Collaborative Sparse Approximation for Multiple-shot Across-camera Person Re-identification

Yang Wu*, Michihiko Minoh*, Masayuki Mukunoki*, Wei Li† and Shihong Lao‡
*Academic Center for Computing and Media Studies, Kyoto University, Kyoto, 606-8501, Japan
Emails: yangwu@mm.media.kyoto-u.ac.jp, {minoh, mukunoki}@media.kyoto-u.ac.jp
†Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University
Email: liwei@mm.media.kyoto-u.ac.jp
‡Omron Corporation, Kyoto 619-0283, Japan, Email: lao@ari.ncl.omron.co.jp

***Abstract*—In this paper we propose a simple and effective solution to the important and challenging problem of across-camera person re-identification. We focus on the common case in video surveillance where multiple images or video frames are available for each person. Instead of exploring new features, the proposed approach aims at making a better use of such images/frames. It builds a collaborative representation over all the gallery images (of known person individuals) to best approximate the query images (containing an unknown person) via affine combinations. The approximation is measured by the nearest point distance between the two affine hulls constructed by the query images and gallery images, respectively. By enforcing the sparsity of the samples used for approximating the two nearest points, the relative importance of the gallery images belonging to different persons has the ability to reveal the identity of the querying person. Extensive experiments on public benchmark datasets demonstrate that the proposed approach greatly outperforms the state-of-the-art methods.**

***Keywords*-person re-identification; set-based recognition; camera network; sparse representation; collaborative representation.**

I. Introduction

The actual video analysis functions been adopted in real video surveillance systems are still limited to people counting, density/occupation estimation, abnormal behavior or event detection, etc. Though such functions are important for some applications, a significant drawback is that when an abnormal event/behavior happens, the system cannot tell the identity of the violater or trouble-maker, let alone tracking and tracing the suspecter in "normal" video footage to aid criminal investigations.

To conquer this, many efforts have been made in the academia towards the goal of human recognition or more precisely identification. An intuitive attempt is to do face recognition. However, despite its great advancement, it is still an unrealistic option due to the usually low-resolution and largely uncontrolled variations (viewpoint, illumination, occlusion, etc.) of the captured faces in real surveillance systems. Therefore, we have to look into another possibility—identifying persons by their body appearance. Such a problem is usually called *person re-identification*, with a common interpretation of identifying an human individual again upon re-entering the scene after some time [1]. Though the re-entering may happen to a single camera surveillance system, a more common case is that it happens across cameras as people traveling under the views of multiple cameras or a camera network. In this paper, we are concerning about the later one, which is relatively more challenging due to the potentially larger human appearance changes (viewpoint, background, illumination, different camera parameters, etc.). Once it is solved, the model should be applicable to the within-camera cases as well.

Person re-identification is commonly classified into two types: single-shot re-identification and multiple-shot re-identification, in which the former concerns the case that both the query and gallery sets for each person have only one single image, while the other means that there are multiple images for some set(s). Within these two, the former has attracted more attention from the community due to its relatively simpler problem definition and modeling requirement, therefore both supervised and unsupervised methods have been largely explored to solve it [1], [2], [3], [4], [5], [6]. As more and more datasets with multiple images per person have been annotated and published, such as the ETHZ datasets [7], [8] and the three different i-LIDS datasets [9], [10], the enthusiasm has gradually shifted to the research on multiple-shot re-identification [11], [12], [10]. In this paper we focus on solving the multiple-shot re-identification problem based on two considerations: it is very common in video surveillance that multiple frames are available for each interested object (i.e. person), and more data generally means a larger potential of a success in re-identification [12], [10].

Multiple-shot re-identification is still in its early stages of research, and the existing methods can be briefly categorized into two groups: a) simple set-based matching, and b) multiple-shot signature generation. The first group usually just use the minimum point distance between two sets as the set-to-set distance metric (though there are other options [13]), and thus many efforts have been made to design the per-image features for a better matching [14], which may integrate multiple cues and various types of features[12].

The second group try to make a condensed representation for the multiple images of a person, so that the distance between a query set and a gallery set is just the distance between their corresponding signatures[11], [10]. Such a signature can also be projected into a lower space as in [8]. A common drawback of these methods is that the performance is directly determined by the goodness of the hand-crafted features/signatures.

