Locality-constrained Collaboratively Regularized Nearest Points for Multiple-shot Person Re-identification

Yang Wu, Masayuki Mukunoki, and Michihiko Minoh*

Multiple-shot person re-identification is not only an important task in itself for video surveillance applications, but also a valuable scientific problem for which a good solution may greatly aid the solving of other problems including multiple-camera tracking/tracing and set-based classification. This paper presents a novel approach called Locality-constrained Collaboratively Regularized Nearest Points (LCRNP) for solving it. It enhances the power of collaborative representation (specifically the collaboratively regularized nearest points model) by introducing locality constraints, following the same idea of the recently proposed Locality-constrained Collaborative Sparse Approximation (LCSA) model. Extensive experiments on three benchmark datasets with various experimental settings demonstrate the effectiveness of locality constraints and the superiority of LCRNP to the most related LCSA models.

Keywords: Person re-identification, set-based classification, locality constraints

1. Introduction

Person re-identification is the task of identifying a person again when he/she reappears in the view of a camera after some time or enters another camera's view. Based on how many images per person are available, it can be categorized into two groups: single-shot person re-identification and multiple-shot person re-identification. Though the single-shot case is of fundamental importance as any remarkable progress on it is directly applicable to the multiple-shot case, the multiple-shot problem is more common in real video surveillance applications and it is also more interesting and promising because of the existence of multiple images for each person. The advances on multiple-shot re-identification will directly benefit the research on people tracking (especially across-camera tracking or tracing) as it can help bridging the gap between broken tracks. Moreover, multiple-shot re-identification can be treated as a set-based classification problem and therefore its solutions may be applicable to other set-based classification tasks like set-based face recognition. This paper focuses on the multiple-shot person re-identification problem.

Existing solutions for multiple-shot re-identification mainly belong to three groups: feature/signature design, metric learning, and sample/set reorganization. A brief review of them can be found in the work of Locality-constrained Collaborative Sparse Approximation (LCSA) [6]. Like LCSA, the approach to be proposed in this paper belongs to the last group. More precisely, both of them are on the specific branch of using all the gallery sets to collaboratively represent the probe set of images, which is different from traditional single set to single set distance finding approaches like the Set Based Discriminative Ranking (SBDR) model.

As far as we are aware, there are four existing collaborative representation based approaches for solving the multiple-shot person re-identification problem, including the Third Party Collaborative Representation (TPCR) model [7], the Collaborative Sparse Approximation (CSA) model [4], LCSA, and the Collaboratively Regularized Nearest Points (CRNP) model. Since TPCR relies on extra data, it is not comparable with the others and the approach to be presented here. The recently proposed CRNP model, which uses l_2-norm based non-sparse constraints for regularizing the representation coefficients, has been proved to be more effective and significantly faster than the CSA model which relies on the computationally expensive l_1-norm based sparse constraints on the approximation coefficients. The work of LCSA has shown that CSA is better to be applied locally but not globally, as the locality constraints help pruning some irrelative sets which may confuse the collaborative representation model. Considering the superiority of CRNP in comparison with CSA, it is attractive to check whether locality constraints can improve CRNP's performance or not, and it is also anticipated to compare the locality constrained CRNP (LCRNP) with LCSA and see whether LCRNP outperforms it as well. This paper is right for answering these questions.

2. Locality-constrained Collaboratively Regularized Nearest Points

Since the proposed approach is an extension of the recently proposed CRNP model, we briefly overview CRNP at first, and then focus on the extension of adding locality constraints to it with two different models. The illustration of these two models in comparison with the related work of LCSA is given in Figure 1.

* Academic Center for Computing and Media Studies, Kyoto University, Yoshida Nihonmatsu-cho, Sakyo-ku, Kyoto, Japan 606-8501,
{yangwu, mukunoki, minoh}@mm.media.kyoto-u.ac.jp

