Collaboratively Regularized Nearest Points for Set Based Recognition

Yang Wu
mm.media.kyoto-u.ac.jp/members/yangwu
Michihiko Minoh
mm.media.kyoto-u.ac.jp/members/minoh
Masayuki Mukunoki
mm.media.kyoto-u.ac.jp/members/mukunoki

Academic Center for Computing and Media Studies
Kyoto University
Kyoto, 606-8501, Japan

Abstract

Set based recognition has been attracting more and more attention in recent years, benefitting from two facts: the difficulty of collecting sets of images for recognition fades quickly, and set based recognition models generally outperform the ones for single instance based recognition. In this paper, we propose a novel model called collaboratively regularized nearest points (CRNP) for solving this problem. The proposal inherits the merits of simplicity, robustness, and high-efficiency from the very recently introduced regularized nearest points (RNP) method on finding the set-to-set distance using the l_2-norm regularized affine hulls. Meanwhile, CRNP makes use of the powerful discriminative ability induced by collaborative representation, following the same idea as that in sparse recognition for classification (SRC) for image-based recognition and collaborative sparse approximation (CSA) for set-based recognition. However, CRNP uses l_2-norm instead of the expensive l_1-norm for coefficients regularization, which makes it much more efficient. Extensive experiments on five benchmark datasets for face recognition and person re-identification demonstrate that CRNP is not only more effective but also significantly faster than other state-of-the-art methods, including RNP and CSA.

1 Introduction

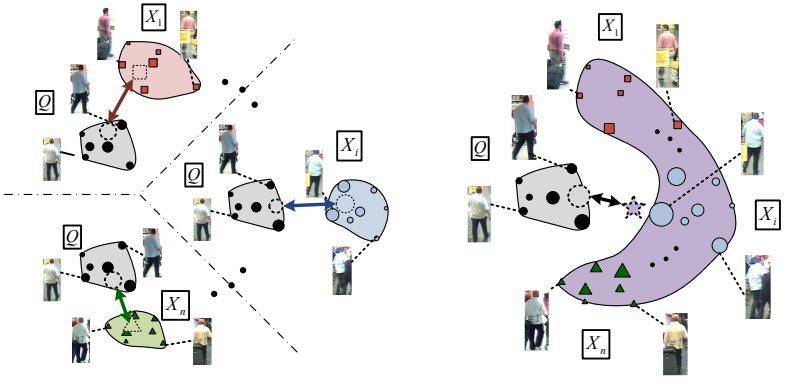
As taking pictures/videos and sharing them over networks get easier and more popularized, collecting a set of images for recognizing an object category/instance becomes increasingly convenient. Therefore, the direction of using a set of instances/images together for recognition, namely set based recognition [10], has got rapidly growing attention in recent years. Though most of the methods proposed for solving this problem have been only tested on face recognition [2, 10, 12, 13] and person re-identification [9, 10], they are generally applicable to any recognition tasks. Set based recognition models have the potential to outperform single instance based recognition approaches under the same conditions, due to that more instances for testing generally means a higher probability to extract discriminative information for correct classification. However, the performance largely depends on the detailed design of the model.

In the past few years various approaches have been proposed, which were reviewed in [10] and [12]. In this paper, we are focusing on the type of non-parametric set-to-set distance finding approaches, which have shown superior performance when compared with other approaches. A simple model using the minimum point-wise distance (referred to as "MPD" for short) as the set-to-set distance has been applied for person re-identification in [4] and got impressive results. However, MPD is sensitive to noises and outliers as the set-to-set distance only depends on a single point from each set. To overcome this issue, AHISD/CHISD (Affine/Convex Hull based Image Set Distance) [2] computes an affine/convex hull for each set and then treats the geometric distance (distance of closest approach) between two hulls as the set-to-set distance. Since such a distance can be viewed as the distance between two virtual points generated from each of the two sets by linear combinations, it has some robustness to noises and outliers.

