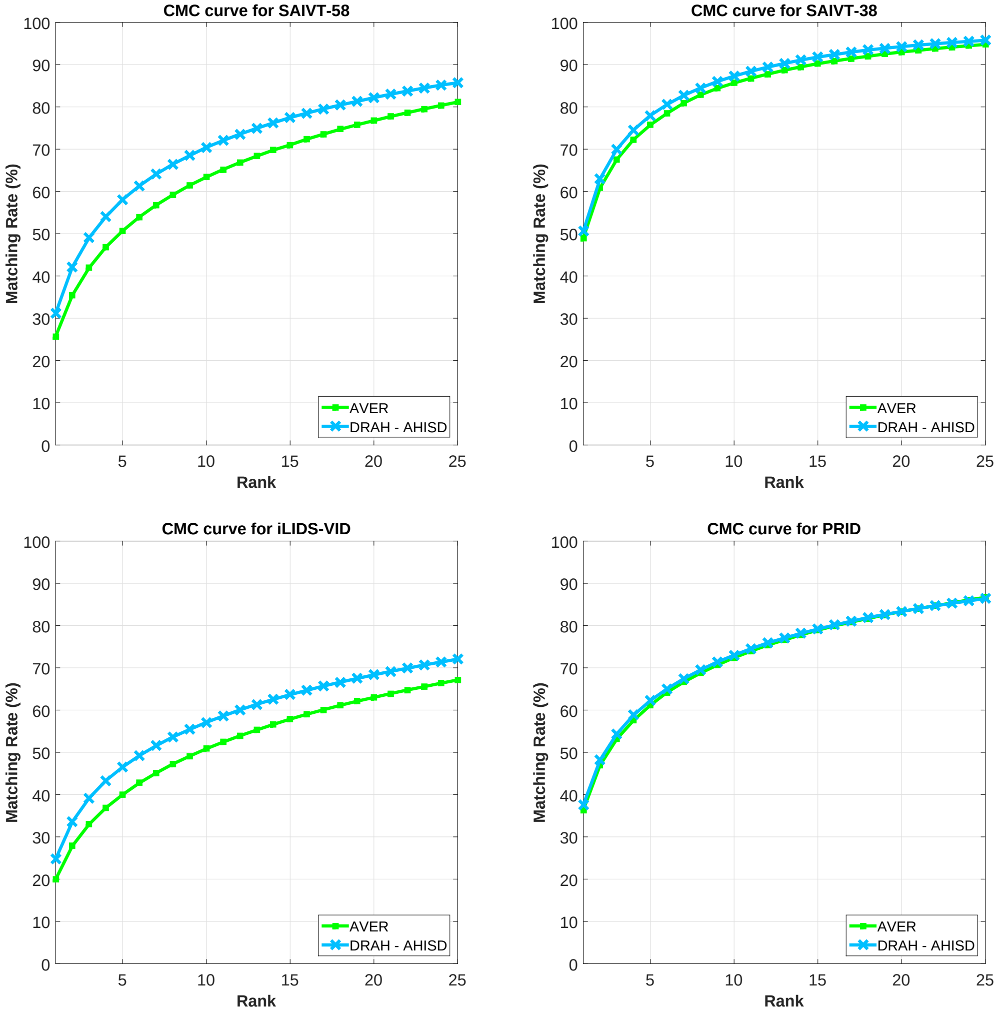


Fig. 6. Evaluating the impact of the DRAH over AVER across all the feature-metric algorithm combinations with the AHISD hull distance algorithm.

TABLE VII
COMPARISON WITH THE BEST PUBLISHED RESULTS TO DATE: RESULTS ON THE PRID AND ILIDS-VID DATASETS.

Dataset	PRID 2011				iLIDS-VID			
Rank	1	5	10	20	1	5	10	20
CNN+XQDA (ECCV 2016, [24])	77.3	93.5	95.7	99.3	53.0	81.4	90.1	95.1
RNN (CVPR 2016, [25])	70.0	90.0	95.0	97.0	58.0	84.0	91.0	96.0
DRAH	**88.7**	**97.9**	**98.9**	**99.7**	**64.0**	**86.0**	**91.7**	**96.3**

TABLE VIII
COMPARISON WITH THE BEST PUBLISHED RESULTS TO DATE: RESULTS ON THE SAIVT-38 AND SAIVT-58 DATASETS.

