Collaborative Mean Attraction for Set Based Recognition

Yang Wu[1,a)] Masayuki Mukunoki[1,b)] Michihiko Minoh[1,c)]

1. Introduction

Recently, set based recognition (recognition with a set of instances of the same object/class) has attracted a lot of attention, especially on the popular tasks like face recognition [1][3][8] and person re-identification [5][11][6]. Existing methods can be classified into two groups based on how they treat the training data in the classification process: independent models and collaborative models. Independent models compute an independent set-to-set distance between the test set and each training set, and then classify the test set by such distances. Representative methods include Minimum Point-wise Distance (MPD) [2], Affine/Convex Hull based Image Set Distance (AHISD/CHISD) [1], Sparse Approximated Nearest Points (SANP) [12] and its kernel version KSANP [3], Set Based Discriminative Ranking (SBDR) [8] which iterates between CHISD or SANP and metric learning, and Regularized Nearest points (RNP) [11]. Collaborative models treat all the training sets together as a large indiscriminate set and compute only one single geometric distance between it and the test set, which is also referred to as set-to-sets distance [6]. Existing collaborative models are so far only Collaborative Sparse Approximation (CSA) [5] and Collaboratively Regularized Nearest Points (CRNP) [6], as well as their extended versions [7][10]. Compared with similar independent models, collaborative ones are not only more effective but also much more efficient.

In this paper we propose a novel collaborative model named Collaborative Mean Attraction (CMA)[*1], which is simpler and more effective than both CSA and CRNP. Unlike them, CMA does not rely on affine/convex hulls but just uses simple l_2-norm based regularization terms to make the linear combinations over both sets (the indiscriminate training set and the test set) to be close to the means of them as much as possible. The model itself has only 2 parameters for balancing the attraction and pulling back, which are much fewer than those in CSA and CRNP. Meanwhile, CMA inherits the efficiency from CRNP.

[1] Academic Center for Computing and Media Studies, Kyoto University, Kyoto, 606-8501, Japan
a) yangwu@mm.media.kyoto-u.ac.jp
b) mukunoki@mm.media.kyoto-u.ac.jp
c) minoh@mm.media.kyoto-u.ac.jp
[*1] Matlab code is available at `http://www.escience.cn/people/yangwu/publication.html`.

2. Collaborative Mean Attraction

2.1 Optimization of the coefficients

Given a test set $\mathbf{Q} \in \mathbb{R}^{m \times N_q}$ and all the training sets $\mathbf{X} \in \mathbb{R}^{m \times N_x}$ with $\mathbf{X} = \cup \mathbf{X}_i, i \in \{1, \ldots, n\}$, where m is the feature dimensionality, N_q and N_x are the number of samples in $\mathbf{Q}$ and $\mathbf{X}$, respectively, and n is the number of classes, CMA solves $\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} f(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with the objective function

$$f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \|\mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \left\| \boldsymbol{\alpha} - \frac{\mathbf{1}_{N_q,1}}{N_q} \right\|_2^2 + \lambda_2 \left\| \boldsymbol{\beta} - \frac{\mathbf{1}_{N_x,1}}{N_x} \right\|_2^2, \quad (1)$$

where λ_1 and λ_2 are two trade-off parameters, and $\mathbf{1}_{i,j}$ denotes the $i \times j$ dimensional matrix of ones. $\|\mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ is the distance between two linear combinations which can be viewed as the distance between two generalized means (with unequal weights), and the other two terms forces them to be not too far away from the actual means.

We follow [6] on alternatively optimizing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which avoids the time-consuming matrix inverse (or pseudo-inverse) operation of an integrated matrix containing both $\mathbf{Q}$ and $\mathbf{X}$ for each test/query set $\mathbf{Q}$. In the alternative optimization, the matrix inverse operation on the training data $\mathbf{X}$ is independent of $\mathbf{Q}$, so it can be pre-computed before testing and reused for each $\mathbf{Q}$. Details are omitted here due to the space limitation.