Later on, SANP (Sparse Approximated Nearest Points) [13] extended the model of CHISD by enforcing the sparsity of samples used for linear combination, and showed better performance on face recognition than CHISD. Similar to CHISD, there are also kernel versions of SANP (KSANP) as presented in [6], which performed slightly better than SANP. The work of SBDR (Set Based Discriminative Ranking for Recognition) [10] opens a door towards integrating the power of unsupervised set-to-set distance finding models with the discriminative ability of supervised learning based approaches. Due to the difficulty of optimizing these two objectives in one step, SBDR adopts an iterative strategy to alternate set-to-set distance finding (either CHISD or SANP) and discriminative metric learning. Though both SANP and SBDR (with SANP) have shown quite impressive results, they have a serious drawback of being time-consuming due to the l_1-norm based sparsity term in their model. Worse than SANP, SBDR needs an additional training stage which takes much time, as shown in the experiments of this paper.

Very recently, the work of regularized nearest points (RNP) [12] has pointed out that there is no need to use the complex formulation with many parameters and variables and the sparse constraint on the representation coefficients as adopted by SANP. Instead, RNP just uses the affine hull formulation (as in AHISD [2]) regularized by the l_2-norm term, resulting in fewer parameters and variables. More attractively, using the l_2-norm makes RNP have a closed-form solution, thus being much more efficient than SANP. Though being simple, RNP performs even better than SANP on face recognition tasks.

Despite their differences in modeling, all of these approaches share the same idea of using the independent set-to-set distances between an arbitrary test/query set $\mathbf{Q}$ and each of the training/gallery sets $\mathbf{X}_i, i \in \{1,\ldots,n\}$ directly for classification, namely, doing nearest-neighbor classification based on these set-level distances. The model of CSA (Collaborative Sparse Approximation) [9], however, explores a new classification model inspired by the success of sparse representation for classification (SRC) [8]. More concretely, CSA finds the set-to-set distance between $\mathbf{Q}$ and all $\mathbf{X}_i$s together (treating them as a large single set), and then it classifies $\mathbf{Q}$ by the reconstruction residuals using only individual gallery sets (i.e. $\mathbf{X}_i$s) and their corresponding coefficients. The collaborative manner in the distance finding coincides with the underlying power of SRC as they both use all the training data to reconstruct a test set/sample. It tends to assign higher weights to the samples belonging to the same class as the test set/sample, because statistically those samples should be closer to the test set/sample than any others in a reasonably good feature space. Therefore, by identifying the set of samples which have higher coefficients and a lower reconstruction residual, we are likely able to get the correct class label for the test set/sample. The differences between the independent distance finding approaches and the collaborative distance finding approach



(a) Set-to-set distances (b) Set-to-sets distance

Figure 1: Independent distance finding v.s. collaborative distance finding, using the problem of person re-identification as an example (other tasks shall look the same). (a) The set-to-set distances generated by traditional independent distance finding approaches; (b) the set-to-sets distance got by collaborative distance finding methods.

CSA are illustrated in Figure 1. For set-to-sets distance finding, CSA utilizes the SANP model as it has the same l_1-norm based constraint on the coefficients as that in the SRC model. Since it only needs to compute the set-to-sets distance once, it is usually more efficient than the SANP approach itself which has to compute n individual set-to-set distances.

In this paper, we propose a novel collaborative distance finding approach called Collaboratively Regularized Nearest Points (CRNP). Unlike CSA, it uses RNP for set-to-sets distance finding, which no longer ensures the sparsity of the coefficients as SANP does, thus being much simpler and computational more efficient than CSA. Compared with RNP, CRNP needs only one-round set-level distance finding so it could be much faster than RNP. Moreover, CRNP enables introducing the discriminative class-specific (or set-specific) coefficients generated by collaborative reconstruction into its classification model, resulting in a further boosting of the performance. The effectiveness of using l_2-norm instead of l_1-norm for classification based on collaborative reconstruction has also been witnessed in the work of CRC (Collaborative Representation for Classification) [14]. However, the proposed CRNP is to the best of our knowledge the first one that deals with set based recognition, and as a set-based model using regularized affine hulls, it is quite different from simply extending CRC to simultaneously handle a set of testing samples, which will be demonstrated in our experiments.