Actually, besides expecting to invent super features destined to work, which is not a trivial task, there are some other ways to make good use of the availability of multiple images. The state-of-the-art set-based recognition methods for face recognition are inspiring as many of them are generally applicable. An example is the work of AHISD/CHISD (Affine/Convex Hull based Image Set Distance)[15], which represents each image set in terms of an affine/convex hull constructed by the feature vectors of the images within it and uses the geometric distance between these two hulls as the between-set distance. It has been claimed to be less overfitting and more robust to outliers than the direct minimum point distance as it can generate new samples on the hull. Another work is SANP (Sparse Approximated Nearest Points) [16] which enforces the sparsity of samples used for point generation via affine combination, and it has been proved to perform better than AHISD/CHISD on several face recognition datasets. These two methods can be viewed as extensions of the simple minimum point distance based set-matching method. However, there is an underlying assumption that both of them assume that the between-set geometric distance between the query set and its relevant gallery set (i.e. the correct match) is small enough, thus with high probability to be smaller than that between the query set and any irrelevant gallery set. More concretely, for SANP, the sparsity constraint actually requires that there are at least some images in the relevant gallery set being similar enough to some image(s) in the query set. Though such a requirement seems to be satisfied in the commonly used benchmark datasets for face recognition, it is hard to be meet in the scenario of across-camera re-identification, especially when there are large viewpoint changes between cameras, which make the image sets from both cameras hard to overlap or stay close enough to each other in the feature space. It can be witnessed from the experimental results hereinafter. In fact, for most of the cases, they are even inferior to the simplest minimum point distance (MPD) which requires no convex model approximation and geometric distance computation.

Therefore, in this paper, we propose to use the geometric distance in another way by collaboratively approximating the query set using all the gallery sets, which is called Collaborative Sparse Approximation (CSA). The remaining of the paper is organized as follows to introduce our method and demonstrate its effectiveness. Section II briefly introduces the framework of set-based recognition by geometric distances, with reviews and comparisons on the two state-of-the-art approaches: AHISD/CHISD and SANP. Section III presents the details of the proposed CSA approach and analyzes its relationship and superiority to other related models—SANP and SRC (Sparse Representation based Classification). Section IV conducts extensive experiments with result analyses and discussions, and finally Section V makes conclusions.

II. Set-based recognition by geometric distances

A common setting for set-based recognition is: given a batch of gallery image sets with instances of a different object individual in each of them (may also be referred to as the training data) and an arbitrary query image set (or test set) containing an object with the same identity as one of the gallery objects, to find out this galley set, or namely to identify the query object. Note that there might be great within-set variations and real world noises in each set. This problem is usually solved by the Nearest-Neighbor based classification approach with some specifically defined between-set distance, therefore, finding a proper between-set distance is the key issue. Instead of using the straightforward closest-point distance which may be overfitting and sensitive to outliers, the state-of-the-art methods use some parametric fitting models (an affine/convex hull) to represent each query and gallery set, and then use the geometric distance (distance of closest approach) [15] between two models as the between-set distance.

Mathematically, let $X_i = [x_{i1}, \ldots, x_{iN_i}]$ and $X_j = [x_{j1}, \ldots, x_{jN_j}]$ denote a query set and an arbitrary gallery set, respectively, where x_{ik} and $x_{jk'}$ with $k \in \{1, \ldots, N_i\}, k' \in \{1, \ldots, N_j\}$ are individual images within them, and suppose there exist some set fitting models H_i and H_j for X_i and X_j, respectively, then the distance between X_i and X_j can be defined as:

$$D(X_i, X_j) = \min_{x_i \in H_i, x_j \in H_j} d(x_i, x_j), \tag{1}$$

where x_i and x_j are arbitrary points on H_i and H_j, respectively, and $d(x_i, x_j)$ is usually just the Euclidean distance. Following this framework, AHISD/CHISD and SANP personalize themselves by choosing different models for Hs and searching for the closest points x_i^* and x_j^* between H_i and H_j (i.e., $D(X_i, X_j) = d(x_i^*, x_j^*)$) in different ways.

