Neurocomputing 135 (2014) 306–312

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Joint sparse representation for video-based face recognition

Zhen Cui [a,b,c], Hong Chang [a,*], Shiguang Shan [a], Bingpeng Ma [c], Xilin Chen [a]

[a] *Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China*
[b] *School of Computer Science and Technology, Huaqiao University, Xiamen 361021, China*
[c] *University of China Academy Science, Beijing 100190, China*

ARTICLE INFO

Article history:
Received 3 July 2013
Received in revised form
9 December 2013
Accepted 19 December 2013
Communicated by Qingshan Liu
Available online 8 January 2014

Keywords:
Face recognition
Sparse representation
Structured sparse representation
Accelerated proximal gradient

ABSTRACT

Video-based Face Recognition (VFR) can be converted into the problem of measuring the similarity of two image sets, where the examples from a video clip construct one image set. In this paper, we consider face images from each clip as an ensemble and formulate VFR into the Joint Sparse Representation (JSR) problem. In JSR, to adaptively learn the sparse representation of a probe clip, we simultaneously consider the class-level and atom-level sparsity, where the former structurizes the enrolled clips using the structured sparse regularizer (*i.e.*, $L_{2,1}$-norm) and the latter seeks for a few related examples using the sparse regularizer (*i.e.*, L_1-norm). Besides, we also consider to pre-train a compacted dictionary to accelerate the algorithm, and impose the non-negativity constraint on the recovered coefficients to encourage positive correlations of the representation. The classification is ruled in favor of the class that has the lowest accumulated reconstruction error. We conduct extensive experiments on three real-world databases: Honda, MoBo and YouTube Celebrities (YTC). The results demonstrate that our method is more competitive than those state-of-the-art VFR methods.

1. Introduction

In traditional face recognition task, face images are identified from only a few samples per subject under controlled environments. Many current algorithms have achieved pretty good performance. However, with the popularization of video cameras such as surveillance cameras and cell phone cameras, we can easily capture large-scale face video clips in the wild, in which face images usually accompany with dramatic appearance changes in lighting, pose, expression, blur, etc. Therefore, the efficient classification of face video clips remains challenging and meaningful in practical applications.

The popular methods are the explosive development of Image Set based Classification (ISC) techniques [1–10]. Generally speaking, these approaches consist of two key steps: representing image set and defining between-set similarity. As image set representation is concerned, popular methods include Gaussian models [1,2], subspaces [3,7,8], nonlinear manifolds [4,11,9], etc. Gaussian model based methods can reasonably extend to unseen data with well-estimated parameters. However, if the data distribution does not follow the Gaussian assumption, the estimated model will not properly fit with the real distribution of image set. Instead, the non-parametric methods revive in more recent years. They usually represent one image set as a linear subspace [3,7,8] or a nonlinear manifold [4,11,9]. Compared with Gaussian based methods, these non-parametric methods have demonstrated many favorable properties (*e.g.*, no assumption on data distribution) with more excellent performance in VFR.

The second concern is how to define between-set distance. Generally, different distance metrics are used for different set representation methods. For Gaussian model, Kullback–Leibler Divergence [2] may be used to define between-set similarity. For subspace model, principal angles between two subspaces are often used as the distance metric. The classic works include Mutual Subspace Method (MSM) [12], Orthogonal Subspace Method (OSM) and their variants [3,13]. To develop more robust distance of two subspaces, specifically, some recent studies attempt to constrain the space of synthetic face images. For example, Cevikalp et al. [7] constrained the subspace spanned from face images of a clip into a convex hull, and then calculate the nearest distance of two convex hulls as the between-set similarity. Hu et al. [8] further extended it and proposed Sparse Approximated Nearest Point (SANP) to make the nearest points between two convex hulls lie on some facets by using the sparse regularizer. In addition, by assuming that face images of a clip lie on a nonlinear manifold, Wang et al. [4] extended Subspace–Subspace Distance (SSD) to Manifold–Manifold Distance (MMD), where a nonlinear manifold is partitioned into several local linear subspaces and then MMD is defined as pair-wise SSDs. However, MMD implicitly suffers a computational bias due to the uncertainty of subspace partitions

[*] Corresponding author.
E-mail address: hong.chang@vipl.ict.ac.cn (H. Chang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.12.004

[9]. To this end, Cui et.al [9] attempted to align all image sets to a pre-specified reference set and then measured the corresponding subspaces, which inevitably leads to the dependence on the choice of the reference set for the classification accuracy.

More recently, the sparse representation based methods [14–16] are developed to address the task of face recognition. Especially, the Sparse Representation-based Classification (SRC) [14] method sparsely represents a probe face image with a dictionary constructed from all gallery examples and then classifies it into the subject with the smallest reconstruction error. If the examples from the same subject construct its own subspace, *i.e.*, the intersection of subspaces spanned from any two subjects is null, the non-zero coefficients of the reconstructed example can ideally focus on the gallery examples from the same subject. Following this assumption, SRC has shown favorable properties in face recognition, especially when face images are partially occluded.