The rest parts of the paper are organized as follows. Section 2 briefly reviews the RNP model, which leads to our proposed model of CRNP to be detailed in section 3. Section 4 shows the experimental results on both face recognition and person re-identification using 5 benchmark datasets. Conclusions and future work are given in section 5.

2 Regularized Nearest Points

RNP models each image set by a regularized affine hull (RAH) formulated as:

$$RAH = \left\{ \mathbf{x} = \mathbf{X}_i \boldsymbol{\alpha} \middle| \sum\nolimits_k \alpha_k = 1, \|\boldsymbol{\alpha}\|_2 \leq \sigma \right\}, \tag{1}$$

where the matrix $\mathbf{X}_i$ is a collection of the samples belong to set i with each column denoting the feature vector of an image and the vector $\boldsymbol{\alpha}$ denotes the combination coefficients. Note that in [12], the regularizer of RAH is defined using a l_p-norm to make the model as general as possible, but only $p = 2$ has been implemented for RNP. Therefore, here we directly use l_2 and keep it consistent throughout the paper. The difference between RAH and AHISD is the additional l_2-norm regularization of $\boldsymbol{\alpha}$, which makes the affine combination only focus on the samples close to the sample mean of the set.

For a given test/query set $\mathbf{Q}$ and a training/gallery set $\mathbf{X}_i$, RNP finds two nearest points from the RAH of $\mathbf{Q}$ and the RAH of $\mathbf{X}_i$, respectively, by solving the following optimization problem:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \|\mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}_i\boldsymbol{\beta}\|_2^2, \quad s.t. \ \sum_k \alpha_k = 1, \sum_j \beta_j = 1, \|\boldsymbol{\alpha}\|_2 \leq \sigma_1, \|\boldsymbol{\beta}\|_2 \leq \sigma_2, \tag{2}$$

whose dual problem is

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ \|\mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}_i\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\}, \quad s.t. \ \sum_k \alpha_k = 1, \sum_j \beta_j = 1, \tag{3}$$

where λ_1 and λ_2 are Lagrangian multipliers, and the affine hull constraints (i.e., $\sum_k \alpha_k = 1, \sum_j \beta_j = 1$) help avoiding the trivial solution ($\boldsymbol{\alpha} = \boldsymbol{\beta} = 0$).

After getting the solution $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ of Equation 3, the set-to-set distance between $\mathbf{Q}$ and $\mathbf{X}_i$ is defined to be

$$d_{RNP}^i = (\|\mathbf{Q}\|_* + \|\mathbf{X}_i\|_*) \cdot \left\|\mathbf{Q}\boldsymbol{\alpha}^* - \mathbf{X}_i\boldsymbol{\beta}^*\right\|_2^2, \tag{4}$$

where $\|\mathbf{Q}\|_*$ is the nuclear norm of $\mathbf{Q}$, i.e. the sum of the singular values of $\mathbf{Q}$. The nuclear norm term reflects the representation ability of a set, and it can remove the possible disturbance unrelated to the class information, to avoid biasing on large sets.

Finally, $\mathbf{Q}$ is classified by

$$C(\mathbf{Q}) = \arg\min_i \left\{ d_{RNP}^i \right\}. \tag{5}$$

3 Collaboratively Regularized Nearest Points

The proposed CRNP model performs collaborative distance finding using all the training/gallery sets instead of the independent set-by-set distance finding in RNP, so the distance finding model and its optimization change according. Besides that, CRNP has a different classification model which makes use of the discriminative coefficients generated by collaborative distance finding. The details of these three aspects are given as follows.

3.1 Collaborative distance finding

Given the test/query set $\mathbf{Q}$ and all the training/gallery sets $\mathbf{X}_i, i \in \{1, \ldots, n\}$, CRNP solves the following optimization problem:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ \|\mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\}, \quad s.t. \ \sum_k \alpha_k = 1, \sum_{i=1}^n \sum_j \beta_i^j = 1, \tag{6}$$

where $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]$ denotes all the training/gallery sets together; $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_n^T]^T$ are the corresponding coefficients for these sets; λ_1 and λ_2 are trade-off parameters.

