Robust Object Recognition via Third-Party Collaborative Representation

Yang Wu, Michihiko Minoh, Masayuki Mukunoki
Academic Center for Computing and Media Studies, Kyoto University, Kyoto 606-8501, Japan
yangwu@mm.media.kyoto-u.ac.jp, {minoh,mukunoki}@media.kyoto-u.ac.jp

Shihong Lao
OMRON Social Solutions Co., LTD, Kyoto 619-0283, Japan
lao@ari.ncl.omron.co.jp

Abstract

A simple and effective method is proposed for object recognition via collaborative representation with ridge regression. Different from existing sparse representation and collaborative representation based approaches, the proposal does not need extensive training samples for each testing class and it is robust to localization errors and large within-class variations, thus being applicable to various real-world object recognition tasks instead of handling only the well-controlled face recognition problem. Its discriminative power is explored from a third-party dataset which can be different from the training and testing datasets, therefore, it enables using an existing dictionary for testing new data without time-consuming data annotation and model re-training. As an example, the proposal is extensively tested on the representative and very challenging task of person re-identification, defining novel state-of-the-art results on widely adopted benchmark datasets using only simple and common features.

1. Introduction

We introduce a novel method for robust generic object recognition leveraging the great discriminative power of l_2-norm based collaborative representation [9] which is simplified from the research on sparse representation for classification [7]. Sparse representation has been widely used for acquiring, representing, compressing and restoring high-dimensional signals, with a recent successful re-exploration for extracting semantic information from images. Despite its striking performance on image content recognition especially face recognition, there are two preconditions for it to work properly: the training images have to be carefully controlled/captured and the number of samples per class should be sufficiently large [7]. In this paper, however, we relax these preconditions, simplify the model, and reduce its dependence on the quality of labeled data, to make it an efficient, robust, and easily implementable technique. It can be applied to various practical object recognition tasks with various tentative industrial applications, including surveillance and safety, image/video search and retrieval, robotics, education and tourism.

Our method collaboratively represents each sample in the query and gallery sets by a common "dictionary" of samples, which can be either from the same classes or different classes with respect to the training (gallery) and testing (query) data as long as they can be represented by the same types of features enabling linear combination and approximation. Moreover, this dictionary may or may not have the same imaging conditions (such as camera parameters, illumination, background, weather, etc.) as those in the training and testing sets. Then, the representation coefficients for each query/gallery sample are summed up for each class in the dictionary, followed by a normalization over all the classes. We will demonstrate that the final representation is in fact discriminative and robust for recognition.

We test our method on person re-identification, which is considered to be one of the most challenging and representative object recognition problems [1]. Though it may not be necessary, a common setting of person re-identification is to re-identify persons traveling across cameras with overlapping or non-overlapping views, thus may contain great within-class variations such as viewpoint, pose, illumination, and background changes, as well as self-occlusions and occlusions. Our method is applicable to both single-shot and multiple-shot scenarios [2] with either a single sample or multiple samples for each person. Therefore, we report experimental results on all the widely-adopted benchmark datasets for person re-identification to demonstrate the priority and properties of the proposed method.

2. Sparse representation for classification

Recently, sparse representation has attracted much attention since it was proven to be very effective for classification, more specifically face recognition [7]. It works as follows. Given a training dataset $X = [X_1, \ldots, X_K] \in \mathbb{R}^{d \times n}$ with totally n images belonging to K classes in the d-dimensional space, where X_i denotes the samples of class i, then a test sample $q \in \mathbb{R}^d$ is classified by seeking a sparse representation of it in terms of the dictionary X and finding the class $y \in \{1, \ldots, K\}$ that can best approximate q by a linear combination of its samples with their corresponding sparse coefficients. It can be formulated as algorithm 1.

Algorithm 1 SPARSE REPRESENTATION-BASED CLASSIFICATION (SRC):

Require: A matrix of training samples $X = [X_1, \ldots, X_K] \in \mathbb{R}^{d \times n}$ belonging to K classes, a test sample $q \in \mathbb{R}^d$, and an error tolerance $\varepsilon > 0$.

Ensure: The identity of q: $y(q) \in \{1, \ldots, K\}$.

1: Normalize the columns of X to have unit l_2-norm.
2: Solve the constrained l_1-minimization problem: $\hat{\alpha} = \arg\min_\alpha \|\alpha\|_1 \ \ s.t. \ \ \|q - X\alpha\|_2^2 < \varepsilon$, where $\hat{\alpha} = [\hat{\alpha}_1, \ldots, \hat{\alpha}_K]$.
3: Compute the representation residuals: $r_i(q) = \|q - X_i\hat{\alpha}_i\|_2^2, \forall i \in \{1, \ldots, K\}$, where $\hat{\alpha}_i$ is the coefficient vector associated with class i.
4: Judge the identity of q: $y(q) = \arg\min_i r_i(q)$.