However, SRC treats every examples in gallery set equally and does not consider the structure of gallery data especially the class label. Intuitively, all examples from the same subject should be treated as an ensemble instead of multiple isolated images, which implies that the dictionary (*i.e.*, gallery set in SRC) may be characterized with the structure of group. For this, Elastic Net [17] and Group Lasso [18–20] are proposed to improve SRC. Specifically, Elhamifar et al. [20] casted the classification task as a structured sparse recovery problem, where the images from the same subject in gallery set form a group, and the sparsity is imposed on these groups, *i.e.*, the class-level or group-level sparsity. However, these methods only address the representation of a singe probe example and do not consider within-class appearance variations of an image set.

In addition, SRC is only designed to encode a single probe image. In a video clip, however, there are multiple frame images of the same subject, *i.e.*, multiple views of a subject. Note that here each clip only contains images of the same subject. Obviously, in a clip there exist strong correlations across different frames because a clip may be regarded as an approximately continuous stream. Therefore, when representing a clip, instead of frame-wise regression, the joint representation of all frames should be more meaningful for resisting the noises and increasing the representation stability. Generally, this problem of jointly estimating models from multiple related images is referred to “multi-view learning” [21,22] in the machine learning literatures. Yuan et al. [21] proposed Multi-Task Joint Sparse Representation (MTJSR), which aims to recover a test sample with multiple features from as few training subjects as possible and simultaneously enforces sparse coefficients on common atoms. However, MTJSR assumes that each sample has the same type of features, which naturally leads to the counterparts between multiple features. In the task of VFR, given any two clips, it is very intractable to obtain the counterparts between two clips (*i.e.*, images with the same poses, expressions, *etc*.). Thus it is impossible to encode one image with the same type of examples (or atoms) across different clips.

In this paper, inspired by recent progresses on sparse learning [14,18,21,23], we formulate VFR into a Joint Sparse Representation (JSR) problem (as shown in Fig. 1). In JSR, two sparse constraints are considered. The first sparse constraint is put on class-level by using $L_{2,1}$ mixed-norm, which assumes that a probe image set (or a clip) can be represented by a few gallery image sets (or gallery clips). The second sparse constraint enforces the sparsity on within-class images by using L_1-norm, with the aim to choose a few related views. Intuitively, different subjects lead to class-level sparsity, while appearance variations cause atom-level sparsity among images of all persons. In addition, in order to make the model more robust, two improvements are further provided: one is to learn a compact dictionary to reduce time cost, and the other is to impose nonnegative constraints on the representation. To solve this model of JSR, the Accelerated Proximal Gradient (APG) [24] optimization strategy is employed with fast convergence rate guaranteed. We conduct extensive experiments on three video databases: Honda [25], MoBo [26] and YouTube Celebrities(YTC) [27]. The results demonstrate that the proposed method is more competitive than the state-of-the-art methods for video-based face recognition.

The remainder of this paper is organized as follows. In Section 2.2, we present the proposed joint sparse representation model. The optimization details along with the final classification rule are stated in Sections 2.3 and 2.4. The applications of our method to face recognition are reported in Section 3. Finally, we reach a conclusion in Section 4.

2. Joint sparse representation

In this section, we first introduce the basic idea of joint sparse representation, then give the mathematical formulation in detail, and finally provide the optimization and the classification rule.

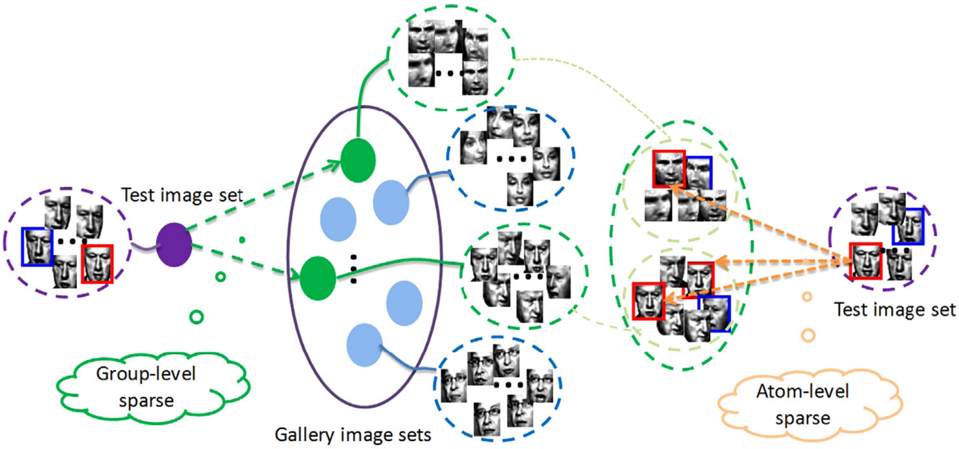

Fig. 1. Illustration of our idea. Given a test image set (or a video clip), the group-level (or class-level) sparse recovery is used to search the most relevant subjects from gallery image sets (or gallery clips), while the atom-level sparse regression is imposed on each image of the test set to find the similar appearance images. Further, such two sparse constrains are jointly imposed on an image set (or a clip) rather than an isolated image, which suppresses noises and leads to more robust representations. Behind that an intuitive explanation is that different subjects lead to the class-level sparsity, while appearance variations cause the atom-level sparsity among images of each person. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

2.1. Basic idea

The basic idea is illustrated in Fig. 1. As shown in this figure, given the test image set (or the probe clip) and the gallery image sets (or the gallery clips), we consider two key priors for image set representation: integrity and sparsity.

