Set Based Discriminative Ranking for Recognition

Yang Wu[1,⋆], Michihiko Minoh[1], Masayuki Mukunoki[1], and Shihong Lao[2]

[1] Academic Center for Computing and Media Studies, Kyoto University
[2] OMRON Social Solutions Co., Ltd., Kyoto 619-0283, Japan
yangwu@mm.media.kyoto-u.ac.jp, {minoh,mukunoki}@media.kyoto-u.ac.jp,
lao@ari.ncl.omron.co.jp

Abstract. Recently both face recognition and body-based person re-identification have been extended from single-image based scenarios to video-based or even more generally image-set based problems. Set-based recognition brings new research and application opportunities while at the same time raises great modeling and optimization challenges. How to make the best use of the available multiple samples for each individual while at the same time not be disturbed by the great within-set variations is considered by us to be the major issue. Due to the difficulty of designing a global optimal learning model, most existing solutions are still based on unsupervised matching, which can be further categorized into three groups: a) set-based signature generation, b) direct set-to-set matching, and c) between-set distance finding. The first two count on good feature representation while the third explores data set structure and set-based distance measurement. The main shortage of them is the lack of learning-based discrimination ability. In this paper, we propose a set-based discriminative ranking model (SBDR), which iterates between set-to-set distance finding and discriminative feature space projection to achieve simultaneous optimization of these two. Extensive experiments on widely-used face recognition and person re-identification datasets not only demonstrate the superiority of our approach, but also shed some light on its properties and application domain.

1 Introduction

The existing research on object recognition mainly focuses on single-image based approaches, i.e., recognizing a single instance at each time. However, in many real applications like recognizing people in video surveillance (either by face or full body), it is not only possible but also easy to get multiple images or video frames for both query and gallery objects. And if allowed, users usually prefer to use as many images of each object as possible, because they think that

[⋆] This work was supported by "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society", Special Coordination Fund for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part III, LNCS 7574, pp. 497–510, 2012.

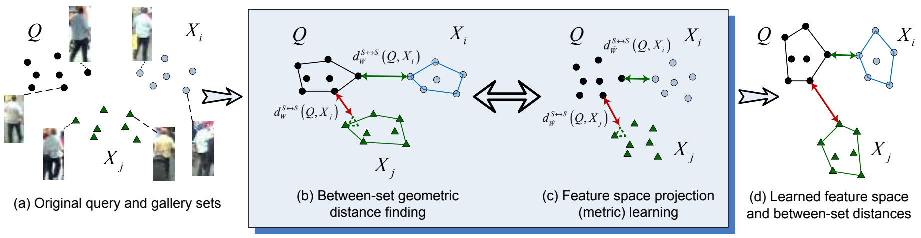

Q
X_i
X_j
(a) Original query and gallery sets
Q
X_i
$d_W^{S\leftrightarrow S}(Q,X_i)$
$d_W^{S\leftrightarrow S}(Q,X_j)$
X_j
(b) Between-set geometric distance finding
Q
$d_W^{S\leftrightarrow S}(Q,X_i)$
X_i
$d_W^{S\leftrightarrow S}(Q,X_j)$
X_j
(c) Feature space projection (metric) learning
Q
X_i
X_j
(d) Learned feature space and between-set distances

Fig. 1. Illustration of the proposed set-based discriminative ranking of a relevant and an irrelevant image sets in the gallery given a query set of images

more candidates generally mean more information and should result in better recognition performance. Such a belief comes from the experiences of human vision, and it has also been proved by recent research efforts [1–4].

Though set-based recognition is believed to be more promising than recognition based on single images, it brings new challenges: the images in a single set may be taken by different cameras under different conditions (illumination, viewpoint, background, etc.), and the object itself may also have large appearance changes (pose, occlusion, etc.). Such a wide coverage increases the chance of finding the correct matchings between sets, but could also increase the probability of making mistakes. How to make the best use of multiple instances while at the same time be discriminative on distinguishing image sets from different classes becomes the key issue. Existing supervised classification methods usually suppose the data to be recognized is a single instance which can be represented by a feature vector in the same space as the training samples stay, so they are not directly applicable to set-based recognition tasks.

In this paper, we propose a novel approach to optimize a global set-based recognition objective function by iteratively optimizing between-set distance finding and feature space projection (or namely metric learning). As shown in Figure 1, given a query set Q and two arbitrary gallery sets X_i and X_j, each of which contains multiple images of an object, our approach finds the distance between each pair of query-gallery sets by computing the geometric distance of their approximated convex hulls (distance of closest approach) [1]. It can be treated as finding the closest points from them, though by convex approximation these points may be virtual as they can be linear combinations of actual sample points. Then a maximum-margin-based ranking algorithm is adopted to learn a good metric, making the closest approach between correct query-gallery pair smaller than that between incorrect ones as much as possible. Such a metric learning can be viewed as feature space projection. After that, the closest approach between each query-gallery pair is updated in the new feature space, and it will activate another metric learning step. The iteration between these two continues until convergence is achieved. Experimental results on both face recognition and person re-identification demonstrate that our approach consistently outperforms state-of-the-art methods without parameter tuning.