This problem inherits the distance finding model from RNP, however, the small change of replacing $\mathbf{X}_i$ in Equation 3 by $\mathbf{X}$ causes a big change from independent distance finding to collaborative distance finding, which may lead to quite different classification results, as illustrated in Figure 1 and demonstrated in Section 4.

3.2 Distance finding optimization

The optimization problem 6 with equality constraints can be transformed to the following unconstrained optimization problem:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ \|\mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \gamma_1 \left(1 - \sum_k \alpha_k\right)^2 + \gamma_2 \left(1 - \sum_{i=1}^n \sum_j \beta_i^j\right)^2 \right\}, \tag{7}$$

where γ_1 and γ_2 are Lagrangian multipliers. It can be rewritten into a simpler form as

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ \left\|\mathbf{z} - \hat{\mathbf{Q}}\boldsymbol{\alpha} - \hat{\mathbf{X}}\boldsymbol{\beta}\right\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\}, \tag{8}$$

where $\mathbf{z} = [\mathbf{0}_{1,m}, \sqrt{\gamma_1}, \sqrt{\gamma_2}]^T$ with m denoting the dimensionality of the image feature space, $\hat{\mathbf{Q}} = [\mathbf{Q}^T, \sqrt{\gamma_1}\mathbf{1}_{N_q,1}, \mathbf{0}_{N_q,1}]^T$ with N_q denoting the number of samples in $\mathbf{Q}$, and $\hat{\mathbf{X}} = [-\mathbf{X}^T, \mathbf{0}_{N_x,1}, \sqrt{\gamma_2}\mathbf{1}_{N_x,1}]^T$ with N_x denoting the number of samples in $\mathbf{X}$. $\mathbf{0}_{i,j}$ and $\mathbf{1}_{i,j}$ denote the $i \times j$ zero matrix and the $i \times j$ dimensional matrix of ones, respectively.

Though the above problem has a closed-form solution, we follow [12] on alternatively optimizing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which avoids the time-consuming matrix inverse operation of an integrated matrix containing both $\mathbf{Q}$ and $\mathbf{X}$ for each test/query set $\mathbf{Q}$. In the alternative optimization of problem 8, however, the matrix inverse operation on the training data $\mathbf{X}$ is independent of $\mathbf{Q}$, so it can be pre-computed before testing.

More concretely, when $\boldsymbol{\alpha}$ is fixed, $\boldsymbol{\beta}$ has a closed-form solution

$$\boldsymbol{\beta}^* = \mathbf{P}_x \left(\mathbf{z} - \hat{\mathbf{Q}}\boldsymbol{\alpha}\right), \tag{9}$$

where

$$\mathbf{P}_x = \left(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \lambda_2\mathbf{I}\right)^{-1}\hat{\mathbf{X}}^T \tag{10}$$

(with $\mathbf{I}$ denoting the identity matrix) only depends on $\mathbf{X}$, so it can be pre-computed. When $\boldsymbol{\beta}$ is fixed, $\boldsymbol{\alpha}$ also has a closed-form solution

$$\boldsymbol{\alpha}^* = \mathbf{P}_q \left(\mathbf{z} - \hat{\mathbf{X}}\boldsymbol{\beta}\right), \tag{11}$$

where $\mathbf{P}_q = \left(\hat{\mathbf{Q}}^T\hat{\mathbf{Q}} + \lambda_1\mathbf{I}\right)^{-1}\hat{\mathbf{Q}}^T$, with $\mathbf{I}$ denoting the identity matrix.

The algorithm of CRNP for computing the nearest points from both training and test sets is summarized in Algorithm 1. As claimed in [12], the objective function in Formula 8 has a lower bound of 0 and it is jointly convex w.r.t. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Since in the alternative optimization, each step on updating $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ decreases the objective, the iteration will converge to the global optimal solution. In our experiments to be presented, the iteration usually terminates in no more than 10 steps.