All face images from a video clip should be regarded as an ensemble in the sparse representation. That is, each clip is conducted as an ensemble at one time, instead of isolated images. Such a joint spare representation on all images of a clip can efficiently suppress the noises in them and further make the probe clip more stable recovery.

To find more meaningful representation, the sparse priors can be employed. In class-level, face images of a subject are usually similar to those of a few of subjects, which cause the concentration of reconstruction energy (or coefficients) on a very few related subjects for a given probe clip, as shown in Fig. 1. We may formulate this observation into the structured sparse norm, an extension of $L_{2,1}$ mixed-norm. In atom-level, based on human vision mechanism of sparsity selection [28], we may sparsely recover each image in the probe clip. An intuitive explanation is that, since real-world face images usually contain complex appearance variations, a given image should exist in a small subspace spanned by those images with the similar appearance. For instance, a face image with yaw 15° pose may be recovered from those images with nearby poses in a larger possibility. For this, we may use the L_1 norm to penalize the reconstruction coefficients.

2.2. Problem formulation

Suppose we have the gallery data $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M] \in \mathbb{R}^{d \times N}$, where $\mathbf{X}_i \in \mathbb{R}^{d \times N_i}$ is the column-stacked feature matrix of all N_i images of the i-th subject, d is the feature dimensionality, $N = \sum_1^M N_i$, and M is the total number of subjects (or classes) in the gallery data. Note that all images of the same subject might come from multiple clips. As a video clip usually contains a lot of redundant information, we may compress each feature set $\mathbf{X}_\mathrm{i}$ into a more compact subdictionary $\mathbf{D}_i \in \mathbb{R}^{d \times l_i}$ by performing the clustering algorithm on $\mathbf{X}_\mathrm{i}$, where l_i is the size of the i-th subdictionary, and i=1,…,M. Therefore, we can obtain the compact dictionary $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_M] \in \mathbb{R}^{d \times l}$ to represent the whole gallery data, where $l = \sum_{i=1}^M l_i$.

Given a probe clip with n different observations from a subject, $\mathbf{Y} \in \mathbb{R}^{d \times n}$, we can recover the probe clip from the gallery dictionary $\mathbf{D}$ as follows:

$$\mathbf{Y} = \sum_{i=1}^{M} \mathbf{D}_i \mathbf{W}^i + \mathbf{E} = \mathbf{D}\mathbf{W} + \mathbf{E}, \tag{1}$$

where $\mathbf{W}^i \in \mathbb{R}^{l_i \times n}$ is the reconstructed coefficient matrix associated with the i-th subdictionary, $\mathbf{E}$ is the residual term, and the matrix $\mathbf{W} = [\mathbf{W}^1; \mathbf{W}^2; \cdots; \mathbf{W}^M]$ is stacked block-wisely in height. Thus, the VFR problem can be formulated to solve the following joint sparse representation model by incorporating two sparse regularization terms as analyzed above:

$$\min_{\mathbf{W}} \quad F(\mathbf{W}) = f(\mathbf{W}) + \lambda_1 \zeta(\mathbf{W}) + \lambda_2 \phi(\mathbf{W}), \tag{2}$$

$$\mathbf{W} \geq 0 \text{ (optional)}, \tag{3}$$

where

$$f(\mathbf{W}) = \frac{1}{2} \| \mathbf{Y} - \mathbf{D}\mathbf{W} \|_F^2, \tag{4}$$

$$\zeta(\mathbf{W}) = \sum_{i=1}^{M} \| \mathbf{W}^i \|_1 = \| \mathbf{W} \|_1, \tag{5}$$

$$\phi(\mathbf{W}) = \sum_{i=1}^{M} \| \mathbf{W}^i \|_F. \tag{6}$$

In the above model, f is the loss function of reconstruction error by Frobenius norm, ζ is the L_1 sparse function of the recovery coefficients $\mathbf{W}$ with $\|\mathbf{W}\|_1 = \sum_{ij} |W_{ij}|$, *i.e.*, sum all the absolute values of each item in the representation matrix. The structured sparse function ϕin Eq. (6) puts L_2 norm on those coefficients of each corresponded class and then sum the coefficients of all classes (*i.e.*, L_1 norm). That is, ϕ only conduct the sparsity on group-level, where one subdictionary may be regarded as one group. To simplify the notation, we still use $L_{2,1}-$norm to mark the structure sparse function ϕ, where the standard $L_{2,1}$ norm imposes L_2-norm on each column/row and L_1-norm on all rows/columns. Besides, the nonnegative constraint may be optional to limit the subspace spanned by face images from the probe clip and further improve the performance.