2 Related Work

Existing methods on set-based object recognition depend much on the object categories they are working on. Generally speaking, research on face recognition focuses much on exploring the distribution [5] or structure of the data sets [6–8] and designing the between-set distance [9, 2], while in the literature of person re-identification, more attention has been paid to feature representation [3, 10, 4]. Such a phenomenon to some extent is due to the differences between these two categories: usually human bodies have greater appearance variations and occlusions than faces, causing difficulties for feature representation. Nevertheless, they are valuable attempts for solving the generic set-based recognition problems. Here we briefly categorize them into three groups: a) set-based signature generation, b) direct set-to-set matching, and c) between-set distance finding.

The first two groups count on exploring new features, in which group (a) aims at a single global and informative representation for each query/gallery set while group (b) looks for good features for individual images. More concretely, in 2006, a spatiotemporal segmentation algorithm was employed to generate signatures that are invariant to variations in illumination, pose, and the dynamic appearance of clothing, followed by clustering and matching methods [11]. Three years later, a new signature called HPE (Histogram Plus Epitome) [10] was proposed to integrate global HSV color histogram with local epitome descriptors (generic and local epitomes). The most recent work in group (a) is called MRCG (Mean Riemannian Covariance Grid) [4] which divides each image into a dense grid structure with overlapping cells, and then describe each cell by the region covariance features [12]. Specially designed mean and variance functions were applied to the cells to form a signature and then a similarity function was proposed to compare two signatures. Though these signatures are compact and directly compatible with image-based matching/classification algorithms, it is hard to make them both representative and discriminative. Unlike them, the feature representation in group (b) does not have to meet such high demands as it seeks for the cooperation with set-to-set matching methods. A typical approach is the one named SDALF (Symmetry-driven Accumulation of Local Features) [3], which explores the symmetry and asymmetry of the human body to partition it into three parts and then uses three types of features to describe each part (except the smallest head part). A weighted combination of distances on these features was defined as the image-to-image distance, and the minimum distance among all possible image-pairs was adopted for set-to-set matching. Though this group takes the advantage of having multiple candidates in each set for matching, the effectiveness of it still largely depends on the goodness of features which requires a case-dependent careful design.

Instead of focusing on features, the third group works on set structure extraction and distance computation. Between-set distance finding approaches were mainly proposed for face recognition. Among these efforts, two recently introduced methods are remarkable. One is AHISD/CHISD (Affine/Convex Hull based Image Set Distance) [1], which characterizes each image set (query/gallery) in terms of an affine/convex hull of the feature vectors of its images, and then finds the geometric distance (distance of closest approach) between two hulls to serve as the set-to-set distance. Such a convex approximation is less overfitting than the models based on sample points because it can generate new samples on the hull, and the approach can also be robust to outliers to some extent. The other work is named SANP (Sparse Approximated Nearest Points) [2] which enforces the sparsity of samples used for point generation via linear combination. Compared with direct set-to-set matching, this group of methods can make a better use of the data sets, however, as they are unsupervised, the feature space where the data sets stay directly determines their effectiveness since all the dimensions are equally weighted for distance computation.

3 Set Based Discriminative Ranking

The problem of set-based object recognition can be formulated as follows. Given a query $Q \in \mathcal{Q}$ where $\mathcal{Q}$ is the query space and Q itself is an image set containing the same object, the goal is to recognize the identity/category of the object by comparing to a gallery of image sets $\mathcal{X}$, each of which ($X \in \mathcal{X}$) is about a specific object or object category[1]. Motivated by both the theoretical and practical advantage of using ranking-based models for recognition [13, 14], here we propose a discriminative ranking method for solving the set-based recognition problem[2]. In this setting, for each query Q, we divide the gallery into two groups I_Q^+ andI_Q^-, where I_Q^+ are the indices of image sets containing the same object as Q, i.e. relevant sets, and I_Q^- denotes irrelevant ones. Then the desired ranking of the gallery sets $\mathbf{y}_Q^* \in \mathcal{Y}$ (where $\mathcal{Y}$ is the space of all the feasible rankings) will be the one that satisfies: $\forall i \in I_Q^+, j \in I_Q^-$, $X_i \prec_{\mathbf{y}_Q^*} X_j$, which denotes that a relevant set should always be ranked before any irrelevant set.

3.1 Set-Based Ranking Model

Given an arbitrary query set $Q \in \mathcal{Q}$ and the whole gallery $\mathcal{X}$ composed by a batch of image sets, we define a joint feature map for a candidate ranking $\mathbf{y}_Q$ of gallery $\mathcal{X}$ as $\psi(Q, \mathcal{X}, \mathbf{y}_Q)$. Then a desired ranking algorithm should be able to learn a model $\mathbf{w}$ that can successfully distinguish the correct ranking $\mathbf{y}_Q^*$ from any other incorrect ranking $\mathbf{y}_Q$. Following the maximum-margin-based formation, it can be approached by optimizing the following objective function:

$$\min_{\mathbf{w}} \{f(\mathbf{w}) + \frac{C}{|\mathcal{Q}|} \sum_{Q} \xi_Q\}, \tag{1}$$

$$s.t. \begin{array}{l} \langle \mathbf{w}, \psi\left(Q, \mathcal{X}, \mathbf{y}_Q^*\right)\rangle \geq \langle \mathbf{w}, \psi\left(Q, \mathcal{X}, \mathbf{y}_Q\right)\rangle + \Delta\left(\mathbf{y}_Q^*, \mathbf{y}_Q\right) - \xi_Q, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall Q \in \mathcal{Q}, \forall \mathbf{y}_Q \neq \mathbf{y}_Q^*; \\ \xi_Q \geq 0, \forall Q \in \mathcal{Q}. \end{array}$$

[1] To be easily understandable, in remainder of the paper "recognition" will simply mean "identification" though our model can also applies to object categorization.