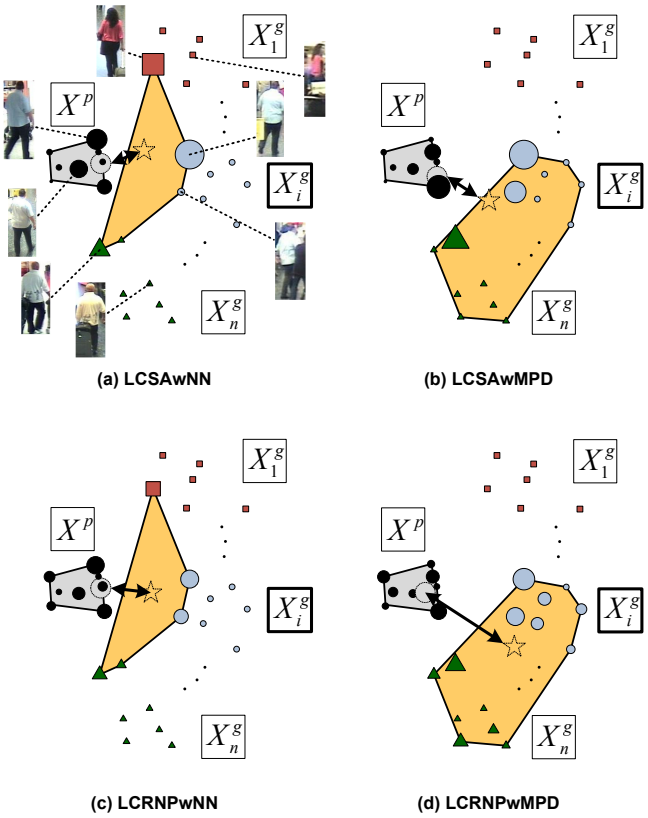

Fig. 1. An illustrative comparison between the proposed LCRNP models (LCRNPwNN and LCRNPwMPD) and the closely related LCSA models (LCSAwNN and LCSAwMPD). X^p denotes the probe set, while $X_i^g, i \in \{1,\ldots,n\}$ denote the gallery sets. Each set contains a number of samples marked by a set-specific color and shape. The colored areas bounded by black lines denote the affine hulls for the sets. The size of a sample marker is set to be proportional to the weight of the sample in the linear combination for set-to-sets dissimilarity, while the double-headed arrows connect the pairs of generated nearest points between sets. Due to that LCRNP models use l_2-norm instead of l_1-norm in LCSA models for coefficients regularization, they have significant denser coefficient values.

2.1 Collaboratively Regularized Nearest Points

Collaboratively Regularized Nearest Points (CRNP) [5] inherits the merits of simplicity, robustness, and high-efficiency from the recently proposed regularized nearest points (RNP) method [8] on finding the set-to-set distance using the l_2-norm regularized affine hulls. Different from RNP, it 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 [3] and collaborative sparse approximation (CSA) for set-based recognition [4]. However, CRNP uses l_2-norm instead of the l_1-norm for coefficients regularization, which makes it much more efficient. Extensive experiments [5] have also shown that CRNP, which is actually a non-sparse representation based model, is more effective than the sparse representation based methods like CSA.

Mathematically, given the test/probe set $\mathbf{X}^p$ and all the training/gallery sets $\mathbf{X}_i^g, i \in \{1,\ldots,n\}$, CRNP solves the following optimization problem:

$$\begin{aligned} &\min_{\alpha,\beta} \left\{ \|\mathbf{X^p}\alpha - \mathbf{X^g}\beta\|_2^2 + \lambda_1 \|\alpha\|_2^2 + \lambda_2 \|\beta\|_2^2 \right\}, \\ &s.t. \quad \textstyle\sum_k \alpha_k = 1,\ \sum_{i=1}^n \sum_j \beta_i^j = 1, \end{aligned} \tag{1}$$

where $\mathbf{X^g} = [\mathbf{X}_1^g,\ldots,\mathbf{X}_n^g]$ 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 can be transformed to the following unconstrained optimization problem:

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

where $\mathbf{z} = [\mathbf{0}_{1,m}, \sqrt{\gamma_1}, \sqrt{\gamma_2}]^T$ with m denoting the dimensionality of the image feature space and γ_1 and γ_2 denoting the Lagrangian multipliers. $\hat{\mathbf{X}}^{\mathbf{p}} = [\mathbf{X^p}^T, \sqrt{\gamma_1}\mathbf{1}_{N_q,1}, \mathbf{0}_{N_q,1}]^T$ in which N_q is the number of samples in $\mathbf{X^p}$, and $\hat{\mathbf{X}}^{\mathbf{g}} = [-\mathbf{X^g}^T, \mathbf{0}_{N_x,1}, \sqrt{\gamma_2}\mathbf{1}_{N_x,1}]^T$ where N_x is the number of samples in $\mathbf{X^g}$. $\mathbf{0}_{i,j}$ 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, alternatively optimizing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is usually a more efficient choice, as it avoids the time-consuming matrix inverse operation of an integrated matrix containing both $\mathbf{X^p}$ and $\mathbf{X^g}$ for each test/query set $\mathbf{X^p}$. More concretely, when $\boldsymbol{\alpha}$ is fixed, $\boldsymbol{\beta}$ has a closed-form solution

$$\boldsymbol{\beta}^* = \mathbf{P}_g \left(\mathbf{z} - \hat{\mathbf{X}}^{\mathbf{p}}\boldsymbol{\alpha} \right), \tag{3}$$

where $\mathbf{P}_g = \left(\hat{\mathbf{X}^{\mathbf{g}}}^T \hat{\mathbf{X}}^{\mathbf{g}} + \lambda_2 \mathbf{I} \right)^{-1} \hat{\mathbf{X}^{\mathbf{g}}}^T$ (with $\mathbf{I}$ denoting the identity matrix) only depends on $\mathbf{X^g}$, so it can be pre-computed. When $\boldsymbol{\beta}$ is fixed, $\boldsymbol{\alpha}$ also has a closed-form solution

$$\boldsymbol{\alpha}^* = \mathbf{P}_p \left(\mathbf{z} - \hat{\mathbf{X}}^{\mathbf{g}}\boldsymbol{\beta} \right), \tag{4}$$

where $\mathbf{P}_p = \left(\hat{\mathbf{X}^{\mathbf{p}}}^T \hat{\mathbf{X}}^{\mathbf{p}} + \lambda_1 \mathbf{I} \right)^{-1} \hat{\mathbf{X}^{\mathbf{p}}}^T$.

As claimed in RNP [8], the objective function in Formula (2) 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 practice, the iteration usually terminates in no more than 10 steps.

For classification, CRNP uses the following dissimilarity between $\mathbf{X^p}$ and $\mathbf{X}_i^g, i \in \{1,\ldots,n\}$:

$$d_{CRNP}^i = \left(\|\mathbf{X^p}\|_* + \|\mathbf{X}_i^g\|_* \right) \cdot \|\mathbf{X^p}\boldsymbol{\alpha}^* - \mathbf{X}_i^g\boldsymbol{\beta}_i^*\|_2^2 \,/\, \|\boldsymbol{\beta}_i^*\|_2^2, \tag{5}$$

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

$$C(\mathbf{Q}) = \arg\min_i \left\{ d_{CRNP}^i \right\}. \tag{6}$$

Like CSA, CRNP expects that the relevant gallery set is able to approximate the probe set better than

any irrelevant gallery set with fewer number of samples. Though this is true for many cases, its reliability actually depends on the number of gallery sets and their set-sizes. As the ratio of set-size to set-number increases, the performance of CSA can significantly decrease, suggesting that too many samples may confuse the model [4]. Similar problem happens to CRNP as well, as it will be shown in the following experimental results. Introducing locality constraints to the collaborative representation is witnessed to be a simple and effective solution (LCSA) to this problem for CSA [6], while leads to significant performance improvement as well, so we would like to do the same thing for CRNP. Hopefully, we can get more promising results than LCSA, considering that CRNP is shown to be superior to CSA [5].

2.2 LCRNPwNN: Representation with Neighboring Samples A simple way to introduce locality constraints is having only those gallery samples close enough to the probe samples (i.e., neighbors) selected for the collaborative representation. For simplicity, we just use the Euclidean distance in the feature space for nearest neighbors selection. More concretely, a certain number of nearest neighbors from X^g are chosen for each sample in X^p, and then all the selected samples together form a new gallery $\hat{X}^g$. We use $\hat{X}^g$ instead of X^g to do CRNP based collaborative representation and classification. This model is called "LCRNPwNN", which means LCRNP with Nearest Neighbors.