A. Affine hull and convex hull approximation

In AHISD, H_i is just the affine hull of an arbitrary image set X_i, i.e., the smallest affine subspace containing all the images in the set:

$$H_i^{\text{aff}} = \left\{ x \;\middle|\; X_i\alpha = \sum_{k=1}^{N_i} \alpha_{ik} x_{ik}, \sum_{k=1}^{N_i} \alpha_{ik} = 1 \right\}. \tag{2}$$

It can also be represented in another form if we choose any reference point μ_i on it (e.g. the sample mean $\mu_i = \frac{1}{N_i}\sum_{k=1}^{N_i} x_{ik}$):

$$H_i^{\text{aff}} = \left\{ x = \mu_i + U_i v_i \;\; | \;\; v_i \in \mathbb{R}^{l_i} \right\}, \tag{3}$$

where U_i stands for the selected l_i orthonormal bases obtained by applying the Singular Value Decomposition (SVD) to $[x_{i1} - \mu_i, \ldots, x_{iN_i} - \mu_i]$ and their corresponding singular values are all significantly larger than zero.

Given two non-intersecting affine hulls $H_i^{\text{aff}} = \{\mu_i + U_i v_i\}$ and $H_j^{\text{aff}} = \{\mu_j + U_j v_j\}$, the closest points x_i^* and x_j^* between them can be found by solving

$$\min_{v_i, v_j} \|(\mu_i + U_i v_i) - (\mu_j + U_j v_j)\|_2^2, \tag{4}$$

and the between-set distance is defined as the distance between x_i^* and x_j^* which has a closed-form solution:

$$D(X_i, X_j) = D\left(H_i^{\text{aff}}, H_j^{\text{aff}}\right) = \|(I - P)(\mu_i - \mu_j)\|_2, \tag{5}$$

where $P = U(U^T U)^{-1} U^T$. When $U^T U$ is not invertible, P can be computed by $\tilde{U}\tilde{U}^T$ where $\tilde{U}$ is an orthonormal basis of U obtained by thin SVD. Such an affine hull model is called AHISD (Affine Hull based Image Set Distance).

Though it is a very simple model with a closed-form solution, AHISD fails when there are hull intersections. The reason is that in such a case the corresponding between-set distances will become zero. It happens when there are outliers. Therefore, a more robust solution is to adopt a tighter approximation model by bounding the coefficients α, e.g. the convex hull model CHISD (Convex Hull based Image Set Distance):

$$\begin{aligned} &D(X_i, X_j) = D\left(H_i^{\text{cvx}}, H_j^{\text{cvx}}\right) = \min_{\alpha_i, \alpha_j} \|X_i\alpha_i - X_j\alpha_j\|_2^2 \\ &s.t. \sum_{k=1}^{N_i} \alpha_{ik} = \sum_{k'=1}^{N_j} \alpha_{jk'} = 1, \;\; \alpha_{ik} \geq 0, \;\; \alpha_{jk'} \geq 0. \end{aligned} \tag{6}$$

B. Sparse approximated nearest points

Instead of moving from loose affine hulls to the tighter convex hulls, the model of sparse approximated nearest points (SANP) [16] still uses the affine hull, while at the same time adds a constraint to the closest point generation: each of them should be able to be approximated by a sparse combination of sample images in the corresponding image set. Such a constraint enforces SANPs to be close to some facet(s) of the affine hull and consequently close to some sample image(s) on those facet(s). It tightens the model and increases its robustness.

The between-set distance based on SANPs is defined as:

$$D(X_i, X_j) = (l_i + l_j) \cdot \left[F_{v_i^*, v_j^*} + \lambda_1 \left(G_{v_i^*, \alpha^*} + Q_{v_j^*, \beta^*} \right) \right], \tag{7}$$

where l_i and l_j are the dimensions of the affine hulls of H_i^{aff} and H_j^{aff}, respectively, and the optimal coefficients $\left(v_i^*, v_j^*, \alpha^*, \beta^*\right)$ are obtained by optimizing

$$\min_{v_i, v_j, \alpha, \beta} \left\{ F_{v_i, v_j} + \lambda_1 \left(G_{v_i, \alpha} + Q_{v_j, \beta} \right) + \lambda_2 \|\alpha\|_1 + \lambda_3 \|\beta\|_1 \right\}, \tag{8}$$

with

$$\begin{aligned} F_{v_i, v_j} &= \|(\mu_i + U_i v_i) - (\mu_j + U_j v_j)\|_2^2 \\ G_{v_i, \alpha} &= \|(\mu_i + U_i v_i) - X_i \alpha\|_2^2 \\ Q_{v_j, \beta} &= \|(\mu_j + U_j v_j) - X_j \beta\|_2^2 . \end{aligned} \tag{9}$$

The three trade-off weights are set as follows: $\lambda_1 = 0.01$, $\lambda_2 = 0.1 \cdot \max(\left|2\lambda_1 \cdot (X_i^T \mu_i)\right|)$, and $\lambda_3 = 0.1 \cdot \max(\left|2\lambda_1 \cdot (X_j^T \mu_j)\right|)$. Details for the meanings of these items and the reasons for the settings of the weights can be found in [16].