Algorithm 1 COLLABORATIVELY REGULARIZED NEAREST POINTS (CRNP):

Require: The training/gallery sets $\mathbf{X} \in \mathbb{R}^{m \times N_x}$, an arbitrary test/query set $\mathbf{Q} \in \mathbb{R}^{m \times N_q}$, the pre-computed $\mathbf{z}$, $\hat{\mathbf{X}}$ and $\mathbf{P}_x$ (using Equation 10), and four trade-off parameters $\{\lambda_1, \lambda_2, \gamma_1, \gamma_2\}$.

Ensure: The representation coefficients for distance finding: $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$.

1: Construct $\hat{\mathbf{Q}} = [\mathbf{Q}^T, \sqrt{\gamma_1}\mathbf{1}_{N_q,1}, \mathbf{0}_{N_q,1}]^T$.
2: Compute the project matrix $\mathbf{P}_q = (\hat{\mathbf{Q}}^T\hat{\mathbf{Q}} + \lambda_1\mathbf{I})^{-1}\hat{\mathbf{Q}}^T$.
3: Initialize $\boldsymbol{\beta}_0 = 1/N_x$.
4: **while** not converged **or** not exceeding the maximum number of iterations **do**
5: Update the representation coefficients:
6: $\boldsymbol{\alpha}_{t+1} = \mathbf{P}_q(\mathbf{z} - \hat{\mathbf{X}}\boldsymbol{\beta}_t)$.
7: $\boldsymbol{\beta}_{t+1} = \mathbf{P}_x(\mathbf{z} - \hat{\mathbf{Q}}\boldsymbol{\alpha}_{t+1})$;
8: **end while**
9: Return $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$.

3.3 Classification

The collaborative distance finding in CRNP implicitly makes $\boldsymbol{\beta}^* = [\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_n^*]$ discriminative. Therefore, we define the dissimilarity between $\mathbf{Q}$ and $\mathbf{X}_i, i \in \{1, \ldots, n\}$ as

$$d_{CRNP}^i = (\|\mathbf{Q}\|_* + \|\mathbf{X}_i\|_*) \cdot \left\|\mathbf{Q}\boldsymbol{\alpha}^* - \mathbf{X}_i\boldsymbol{\beta}_i^*\right\|_2^2 / \left\|\boldsymbol{\beta}_i^*\right\|_2^2, \tag{12}$$

where $\|\mathbf{Q}\|_*$ is the nuclear norm of $\mathbf{Q}$, i.e. the sum of the singular values of $\mathbf{Q}$. Then, $\mathbf{Q}$ is classified by

$$C(\mathbf{Q}) = \arg\min_i \left\{d_{CRNP}^i\right\}. \tag{13}$$

Compared with d_{RNP}^i, d_{CRNP}^i benefits from the discriminative power of $\boldsymbol{\beta}^*$, which tends to make the class-specific reconstruction residual $\left\|\mathbf{Q}\boldsymbol{\alpha}^* - \mathbf{X}_i\boldsymbol{\beta}_i^*\right\|_2^2$ smaller and $\left\|\boldsymbol{\beta}_i^*\right\|_2^2$ larger for the ground-truth label i than any other labels $j \in \{1, \ldots, n\}, j \neq i$. The effectiveness of CRNP's classification model will be demonstrated in the following section.

4 Experiments and Results

To make the comparison with the state-of-the-art approaches for set based recognition fair and sufficient, we perform experiments on both face recognition and person re-identification, and report results of all the related methods when applicable. Detailed experimental settings are given in section 4.1, followed by the result comparison and analysis shown in section 4.2. Finally, we report the computational cost for the compared methods in section 4.3.