[2] Classification models (e.g. Structured SVM) can also be adopted here. We choose ranking for a potential application of our model to search and retrieval applications.

where ξ_Q is the slack variable for Q and C is the trade-off parameter. $\Delta\left(\mathbf{y}_Q^*, \mathbf{y}_Q\right)$ is the loss function measuring the penalty of predicting $\mathbf{y}_Q$ instead of $\mathbf{y}_Q^*$. And in the test stage, given an input query Q, the model will infer the best ranking of $\mathcal{X}$ from the compatibility score:

$$\mathbf{y}_Q^* = \arg\max_{\mathbf{y}_Q \in \mathcal{Y}} \langle \mathbf{w}, \psi\left(Q, \mathcal{X}, \mathbf{y}_Q\right)\rangle. \tag{2}$$

We decompose the feature map by the so-called partial order features [15]:

$$\psi_{po}(Q, \mathcal{X}, \mathbf{y}_Q) = \sum_{i \in I_Q^+} \sum_{j \in I_Q^-} y_{ij} \left(\frac{\phi\left(Q, X_i\right) - \phi\left(Q, X_j\right)}{\left|I_Q^+\right| \cdot \left|I_Q^-\right|} \right), \tag{3}$$

where

$$y_{ij} = \begin{cases} 1 & X_i \prec_{\mathbf{y}_Q} X_j \\ -1 & X_i \succ_{\mathbf{y}_Q} X_j \end{cases},$$

and $\phi(Q, X_i)$ is a feature map characterizing the relationship between query set Q and gallery set X_i. Like what has been studied in single-instance-based ranking [16], it can be proved that such a definition has a very attractive property: given the model $\mathbf{w}$, the ranking $\mathbf{y}_Q^*$ that maximizes $\langle \mathbf{w}, \psi\left(Q, \mathcal{X}, \mathbf{y}_Q^*\right)\rangle$ is the one that sorts $\mathcal{X}$ by descending $\langle \mathbf{w}, \phi(Q, X_i)\rangle$, where X_i is an arbitrary set in $\mathcal{X}$. This property transfers the ranking problem to a weighted set-to-set similarity scoring problem.

Now the problem becomes how to define a proper set-to-set joint feature map $\phi(Q, X_i)$. It looks like an extension of the traditional relative feature map or "relative distance" as called in [17], however, for two sets which may contain different numbers of points it is nontrivial to design such a relative feature map as simple feature subtraction cannot be directly applied to point sets.

3.2 Set-to-Set Distance Metric

Let $W \succeq 0$ denote a symmetric, positive semi-definite matrix in $\mathbb{R}^{d \times d}$, then the distance between two arbitrary samples $\mathbf{x}_k$ and $\mathbf{x}_l$ in the d dimensional feature space under the metric defined by W can be denoted as $\|\mathbf{x}_k - \mathbf{x}_l\|_W = \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^T W (\mathbf{x}_k - \mathbf{x}_l)}$. To avoid the square root operation, we can just use the squared distance $\|\mathbf{x}_k - \mathbf{x}_l\|_W^2$ instead:

$$\begin{aligned} d_W(\mathbf{x}_k, \mathbf{x}_l) &\overset{\Delta}{=} (\mathbf{x}_k - \mathbf{x}_l)^T W (\mathbf{x}_k - \mathbf{x}_l) && (4) \\ &= tr(W(\mathbf{x}_k - \mathbf{x}_l)(\mathbf{x}_k - \mathbf{x}_l)^T) && (5) \\ &= \langle W, (\mathbf{x}_k - \mathbf{x}_l)(\mathbf{x}_k - \mathbf{x}_l)^T \rangle_F, && (6) \end{aligned}$$

where $tr(\cdot)$ means the trace of a matrix and $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product. Eqn. 6 presents a way to separate the metric matrix W from the operation of feature vectors. Therefore, as stated in [16], $(\mathbf{x}_k - \mathbf{x}_l)(\mathbf{x}_k - \mathbf{x}_l)^T$ can be used to define the relative feature map between $\mathbf{x}_k$ and $\mathbf{x}_l$:

$$\phi(\mathbf{x}_k, \mathbf{x}_l) \overset{\Delta}{=} -(\mathbf{x}_k - \mathbf{x}_l)(\mathbf{x}_k - \mathbf{x}_l)^T. \tag{7}$$

Therefore, if we can define our set-to-set relative feature map similar to that, then by replacing the vector style model parameter $\mathbf{w}$ with the metric matrix W and changing the normal inner product $\langle \mathbf{w}, \phi(Q, X_i)\rangle$ to the Frobenius inner product $\langle W, \phi(Q, X_i)\rangle_F$, the large-margin-based objective function for ranking will directly optimize the set-to-set distance metric. About the embedded function $f(W)$ in the objective function, there may be different options such as $tr(W)$, $\frac{1}{2}tr(W^TW)$, etc. $tr(W)$ is a good choice for us as it prefers sparse solutions.

