With One Look: Robust Face Recognition Using Single Sample Per Person

De-An Huang
Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
andrew800619@gmail.com

Yu-Chiang Frank Wang
Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
ycwang@citi.sinica.edu.tw

ABSTRACT

In this paper, we address the problem of robust face recognition using single sample per person. Given only *one* training image per subject of interest, our proposed method is able to recognize query images with illumination or expression changes, or even the corrupted ones due to occlusion. In order to model the above intra-class variations, we advocate the use of external data (i.e., images of subjects *not* of interest) for learning an exemplar-based dictionary. This dictionary provides auxiliary yet representative information for handling intra-class variation, while the gallery set containing one training image per class preserves separation between different subjects for recognition purposes. Our experiments on two face datasets confirm the effectiveness and robustness of our approach, which is shown to outperform state-of-the-art sparse representation based methods.

Categories and Subject Descriptors

I.4.9 [**Image Processing & computer vision**]: Applications; I.5.4 [**Pattern Recognition**]: Applications—*Computer Vision*; H.3.3 [**Information Storage & Retrieval**]: Information Search & Retrieval

General Terms

Algorithms, Experimentation, Performance

Keywords

Face recognition, sparse representation, low-rank matrix decomposition, affinity propagation

1. INTRODUCTION

Recognizing faces in real-world scenarios not only requires one to deal with illumination or expression changes, the face images to be recognized might also be corrupted due to occlusion or disguise. Solving the above task is typically known as *robust face recognition* [10]. Although very promising performance has been reported in recent works like [10, 2], their requirement of collecting a large number of training data might not be practical. If a sufficient amount of training face images cannot be obtained for modeling intra and inter-class variations, one cannot expect satisfactory recognition performance. For real-world face recognition, one might have only *one* training image for each subject of interest (e.g., a mug shot). Therefore, how to address single-sample robust face recognition has been a challenging problem for researchers in related fields.

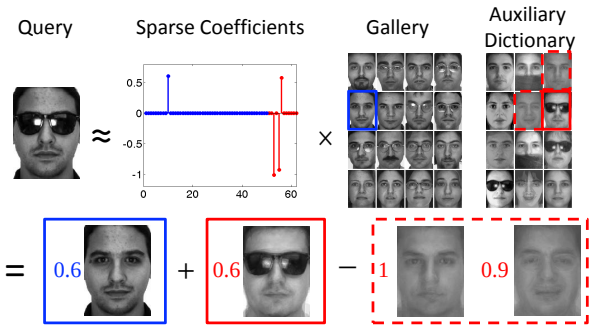

Figure 1: Illustration of our proposed method. Note that the gallery set contains only one training image per subject of interest, while the auxiliary dictionary to be learned utilizes external data for observing possible image variants (including occlusion).

For single-sample face recognition, Zhu *et al.* [12] proposed a multi-scale patch based collaborative representation (PCRC) approach. Since PCRC performs recognition based on patch-wise reconstruction error, it would be sensitive to corrupted face images (e.g., images with sunglasses). Recently, the use of *external* data (i.e., images collected from subjects *not* of interest) is considered as an alternative for solving single-sample face recognition problems. For example, Su *et al.* [9] proposed adaptive generic learning (AGL) for deriving a discriminant model using external data, while one training image per subject is available. Inspired by sparse representation based classification (SRC) [10], Deng *et al.* [4] presented extended SRC (ESRC), which considered external face data as an additional dictionary for modeling intra-class variations. Although both AGL and ESRC utilized external face data for modeling inter or intra-class variations, their direct use of external data might *not* be preferable, since such data might contain noisy, redundant, or undesirable information. Without proper selection or processing of external data, the direct use of such data does not necessarily improve the performance.

In this paper, we present a novel approach for solving single-sample face recognition. In order to model *intra*-class

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502158.