III. Collaborative sparse approximation

A. Collaborative representation

In face recognition, recently sparse representation based classification (SRC) has attracted much attention due to its impressive performance on some experiments [17]. It has just been revealed [18] that the component that truly improves the performance is the collaborative representation mechanism (i.e. representing the query image by the whole gallery set of images from all the classes collaboratively), but not the l_1-norm sparsity constraint. After the collaborative sparse representation, the query image is classified to the class whose gallery images contributes most to the representation, i.e., using only images from this class leads to the minimum representation error. This strategy has an intrinsic discrimination ability: identifying the backbone through indistinctive collaboration. Though collaborative representation has so far only been proved to be effective for single-shot recognition, in this paper we will show that it is also applicable to multiple-shot cases.

B. Sparse approximation with collaboration

As pointed out in [18], though collaborative representation is the key for a good performance, it has to work with some sparsity constraint. SANP is an effective model to encode sparse representation in geometric distance for set-based recognition, therefore, in this paper we use the same method to build our model with collaborative representation, which is named Collaborative Sparse Approximation (CSA). It operates as follows.

For a given query set X^q and all the gallery sets $\{X_1^g, \ldots, X_n^g\}$ of n different person individuals, build a unified gallery set $X^g = X_1^g \cup \cdots \cup X_n^g$, and then find the sparse approximated nearest points between X^q and X^g by optimizing

$$\min_{v_q, v_g, \alpha, \beta} \left\{ F_{v_q, v_g} + \lambda_1 \left(G_{v_q, \alpha} + Q_{v_g, \beta} \right) + \lambda_2 \|\alpha\|_1 + \lambda_3 \|\beta\|_1 \right\}, \tag{10}$$

with

$$\begin{aligned} F_{v_q,v_g} &= \|(\mu_q + U_q v_q) - (\mu_g + U_g v_g)\|_2^2 \\ G_{v_q,\alpha} &= \|(\mu_q + U_q v_q) - X^q \alpha\|_2^2 \\ Q_{v_g,\beta} &= \|(\mu_g + U_g v_g) - X^g \beta\|_2^2 . \end{aligned} \tag{11}$$

The trade-off weights are just set to be the same as those in SANP: $\lambda_1 = 0.01$, $\lambda_2 = 0.1 \cdot \max(\left|2\lambda_1 \cdot (X^{q^T} \mu_q)\right|)$, and $\lambda_3 = 0.1 \cdot \max(\left|2\lambda_1 \cdot (X^{g^T} \mu_g)\right|)$.

After getting the optimal (α^*, β^*), in which $\beta^* = [\beta_1^*; \ldots; \beta_n^*]$, we use the following distance to measure the dissimilarity of X^q and an arbitrary gallery set $X_k^g, k \in \{1, \ldots, n\}$:

$$D\left(X^q, X_k^g\right) = \|X^q \alpha - X_k^g \beta_k\|_2^2 / \|\beta_k\|_1. \tag{12}$$

This dissimilarity definition makes use of both the reconstruction error (i.e. the representation residual) and the "sparsity" of the reconstruction coefficients because both of them have been proved to be discriminative for recognition [18]. We directly follow the formulation in [18] except using the l_1-norm instead of the l_2-norm which doesn't have significant impact on the performance.

The relevant gallery set is expected to be able to approximate the query set (actually its sparsely approximated point) better with fewer number of samples than any irrelevant gallery set. Since the optimization objective function has the same form as that in SANP, the same gradient-based optimization algorithm can be adopted (see [16] for details). However, as our model requires only one round of optimization for each query set instead of n rounds for SANP, the computation time can be much shortened. Moveover, as it will be shown in Section IV, the proposed CSA approach greatly outperforms SANP for across-camera person re-identification.

IV. Experiments and results

In this section we present the experimental results and discuss some important findings.