2.3 LCRNPwMPD: Representation with Neighboring Sets Another way for implementing LCRNP is preselecting the neighboring gallery sets of the probe set from all the candidates $\{X_1^g, \ldots, X_n^g\}$ using certain set-to-set distance, followed by doing CRNP with only these selected gallery sets (i.e., local sets). In this paper, without loss of generality, we use the simple and fast Minimum Point-wise Distance (MPD) model as the set-to-set distance measurement and thus name this LCRNP model "LCRNPwMPD" (i.e., LCRNP with MPD). Note that both LCRNPwNN and LCRNPwMPD follow the same ways for introducing locality constraints as LCSAwNN and LCSAwMPD [6], allowing a fair comparison with them.

3. Experiments and Results

We verify the effectiveness of the proposed LCRNP models (LCRNPwNN and LCRNPwMPD) on the task of multiple-shot person re-identification, comparing with the recently proposed CRNP, CSA and LCSA models (LCSAwNN and LCSAwMPD).

In this section, we first briefly introduce the benchmark datasets adopted and the features extracted for our experiments, and then present an experimental study on the impact of the key parameter (neighborhood size) within our models. Finally, we show all the experimental results with our discussions.

3.1 Datasets and Features We follow the work of LCSA on using the iLIDS-MA, iLIDS-AA [1] and CAVIAR4REID [2] datasets for our experiments. The first two are extracted from the multiple-camera tracking task of the i-LIDS video surveillance data released by the Home Office of UK, while the last one is chosen from the original CAVIAR videos. All of them contain multiple frames for each person captured by two non-overlapping cameras with large viewpoint changes.

In greater details, there are 40 persons with exactly 46 instances per camera for each person in the iLIDS-MA dataset (resulting in 3680 images), while 100 persons with averagely about 54 cropped images per camera for each person (totally 10754 images) are involved in the iLIDS-AA dataset. It is worth noticing that iLIDS-MA has the bounding boxes manually annotated while iLIDS-AA has its samples automatically generated by a tracking algorithm with a HOG-based detector. Therefore, iLIDS-MA is suitable for evaluating the performance on solving re-identification in its narrow sense, and iLIDS-AA is valuable for verifying the robustness of re-identification algorithms w.r.t. localization errors in real systems. In the CAVIAR4REID dataset, 72 persons were manually annotated and selected, of which 50 have both camera views while the other 22 have only one camera view. For each person, a set of manually cropped images were carefully selected for each camera view. Comparing with iLIDS-MA and iLIDS-AA, it has broader changes in resolution, severer images variations (pose, illumination, occlusion, etc.) and lower redundancy (due to sparse frame sampling).

An interesting phenomenon is that the performance of CSA doesn't vary monotonously w.r.t. the number of samples per set (denoted by "N" in this paper), while LCSA does not have this problem, as more samples generally lead to a better performance. To study how LCRNP models perform as N changes, we follow LCSA in doing experiments with $N = 10$, $N = 23$, and $N = 46$ for iLIDS-MA and iLIDS-AA datasets, and $N = 5$and10 for the CAVIAR4REID dataset (both specified and unspecified training/test identities are tested for the case of $N = 10$ [6]). For iLIDS-AA dataset, when the total number of available images for some people are not enough for large Ns, we use the largest possible numbers.

The same color and texture histograms based features as mentioned in [7] are adopted, followed by feature dimension reduction to 400 dimensions using Principal Component Analysis (PCA). Furthermore, we standardize the feature vectors to make them have a 0-mean and a 1-variance in each dimension. Following the tradition, in all of our experiments, the probe and gallery sets are randomly sampled (except for the $N = 46$ case in iLIDS-MA), and such random sampling is repeated 10 times for result averaging.