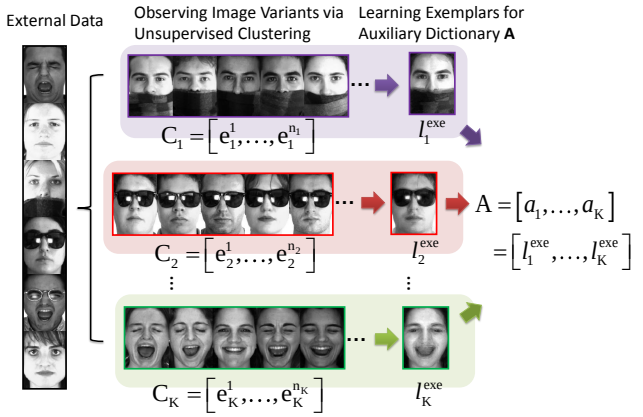

Figure 2: Learning of an exemplar-based auxiliary dictionary A from external face images for modeling intra-class variations.

variations, we propose the learning of an exemplar-based dictionary using external data. Based on the techniques of affinity propagation [5] and low-rank matrix decomposition [1], the derived dictionary is able to extract representative information from external data in terms of *image variants* instead of subject identities. Thus, each dictionary atom corresponds to a particular image variant like illumination, expression, or occlusion type. As illustrated in Fig. 1, together with gallery set images (only one per subject of interest), we apply the observed auxiliary dictionary for performing recognition via ESRC [4]. One of the advantages of our approach is that our dictionary size only depends on the number of types of image variants, *not* the size of external data. Our experiments will verify both the effectiveness and robustness of our method over state-of-the-art recognition approaches.

2. A BRIEF REVIEW OF SRC AND ESRC

Proposed by Wright *et al.* [10], sparse representation based classification (SRC) performs recognition by taking each test image $\mathbf{y}$ as a sparse linear combination of atoms in an overcomplete dictionary $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_k]$, where $\mathbf{D}_i$ contains the training images of class i. SRC calculates the sparse coefficient $\boldsymbol{\alpha}$ of $\mathbf{y}$ by solving:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1. \tag{1}$$

Once (1) is solved, the label j of $\mathbf{y}$ is determined by

$$j = \arg\min_i \|\mathbf{y} - \mathbf{D}\delta_i(\boldsymbol{\alpha})\|_2, \tag{2}$$

where $\delta_i(\boldsymbol{\alpha})$ is a vector whose nonzero entries are those associated with class i only. Thus, SRC performs recognition based on the minimum class-wise reconstruction error, which implies the query $\mathbf{y}$ approximately lies in the column subspace spanned by the training images of the associated class.

In practice, the use of SRC is limited due to its need to collect of a large amount of training data as the over-complete dictionary $\mathbf{D}$. To address this concern, Deng *et al.* [4] proposed Extended SRC (ESRC) by solving:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\| \mathbf{y} - [\mathbf{D},\ \mathbf{A}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_2^2 + \lambda \left\| \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_1, \tag{3}$$

where $\mathbf{A}$ is an auxiliary dictionary modeling the intra-class variants, and $\boldsymbol{\beta}$ is the associated sparse coefficient. Different from $\mathbf{D}$, the auxiliary dictionary $\mathbf{A}$ in [4] consists of images collected from external data (i.e., subjects *not* of interest). Similar to SRC, ESRC performs recognition by:

$$j = \arg\min_i \left\| \mathbf{y} - [\mathbf{D},\ \mathbf{A}] \begin{bmatrix} \delta_i(\boldsymbol{\alpha}) \\ \boldsymbol{\beta} \end{bmatrix} \right\|_2. \tag{4}$$

Although modeling intra-class variants by utilizing external data has been shown to achieve improved performance for undersampled face recognition problems, applying all images from an external dataset might not be preferable. This is not only because that redundant or noisy information might be taken into account, the size of the auxiliary dictionary $\mathbf{A}$ will linearly increase with that of external data, which would make (4) very computationally expensive to solve.

3. AUXILIARY DICTIONARY LEARNING

3.1 Observing Image Variants

As depicted in Fig. 2, we propose to learn an exemplar-based dictionary from external face images. To model intra-class variations during recognition, each dictionary atom is expected to correspond to a particular type of image variant. As a result, our goal is to automatically identify different types of image variants from external data which contains possible variations (e.g., illumination, expression, or occlusion changes), so that we can extract representative information and derive the corresponding dictionary atoms.