A. Benchmark Datasets

There are two publicly available datasets widely used as benchmarks for multiple-shot person re-identification : ETHZ datasets [7] (containing three subsets) and the i-LIDS dataset [9]. Unfortunately, none of them is suitable for testing our model. The ETHZ datasets was captured by a single camera, and it is so simple that the performance on it has already got saturated [10]. The i-LIDS dataset is much more difficult (the performance on which is still far from getting saturated) and it has averagely 4 images per person taken by two non-overlapping cameras. However, there are still 22 human individuals having only 2 images, so that a common setting is to have only 1 image per person for the gallery set [5] which is actually no longer multiple-shot. Though sample generation methods like affine transformations can be applied to enrich the samples as in [10], such pseudo samples might mislead the model and it is also unfair to compare with the methods which don't use these generated samples.

Therefore, we choose to experiment on the two newly built datasets i-LIDS-MA and i-LIDS-AA [19] from the i-LIDS video surveillance data released by the Home Office of UK. Both of them contain multiple images for each human individual captured by two cameras, and there are large viewpoint changes. The i-LIDS-MA dataset contains 40 individuals with exactly 46 manually annotated frames per camera for each individual, resulting 3680 images in all. In order to take into account of the imperfectness of human detection and tracking in real systems, human images in the i-LIDS-AA dataset were extracted automatically using HOG-based detector instead of manual annotation, so that there are significant noises in the bounding box localization. This dataset is also much bigger, containing as many as 100 individuals with totally 10754 images (averagely 54 images per camera for each person), and the number of images for each person varies greatly due to the variations of its appearing time and tracking/detection recall.

B. Methods for comparison

As mentioned in section I, the proposed CSA approach is not focusing on feature representation, but on exploring set-based matching/recognition. Therefore here we do not compare it with the papers proposing new features (e.g. HPE[11] and MRCG[10]), but evaluate it against both the standard method and the state-of-the-art approaches on set-based recognition using the widely adopted features. Classification by the minimum point-wise distance between two sets is a standard and widely-used method (see [11], [13], [12], [10]), despite its simplicity. We use "MPD" to denote it for brevity. We also compare with the closely related approaches: CHISD and SANP, which currently lead the performance on set-based face recognition and their models are directly applicable to person re-identification. As mentioned later, for all the experiments we simply adopted the same basic features as proposed in some other approaches ([3], [2]) which utilize learning-based classifiers. Though they are still the state-of-the-art approaches for person re-identification, they were designed for single-shot cases only and not directly applicable to our multiple-shot settings. Therefore, they are not involved in our comparison.

C. Experimental settings

Color is considered to be the most informative cue for person re-identification[3]. Among all the different designs, the feature of densely sampled color histograms[2] has been proved to be a good choice[6], so we use it here as the color descriptor. The parameter settings for it are the same as those in [2] and similarly the feature dimension is reduced by PCA (keeping 90% cumulative energy). Besides that, the Schmid and Gabor texture filters are also adopted as complementary

features with the settings exactly following[3]. We found that the Gabor texture filter response has a relatively flat distribution during histogram binning while that of the Schmid texture is largely biased on low values, so we used linear bin ranges $\{[0, \frac{1}{15}), [\frac{1}{15}, \frac{2}{15}), \ldots, [\frac{14}{15}, 1), [1, 1]\}$ for Gabor texture, and exponential ones $\{\{[\frac{1}{2} - \frac{1}{2^k}, \frac{1}{2} - \frac{1}{2^{k+1}})\}, k \in \{1, 2, \ldots, 7\}\} \cup \{[\frac{1}{2} - \frac{1}{2^8}, \frac{1}{2} + \frac{1}{2^8})\} \cup \{\{[\frac{1}{2} + \frac{1}{2^k}, \frac{1}{2} + \frac{1}{2^{k-1}})\}, k \in \{8, 7, \ldots, 2\}\} \cup \{[1, 1]\}$ for Schmid texture, which might be different from the original ones.