4.1 Experimental settings

Datasets. For face recognition, we follow most of the related approaches on using the Honda/UCSD dataset [7] and the CMU MoBo dataset [5] with the same settings of using the first 50 or 100 frames for recognition as in [13], [10] and [12]. The sizes of the images and the feature representations exactly follow [10] (namely, raw pixels for Honda/UCSD dataset and LBP features for CMU MoBo dataset). Following the common settings, for

Honda/UCSD dataset, the specified 20 sequences of 20 subjects are used for training/gallery and the rest 39 sequences are left for testing/querying; for CMU MoBo dataset, we perform a 10-time cross-validation by randomly choosing one sequence from the four candidates for each of the 24 subjects and have the unselected sequences left for testing. The properties of the these two datasets can be found in [10] and [12].

For person re-identification, we follow [9] on using the iLIDS-MA and iLIDS-AA datasets [1] for evaluation, along with an additional dataset CAVIAR4REID [3]. These datasets have interesting and complementary properties, thus being informative for comparison. "iLIDS-MA" and "iLIDS-AA" contain multiple images (92 in iLIDS-MA, and 21 to 243 in iLIDS-AA) for each human individual captured by two non-overlapping cameras (camera 1 and camera 3 in their original setting) at an airport with large viewpoint changes. The iLIDS-MA dataset has 40 persons with manually cropped images, while the iLIDS-AA dataset contains as many as 100 individuals collected by an automatic tracking algorithm. Therefore, compared with iLIDS-MA, iLIDS-AA is not only larger, but also noisier with localization errors and unequal class sizes. For both datasets, we perform 10-time cross-validation by randomly choosing 10 images for each gallery/query set. Unlike iLIDS-MA and iLIDS-AA, CAVIAR4REID consists of several sequences filmed in a shopping centre. Besides viewpoint changes, it has broad resolution changes and severe pose variations. We follow [3] on training with 22 specified subjects and testing on the other 50 subjects. Each set (either for gallery or query) contains 5 randomly sampled images, and 10-time cross-validation is performed for result averaging. We use the same 400-dimensional color and texture histograms based features as mentioned in [11] for all the three datasets.

Methods for comparison. We compare CRNP with all the related state-of-the-arts methods as introduced in section 1, including MPD[4], SRC[8], CRC[14], CHISD[2], SANP[13], KSANP[6], SBDR[10], CSA[9] and RNP[12]. Note that SRC and CRC in this paper stand for the extended versions from their original models. We simply replace the single test image in these two models by a set of images to do an overall reconstruction and classification. For KSANP and SBDR, we only report the results that have been listed for the same tasks in their original publications, while for the others, we have their codes either from their authors (such as SRC, CHISD, SANP, and CSA) or implemented by ourselves (such as MPD, CRC and RNP) run on the same data splits using the same features.

Parameters. For CRNP, we have $\lambda_1 = \lambda_2 = 4, \gamma_1 = \gamma_2 = 1$ fixed for all the experiments. For the other methods, we used the recommended parameters given in their original papers.

4.2 Experimental results and analysis

Honda/UCSD dataset. The results for all the concerned methods are listed in Table 1 for both the 50 frames/set and 100 frames/set settings. Besides the referred results for KSANP and SBDR, we also refer the originally reported results for some other methods when they are significantly different from their results in our experiments. It is clear that the proposed CRNP greatly outperforms all the other methods (over 2.5%). For both experimental settings, SRC, KSANP and RNP are relatively better than the others.

CMU MoBo dataset. The results for the CMU MoBo dataset are shown in Table 2. Unfortunately, the referred results for SBDR were generated with the training set fixed to be the frontal sequence, which is unlike the completely random sequence sampling in this paper and also the other papers. Therefore, they are not counted for competing the best performance. In can be seen that CRNP still significantly outperforms all the others when 50 frames/set is used, while it is only slightly worse than RNP when 100 frames are available for each set.

Table 1: Face recognition accuracy (%) comparison on the Honda/UCSD dataset. The results with stars are directly copied from their original papers for reference, while those without stars are got from our experiments. The best results are shown in bold.