2.3. Efficient solution

In the above model, since the function f is a smooth and convex function, and two sparse regularization terms are convex, thus we can employ the popular APG method [24,29], which uses "optimal" first order gradient to solve the objective value with the convergent rate $O(1/t^2)$. First, we can define the generalized gradient update step of Eq. (2) as follows:

$$Q_L(\mathbf{W}, \mathbf{W}_t) = f(\mathbf{W}_t) + \langle \mathbf{W} - \mathbf{W}_t, \nabla f(\mathbf{W}_t) \rangle + \frac{L}{2} \| \mathbf{W} - \mathbf{W}_t \|_F^2 + \lambda_1 \zeta(\mathbf{W}) + \lambda_2 \phi(\mathbf{W}), \tag{7}$$

where $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{trace}(\mathbf{A}^T \mathbf{B})$ denotes the matrix inner product. Given the solution $\mathbf{W}_t$, thus we may iteratively update $\mathbf{W}$ according to the above Eq. (7). We can further simplify it into the following equation:

$$Q_L(\mathbf{W}, \mathbf{W}_t) = \frac{L}{2} \| \mathbf{W} - \mathbf{A} \|_F^2 + C + \lambda_1 \zeta(\mathbf{W}) + \lambda_2 \phi(\mathbf{W}), \tag{8}$$

$$\mathbf{A} = \mathbf{W}_t - \frac{1}{L} \nabla f(\mathbf{W}_t), \tag{9}$$

$$C = f(\mathbf{W}_t) - \frac{1}{2L} \| \nabla f(\mathbf{W}_t) \|_F^2. \tag{10}$$

Algorithm 1. Joint sparse representation.

```
Input: Dictionary $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_M]$ of $M$ subjects in the gallery
    data, a probe clip (or image set) $\mathbf{Y}$, balance parameters $\lambda_1$
    and $\lambda_2$, and group indices.
Output: Regression coefficients $\mathbf{W}$.
1:  Initialization: $L > 0, \eta > 1, \mathbf{W}_0 \in \mathbb{R}^{d \times n}, \mathbf{Z}_0 = \mathbf{W}_0, p = 1, t = 1$.
2:  repeat
3:    Calculate $\mathbf{A}, C$ using Eqs. (9) and (10).
4:    if the nonnegative constraint then
5:      update $\mathbf{A}$ by $\mathbf{A}^+$.
6:    end if
7:    Calculate the sparse matrix $\mathbf{B}$ by Eq. (14).
8:    Project $\mathbf{W}$ to $\mathbf{B}$ by Eq. (15).
9:    if $F(\mathbf{W}) > Q_L(\mathbf{W}, \mathbf{Z}_t)$ then
10:     extend the step $L = \eta L$, and go to step 3.
11:   end if
12:   Update variables: $\mathbf{Z}_{t+1} = \mathbf{W}$; $p_{t+1} = 2/(t+3)$;
    $\mathbf{W}_{t+1} = \mathbf{Z}_{t+1} - \frac{(1-p_t)p_{t+1}}{p_t}(\mathbf{Z}_{t+1} - \mathbf{Z}_t)$.
13:   $t = t + 1$.
14: until the convergence criterion is reached.
15: return $\mathbf{W}$.
```

Next we focus on how to solve the generalized gradient mapping step (8), where $\mathbf{A}$ and C are known if $\mathbf{W}_t$ is fixed. Therefore, we only need to solve the following minimization problem:

$$\min_{\mathbf{W}} \quad Q_L(\mathbf{W},\mathbf{W}_t)=\frac{L}{2}\|\mathbf{W}-\mathbf{A}\|_F^2+\lambda_1\zeta(\mathbf{W})+\lambda_2\phi(\mathbf{W}), \tag{11}$$

where the nonnegative constraint on $\mathbf{W}$ may be added into this model, *i.e.*, $\mathbf{W}\geq 0$. To address the nonnegative constraint, we can directly decompose $\mathbf{A}$ into two nonnegative matrices $\mathbf{A}^+$ and $\mathbf{A}^-$ so that $\mathbf{A}=\mathbf{A}^+-\mathbf{A}^-$. Then the optimal solution of Eq. (11) with the nonnegative constraint $\mathbf{W}\geq 0$ may be obtained by simply replacing $\mathbf{A}$ with $\mathbf{A}^+$, *i.e.*, minimizing the following model:

$$\min_{\mathbf{W}} \quad Q_L(\mathbf{W},\mathbf{W}_t)=\frac{L}{2}\|\mathbf{W}-\mathbf{A}^+\|_F^2+\lambda_1\zeta(\mathbf{W})+\lambda_2\phi(\mathbf{W}). \tag{12}$$

After a similar derivation with Theorem 1 in [30], the above optimization problem (11) or (12) is equal to the following optimization model:

$$\min_{\mathbf{W}} \quad Q_L(\mathbf{W},\mathbf{W}_t)=\frac{L}{2}\|\mathbf{W}-\mathbf{B}\|_F^2+\lambda_2\phi(\mathbf{W}), \tag{13}$$

$$\mathbf{B}=\mathrm{sgn}(\mathbf{A})\odot\max(|\mathbf{A}|-\lambda_1,0). \tag{14}$$