Inspired by the closest points based set-to-set distance and the nearest neighbor based set-to-set matching, we define our set-to-set distance metric as follows:

$$d_W^{S\leftrightarrow S}(X_i, X_j) \overset{\Delta}{=} \min_{\mathbf{x}_k^i \in X_i, \mathbf{x}_l^j \in X_j} d_W(\mathbf{x}_k^i, \mathbf{x}_l^j) \tag{8}$$

$$= \min_{\mathbf{x}_k^i \in X_i, \mathbf{x}_l^j \in X_j} (\mathbf{x}_k^i - \mathbf{x}_l^j)^T W (\mathbf{x}_k^i - \mathbf{x}_l^j) \tag{9}$$

$$= \min_{\mathbf{x}_k^i \in X_i, \mathbf{x}_l^j \in X_j} \langle W, (\mathbf{x}_k^i - \mathbf{x}_l^j)(\mathbf{x}_k^i - \mathbf{x}_l^j)^T \rangle_F. \tag{10}$$

Then, we can choose

$$\langle W, \phi(Q, X_i)\rangle_F = -d_W^{S\leftrightarrow S}(Q, X_i) \tag{11}$$

for our ranking model. However, as minimization cannot be exchanged with the Frobenius inner product, we do not have an explicit form for $\phi(Q, X_i)$. Nevertheless, it doesn't influence the usage of distance metric in our model.

3.3 Geometric Distance between Convex Models

The definition of the set-to-set distance metric in Eqn. 9 has two limitations. On one side, it needs to compute the minimum pair-wise distance over two sets which is computationally expensive as it is a quadratic form of the number of set points. On the other side, the distance highly depends on the actual positions of sample points which indicates low generalization ability. Similar to the recently proposed approaches on set-based distance finding [1, 2], we break these two limitations by using convex approximations of the point sets and then measure their dissimilarity by the geometric distance between them. For example, when the affine hull is adopted for convex approximation, $d_W^{S\leftrightarrow S}(X_i, X_j)$ can be rewritten as the minimum distance between two affine hulls of X_i and X_j:

$$d_W^{S\leftrightarrow S}(X_i, X_j) = (X_i\boldsymbol{\alpha}_i^* - X_j\boldsymbol{\alpha}_j^*)^T W (X_i\boldsymbol{\alpha}_i^* - X_j\boldsymbol{\alpha}_j^*),$$

in which the linear combination coefficient vectors $\boldsymbol{\alpha}_i^*$ and $\boldsymbol{\alpha}_j^*$ can be found by

$$(\boldsymbol{\alpha}_i^*, \boldsymbol{\alpha}_j^*) = \arg\min_{\alpha_i, \boldsymbol{\alpha}_j} (X_i\boldsymbol{\alpha}_i - X_j\boldsymbol{\alpha}_j)^T W (X_i\boldsymbol{\alpha}_i - X_j\boldsymbol{\alpha}_j),$$

$$s.t. \sum_{k=1}^{n_i} \alpha_{ik} = 1 = \sum_{k'=1}^{n_j} \alpha_{jk'},$$

where n_i and n_j are the numbers of points in X_i and X_j, respectively.

Recall that the metric matrix W is positive semi-definite, so that we can perform eigen decomposition to it:

$$W = A\Lambda A^T = PP^T, P = A\Lambda^{\frac{1}{2}},$$

where the columns of A are eigenvectors of W and Λ is a diagonal matrix whose diagonals are the corresponding eigenvalues. Therefore, the geometric distance can be transformed to

$$d_W^{S\leftrightarrow S}(X_i, X_j) = \|P^T X_i \boldsymbol{\alpha}_i^* - P^T X_j \boldsymbol{\alpha}_j^*\|^2 = \|X_i^P \boldsymbol{\alpha}_i^* - X_j^P \boldsymbol{\alpha}_j^*\|^2, \qquad (12)$$

where $X_i^P = P^T X_i$ and $X_j^P = P^T X_j$ can be viewed as feature space projections of the original point sets X_i and X_j. After the projection by P, it becomes a traditional geometric distance between affine hulls. By bounding the coefficients α_{ik} and $\alpha_{jk'}$ within a predefined range $[L, U]$, we can constrain the hulls to be the reduced affine hulls or even convex hulls $(L = 0, U \geq 1)$, so that AHISD and CHISD methods [1] can be directly used here, along with their extended kernel versions. If we put sparsity constraints on $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$, then it becomes the SANP model. Therefore, both the two latest set-based distance finding models can be used in our set-to-set distance metric.

3.4 Simultaneous Optimization

As the above formulation shows, our set-based ranking model directly embeds the set-based geometric distance finding in the maximum-margin based model learning. Therefore, the optimization of the convex approximation coefficients $\boldsymbol{\alpha}$s and the optimization of the ranking model (namely W) are mutually dependent which demands simultaneous optimization. However, it is hard to directly optimize them in a single objective function because they are optimizing over different data (two sets vs. multiple sets) with different objectives and constraints. Therefore, we use an iterative algorithm in this work to do the simultaneous optimization.