3.2 Parameter Analysis The most important factor for LCRNP models (including both LCRNPwNN and LCRNPwMPD) is the neighborhood size, because it determines how many samples/sets are to be included in the collaborative representation, and thus may greatly influence the final re-identification performance. Therefore, we first study the impact of this parameter for both models. For dealing with different datasets and different Ns (which means different scales of the experimental data), we use the same relative neighborhood size measure called *locality ratio* as introduced for LCSA. More

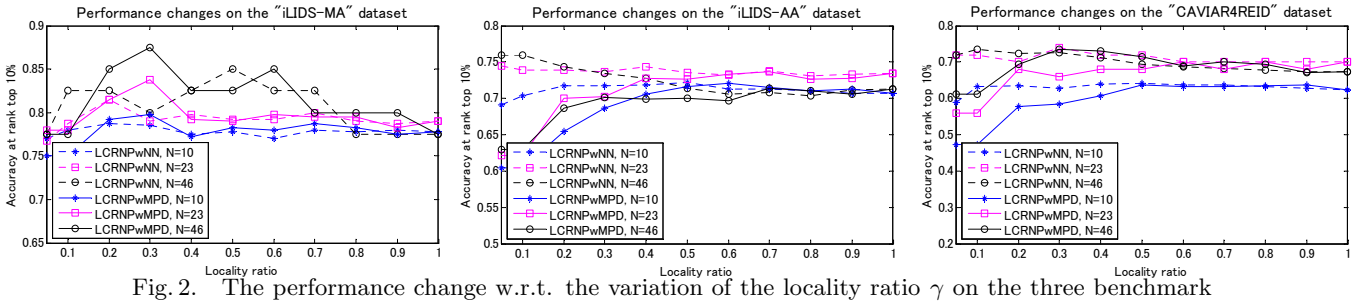

Fig. 2. The performance change w.r.t. the variation of the locality ratio γ on the three benchmark datasets for person re-identification. For each dataset, the three experimental settings with different set sizes or data sampling strategies are adopted. For a clear comparison, results for the same experimental setting are plotted using the same color and mark but different line styles. Note that when $\gamma = 1$, the locality constraint vanishes, both LCRNPwNN and LCRNPwMPD become the original CRNP model.

concretely, the *locality ratio* (denoted by γ) is the ratio of the number of closest gallery samples/sets for each probe image/set to the total number of gallery samples/sets available. For all the experiments, we have γ varies from 0.05 to 1 (mainly with a 0.1 step) for analyzing its influence on the performance, and present the results in Figure 2.

For the person re-identification problem, which is usually evaluated by CMC (Cumulative Matching Characteristic) curves, we report the rank top 10% recognition accuracy (an indicative value) for studying the performance change w.r.t. the variation of the locality ratio γ. As the results shows, with a proper locality ratio, LCRNP models perform better than unconstrained CRNP model (when locality ratio $\gamma = 1$). However, improper locality ratios, especially the small ones, may lead to even worse performance. This is unlike the cases of LCSA, which generally favors small locality ratios. It is probably because the non-sparse representation based CRNP model needs a proper balance of the number of irrelative samples and the number of relative samples. When γ gets smaller, the number of irrelative samples decreases, but it may also lead to a significant loss of the few relative samples (due to the possible mistakes in the pre-selection of samples/sets), and thus increase the risk that the remaining relative samples may be overrode by the irrelative ones. For LCSA, however, the sparsity constraints in the collaborative approximation to some extent highlights role of relative samples and suppresses the irrelative ones, so its performance doesn't decrease much (if not increasing) when γ diminishes.

Within the two models of LCRNP, the relative superiority is not as significant as that in LCSA models. Generally speaking, LCRNPwMPD performs slightly better on iLIDS-MA, while LCRNPwNN has a small superiority on the iLIDS-AA and CAVIAR4REID datasets. Compared with LCSA, it is harder to give a recommendable choice of γ for both LCRNPwNN and LCRNPwMPD. The results reveal that $[0.3, 0.6]$ seems to be good for LCRNPwMPD, while $[0.1, 0.2] \cup [0.4, 0.6]$ looks better than other choices for LCRNPwNN.

3.3 Performance Comparison As mentioned before, we compare LCRNPwNN and LCRNPwMPD with the most related state-of-the-art methods, including CRNP, CSA and LCSA models (LCSAwNN and LCSAwMPD) in all the experiments. Since LCSA models have already outperformed other unrelated state-of-the-art methods on the same datasets, they are not compared with in this paper. Similar to LCSA, we use the CMC curves for performance comparison, with the recognition rates at rank top 10% highlighted in the legend. For the proposed LCRNP models, we show the results generated with the most suitable locality ratio.