2.2 Classification

The collaborative distance finding implicitly makes $\boldsymbol{\beta}^* = [\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_n^*]$ discriminative. Following [6], we define the dissimilarity between $\mathbf{Q}$ and $\mathbf{X}_i, i \in \{1, \ldots, n\}$ as

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

where $\|\mathbf{Q}\|_*$ and $\|\mathbf{X}_i\|_*$ are the nuclear norms (i.e. the sum of the singular values) of $\mathbf{Q}$ and $\mathbf{X}_i$, respectively. Then, $\mathbf{Q}$ is classified by

$$C(\mathbf{Q}) = \arg\min_i \left\{ d_{CMA}^i \right\}. \quad (2)$$

This classification model benefits from the discriminative power of $\boldsymbol{\beta}^*$, which tends to make the class-specific reconstruction residual $\|\mathbf{Q}\boldsymbol{\alpha}^* - \mathbf{X}_i\boldsymbol{\beta}_i^*\|_2^2$ smaller and $\|\boldsymbol{\beta}_i^*\|_2^2$ larger for the ground-truth label i than any other labels $j \in \{1, \ldots, n\}, j \neq i$.

3. Experiments and Results

3.1 Experimental settings

We make our experiments exactly follow those presented in [6] as they are so far most comprehensive and up-to-date. All the related state-of-the-art methods are compared with using exactly the same experimental settings if possible. For KSANP and SBDR, only the results listed for the same tasks in their original papers are included, while for the others, we get the results by running their codes from their authors (for SRC, CHISD, SANP, CSA, and CRNP) or implemented by ourselves (for MPD, CRC and RNP). We used the parameters recommended in their original papers for all the other methods, while for CMA, the detailed settings are shown in Table 1. Values in Table 1 are just found to be good enough without fine tuning, so there might be better choices. The actual performances are not sensitive to these parameters in large ranges, and it should be easy to find good values for them given any new dataset or task.

Table 1 Parameter setting of CMA for all the experiments.

Experiment	λ_1	λ_2	R_{th}	T_{th}
Honda/UCSD (50 frames)	8	16	0.01	15
Honda/UCSD (100 frames)	12	24	0.01	15
CMU MoBo (50 frames)	0.1	0.2	0.01	15
CMU MoBo (100 frames)	0.4	0.8	0.01	15
iLIDS-MA	4	20	0.01	15
iLIDS-AA	10	2	0.01	15
CAVIAR4REID	2	10	0.01	15

3.2 Experimental results and analysis

The results for all the concerned methods on seven experiments are listed in Table 2. To make the comparison as fair as possible, we tried our best to generate the results for other methods with exactly the same data splits as those for CMA, when it is doable. Meanwhile, we also cite the originally reported results for those methods which have been tested on the same datasets (though the experimental data may be different). It is worth noticing that the referred results for SBDR on CMU MoBo are not counted for performance competition as they 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.

It is clear that the proposed CMA method generally outperforms all the other methods in terms of recognition accuracy. The only weakness (more precisely it should be just not a superiority) is that it cannot make the best out of rich samples for each class/set (such as 100 frames for Honda/UCSD and CMU MoBo). Like CRNP, it indicates that too large set size may weaken the discriminative power of the l_2-norm based collaborative distance finding as it is more likely that some samples from different classes may replace the correct class in collaborative representation [6] with a smaller distance to the test set. Therefore, a pre-selection step may be helpful for reducing such a risk, as suggested in [7] and [10]. It could be an interesting future work.

Table 2 Recognition accuracy (%) comparison. The results with stars are directly copied from their original papers, while those without stars are got from our experiments. The best ones are shown in bold (results from our experiments and results from cited papers are compared separately, and if the best ones from cited results are better than those from our experiments they are marked in bold as well). The ones which do not exist in cited papers are stated as "not available (N/A)".