We observe that face images with the same type of corruption/variation have similar distributions in terms of intensity gradients. Thus, we consider the use of Histogram of Oriented Gradients (HOG) [3] features for describing each image. Since it is *not* practical to assume the exact number of variant types to be known in advance, we need an *automatic* and *unsupervised* learning algorithm for solving this task. In our work, we advance affinity propagation (AP) [5], which is a unsupervised clustering technique and not requires the prior knowledge of the cluster numbers. To automatically identify different types of image variants, we solve the following problem which minimizes the net-similarity (NS) between different external face images:

$$\begin{aligned} NS \quad = \quad & \sum_{i=1}^{N}\sum_{j=1}^{N} c_{ij}\, s(\mathbf{e}_i, \mathbf{e}_j) \\ & -\gamma \sum_{i=1}^{N}(1 - c_{ii})(\sum_{j=1}^{N} c_{ij}) - \gamma \sum_{i=1}^{N} |(\sum_{j=1}^{N} c_{ij}) - 1|. \end{aligned} \tag{5}$$

In (5), $s(\mathbf{e}_i, \mathbf{e}_j) = exp(-\|HOG(\mathbf{e}_i) - HOG(\mathbf{e}_j)\|^2)$ measures the similarity between images $\mathbf{e}_i$ and $\mathbf{e}_j$ in terms of HOG features. The coefficient $c_{ij} = 1$ indicates that $\mathbf{e}_i$ is the cluster representative of $\mathbf{e}_j$ (i.e., $\mathbf{e}_j$ is assigned to cluster i). Thus, $c_{ii} = 1$ means that $\mathbf{e}_i$ is the representative and belongs to its own cluster i. The first term in (5) is to calculate the sum of similarity between images within each cluster, while the second term penalizes the case when images are assigned to an empty cluster (i.e. $c_{ii} = 0$ but with $\sum_{j=1}^{M} c_{ij} \geq 1$). The last term in (5) penalizes the cases when images are assigned to more than one cluster, or not belong to any of them. We set the parameter γ to $+\infty$ to strictly avoid the above problems.

It is necessary to verify the effectiveness and practicability of this unsupervised strategy for automatically dividing different image variants into distinct groups (instead of separating images of different subjects into different clusters).

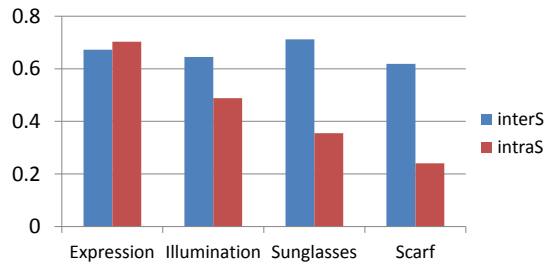

Figure 3: Average inter/intra-class similarity (*interS*/*intraS*) for the AR database using HOG features. The x-axis indicates four types of image variants, and the y-axis is the similarity value. Note that *intraS* measures the similarity between the neutral image and a particular image variant of the same subject, and *interS* is that between a face image and the most similar one of the same variation but from a different subject (chosen by Nearest Neighbor).

We statistically verified this observation on the images of the first session in the AR database [8], and the results are shown in Fig. 3. From this figure, it can be seen that images of the *same* variant type but from *different* subjects generally achieved *higher* similarities than those of different variations but from the same subject (i.e., $interS > intraS$ in Fig. 3, except for expression changes in which the two values are comparable). This supports the use of our approach for identifying and separating face images into different groups in terms of their variant types instead of identities.

3.2 Learning Exemplars for Image Variants

After dividing external face images into different groups in terms of variant types, we need to extract representative information from each cluster, so that such information (and the derived auxiliary dictionary) can be utilized for modeling intra-class variations. To solve this task, we advance low-rank matrix decomposition (LR) [1] to learn exemplars for representing each cluster (and thus variant type).

As discussed in [1], LR seeks to decompose an input data matrix into a low-rank version and a matrix containing the associated sparse error. Since the resulting low-rank matrix can be considered as a compact and representative version of the original input, we advocate the use of LR for learning exemplars for face image variants in Fig. 2. In our work, we solve the following optimization problem:

$$\min_{\mathbf{L}_i,\mathbf{S}_i} \|\mathbf{L}_i\|_* + \lambda \|\mathbf{S}_i\|_1 \quad \text{s.t.} \quad \mathbf{C}_i = \mathbf{L}_i + \mathbf{S}_i. \tag{6}$$

In (6), $\mathbf{C}_i = [\mathbf{e}_i^1, \ldots, \mathbf{e}_i^{n_i}]$ indicates the set of external face images grouped in cluster i (n_i is the number of images in it). The nuclear norm $\|\mathbf{L}_i\|_*$ (i.e., the sum of singular values) approximates the rank of $\mathbf{L}_i$, which makes the optimization problem of (6) convex and thus can be solved by techniques of augmented Lagrange multipliers (ALM) [1, 7]. As a result, we choose to decompose $\mathbf{C}_i$ into a low-rank matrix $L_i = [\mathbf{l}_i^1, \ldots, \mathbf{l}_i^{n_i}]$ and a sparse error matrix $\mathbf{S}_i = [\mathbf{s}_i^1, \ldots, \mathbf{l}_s^{n_i}]$. It is worth noting that, unlike [2] in which LR was applied to *remove* intra-class variations of each subject, our approach aims at extracting representative intra-class information of face images by disregarding their *inter-class* variations.