The results shown in Figure 3 clearly shows that LCRNPwNN and LCRNPwMPD perform better than CRNP, demonstrating the advantage of introducing locality constraints to it. The difference on the iLIDS-AA dataset is not as significant as those for the other two datasets due to the great within-class variations of the data, which makes the assumption that relative samples naturally distributed close to the probe set fails. Similar to CSA, the performance of CRNP on the iLIDS-MA and iLIDS-AA datasets decreases significantly when N increases from 23 to 46, showing that too many samples may confuse the model collaborative representation model. However, the LCRNP models do not have such a problem. In contrast, more samples generally results in a better performance, especially for LCRNPwNN. The LCRNP models are also much superior to CSA and LCSA models in all the experiments expect the last one, in which LCSAwNN is slightly better.

It is worth mentioning that besides its significantly better effectiveness than that of LCSA, LCRNP is also much more efficient than LCSA. This can be witnessed by the run time comparison between CRNP and CSA which shows a up to 2-orders speed improvement [(5)].

4. Conclusion

We experimentally show that introducing locality constraints for CRNP can improve its performance on multiple-shot person re-identification tasks, especially for the cases where redundance exists in the data. The proposed LCRNP models are not only significantly more effective than the recently introduced LCSA models, but also much more efficient than them. The paper has also given extensive experimental analysis on the model parameters, which may be inspiring for properly applying the models to new datasets and to solve new problems.