The procedure of our Set-based Discriminative Ranking (SBDR) model is briefly described in Algorithm 1. The framework is based on the 1-Slack margin-rescaling cutting-plane algorithm as described in [18]. We have also extended the efficient computation strategy proposed in [16] for the implementation of our set-to-set distances $d_W^{S\leftrightarrow S}(Q_i, X_i)$ and adopting them for both constraint updating and metric optimization. Readers are referred to [16] for implementation details.

In the iteration, on one hand, the set-to-set distance finding generally decreases all the weighted joint features (i.e., the set-to-set distances), thus indirectly decrease ξ, resulting a reduction of the objective function value, while on the other hand, metric learning also approaches the global optimal by optimizing W. Therefore, it will finally converge. In our experiments to be presented as below, it always converged within 30 iterations.

Algorithm 1. Learning the SBDR model:

```
Require:
    Query set collection $\mathcal{Q}$, gallery set collection $\mathcal{X}$, desired rankings $\mathbf{y}_1^*, \ldots, \mathbf{y}_{|\mathcal{Q}|}^*$ of $\mathcal{X}$ for each query set, slack trade-off $C > 0$, termination threshold $\epsilon > 0$.
Ensure: Metric matrix $W$.
Initialize $W$ with a diagonal matrix whose diagonal values are standard deviations of the whole dataset on each feature dimension.
Initialize working set of constraints: $\mathcal{C} \leftarrow \emptyset$.
repeat
    Compute the currently optimal metric in Eqn. 1.
Compute set-to-set distances:
    $\forall Q_i \in \mathcal{Q}, \forall X_i \in \mathcal{X}, d_W^{S \leftrightarrow S}(Q_i, X_i)$.
Update the working set of constraints:
for $i = 1 \rightarrow |\mathcal{Q}|$ do
    $\mathbf{y}_{Q_i} \leftarrow \arg\max_{\mathbf{y} \in \mathcal{Y}} \left\{ \Delta\left(\mathbf{y}_{Q_i}^*, \mathbf{y}\right) + \langle W, \psi_{po}(Q_i, \mathcal{X}, \mathbf{y}) \rangle_F \right\}$
end for
$\mathcal{C} \leftarrow \mathcal{C} \cup \left\{ (\mathbf{y}_1, \ldots, \mathbf{y}_{|\mathcal{Q}|}) \right\}$.
until
    $$\frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \left\{ \Delta\left(\mathbf{y}_{Q_i}^*, \mathbf{y}_{Q_i}\right) - \left\langle W, \psi_{po}(Q_i, \mathcal{X}, \mathbf{y}_{Q_i}^*) - \psi_{po}(Q_i, \mathcal{X}, \mathbf{y}_{Q_i}) \right\rangle_F \right\} \leq \xi + \epsilon.$$
```

4 Experiments and Results

We evaluate the proposed approach "SBDR" with CHISD (linear) and SANP as its geometric distance finder (denoted by "SBDR$_{CHISD}$" and "SBDR$_{SANP}$" respectively) on two representative set-based recognition tasks: face recognition and person re-identification. They are not only important in many real applications, but are also representative of two typical cases for research: faces are relatively more rigid with less appearance variations but the between-class differences are also very subtle, while person re-identification has greater appearance variations (caused by both the object itself and its surrounding environment) which challenge both feature representation and the recognition model. Therefore, we present comparisons with state-of-the-art methods on widely used databases for these two different tasks, with experimental details and results presented separately for clarity.

4.1 Experiments on Face Recognition

The well-known Honda/UCSD [19] and the CMU MoBo [20] datasets are used for our experiments. The Honda/UCSD dataset was collected for video-based face recognition, containing 20 individuals in 59 video sequences. We use the version from [2] which has the faces detected, resized to gray-scale images of size 20×20 and histogram equalized. The length of each sequence varies from 13 to 782. Within each sequence there are large pose variations with moderate expression changes. The raw pixels of images were used as features. The CMU MoBo dataset was originally built for pose recognition, but recently the human faces have been detected and resized to 40×40 for face recognition. There are 24 individuals appearing in 96 sequences, which were captured from multiple cameras with four different walking styles: slow, fast, inclined, and carrying a ball. Each individual appears in 4 sequences which cover different walking styles.

We used the exact LBP features as adopted by CHISD and SANP for CMU MoBo dataset, and followed their original experimental setting: select one sequence for each individual as the gallery set and use the other 3 as query sets. The gallery-query splitting on Honda/UCSD data exactly follows [2], i.e., 20 sequences for query and the remaining 39 for gallery. Since the performance on the whole sequences of the Honda/UCSD dataset has got saturated (SANP claimed 100% accuracy), we sampled 50 or 100 frames per sequence instead. This strategy has also been applied to the CMU MoBo dataset for in-depth comparison.