As remarked in [7], SRC has two important preconditions for ensuring a good performance: a) the training images have been carefully controlled, and b) the number of samples per class is sufficiently large. Therefore, it cannot be applied to recognition tasks with unconstrained training data which is actually quite common in real applications. Though it has been followed by quite a few publications, as far as we are aware, none of them has managed to significantly relax these preconditions for solving practical object recognition problems.

3. Third-party collaborative representation

The two preconditions of SRC actually aim at making the dictionary for every class (i.e., $X_i, i \in \{1, \ldots, K\}$) over-complete and covering as many within-class variations as possible, so that for any query/test sample q belonging to class y, $q \approx X_y\hat{\alpha}_y$. By doing so, it is expected that due to the l_1-minimization objective, most coefficients in $\hat{\alpha}_k, k \neq i$ are close to zero while only $\hat{\alpha}_i$ has significant entries.

Though it sounds true, we believe that for many real-world tasks, enforcing the over-completeness and well within-class coverage of X_i is not only impractical but also not necessarily leading to the expected highly selective $\hat{\alpha}$. The reason is that the two preconditions of SRC only focus on the quality of X_i for class i, but not on the discrimination between X_i and $X_k, k \neq i$ during sparse representation. Namely, though you may find a very good X_i, there may be another $X_k, k \neq i$ that can also sparsely and faithfully represent the test sample q from class i with $r_k(q) < r_i(q)$. It is argued in [9] that collaborative representation but not the sparsity itself is the key for effectiveness of sparse representation. Though we agree with it, we still think the simplified CRC_RLS (collaborative representation based classification with regularized least square) algorithm hasn't well explored the power of $\hat{\alpha}$ as it still sticks to classification by representation residuals $r_i(q), i \in \{1, \ldots, K\}$. Instead, we propose to compute $\hat{\alpha}$ on a third-party dictionary and use it as a feature descriptor for recognition. Therefore, the method is named Third-Party Collaborative Representation (TPCR) in contrast to the traditional representation on the training data.

3.1. Third-party dictionary

Instead of using the training dataset as the dictionary for collaborative representation, we propose to adopt a third-party dictionary which may have similar but different data with respect to the training and testing data. Taking person re-identification for example, we still follow the standard setting of splitting a benchmark dataset D for training and testing, but use another dataset D_{tp} which contains images of different persons as our dictionary. A good dictionary is expected to have a certain amount of persons with enough instances for each of them covering the within-class variations of the testing dataset. This is to make it possible that for any given test sample, there exists a collaborative representation of it by a few samples of the dictionary with alike visual aspects (e.g. viewpoint and pose) from few similar persons. However, as it will be witnessed in our experiments, such a preference is not mandatory, and TPCR has some robustness to the quality of the dictionary.

3.2. Within-class summarization

After the collaborative representation of a query q over the third-party dictionary $D_{tp} = [D_{tp}^1, \ldots, D_{tp}^L]$ containing samples from L different classes, and suppose $\hat{\alpha} = [\hat{\alpha}_1, \ldots, \hat{\alpha}_L]$ are the corresponding linear combination coefficients, we sum up each $\hat{\alpha}_i, i \in \{1, \ldots, L\}$ to a scalar $\hat{\beta}_i = \sum_{j=1}^{n_i} \hat{\alpha}_{ij}$ with n_i denoting the number of samples in class i of D_{tp}. Then we normalize $\hat{\beta} = [\hat{\beta}_1, \ldots, \hat{\beta}_L]$ by $\sum_{i=1}^{L} \left|\hat{\beta}_i\right|$ and use the normalized $\hat{\beta}$ as a feature descriptor for q. The within-class summarization shares the power of being robust

to within-range changes with the strategy of using histograms for feature description. Since the range here is a class, it can handle within-class variations without explicit modeling of them, which is not a trivial task.

3.3. The TPCR algorithm

Inspired by CRC_RLS [9], we use l_2-minimization for collaborative representation as it has the same discrimination power as that of l_1-minimization but has a much more efficient close-form solution. The detailed algorithm of TPCR is presented in algorithm 2. Unlike SRC, it serves as a novel feature descriptor but not a classification model, therefore any existing classification methods can be used for the final recognition.