We followed the experimental setting as in [10] by randomly sampling images for each set and repeated the experiments for 10 times to average the results. For the i-LIDS-MA dataset, we used the CMC (Cumulative Matching Characteristic) values of the rank top 20% (i.e., 1 to 8) for performance comparison, with set size $N = 10$ and $N = 23$, respectively, as shown in Figure 1(a) and Figure 1(b). Since N is an important factor for all the four methods being evaluated and it can be set up to 46, we also repeated the experiments by varying N in $\{3, 5, 10, 20, 23, 46\}$, and utilized the condensed ranking performance criterion "MRR (Mean Reciprocal Rank)"[6] to show its influence on different methods (see Figure 1(c)). Besides that, we evaluated the influence of the number of persons (denoted by P) as well, by choosing P within $\{10, 20, 30, 40\}$. For a fair comparison, we used the recognition rate at the rank top 20% point on the CMC curve as performance measurement for different Ps (see Figure 1(d)). For the i-LIDS-AA dataset, CMC curves with $N = 23$ (which is more representative than $N = 10$) and $P = 40$ were generated (Figure 1(e)), and we varied P within $\{20, 40, 70, 100\}$ to show its influence (Figure 1(f)). Since the actual query and gallery images for each person varies greatly in the i-LIDS-AA dataset and these images are quite noisy on person localization, we didn't do experiments on different Ns as those for the i-LIDS-MA dataset.

D. Results and discussions

As it clearly shows, SANP performs slightly better than MPD when N is large enough (≥ 5), which is necessary for both SANP and CHISD since they need samples for affine reconstruction. However, the superiority of SANP is so weak that even much worse than CHISD (see chart (c) in Figure 1), which indicates that the within-person variations across cameras are very large compared to the between-person differences and thus SANP cannot well differentiate these human individuals. Though the main difference between our CSA approach and SANP is just collaborative representation vs. individual representation, it makes very big difference on the performance. CSA performs much better than SANP for all the cases, and it is also significantly better than CHISD and MPD. Therefore, compared with existing methods, CSA is much more robust to large within-class changes and error-prone human detection and tracking results.

The experimental results also reveal some other important findings. First, the performance of all the four methods doesn't increase monotonously as the ratio of set-size over class-number N/P grows (Actually such a ratio can be increased by either enlarging N as shown in the chart (c) of Figure 1 or decreasing P whose results are presented in chart (d)). Such a phenomenon tells us that larger N increases the probability of a better reconstruction of a query image (or a virtual one on the convex hull) using fewer relevant gallery images while at the same time it also increases the risk of having an occasional sparse set of samples that can better represent some query image(s). The performance of CSA varies more greatly than that of the others as the ratio changes, and the optimal ratio for CSA is smaller than those for the others. Therefore, it's important to adjust the ratio (N/P) when using CSA and it's relatively easier to find the optimal number for it. Second, detection and tracking errors greatly influence the performance of CHISD, SANP and CSA, but it has little impact on MPD (derived by comparing chart (b) with chart (e)). Even though, CSA is still much superior to MPD in terms of the absolute performance.

V. Conclusion

This paper presents a novel approach named Collaborative Sparse Approximation (CSA) for multiple-shot across-camera person re-identification. It incorporates the discriminative power of collaborative representation into the geometric distance based modeling framework, and thus increases the robustness to across-camera appearance changes. Extensive experiments on public benchmark datasets demonstrate its superiority to existing state-of-the-art methods when the same features are used.

Acknowledgment

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.

References

[1] D. Gray and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007, pp. 41–47.

[2] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *ACCV (4)*, 2010, pp. 501–512.

[3] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European Conference on Computer Vision*, 2008, pp. 262–275.

[4] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," *Advanced Video and Signal Based Surveillance, IEEE Conference on*, pp. 435–440, 2010.