The frame sampling strategy was originally proposed by [2], however, it simply treats the first 50/100 frames in the beginning of each sequence as testing samples, which may be improper for long sequences with probably more variations in the remaining frames. Therefore, we propose to experiment on two different types of sampling strategies to better show the properties of the data and recognition algorithms: **Type I** refers to sampling from the beginning of each sequence for testing and randomly sampling images from the rest for training; while **Type II** stands for doing random sampling for both training and testing without overlapping. For those short sequences which do not have enough frames, we sampled as many as we can while making the training and test datasets balanced. For each setting, the results are averaged over 10 trials to eliminate the uncertainty of random sampling.

As the recently proposed CHISD and SANP models present so far the best performance on these two datasets and our model can directly utilizes them, in this paper we just compare our "SBDR" model with these two approaches. We used $Prec@k$ with $k = 1$ as our loss function since the evaluation is on the recognition rate, and we set $C = 10$ without further tuning.

Table 1. Average recognition rate (%) on faces. The stars indicate that the results are different from the ones presented in [2] due to the small change of experimental settings. Detailed explanations are given in the text.

		50 frames		100 frames	
Dataset	Sampling Strategy	Type I	Type II	Type I	Type II
Honda/UCSD	CHISD*	80.51	93.85	78.97	94.02
	SBDR$_{CHISD}$	83.08	**96.41**	84.10	95.73
	SANP*	81.03	91.03	83.59	92.31
	SBDR$_{SANP}$	**87.69**	95.64	**89.23**	**97.95**
CMU MoBo	CHISD*	90.28	96.11	93.06	96.67
	SBDR$_{CHISD}$	93.89	97.78	95.56	98.61
	SANP*	88.89	93.06	93.06	96.67
	SBDR$_{SANP}$	**95.00**	**98.61**	**96.11**	**98.89**

Experimental results are shown in Table 1. They not only clearly demonstrate the superiority of the proposed SBDR approach but also provide important insights on how to use it properly (data sampling and set-distance finder selection). We explain the results from the following perspectives.

- ***SBDR significantly outperforms both CHISD and SANP, and SANP fits SBDR better than CHISD.*** All the results consistently show that embedding discriminative metric learning using the proposed SBDR model can significantly improve the performance of unsupervised set-based distance-finding methods (CHISD and SANP), and such an improvement is much greater over SANP. For almost all the cases (with only one exception), SBDR_{SANP} performs better than SBDR_{CHISD} even when SANP itself performs worse than CHISD. A positive reason for such an interesting phenomenon might be that SANP provides SBDR with between-set distances determined by sparse but representative samples which could be more informative for discriminative learning than those distances computed over all the samples in the CHISD model.
- ***Wider within-class variation coverage results in better recognition performance.*** For the same set size, random sampling is more likely results in larger within-set variations than sequential sampling from the beginning of the video sequences, which can be witnessed by the difference between training sets and testing set for sampling strategy Type I as presented in Fig. 2. As expected, experimental results clearly show that for all the methods the performance on randomly sampled testing sets is much better than that on sequentially sampled ones. This is more significant on the Honda/UCSD dataset, which may due to that the facial appearance changes more gradually and sequentially in the Honda/UCSD dataset than in the CMU Mobo dataset.
- ***CHISD vs. SANP: CHISD performs better on sparse data, but SANP exceeds as the within-set data density increases.*** For sampling strategy Type II, CHISD performs considerably better than SANP on both datasets when the set size equals 50, but the superiority is weakened or even eliminated when the set size goes up to 100. Such a change over the growing of set size can be observed for sampling strategy Type I as well. Results on the Honda/UCSD dataset demonstrate that more significantly, which coincides with the experimental results in [2], though the exact numbers there are different from the ones shown here due to two reasons: we have much smaller testing sets for persons with less than 50/100 images (in such cases only half of them are used for testing while the other half are reserved for training), and our results are averaged over 10 trails instead of only one trail. With the same set size, comparing to CHISD, SANP performs better on Type I than Type II, this is because Type I has smaller within-set variations so that the density of data is relatively higher than that of Type II.

Note that even with only 100 frames per set for testing, the recognition rate of SBDR_{SANP} has already passed the ones ever reported by other state-of-the-art methods (excluding CHISD and SBDR as they have already been compared