We now discuss how we learn the auxiliary dictionary by solving the above LR problems. Supposed that the jth image $\mathbf{e}_i^j$ is identified as the centroid of cluster i during the aforementioned unsupervised clustering process (see Sec. 3.1), we consider the corresponding low-rank component $\mathbf{l}_i^j$ as the exemplar $\mathbf{l}_i^{exe}$ for representing that particular type of image variant. Once all exemplars for all clusters (variant types) are obtained, we have the auxiliary dictionary $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K] = [\mathbf{l}_1^{exe}, \ldots, \mathbf{l}_K^{exe}]$.

The advantage of our proposed auxiliary dictionary learning strategy is two-fold. Firstly, we derive the dictionary $\mathbf{A}$ in an automatic and unsupervised way. Since $\mathbf{A}$ contains exemplars describing each type of image variants, it can be applied to model intra-class variations during recognition. Secondly, the size of $\mathbf{A}$ does not grow linearly with the amount of external data; otherwise it would significantly increase the computation time for SRC or ESRC. We also note that, when utilizing external data for learning the auxiliary dictionary, we do *not* require any label/identity information from such data. In other words, our method allows one subject from external data to provide some particular image variants (e.g., illumination changes), while another subject for other types of image variants (e.g., occlusion). This provides additional flexibility and practicability for the use of external data.

Figure 4: Example images of the Extended Yale database (only 16 out of 64 illuminations are shown).

Table 1: Recognition performance on the Extended Yale B dataset. * denotes that the auxiliary dictionary size of ESRC is the same as ours.

SRC [10]	RSC [11]	AGL [9]	ESRC* [4]	Ours
48.4	38.1	50.3	54.7	64.9

3.3 Performing Recognition

Once the auxiliary dictionary $\mathbf{A}$ is learned from external data, we perform recognition of queries $\mathbf{y}$ using ESRC (i.e., (3) and (4)). Recall that the gallery set $\mathbf{D}$ in (3) contains only one training image from each subject of interest. Thus, similar to SRC/ESRC, our recognition rule is also based on the minimum class-wise reconstruction error.

4. EXPERIMENTAL RESULTS

4.1 Extended Yale B Database

We first consider the Extended Yale B database [6] for our experiments. This database contains frontal-face images of 38 subjects, each with about 64 images taken under various illumination conditions. In our experiment, all images are converted into grayscale and cropped to 192×168 pixels. Examples of the images are shown in Fig. 4.

We randomly select 19 from the 38 people as the subjects of interest (i.e., to be recognized), and the rest as external data for learning the auxiliary dictionary. For the 19 subjects of interest, we select the neutral face (A+000 E+00) of each for training (as the gallery), and the remaining 63 images as query images for testing. For the 19 subjects *not* of interest), we randomly select 5 images out of the total 64 images for each subject as external data, and thus $\mathbf{E}$ con-

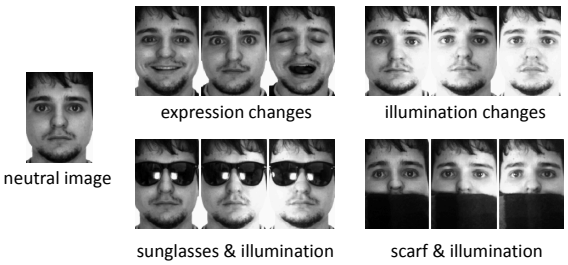

Figure 5: Example images of the AR database. Only the neural image of each subject is in the gallery set, and the rest are the queries to be recognized.

Table 2: Recognition results of the AR database for queries under different scenarios. * denote the use of the same auxiliary dictionary size.