Note, when $\mathbf{W}\geq 0$, $\mathbf{A}$ in Eq. (14) should be replaced with $\mathbf{A}^+$ from the above analysis. Now we convert the joint sparse model into the regression model (13) only with the $L_{2,1}$ mixed-norm penalty item, *i.e.*, the notable structure sparse model. Following [31], we can obtain the solution of the model (13) as follows:

$$\mathbf{W}^i=\begin{cases} 0 & \text{if } \lambda_2\geq L\|\mathbf{B}^i\|_F \\ \dfrac{L\|\mathbf{B}^i\|_F-\lambda_2}{L\|\mathbf{B}^i\|_F}\mathbf{B}^i & \text{if } \lambda_2< L\|\mathbf{B}^i\|_F \end{cases}, \tag{15}$$

where the submatrix $\mathbf{B}^i$ of $\mathbf{B}$ corresponds to the i-th subdictionary.

The whole process of solving $\mathbf{W}$ is summarized in Algorithm 1. In each iteration, the computational time mainly contains the computation of the gradient of the loss function and the solution of the minimization problem (11) or (12). The computation of gradient depends on the computation of inner product. Given n face images from a probe clip, the time complexity of gradient calculation is $O(ldn)$ while the computation of $\mathbf{W}$ in Eq. (15) takes about $O(ln)$. If a proper step L is chosen with the search of t_L times, the time complexity from step (4) to step (11) is $O(ldnt_L)$ in one iteration. Suppose the algorithm terminates after t_W iterations, the total time complexity is $O(ldnt_Lt_W)$. Fortunately, Nesterov [24] proves that APG is able to reach the convergent rate of $O(1/t^2)$, which makes the algorithm iterate a few times to reach a feasible solution.

2.4. Classification rule

Our basic idea is to use the reconstruction error as the decision rule. Given a probe clip with n frames, denoted as $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_n]$, we accumulate the reconstruction errors of all n frames on the i-th subdictionary as the similarity between the probe clip and the i-th class, and then assign the class label with minimal reconstruction error to the label of this probe clip. Formally, the decision label can be defined as

$$i^*=\arg\min_i \quad S(\mathbf{Y},\mathbf{D}_i), \tag{16}$$

where

$$S(\mathbf{Y},\mathbf{D}_i)=\sum_{k=1}^{n}e_k^i, \tag{17}$$

$$e_k^i=\|\mathbf{y}_k-\mathbf{D}_i\mathbf{W}_k^i\|^2, \tag{18}$$

where i indicates the i-th class, k means the k-th frame, and e_k^i is the reconstruction error of the k-th frame on the i-th class.

3. Experiments

We conduct extensive experiments on three real-world video face datasets, in which face images contain complex appearance variations in poses, expressions, illuminations, etc. Below we first introduce three datasets and the experimental setup, and then evaluate our method by comparing several state-of-the-art methods.

3.1. The databases

We use three public datasets: Honda/UCSD [25], CMU MoBo [26] and YouTube Celebrities (YTC) [27], from which some examples are shown in Fig. 2.

Honda/UCSD [25] was collected for video-based face recognition. In this paper, we use its first subset, which contains 59 videos of 20 subjects and each subject has at least 2 videos. The length of video clips varies from 12 to 645 frames. Different poses and expressions usually appear across different clips of each subject. We apply a cascaded face detector [32] to detect face from each video clip, and then resize all face images to gray-scale images with 20 × 20 pixels as used in [9]. To eliminate the lighting effects, histogram equalization is employed in the pre-processing step.

MoBo [26] was originally created for human pose identification. There are 96 sequences of 24 different subjects walking on a treadmill. Each subject contains 4 clips, about 300 frames per clip. We detect face images with the same way as we do in Honda, and resize them into 30 × 30 pixels.

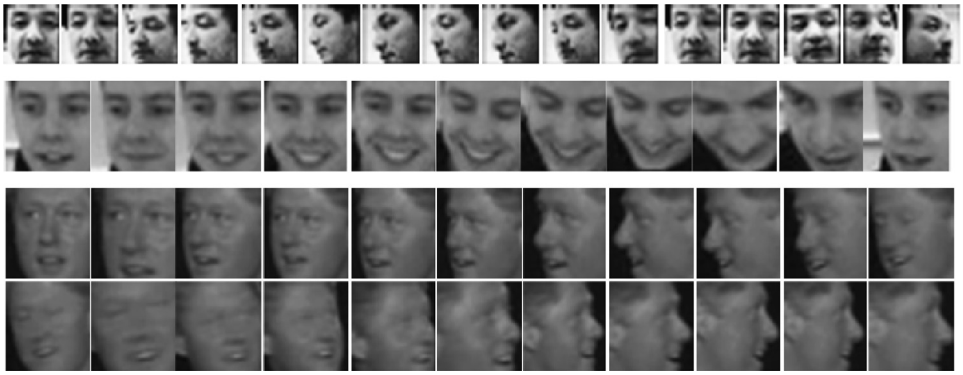

Fig. 2. Examples from three video face databases: Honda [25] (the first row), MoBo [26] (the second row), and YouTube Celebrities [27] (the last two rows).