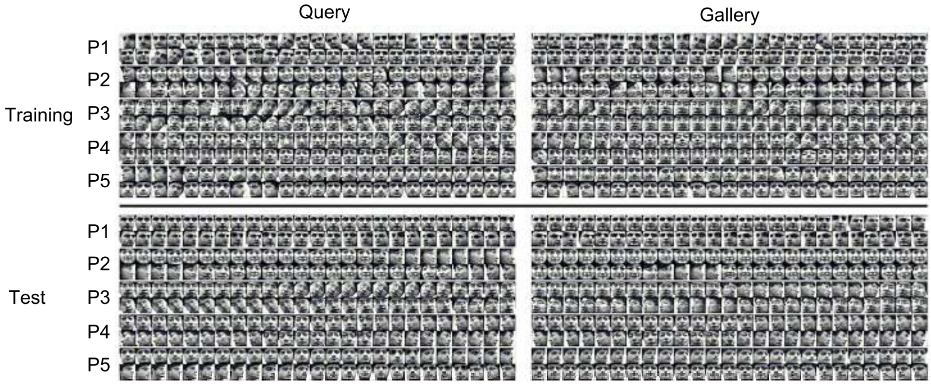


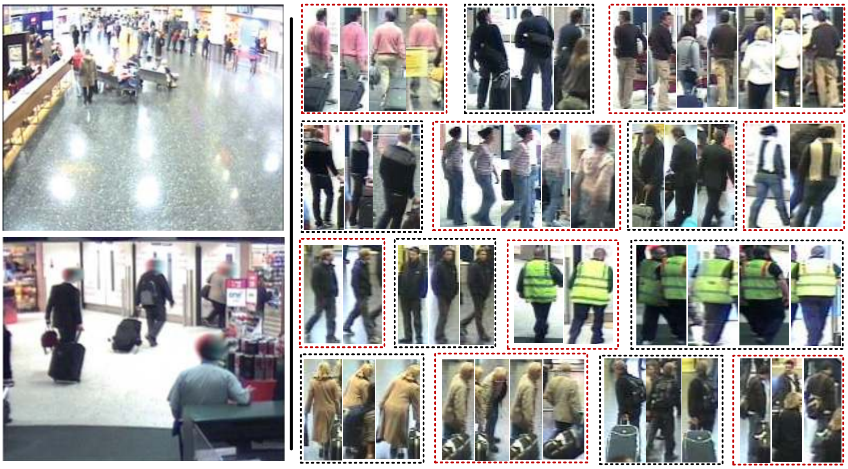
Fig. 2. Different sampling styles (random and sequential) result in different within-class variation coverage for each set. Without loss of generality, the training and testing sets for the first 5 persons (P1 to P5) in the Honda/UCSD dataset are presented using sampling strategy **Type I** with a set size of 50. Therefore, both the test sets for query and gallery are sequentially sampled from the beginning, while the training sets are randomly sampled from the remaining frames of each sequence. It can be clearly seen that the training sets have larger within-set variations.

with) using full length sequences. More concretely, the best ever reported rate on the Honda/UCSD dataset is 97.44% with averagely 267 frames per set, while that on the CMU Mobo dataset is 95.97% with averagely 496 frames per set [2].

4.2 Experiments on Person Re-identification

Datasets and Settings: There are two publicly available and commonly used datasets for set-based (or namely multiple-shot) person re-identification: the ETHZ dataset [21] and the i-LIDS dataset [22]. However, due to the recently saturated performance on the ETHZ dataset [4] which suggests its unsuitability for further comparison, we only experiment on the i-LIDS dataset in this paper.

The i-LIDS dataset for person re-identification was adjusted from the 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS), released by the Home Office of UK for the research on cross-camera human tracking. Since its introduction, it has been tested by almost all of the approaches proposed for person re-identification. It contains 476 images of 119 unique individuals. As the images are automatically extracted, the width-height-ratio varies and misalignment happens. Though it has an average of 4 images for each individual, the actual number varies from 2 to 8. This dataset is very challenging due to that it was collected by two non-overlapping cameras from two different view angles at a busy airport and there are many occlusions and truncations, as well as large illumination changes, as shown in Fig. 3. Since there are too few images for each person, we follow the setting in [17] to have only one image per person for gallery sets (actually not sets any more), and the other images for query sets.



(a) Two non-overlapping views

(b) Images of sampled individual persons

Fig. 3. The i-LIDS dataset for person re-identification. (a) shows the two non-overlapping views selected for the dataset. (b) lists the cropped images for several persons, of which the first two rows show the great intra-class variations including viewpoint, pose, illumination, occlusion and background changes, while the last two rows present the rather small within-class variations between some pairs of persons. Red/black bounding boxes are used to separate different persons.

The same color and texture features as used in [17] and [23] were adopted in our experiments.

Comparison with Existing Methods: We compare our approach with the state-of-the-art methods from each of the three groups mentioned in Section 2. Concretely, they are: MRCG [4] for set-based signature generation, SDALF [3] for direct set-to-set matching, CHISD (linear) for set-based distance finding. Note that the sparsity-based SANP model is not suitable here due to the extremely small set size ("N=1" for gallery sets and averagely "N=3" for query sets). They represent the latest and so far the most powerful methods in each specific group. Unfortunately these three groups of methods are all unsupervised and as far as we are aware there is no existing work on using supervised learning to directly optimizes multiple-shot (i.e. set-based) re-identification. Therefore, to show the advantage of set-based recognition we also present here the best results from two recently proposed single image based classification and ranking algorithms: RankSVM [24] and PRDC [17], with exactly the same feature representation and experimental setting as ours. For the loss function in our model, we just used the Mean Reciprocal Rank (MRR) as in [14]. The trade-off parameter C was set to 10 without cross-validation. Since we randomly sampled the data for training and test, all the experiments are repeated 10 times for averaging.