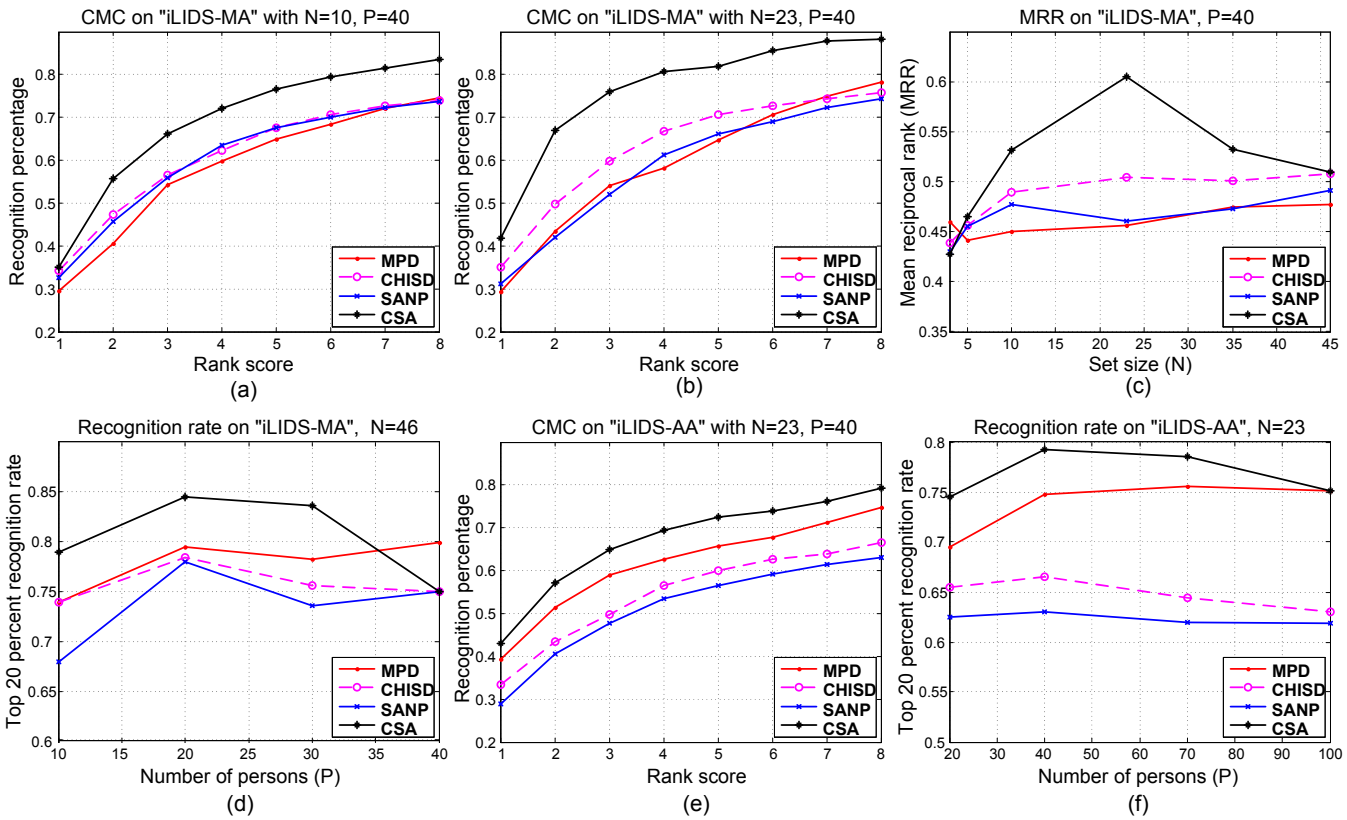

Figure 1. Experimental results. (a),(b) and (e) are CMC (Cumulative Matching Characteristic) curves with $P = 40$. (c) is MRR (Mean Reciprocal Rank) value w.r.t set size variations. (d) and (f) are the recognition rate at rank top 20% vs. the number of persons (P).

[5] Wei-Shi Zheng and Shaogang Gong and Tao Xiang, "Person Re-identification by Probabilistic Relative Distance Comparison," in *CVPR*, 2011, pp. 649–656.

[6] Y. Wu, M. Mukunoki, T. Funatomi, M. Minoh, and S. Lao, "Optimizing Mean Reciprocal Rank for Person Re-identification," in *2011 8th IEEE International Conference on Advanced Video and Signal-Based Surveillancei (AVSS)*, Aug. 2011, pp. 408–413.

[7] A. Ess, B. Leibe, and L. van Gool, "Depth and appearance for mobile scene analysis," in *ICCV*, 2007, pp. 1–8.

[8] W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *SIBGRAPI*, 2009, pp. 322–329.

[9] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *BMVC*, 2009.

[10] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid," in *AVSS*, 2011. [Online]. Available: http://hal.inria.fr/inria-00620496/en/

[11] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by hpe signature," in *ICPR*, 2010, pp. 1413–1416.

[12] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010.

[13] M. J. Metternich and M. Worring, "Semi-interactive tracing of persons in real-life surveillance data," in *MiFOR*, 2010.

[14] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *BMVC*, 2011.

[15] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *CVPR*, 2010, pp. 2567–2573.

[16] Yiqun Hu and Ajmal S. Mian and Robyn Owens, "Sparse Approximated Nearest Points for Image Set Classification," in *CVPR*, 2011, pp. 121–128.

[17] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[18] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: which helps face recognition?" in *ICCV*, 2011.

[19] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Boosted human re-identification using riemannian manifolds," *Image and Vision Computing*, 2011.