YouTube Celebrities [27] was collected for face tracking and recognition in real-world applications. The dataset contains 1910 video clips of 47 celebrities (actors, actresses, politicians, etc.). Each clip contains hundreds of frames, which are mostly low resolution and recorded at high compression rates. Compared with Honda and MoBo, this database is much more challenging due to dramatic appearance variations. We crop face images as the above MoBo.

On all of three datasets, we conduct five random experiments, *i.e.*, five randomly selected training and testing combinations, and then report the average rate as well as the standard deviation. For Honda and MoBo, one clip of each subject is chosen for training and the rest clips for testing. For YTC, each person has 41 clips on average across 3 sessions. We randomly choose 3 clips, one clip per session, for training and 6 clips per subject for testing.

3.2. *The comparison methods*

We compare our proposed method with several image set based methods proposed in recent years. They include Mutual Subspace Method (MSM) [12], Discriminant Canonical Correlation Analysis (DCC) [3], Manifold-to-Manifold Distance (MMD) [4], Manifold Discriminant Analysis (MDA) [5], and Sparse Approximated Nearest Points (SANP) [8]. Recent literatures [3,4,8] have shown that image set based methods generally outperform exemplar based methods, so here we do not provide the comparison with those exemplar based methods. Besides, we also compare with the baseline SRC [14] by using all training data as the dictionary.

For MSM, we adopt the technique in accordance with [3]. The source codes of MMD, MDA and SANP are downloaded from the original author websites, and the referred parameters are respectively followed by their papers [4,5,8]. For SRC [14], we follow their protocols with the dictionary constructed from all gallery samples, which thus lead to a huge dictionary due to hundreds of frames per clip. For this, we use the Orthogonal Matching Pursuit (OMP) algorithm [33,34] to accelerate the solution of SRC in our experiments.

In our method, we use PCA to reduce original feature to $d=80$ dimensions on three databases in order to speed up the algorithm. The i-th subdictionary size $l_i (i=1,2,\ldots,M)$ corresponding to the i-th subject is set to 20, 20, 60 for Honda, MoBo and YTC respectively.

3.3. *Parameter tuning*

The key parameters in our method are the balance parameters λ_1 and λ_2. We search the two parameters on MoBo and the results are reported in Table 1. As shown in this table, JSR first improves and then degrades the performance with the increase of one parameter value by fixing the other parameter. When the larger parameter values (*e.g.*, $\lambda_2=6$ or $\lambda_1=0.8$) are imposed on their corresponding regularization items, the learnt coefficient matrix might contain most zero items, which naturally makes the discriminability of this model weaken. In contrast, if the penalties become too light, *i.e.*, small values on λ_1, λ_2, the learnt model might be lack of discriminability because the reconstruction error is overly emphasized with less selectivity. In the following experiments, we choose $\lambda_1=0.01$ and $\lambda_2=0.1$ as default parameters in our method.

Table 1
The performance of JSR with different λ_1 and λ_2 on MoBo. The performance trends of JSR with $\lambda_1=0$ or $\lambda_2=0$ are respectively reported in (a) and (b).

(a) $\lambda_1=0$								
λ_2	0	0.001	0.01	0.1	1	2	4	6
Accuracy	0.931	0.931	0.931	0.937	0.937	0.950	0.916	0.878
(b) $\lambda_2=0$								
λ_1	0	0.001	0.01	0.1	0.2	0.4	0.6	0.8
Accuracy	0.931	0.937	0.946	0.934	0.938	0.916	0.916	0.788

In fact, our method generalizes those classic sparse representation methods. As the special cases of our method, we can easily have the following observations:

- Least square regression (LSR): $\lambda_1=0$ and $\lambda_2=0$.
- Sparse regression (SR): $\lambda_2=0$. From Table 1(b), we choose $\lambda_1=0.01$ and $\lambda_2=0$ as the default values.
- Group sparse regression (GSR): $\lambda_1=0$. From Table 1(a), we choose $\lambda_1=0$ and $\lambda_2=0.1$ as the default values.

3.4. *Experimental results and analysis*

In Table 2, we compare our method with those competitive methods, and report their classification accuracies with standard deviations. Overall, our proposed method can achieve better performances in most cases. From this table, we reach the following conclusions:

- The image set based methods show distinct performances according to their properties. Among them, MSM, MMD and SANP directly use image data in original space, which makes them less appealing than the supervised methods, DCC and MDA. Further, from the view of manifold, MMD and MDA are superior to MSM and SANP because they partition a clip into multiple local linear models. Compared with these classic image set based methods, the regression (or reconstruction) base models are more efficient, which might be attributed to the use of reconstruction error.
- Joint sparse constraints improve the performance. Compared with the baseline SRC, our proposed method can achieve better performance. The reason lies in two folds: one is to use the compact dictionary to reduce noises in gallery data, the other is to jointly use two sparse priors in the regression model to recover the probe clip more credibly. Generally, group sparse regression (*i.e.*, $\lambda_1=0$) and sparse regression (*i.e.*, $\lambda_2=0$) are superior to least square regression (*i.e.*, $\lambda_1=0$ and $\lambda_2=0$), which indicates that face recognition can benefit from the use of sparsity and structuration. Moreover, combining the two priors into the regression model (*i.e.* $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$) can further improve the performance. In addition, when only considering positive responses, *i.e.*, non-negative constraint on coefficients, JSR can achieve the best performance, where one possible explanation is that non-negativity eliminates some uncorrelated components in the regression model.