Experiment	Honda/UCSD		CMU MoBo		iLIDS-	iLIDS-	CAVIAR-
	50fs	100fs	50fs	100fs	MA	AA	4REID
MPD[2]	79.5	87.2	92.2	94.3	50.0	23.8	19.0
SRC[4]	76.9	**94.9**	88.9	92.4	57.3	36.0	25.4
CRC[13]	76.9	82.1	89.7	93.1	28.5	24.7	16.6
CHISD	79.5	79.5	90.8	94.2	52.5	24.6	25.4
CHISD*[1]	82.1*	84.6*	N/A	N/A	N/A	N/A	N/A
SANP	84.6	89.7	90.1	93.6	46.8	19.2	25.2
SANP*[12]	84.6*	92.3*	N/A	N/A	N/A	N/A	N/A
KSANP*[3]	87.2*	94.9*	N/A	N/A	N/A	N/A	N/A
SBDR*[8]	87.7*	89.2*	*95.0**	*96.1**	N/A	N/A	N/A
CSA[5]	84.6	92.3	86.3	94.4	59.0	22.5	24.6
RNP	66.7	92.3	91.8	**94.6**	53.3	25.5	24.0
RNP*[11]	87.2*	94.9*	91.9*	**94.7***	N/A	N/A	N/A
CRNP	89.7	**94.9**	93.5	94.4	59.3	35.5	27.0
CRNP*[6]	89.7*	**97.4***	93.3*	94.4*	59.0*	35.4*	26.8*
CMA	**92.3**	**94.9**	**94.4**	**94.6**	**61.3**	**36.1**	**28.4**

It is worth mentioning that when there is only one concerned object in a single frame of video records, CMA itself can finish recognizing more than 370 frames per second with feature vectors of 400 dimensions or even more, which is about 15 times faster than real time (suppose the video has a 25 fps frame rate). This is a good news for people who care about processing speed in real applications.

References

[1] Cevikalp, H. and Triggs, B.: Face recognition based on image sets, *CVPR*, pp. 2567–2573 (2010).

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

[3] Hu, Y., Mian, A. S. and Owens, R.: Face Recognition Using Sparse Approximated Nearest Points between Image Sets, *IEEE TPAMI*, Vol. 34, No. 10, pp. 1992–2004 (2012).

[4] Wright, J., Yang, A., Ganesh, A., Sastry, S. and Ma, Y.: Robust Face Recognition via Sparse Representation, *IEEE TPAMI*, Vol. 31, No. 2, pp. 210–227 (2009).

[5] Wu, Y., Minoh, M., Mukunoki, M., Li, W. and Lao, S.: Collaborative Sparse Approximation for Multiple-Shot Across-Camera Person Re-identification, *AVSS*, pp. 209 –214 (2012).

[6] Wu, Y., Minoh, M. and Mukunoki, M.: Collaboratively Regularized Nearest Points for Set Based Recognition, *BMVC*, pp. 1–10 (2013).

[7] Wu, Y., Minoh, M. and Mukunoki, M.: Locality-constrained Collaborative Sparse Approximation for Multiple-shot Person Re-identification, *ACPR*, pp. 140–144 (2013).

[8] Wu, Y., Minoh, M., Mukunoki, M. and Lao, S.: Set Based Discriminative Ranking for Recognition, *ECCV* , pp. 497–510 (2012).

[9] Wu, Y., Mukunoki, M., Funatomi, T., Minoh, M. and Lao, S.: Optimizing Mean Reciprocal Rank for person re-identification, *AVSS*, pp. 408 –413 (2011).

[10] Wu, Y., Mukunoki, M. and Minoh, M.: Locality-constrained Collaboratively Regularized Nearest Points for Multiple-shot Person Re-identification, *FCV* (2014).

[11] Yang, M., Zhu, P., Gool, L. V. and Zhang, L.: Face Recognition based on Regularized Nearest Points between Image Sets, *FG* (2013).

[12] Yiqun Hu and Ajmal S. Mian and Robyn Owens: Sparse Approximated Nearest Points for Image Set Classification, *CVPR*, pp. 121–128 (2011).

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