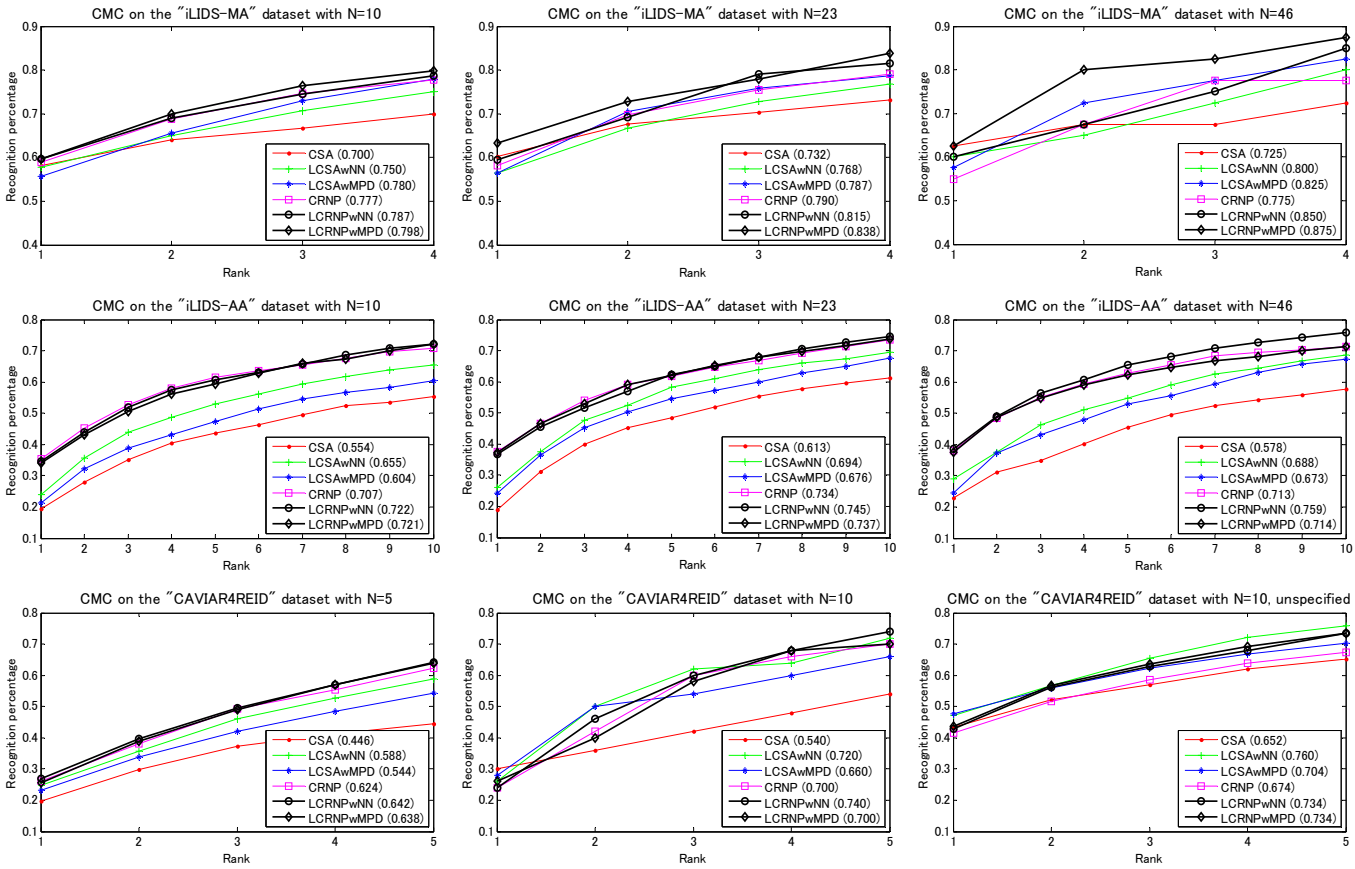

Fig. 3. Performance comparison with related state-of-the-art methods on three standard benchmark datasets with three different experimental settings. Only the records within rank top 10% are shown due to their relative importance and the space limitation. For a clear comparison, the recognition rate of each method at rank top 10% is also presented in the legend right after the name of the method.

References

(1) S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Boosted human re-identification using riemannian manifolds," Image and Vision Computing, vol.30, no.6-7, pp.443 – 452, 2012.

(2) D.S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," British Machine Vision Conference (BMVC), pp.68.1–8.11, 2011.

(3) J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," IEEE TPAMI, vol.31, no.2, pp.210–227, 2009.

(4) Y. Wu, M. Minoh, M. Mukunoki, W. Li, and S. Lao, "Collaborative sparse approximation for multiple-shot across-camera person re-identification," AVSS, pp.209 –214, sept. 2012.

(5) Y. Wu, M. Minoh, and M. Mukunoki, "Collaboratively regularized nearest points for set based recognition," Proc. of The 24th British Machine Vision Conference (BMVC), 2013.

(6) Y. Wu, M. Minoh, and M. Mukunoki, "Locality-constrained collaborative sparse approximation for multiple-shot person re-identification," The 2nd IAPR Asian Conference on Pattern Recognition (ACPR2013), Nov. 2013.

(7) Y. Wu, M. Minoh, M. Mukunoki, and S. Lao, "Robust object recognition via third-party collaborative representation," ICPR, November 2012.

(8) M. Yang, P. Zhu, L.V. Gool, and L. Zhang, "Face recognition based on regularized nearest points between image sets," FG, 2013.

Yang Wu is currently a post-doctoral researcher of Academic Center for Computing and Media Studies, Kyoto University. He received a BS degree in information engineering and a Ph.D degree in pattern recognition and intelligent systems from Xi'an Jiaotong University in 2004 and 2010, respectively. His research is in the fields of computer vision and pattern recognition, with particular interests in the detection, tracking and recognition of humans and also generic objects. He is also interested in image/video search and retrieval, along with machine learning techniques.

Masayuki Mukunoki received the BS, MS and PhD degrees in information engineering from Kyoto University. He is now an associate professor in the Academic Center for Computing and Media Studies and a faculty member in the Graduate School of Informatics, in Kyoto University. His research interests include computer vision, video media processing, lecture video analysis, and human activity sensing with camera.

Michihiko Minoh is a professor at Academic Center for Computing and Media Studies, Kyoto University, Japan. He received the BEng, MEng, and DEng degrees in information science from Kyoto University, in 1978, 1980, and 1983, respectively. He served as director of ACCMS from April 2006 to March 2010 and concurrently served as vice director in the Kyoto University Presidents Office from October 2008 to September 2010. Since October 2010, he has been vice-president, chief information officer at Kyoto University, and director-general at Institute for Information Management and Communication, Kyoto University. His research interests include a variety area of Image Processing, Artificial Intelligence and Multimedia Applications, particularly, model centered framework for the computer system to help visual communication among humans and information media structure for human communication. He is a member of Information Processing Society of Japan, Institute of Electronics, Information and Communication Engineers of Japan, the IEEE Computer Society and Communication Society, and ACM.