Table 2
Identification rates on three databases (mean ± standard deviation)

Method	Honda	MoBo	YTC
lMSM [12]	0.923 ± 0.04	0.886 ± 0.03	0.616 ± 0.04
MMD [4]	0.969 ± 0.02	0.897 ± 0.01	0.634 ± 0.02
DCC [3]	0.980 ± 0.01	0.903 ± 0.05	0.673 ± 0.03
MDA [5]	0.989 ± 0.01	0.947 ± 0.01	0.676 ± 0.02
SANP [8]	0.959 ± 0.01	0.900 ± 0.02	0.634 ± 0.03
SRC [14]	0.969 ± 0.02	0.941 ± 0.03	0.705 ± 0.02
JSR			
$\lambda_1=0,\ \lambda_2=0$	0.994 ± 0.01	0.931 ± 0.01	0.633 ± 0.02
$\lambda_1=0,\ \lambda_2\neq 0$	$\mathbf{1.000} \pm 0.00$	0.937 ± 0.02	0.673 ± 0.02
$\lambda_1\neq 0,\ \lambda_2=0$	$\mathbf{1.000} \pm 0.00$	0.946 ± 0.03	0.712 ± 0.02
$\lambda_1\neq 0,\ \lambda_2\neq 0$	$\mathbf{1.000} \pm 0.00$	0.956 ± 0.02	0.722 ± 0.02
$\lambda_1\neq 0,\ \lambda_2\neq 0,\ W\geq 0$	$\mathbf{1.000} \pm 0.00$	$\mathbf{0.965} \pm 0.01$	$\mathbf{0.737} \pm 0.02$

- Among three databases, the performance on YTC is worst for all methods, because the clips in YTC are captured under more complex unconstrained environments. Even so, our method outperforms the other classic VFR methods with an improvement of more than 5 percent. Compared with the baseline SRC, our method has also an improvement of about 3 percent. For Honda, since the data has good separability, most methods win a very high accuracy. Note that the result on MoBo is relatively lower than that reported in [8], because LBP is used as the feature representation in their work.

4. Conclusion

In this paper, we propose a Joint Sparse Representation method to handle the video-based face recognition problem. JSR treats multiple frames of a probe clip as an ensemble, and jointly recovers those face images in the clip. In JSR, we introduce two sparse regularization terms to make full use of the sparse and structure priors of data, which makes learnt model better discriminative in the task of video-based face recognition. Moreover, the pre-trained compact dictionary can partly remove noises of the gallery data while speeding up the optimization. We propose a fast and efficient gradient-based algorithm to solve JSR. Experimental results on three public video face databases, especially the most challenging YTC database, demonstrate that our proposed method is more competitive than those state-of-the-art methods for video-based face recognition. In the future work, we will extend this method to a kernel version and try to apply it to other tasks in computer vision.

Acknowledgments

The work is partially supported by Natural Science Foundation of China under Contracts nos. 61025010, 61222211, 61202297 and 61003103; Natural Science Foundation of Fujian Province (No. 2013J01239).