Method	MPD[4]	SRC[8]	CRC[14]	CHISD[2]	SANP[13]	KSANP[6]	SBDR[10]	CSA[9]	RNP[12]	CRNP
50 frames	79.49	76.92	76.92	79.49/82.05*	84.62/84.62*	87.18*	87.69*	84.62	66.67/87.18*	**89.74**
100 frames	87.18	94.87	82.05	79.49/84.62*	89.74/92.31*	94.87*	89.23*	92.31	92.31/94.87*	**97.44**

Table 2: Face recognition accuracy (%) comparison on the CMU MoBo dataset. The results with stars are directly copied from their original papers for reference, while those without stars are got from our experiments. All these results are averaged over 10 random trials, with the best ones shown in bold. Note that the referred results for SBDR were generated with a different experimental setting, so they are not counted for performance ranking.

Method	MPD[4]	SRC[8]	CRC[14]	CHISD[2]	SANP[13]	SBDR[10]	CSA[9]	RNP[12]	CRNP
50 frames	92.22	88.89	89.72	90.83	90.14	*95.00**	86.25	91.81/91.9*	**93.33**
100 frames	94.31	92.36	93.06	94.17	93.61	*96.11**	94.44	**94.58/94.7***	94.44

It indicates that too large set size may weaken the discriminative power of the collaborative distance finding as it is more likely that some samples from different classes may confuse the correct class in collaborative representation. Even though, such a risk is still very low, which demonstrates the robustness of CRNP.

Person re-identification datasets. For person re-identification, since it is widely treated as a ranking problem and people usually care about the recognition accuracy in the top few ranks, we report both the rank-1 accuracy and the rank top 10% accuracy for a considerate comparison. As Table 3 shows, CRNP has significant superiority in both measures comparing with the others, especially on the most challenging CAVIAR4REID dataset. The results of CRNP on CAVIAR4REID also greatly outperform the state-of-the-art results in [3].

Table 3: Performance comparison for person re-identification with all the related methods on three benchmark datasets. Both the rank-1 accuracy and the rank top 10% accuracy (shown in parenthesis). The results are averaged over 10 random trials, with the best ones marked in bold.

Dataset	MPD[4]	SRC[8]	CRC[14]	CHISD[2]	SANP[13]	CSA[9]	RNP[12]	CRNP
iLIDS-MA	50.0(75.0)	57.3(74.8)	28.5(50.0)	52.5(72.8)	46.8(74.8)	**59.0**(71.3)	53.3(76.0)	**59.0**(**78.3**)
iLIDS-AA	23.8(60.4)	**36.0**(68.9)	24.7(54.1)	24.6(58.2)	19.2(57.3)	22.5(59.6)	25.5(59.9)	35.4(**71.6**)
CAVIAR4REID	19.0(47.2)	25.4(50.8)	16.6(37.6)	25.4(51.2)	25.2(52.4)	24.6(48.8)	24.0(50.2)	**26.8**(**63.6**)

4.3 Computational cost

In addition to accuracy, we also evaluate the efficiency of CRNP, in comparison with the others. All the methods compared are implemented in Matlab. We implemented MPD, CRC, RNP and CRNP by ourselves without specific code optimization. All the experiments were conducted on a 2.67 GHz dual-core machine with 20GB memory (actually no more than 2GB were used). Since some of the methods can have (parts of) their models pre-computed before testing, we report the pre-computation time for them in Table 4. The results show that the training phase of SBDR is very time-consuming, while the pre-computation time for other three methods are ignorable. In greater detail, CRNP needs a bit more time than RNP, but less time than CSA.

Table 4: For those methods which can have (parts of) their models pre-computed using the training data, the total pre-computation time (in seconds) is listed for comparison.

Dataset	Honda/UCSD		CMU MoBo		iLIDS-MA	iLIDS-AA	CAVIAR4REID
	50 frames	100 frames	50 frames	100 frames			
SBDR[10]	9.23×10^3	1.46×10^4	1.23×10^4	3.14×10^4	N/A	N/A	N/A
CSA[9]	0.59	0.74	28.7	50.2	0.39	0.62	0.26
RNP[12]	0.06	0.20	0.17	0.64	0.02	0.05	0.02
CRNP	0.22	0.87	0.64	2.66	0.04	0.22	0.02

Table 5: Computational cost comparison with all the related methods on all of the recognition tasks. The results are averaged over 10 random trials if applicable, and we report them in the "milliseconds per sample" manner to eliminate the influence of dataset size variation. The best results for the methods excluding "CRC" are shown in bold.