Dataset	SAIVT-38				SAIVT-58			
Rank	1	5	10	20	1	5	10	20
GOG (CVPR 2016, [23])+ITML [30]	85.4	98.5	99.7	99.9	57.0	81.0	89.9	94.1
GOG (CVPR 2016, [23])+XQDA [22]	86.0	98.8	**100**	**100**	56.7	83.7	92.1	**96.1**
DRAH	**92.0**	**99.7**	**100**	**100**	**66.7**	**87.4**	**92.3**	95.9

modeling paradigm. In the scenario involving occlusions and background distractions, we can use person motion and gait information [58] to construct the feature space. Once we have a well-represented feature space, the algorithm proposed in this paper can readily be applied in conjunction with metric learning algorithms.

Acknowledgments

This material is based upon work supported by the U.S. Department of Homeland Security, Science and Technology Directorate, Office of University Programs, under Grant Award 2013-ST-061-ED0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used in this research.

References

[1] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking." in *Proc. Brit. Mach. Vision Conf. (BMVC)*, 2010.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, San Francisco, CA, 2010, pp. 2360–2367.

[3] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Eur. Conf. Comput. Vision (ECCV)*, Florence, Italy, 2012, pp. 780–793.

[4] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat, "Learning to match appearances by correlations in a covariance metric space," in *Eur. Conf. Comput. Vision (ECCV)*, Florence, Italy, 2012, pp. 806–820.

[5] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2013.

[6] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Eur. Conf. Comput. Vision (ECCV)*, 2014.

[7] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *IEEE Conf. Comput. Vision and Pattern Recognition*, Boston, MA, 2015, pp. 3908–3916.

[8] Z. Wu, Y. Li, and R. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1095–1108, 2015.

[9] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2015.

[10] Y. Li, Z. Wu, S. Karanam, and R. J. Radke, "Real-world re-identification in an airport camera network," in *Proc. Int. Conf. Distributed Smart Cameras (ICDSC)*, 2014.

[11] O. Camps, M. Gou, T. Hebble, S. Karanam, O. Lehmann, Y. Li, R. Radke, Z. Wu, and F. Xiong, "From the lab to the real world: Re-identification in an airport camera network," *IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT)*, vol. 27, no. 3, 2017.

[12] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2012.

[13] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2012.

[14] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Annu. Conf. Neural Inform. Process. Syst. (NIPS)*, 1998.

[15] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2010.

[16] Y. Hu, A. S. Mian, and R. Owens, "Sparse approximated nearest points for image set classification," in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2011.

[17] M. Yang, P. Zhu, L. Van Gool, and L. Zhang, "Face recognition based on regularized nearest points between image sets," in *IEEE Int. Conf. Automatic Face and Gesture Recognition (FG)*, 2013.

[18] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Eur. Conf. Comput. Vision (ECCV)*, 2008.

[19] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Comput. Vision and Image Understanding (CVIU)*, vol. 117, no. 2, pp. 130–144, 2013.

[20] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by Fisher vectors for person re-identification," in *ECCV Workshops*, 2012.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision (IJCV)*, vol. 60, no. 2, pp. 91–110, 2004.

[22] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2015.

[23] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2016.

[24] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "MARS: A video benchmark for large-scale person re-identification," in *Eur. Conf. Comput. Vision (ECCV)*, 2016.

[25] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2016.

[26] R. Satta, "Appearance descriptors for person re-identification: a comprehensive review," *arXiv preprint arXiv:1307.5748*, 2013.

[27] N. Martinel, C. Micheloni, and G. L. Foresti, "Kernelized saliency-based person re-identification through multiple metric learning," *IEEE Trans. Image Process. (T-IP)*, vol. 24, no. 12, pp. 5645–5658, 2015.

[28] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[29] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Re-identification in the function space of feature warps," *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, vol. 37, no. 8, pp. 1656–1669, 2015.

[30] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Int. Conf. Mach. Learning (ICML)*, 2007.

[31] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," vol. 10, pp. 207–244, 2009.

[32] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2011.

[33] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2013.

[34] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets," *arXiv preprint arXiv:1605.09653*, 2016.

[35] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Eur. Conf. Comput. Vision (ECCV)*, 2014.

[36] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *Proc. Brit. Mach. Vision Conf. (BMVC)*, 2008.

[37] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in *IEEE Int. Conf. Comput. Vision (ICCV)*, 2015.

[38] Y. Li, Z. Wu, and R. J. Radke, "Multi-shot re-identification with random-projection-based random forests," in *IEEE Winter Conf. Applicat. Comput. Vision (WACV)*, 2015.

[39] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, vol. 37, no. 8, pp. 1629–1642, 2015.

[40] Y. Li, Z. Wu, S. Karanam, and R. J. Radke, "Multi-shot human re-identification using adaptive Fisher discriminant analysis," in *Proc. Brit. Mach. Vision Conf. (BMVC)*, 2015.