Algorithm 2 THIRD-PARTY COLLABORATIVE REPRESENTATION (TPCR):

Require: A dataset $D \in \mathbb{R}^{d\times n}$ for training and test, a third-party dataset $D_{tp} = [D_{tp}^1, \ldots, D_{tp}^L] \in \mathbb{R}^{d\times m}$ of L classes, and a regularization parameter λ.

Ensure: A collaborative representation based description $\hat{\beta}(q)$ for each sample $q \in D$ over D_{tp}.

1: Normalize the columns of D and D_{tp} to have unit l_2-norm.

2: Solve the constrained l_2-minimization problem: $\hat{\alpha} = \arg\min_\alpha\{\|q - D_{tp}\alpha\|_2^2 + \lambda\|\alpha\|_2^2\}$, with a close form solution $\hat{\alpha} = Pq$, where $P = \left(D_{tp}^T D_{tp} + \lambda \cdot I\right)^{-1} D_{tp}^T$. Note that P can be precomputed once D_{tp} is given.

3: Compute the summed coefficients: $\hat{\beta}_i(q) = \sum_{j=1}^{n_i} \hat{\alpha}_{ij}, \forall i \in \{1, \ldots, L\}$.

4: Normalize $\hat{\beta}(q)$: $\hat{\beta}'(q) = \hat{\beta}(q)/\sum_{i=1}^{L} \left|\hat{\beta}_i(q)\right|$.

5: Return $\hat{\beta}(q) = \hat{\beta}'(q)$.

4. Experimental results

Though TPCR is generically applicable, we choose the challenging task of person re-identification as an example to demonstrate its superiority.

Table 1. Dataset properties.

Dataset	NS	NP	NSPP	Seq?	WCV
$VIPeR$ [5]	1264	632	1×2	No	CVIPOB
$iLIDS$ [6]	476	119	2 to 8	Partly	CVIPOB
$iLIDS-MA$ [1]	3680	40	46×2	Yes	CVIPOB
$iLIDS-AA$ [1]	10329	100	21 to 243	Yes	CVIPOBL
$ETHZ-Seq1$ [2]	4857	83	7 to 226	Yes	POB
$ETHZ-Seq2$ [2]	1961	35	6 to 206	Yes	POB
$ETHZ-Seq3$ [2]	1762	28	5 to 356	Yes	POB

Datasets. Seven datasets with different properties are adopted, as shown in Table 1. The meanings of the abbreviations in it are as follows. NS–number of samples, NP–number of persons, NSPP–number of samples per person, Seq?–sequential or not, WCV–within-class variations. Abbreviations for WCV are: C–camera parameters, V–viewpoint, I–illumination, P–pose, O–occlusion, B–background, and L–localization.

Method comparison. To show the effectiveness of TPCR, we just use the minimum point-wise Euclidean distance between a query set and a gallery set (denoted by "MPD") as the distance measurement for them without discriminative learning. Based on "MPD", we rank all the gallery sets for the person re-identification problem. "MPD" has been widely used for multiple-shot person re-identification [4], and when it is applied to single-shot cases it degrades to the Euclidean distance. For simplicity, we use "TPCR" to denote our method without mentioning the distance measurement, and use "MPD" to stand for the method working with raw image features instead of "TPCR". To show the power of the third-party dictionary, we also compare with "MCRC_RLS", which denotes the extended CRC_RLS algorithm for handling mutiple-shot cases by accumulating the individual residuals of samples in a query set for identification. The recent work of "CHISD" [3] with good performance on set-based face recognition is also compared with here, because it was designed to be less over-fitting and more robust to outliers than MPD. Besides of these approaches for algorithmic comparison, we also compare with the methods which are quite different but currently hold the state-of-the-art performance on each dataset. They include "OMRR"[8], "PRDC" and "MCC"[6] which are discriminative learning based methods, and "MRCG"[1] which is an sophisticated hand-crafted descriptor.

Experimental settings. We normalized all the images to 128×48 pixels as in [5] and adopted two commonly used features as the raw image features: densely sampled color histograms[8] and texture descriptor with Schmid and Gabor filters[6]. The features are mapped by PCA into a 400-dimensional space. For TPCR, since there are six third-party datasets when working on any one of the seven datasets, we select some typical combinations for D_{tp} based on the dataset properties, and add the information to the name of "TPCR" for comparison. "TPCR" on its own denotes the case that the dictionary is just a part of working dataset, which also serves as the training set for other learning-based methods. Please refer to the legends in Figure 1. Whenever random sampling is involved, the experiments were repeated 10 times to average the results.