Results and Analysis: As shown in Table 2, using simple features for representation and training with about the same amount of data as that for testing

Table 2. Top ranked matching rate (%) on i-LIDS dataset, where "r" means rank

	Unsupervised		$p = 50$			
Methods	SDALF	MRCG	RankSVM	PRDC	CHISD	SBDR$_{CHISD}$
r = 1	34.96	45.80	37.41	37.83	41.80	**46.60**
r = 5	60.92	66.81	63.02	63.70	68.80	**71.60**
r = 10	73.36	75.21	73.50	75.09	**83.60**	80.40
r = 20	83.78	83.61	88.30	88.35	**94.40**	90.40

(the number of individuals for testing $p = 50$), SBDR$_{CHISD}$ performs significantly better than SDALF and MRCG which have paid great efforts on feature design. For comparing with other learning-based methods, we choose different training-test-ratios by changing p following PRDC [17]. Though in that paper PRDC was claimed to be less overfitting than other learning based approaches, the experimental results presented in Table 3 show that our proposed SBDR model is even more promising when the training set is small (decreasing less as the training set shrinks). Overall, SBDR$_{CHISD}$ can get a better performance (lower ranks matter more) than CHISD, indicating that SBDR$_{CHISD}$ learns a better feature space for set-based distance metric than the original space.

Table 3. Top ranked matching rate (%) on i-LIDS dataset, where "r" means rank

	$p = 30$				$p = 80$			
Methods	RankSVM	PRDC	CHISD	SBDR$_{CHISD}$	RankSVM	PRDC	CHISD	SBDR$_{CHISD}$
r = 1	42.96	44.05	50.67	**53.67**	31.73	32.60	**37.87**	37.75
r = 5	71.30	72.74	76.00	**79.00**	55.69	54.55	61.75	**64.13**
r = 10	85.15	84.69	**90.00**	88.67	67.02	65.89	75.37	**76.38**
r = 20	96.99	96.29	**98.67**	97.67	77.78	78.30	84.13	**84.63**

5 Conclusions and Future Work

In this paper we presents a novel model called "set based discriminative ranking" for set-based recognition. It simultaneously optimizes the set-to-set geometric distance finding and the feature space projection, resulting in a discriminative set-distance-based model. As far as we are aware, this is the first time a global optimal learning-based model is proposed for solving this challenging problem. We demonstrate its superiority by comparing with the state-of-the-art methods on two representative object recognition tasks: face recognition and person re-identification. Since our model is a general approach which makes no assumptions on the data, future work can be done on applying it to other object recognition tasks, along with other related problems like image search and retrieval.

References

1. Cevikalp, H., Triggs, B.: Face recognition based on image sets. In: CVPR, pp. 2567–2573 (2010)
2. Hu, Y., Mian, A.S., Owens, R.: Sparse Approximated Nearest Points for Image Set Classification. In: CVPR, 121–128 (2011)

3. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR (2010)
4. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid. In: Advanced Video and Signal-Based Surveillance (2011)
5. Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R., Darrell, T.: Face recognition with image sets using manifold density divergence. In: CVPR, pp. 581–588 (2005)
6. Yamaguchi, O., Fukui, K., Maeda, K.: Face recognition using temporal image sequence. In: IEEE International Conference on Automatic Face and Gesture Recognition, p. 318 (1998)
7. Li, X., Fukui, K., Zheng, N.: Image-Set Based Face Recognition Using Boosted Global and Local Principal Angles. In: Zha, H., Taniguchi, R.-i., Maybank, S. (eds.) ACCV 2009, Part I. LNCS, vol. 5994, pp. 323–332. Springer, Heidelberg (2010)
8. Kim, T.K., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE Trans. Pattern Anal. Mach. Intell. 29, 1005–1018 (2007)
9. Wang, R., Shan, S., Chen, X., Gao, W.: Manifold-manifold distance with application to face recognition based on image set. In: CVPR (2008)
10. Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V.: Multiple-shot person re-identification by hpe signature. In: ICPR, pp. 1413–1416 (2010)
11. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: CVPR, pp. 1528–1535 (2006)
12. Tuzel, O., Porikli, F., Meer, P.: Region Covariance: A Fast Descriptor for Detection and Classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part II. LNCS, vol. 3952, pp. 589–600. Springer, Heidelberg (2006)
13. Prosser, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: BMVC (2010)
14. Wu, Y., Mukunoki, M., Funatomi, T., Minoh, M., Lao, S.: Optimizing Mean Reciprocal Rank for Person Re-identification. In: 2011 8th IEEE International Conference on Advanced Video and Signal-Based Surveillancei (AVSS), p. 6 (2011)
15. Joachims, T.: A support vector method for multivariate performance measures. In: ICML, pp. 377–384 (2005)
16. McFee, B., Lanckriet, G.: Metric learning to rank. In: ICML, pp. 775–782 (2010)
17. Zheng, W.-S., Gong, S., Xiang, T.: Person Re-identification by Probabilistic Relative Distance Comparison. In: CVPR, pp. 649–656 (2011)
18. Joachims, T., Finley, T., Yu, C.N.: Cutting-plane training of structural svms. Machine Learning 77, 27–59 (2009)
19. Lee, K., Ho, J., Yang, M.H., Kriegman, D.: Video-based face recognition using probabilistic appearance manifolds. In: CVPR, pp. 313–320 (2003)
20. Gross, R., Shi, J.: The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (2001)
21. Schwartz, W.R., Davis, L.S.: Learning discriminative appearance-based models using partial least squares. In: SIBGRAPI, pp. 322–329 (2009)
22. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: BMVC (2009)
23. Gray, D., Tao, H.: Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008)
24. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 133–142 (2002)