	Expression	Illumination	Sunglasses	Scarf	Avg
SRC [10]	77.3	75.6	47.7	17.5	55.4
RSC [11]	**81.4**	32.7	59.3	34.3	53.1
AGL [9]	62.6	77.0	43.1	27.5	53.0
ESRC* [4]	77.0	82.7	60.2	25.8	62.1
Ours*	78.9	**89.1**	**71.9**	**36.9**	**69.6**

tains a total of $5 \times 19 = 85$ images. The parameter λ in (3) is set to 0.15, and we consider the default parameter choices for AP and LR as their original works do. We perform five random trials, and compare our method with SRC [10], Robust Sparse Coding (RSC) [11], AGL [9] and ESRC [4]. For AGL and ESRC, their gallery and external data selections are the same as ours. We perform five random trials, and list the average recognition rates of each in Table 1.

From Table 1, it can be seen that SRC and RSC were not able to perform recognition well for single-sample recognition without the use of external data. Compared to AGL and ESRC which applied the same size of the external/auxiliary data, our proposed method achieved the highest accuracy. This supports the use of our derived auxiliary dictionary for better modeling intra-class variations, and our proposed method for single-sample face recognition.

4.2 AR Database

The AR database [8] contains over 4,000 frontal images of 126 individuals taken under different variations, including illumination, expression, and facial occlusion. The AR images were taken in two separate sessions. For each session, thirteen images are available for each subject, see Fig. 5 for example. In our experiment, we choose a subset of AR consisting of 50 men and 50 women as [10] did, and all images are cropped to 165×120 pixels and converted to grayscale.

Among the 100 subjects, we randomly select 50 from them to be recognized (25 men and 25 women), and the remaining 50 as external data for auxiliary dictionary learning. For the subjects of interest, only the first *neutral* image is used for training, and the rest 25 images for testing (12 remaining from the first session and 13 from the second). To learn the auxiliary dictionary $\mathbf{A}$, we randomly select 2 images from each subject not of interest, and thus $\mathbf{E}$ contains a total of $2 \times 50 = 100$ images.

Similar to our experiments on the Extended Yale B dataset, we consider SRC, RSC, AGL, and ESRC for comparisons. We also perform five random trials (for external data selection) and compare the average recognition results of different approaches (see Table 2). From this table, we see that our approach outperformed others for most cases, except for the case of image variants with expression changes. This is consistent with our observation in Fig. 3, in which inter-class similarity for facial expression is actually lower than that for other types of facial variants. Although RSC is particularly designed for handle such expression variations [11]), it cannot be generalized well to other variant types such as illumination or occlusion variations (like ours does). From the above experiments, the robustness and effectiveness of our proposed method for single-image face recognition can be successfully verified.

5. CONCLUSION

We presented a novel ESRC-based approach for solving single-sample face recognition problems. In order to handle face images of different variations or occlusions, we learned an exemplar-based dictionary as the auxiliary dictionary. By observing external face data, this dictionary was able to automatically identify and thus model intra-class variations and corruptions. Together with the gallery set containing only one training image per subject of interest, our proposed method was shown to preserve inter-class variations, while intra-class variations can be well observed. Experimental results on two face databases confirmed the effectiveness of our method, which was shown to outperform state-of-the-art with or without using external face data.

Acknowledgement This work is supported in part by National Science Council of Taiwan via NSC100-2221-E-001-018-MY2.

6. REFERENCES

[1] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Jour. ACM*, 58, 2011.

[2] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang. Low-rank matrix recovery with structural incoherence for robust face recognition. In *Proc. IEEE CVPR*, 2012.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection.

[4] W. Deng, J. Hu, and J. Guo. Extended SRC: Undersampled face recognition via intraclass variant dictionary. *IEEE PAMI*, 2012.

[5] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 2007.

[6] A. S. Georghiades et al. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE PAMI*, 2001.

[7] Z. Lin et al. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC Tech. Rep. UILU-ENG-09-2215*, 2010.

[8] A. M. Martinez and R. Benavente. The AR face database. *CVC Technical Report*, 1998.

[9] Y. Su, S. Shan, X. Chen, and W. Gao. Adaptive generic learning for face recognition from a single sample per person. In *Proc. IEEE CVPR*, 2010.

[10] J. Wright et al. Robust face recognition via sparse representation. *IEEE PAMI*, 2009.

[11] M. Yang et al. Robust sparse coding for face recognition. In *Proc. IEEE CVPR*, 2011.

[12] P. Zhu et al. Multi-scale patch based collaborative representation for face recognition with margin distribution optimization. In *Proc. ECCV*, 2012.