[41] S. Karanam, Y. Li, and R. J. Radke, "Sparse Re-Id: Block sparsity for person re-identification," in *IEEE/ISPRS 2nd Joint Workshop on Multi-Sensor Fusion for Dynamic Scene Understanding*, 2015.

[42] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with block sparse recovery," *Image and Vision Computing*, 2017.

[43] Y. Wu, M. Minoh, and M. Mukunoki, "Collaboratively regularized nearest points for set based recognition," in *Proc. Brit. Mach. Vision Conf. (BMVC)*, 2013.

[44] P. Zhu, L. Zhang, W. Zuo, and D. Zhang, “From point to set: Extend the learning of distance metrics,” in *IEEE Int. Conf. Comput. Vision (ICCV)*, 2013.

[45] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[46] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, “Person re-identification by descriptive and discriminative classification,” in *Proc. Scandinavian Conf. Image Analysis (SCIA)*, 2011.

[47] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, and P. Lucey, “A database for person re-identification in multi-camera surveillance networks,” in *Int. Conf. Digital Image Computing Techniques and Applicat. (DICTA)*, 2012.

[48] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2014, pp. 1701–1708.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Annu. Conf. Neural Inform. Process. Syst. (NIPS)*, 2012.

[50] B. Ma, Y. Su, and F. Jurie, “BiCov: a novel image representation for person re-identification and face verification,” in *Proc. Brit. Mach. Vision Conf. (BMVC)*, 2012.

[51] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, “Graph embedding and extensions: a general framework for dimensionality reduction,” *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, vol. 29, no. 1, pp. 40–51, 2007.

[52] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, “Local Fisher discriminant analysis for pedestrian re-identification,” in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2013.

[53] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in *IEEE Int. Conf. Comput. Vision (ICCV)*, 2015.

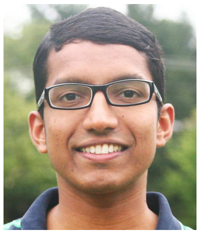
[54] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: towards good practices for deep action recognition,” in *Eur. Conf. Comput. Vision (ECCV)*, 2016.

[55] E. Elhamifar, G. Sapiro, and R. Vidal, “Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery,” in *Annu. Conf. Neural Inform. Process. Syst. (NIPS)*, Lake Tahoe, NV, 2012, pp. 19–27.

[56] E. Elhamifar, G. Sapiro, and R. Vidal, “See all by looking at a few: Sparse modeling for finding representative objects,” in *IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2012.

[57] P. Hu and D. Ramanan, “Finding tiny faces,” *arXiv preprint arXiv:1612.04402*, 2016.

[58] M. Gou, X. Zhang, A. Rates-Borras, S. Asghari-Esfeden, M. Sznaier, and O. Camps, “Person re-identification in appearance impaired scenarios,” in *Proc. Brit. Mach. Vision Conf. (BMVC)*, 2016.



Srikrishna Karanam Srikrishna Karanam is a Research Scientist in the Vision Technologies and Solutions group at Siemens Corporate Technology, Princeton, NJ. He has a B.Tech. degree in Electronics and Communication Engineering from the National Institute of Technology Warangal, and M.S. degree in Electrical Engineering and Ph.D. degree in Computer & Systems Engineering from Rensselaer Polytechnic Institute. His research interests include computer vision, machine learning, and data analytics with focus on all aspects of image indexing, search, and retrieval for object recognition applications.

Ziyan Wu Ziyan Wu received a Ph.D. degree in Computer and Systems Engineering from Rensselaer Polytechnic Institute in 2014. He has B.S. and M.S. degrees in Engineering from Beihang University. He joined Siemens Corporate Research as a Research Scientist in 2014. His current research interests include 3D object recognition and autonomous perception.

Richard J. Radke Richard J. Radke joined the Electrical, Computer, and Systems Engineering department at Rensselaer Polytechnic Institute in 2001, where he is now a Full Professor. He has B.A. and M.A. degrees in computational and applied mathematics from Rice University, and M.A. and Ph.D. degrees in electrical engineering from Princeton University. His current research interests involve computer vision problems related to human-scale, occupant-aware environments, such as person tracking and re-identification with cameras and range sensors. Dr. Radke is affiliated with the NSF Engineering Research Center for Lighting Enabled Service and Applications (LESA), the DHS Center of Excellence on Explosives Detection, Mitigation and Response (ALERT), and Rensselaer’s Experimental Media and Performing Arts Center (EMPAC). He received an NSF CAREER award in March 2003 and was a member of the 2007 DARPA Computer Science Study Group. Dr. Radke is a Senior Member of the IEEE and a Senior Area Editor of *IEEE Transactions on Image Processing*. His textbook *Computer Vision for Visual Effects* was published by Cambridge University Press in 2012.