References

[1] G. Shakhnarovich, J. Fisher, T. Darrell, Face recognition from long-term observations, in: European Conference on Computer Vision (ECCV), 2002.
[2] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, T. Darrell, Face recognition with image sets using manifold density divergence, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005.
[3] T. Kim, J. Kittler, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 1005–1018.
[4] R. Wang, S. Shan, X. Chen, W. Gao, Manifold-manifold distance with application to face recognition based on image set, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[5] R. Wang, X. Chen, Manifold discriminant analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[6] C. Wang, Y. Li, Combine image quality fusion and illumination compensation for video-based face recognition, Neurocomputing 73 (7) (2010) 1478–1490.
[7] H. Cevikalp, B. Triggs, Face recognition based on image sets, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[8] Y. Hu, A. Mian, R. Owens, Sparse approximated nearest points for image set classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[9] Z. Cui, S. Shan, H. Zhang, S. Lao, X. Chen, Image sets alignment for video-based face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[10] Z. Cui, W. Li, D. Xu, S. Shan, X. Chen, Fusing robust face region descriptors via multiple metric learning for face recognition in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[11] M. Harandi, C. Sanderson, S. Shirazi, B. Lovell, Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
[12] O. Yamaguchi, K. Fukui, K. Maeda, Face recognition using temporal image sequence, in: Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998.
[13] X. Li, K. Fukui, N. Zheng, Boosting constrained mutual subspace method for robust image-set based object recognition, in: Proceedings of the 21st International Joint conference on Artificial intelligence, 2009.
[14] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[15] Z. Cui, S. Shan, L. Zhang, X. Chen, Sparsely encoded local descriptor for face recognition, in: 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), IEEE, 2011, pp. 149–154.
[16] Z. Cui, S. Shan, H. Zhang, S. Lao, X. Chen, Structured sparse linear discriminant analysis, in: 2012 19th IEEE International Conference on Image Processing (ICIP), IEEE, 2012, pp. 1161–1164.
[17] A. Majumdar, R. Ward, Classification via group sparsity promoting regularization, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009.
[18] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 68 (1) (2006) 49–67.
[19] R. Jenatton, J. Audibert, F. Bach, Structured Variable Selection with Sparsity-Inducing Norms, Arxiv preprint arXiv:0904.3523.
[20] E. Elhamifar, R. Vidal, Robust classification using structured sparse representation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
[21] X. Yuan, S. Yan, Visual classification with multi-task joint sparse representation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.
[22] H. Zhang, N. Nasrabadi, Y. Zhang, T. Huang, Multi-observation visual recognition via joint dynamic sparse representation, in: IEEE International Conference on Computer Vision (ICCV), 2011.
[23] P. Sprechmann, I. Ramírez, G. Sapiro, Y. Eldar, C-HiLasso: a collaborative hierarchical sparse modeling framework, IEEE Trans. Signal Process. 59 (9) (2011) 4183–4198.
[24] Y. Nesterov, I. Nesterov, Introductory Lectures on Convex Optimization: a Basic Course, vol. 87, Springer, 2004.
[25] K. Lee, J. Ho, M. Yang, D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003.
[26] R. Gross, J. Shi, The CMU motion of body (MoBo)database, Technical Report, CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University.
[27] M. Kim, S. Kumar, V. Pavlovic, H. Rowley, Face tracking and recognition with visual constraints in real-world videos, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.
[28] R. Rao, B. Olshausen, M. Lewicki, Probabilistic Models of the Brain: Perception and Neural Function, The MIT Press, 2002.
[29] W. Zuo, Z. Lin, A generalized accelerated proximal gradient approach for total variation-based image restoration, IEEE Trans. Image Process. 20 (99) (2011) 2748–2759.
[30] J. Liu, J. Ye, Fast Overlapping Group Lasso, Arxiv preprint arXiv:1009.0306.
[31] D. Luo, C. Ding, H. Huang, Towards structural sparsity: an explicit l2/l0 approach, in: IEEE 10th International Conference on Data Mining (ICDM), 2010.
[32] P. Viola, M. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[33] G. Davis, S. Mallat, M. Avellaneda, Adaptive greedy approximations, Constr. Approx. 13 (1) (1997) 57–98.
[34] R. Rubinstein, M. Zibulevsky, M. Elad, Efficient Implementation of the k-SVD Algorithm Using Batch Orthogonal Matching Pursuit, CS Technion.

Zhen Cui received the B.S. degree from Shandong Normal University, Jinan, China, in 2004, and then the M.S. degree from Sun Yatsen University, Guangzhou, China, in 2006. Currently, he is a Ph.D Candidate in Institute of Computing Technology, Chinese Academy of Science since 2009, and also a Lecturer in Huaqiao University since 2006. From June 2012 to Dec 2012, he was a Research Associate in Nanyang Technological University, Singapore. His research interests cover pattern recognition and computer vision, especially face recognition and image super resolution based on recently emerged theories, such as Sparse Coding, Manifold Learning, and Deep Learning.

Hong Chang received the Bachelors degree from Hebei University of Technology, Tianjin, China, in 1998; the M. S. degree from Tianjin University, Tianjin, in 2001; and the Ph.D. degree from Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2006, all in computer science. She was a Research Scientist with Xerox Research Centre Europe. She is currently an Associate Researcher with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Her main research interests include algorithms and models in machine learning, and their applications in pattern recognition, computer vision, and data mining.

Shiguang Shan received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences, Beijing, in 2004. He has been with ICT, CAS since 2002 and has been a Professor since 2010. He is also the Vice Director of the Key Lab of Intelligent Information Processing of CAS. His research interests cover image analysis, pattern recognition, and computer vision. He is focusing especially on face recognition related research topics. He received the China′s State Scientific and Technological Progress Awards in 2005 for his work on face recognition technologies.

Bingpeng Ma received the B.S. degree in Mechanics, in 1998, and the M.S. degree in Mathematics, in 2003, from HuaZhong University of Science and Technology. He received Ph.D. degree at the Institute of Computing Technology, Chinese Academy of Sciences, PR China, in 2009. He was a Post-doctorial researcher in University of Caen, France, from 2011 to 2012. He joined the University of Chinese Academy of Sciences, Beijing, in March 2013. His current interests are in the areas of pattern recognition and machine learning.

Xilin Chen received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 1988, 1991, and 1994 respectively. He was a Professor with the HIT from 1999 to 2005 and was a Visiting Scholar with Carnegie Mellon University from 2001 to 2004. He has been a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, since August 2004. His research interests include image processing, pattern recognition, computer vision, and multimodal interface. He has received several awards, including the China′s State Scientific and Technological Progress Award in 2000, 2003, 2005, and 2012 for his research work.