Dataset	MPD[4]	SRC[8]	CRC[14]	CHISD[2]	SANP[13]	SBDR[10]	CSA[9]	RNP[12]	CRNP
Honda/UCSD (50)	3.2	1.2×10^3	*0.28*	77.7	19.6	259	17.4	11.5	**0.32**
Honda/UCSD (100)	6.4	4.1×10^3	0.55	330	17.3	97.8	32.6	14.5	**0.46**
CMU MoBo (50)	12.4	7.6×10^3	*0.94*	89.0	47.2	85.0	29.0	3.5	**2.1**
CMU MoBo (100)	71.4	2.7×10^4	*1.8*	394	53.0	79.3	39.1	5.9	**2.5**
iLIDS-MA	3.9	741	*0.51*	58.7	121	N/A	9.6	24.5	**3.3**
iLIDS-AA	9.9	2337	*1.2*	150	344	N/A	36.8	83.4	**7.2**
CAVIAR4REID	**3.8**	214	*0.35*	55.3	249	N/A	15.8	30.8	8.0

The testing time for all the methods is listed in Table 5, showing that CRNP is generally more efficient than all the others except CRC as it just performs MPD in a low-dimensional projected space.

5 Conclusion and Future Work

We have proposed a novel set based recognition model which is generally more effective and significantly faster than other related methods. Experimental results on five different benchmark datasets have demonstrated its superiority. A possible future work will be introducing dictionary learning into the model to further improve its performance and robustness.

Acknowledgment

This work was supported by "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society", Funds for integrated promotion of social system reform and research and development of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government.

References

[1] Slawomir Bak, Etienne Corvee, François Bremond, and Monique Thonnat. Boosted human re-identification using riemannian manifolds. *Image and Vision Computing*, 30 (6-7):443 – 452, 2012. ISSN 0262-8856. doi: 10.1016/j.imavis.2011.08.008.

[2] Hakan Cevikalp and Bill Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573, 2010.

[3] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference (BMVC)*, pages 68.1–8.11, 2011. ISBN 1-901725-43-X. http://dx.doi.org/10.5244/C.25.68.

[4] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[5] Ralph Gross and Jianbo Shi. The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, June 2001.

[6] Yiqun Hu, Ajmal S. Mian, and Robyn Owens. Face recognition using sparse approximated nearest points between image sets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1992–2004, 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.283.

[7] K.C. Lee, Jeff Ho, M.-H. Yang, and David Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*, page 313–320, 2003.

[8] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE TPAMI*, 31(2):210–227, 2009.

[9] Yang Wu, M. Minoh, M. Mukunoki, Wei Li, and Shihong Lao. Collaborative sparse approximation for multiple-shot across-camera person re-identification. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 209 –214, sept. 2012. doi: 10.1109/AVSS.2012.21.

[10] Yang Wu, Michihiko Minoh, Masayuki Mukunoki, and Shihong Lao. Set based discriminative ranking for recognition. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision - ECCV 2012*, volume 7574 of *Lecture Notes in Computer Science*, pages 497–510. Springer Berlin Heidelberg, 2012.

[11] Yang Wu, Michihiko Minoh, Masayuki Mukunoki, and Shihong Lao. Robust object recognition via third-party collaborative representation. In *21st International Conference on Pattern Recognition (ICPR), 2012*, November 2012.

[12] Meng Yang, Pengfei Zhu, Luc Van Gool, and Lei Zhang. Face recognition based on regularized nearest points between image sets. In *The 10th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2013.

[13] Yiqun Hu and Ajmal S. Mian and Robyn Owens. Sparse Approximated Nearest Points for Image Set Classification. In *CVPR*, pages 121–128, 2011.

[14] Lei Zhang, Meng Yang, and Xiangchu Feng. Sparse representation or collaborative representation: which helps face recognition? In *ICCV*, 2011.