Results and discussions. Since the results on the three ETHZ datasets are getting saturated [1], we omit the details of comparisons here. Briefly, TPCR with either the other two ETHZ datasets or the significantly different and noisy iLIDS-AA dataset as its dictionary achieves

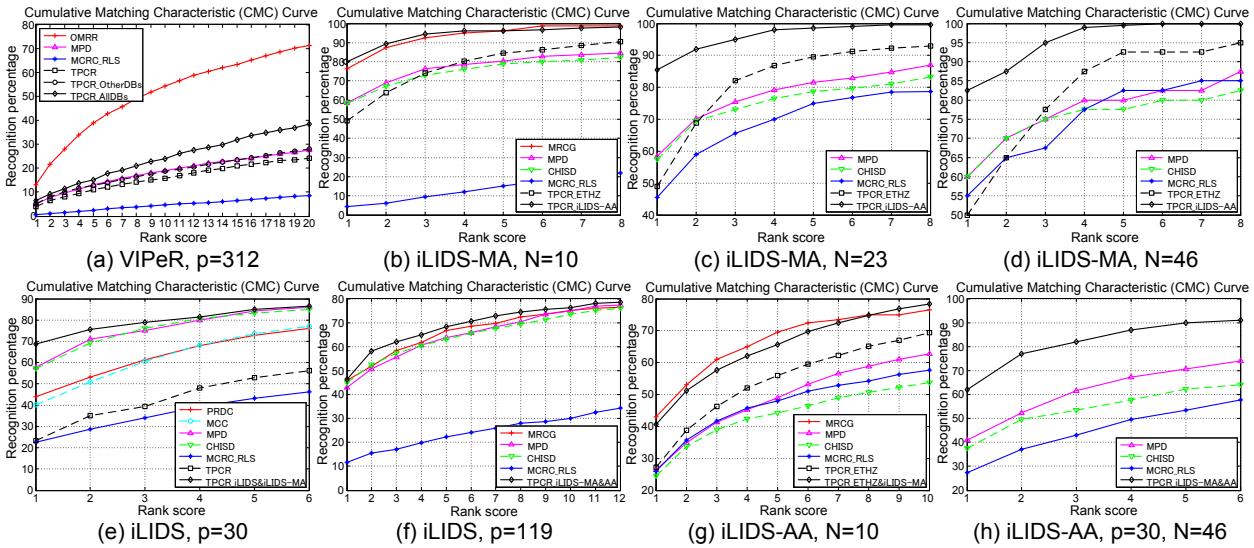

Figure 1. Experimental results in CMC curves. p is the number of randomly selected persons for testing. N is the number of randomly selected samples for each query/gallery set for multiple-shot re-identification. When p or N is not specified, it takes the maximum feasible value.

almost perfect results (over 99.2% on rank 1 for all 3 datasets, and 100% for ETHZ-Seq2 and ETHZ-Seq3), superior to any other methods including MRCG[1]. The results for other datasets are shown in Figure 1. Overall, TPCR with both similar and non-similar datasets as its dictionary gets higher performance than MPD, CHISD and MCRC_RLS, and it also outperforms the state-of-the-art results of methods with much high complexity. One exception is the VIPeR dataset as it has too few samples per person (only 2) with very large within-class variations so that it is very hard to find a good third-party dataset to cover all these variations. Another one is the iLIDS-AA dataset which has significant localization errors caused by automatic tracking, even though, with only ETHZ and iLIDS-MA datasets as its dictionary and no specific treatment for handling misalignment, the result of TPCR is still comparable to MRCG which explicitly deals with localization errors.

5. Conclusions and future work

In this paper, we address the object recognition problem via collaborative representation on a third-party dictionary. This novel method properly utilizes and extends the power of sparse representation while at the same time alleviates its demanding preconditions, thus making it applicable to real-world object recognition tasks. Without using any learning-based classifier or sophisticated image features, the proposed method already significantly outperforms other methods on person re-identification in real scenarios. Future work includes finding the optimal dictionary for a given data and exploring learning to further boost the performance.

Acknowledgments 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] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid. In *AVSS*, 2011.

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

[3] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573, 2010.

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

[5] D. Gray and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on PETS*, pages 41–47, 2007.

[6] Wei-Shi Zheng and Shaogang Gong and Tao Xiang. Person Re-identification by Probabilistic Relative Distance Comparison. In *CVPR*, pages 649–656, 2011.

[7] J. Wright, Y. Ma, J. Mairal, G. Spairo, T. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 2010.

[8] Y. Wu, M. Mukunoki, T. Funatomi, M. Minoh, and S. Lao. Optimizing Mean Reciprocal Rank for Person Re-identification. In *AVSS*, 2011.

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