Robust and Practical Face Recognition via Structured Sparsity

Kui Jia[1], Tsung-Han Chan[1], and Yi Ma[2,3]

[1]Advanced Digital Sciences Center, Singapore
[2] Microsoft Research Asia, Beijing, China
[3] Dept. Elec. and Comp. Eng., University of Illinois at Urbana-Champaign

Abstract. Sparse representation based classification (SRC) methods have recently drawn much attention in face recognition, due to their good performance and robustness against misalignment, illumination variation, and occlusion. They assume the errors caused by image variations can be modeled as pixel-wisely sparse. However, in many practical scenarios these errors are not truly pixel-wisely sparse but rather sparsely distributed with structures, i.e., they constitute contiguous regions distributed at different face positions. In this paper, we introduce a class of structured sparsity-inducing norms into the SRC framework, to model various corruptions in face images caused by misalignment, shadow (due to illumination change), and occlusion. For practical face recognition, we develop an automatic face alignment method based on minimizing the structured sparsity norm. Experiments on benchmark face datasets show improved performance over SRC and other alternative methods.

1 Introduction

Face recognition is a long-standing problem in computer vision. It has broad applications ranging from less-demanding ones such as family photo album organization (e.g., Apple iPhoto), to the most challenging applications of mass surveillance and terrorist watchlist that require high recognition performance but good training images are difficult to be obtained. In this work, we consider an application scenario that falls between these two extremes, where high recognition performance is desired but a rich set of training face images can be pre-captured in controlled conditions. Notable applications of this kind are access control for secure facilities, computer systems, automobiles, etc. Among face recognition methods targeting for this scenario, the classical subspace methods such as Eigenfaces [1], Fisherfaces [2] and nearest subspace (NS) [7] have been extensively studied. They generally work well in laboratory conditions. Under practical working or testing conditions their performance is very sensitive to illumination change, occlusion, or misalignment (due to scale or pose changes).

Recently, sparse representation based classification (SRC) methods have been proposed [3, 13, 11] and shown their promise in handling these variabilities in face recognition. In particular, Wright *et al.* [3] proposed to use an extended ℓ_1-norm minimization for robust face recognition. Assuming access to a face

database with each subject having multiple registered training images taken under varying illuminations, [3] casts face recognition as the problem of finding a sparse representation of a test image in terms of the training ones, plus a sparse error image compensating for possible occlusion or corruption. Denote the set of training images as $\{\mathbf{A}_k\}_{k=1}^K$ for K subjects. $\mathbf{A}_k \in \mathbb{R}^{m\times n_k}$ contains images of subject k, with each image being concatenated as a column vector of $\mathbf{A}_k$. We can put images of all subjects together to form a large matrix $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K] \in \mathbb{R}^{m\times n}$. The sparse representation $\mathbf{x}$ and sparse error $\mathbf{e}$ are recovered in [3] by solving the extended ℓ_1-norm minimization problem

$$(\ell_1\_\ell_1): \qquad \min_{\mathbf{x},\mathbf{e}} \|\mathbf{x}\|_1 + \|\mathbf{e}\|_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^m$ is the given test face image. A key component in their method leading to the above robustness is to enforce sparsity by ℓ_1-norm on the residual or error image $\mathbf{e}$. By leveraging the same sparsity assumption using ℓ_1-norm minimization, an automatic face alignment algorithm was developed in [13]. Suppose $\mathbf{y}'$ is an observed test face that is not in register with the training images $\{\mathbf{A}_k\}_{k=1}^K$. To recover a well aligned image $\mathbf{y} = \mathbf{y}' \circ \tau$ so that it can be readily used for robust face recognition, where τ represents some transformation acting on the image domain (e.g., 2D similarity transformation), [13] proposed to solve the following optimization problem to seek the correct transformation τ and sparse error $\mathbf{e}$

$$\min_{\mathbf{e},\tau_k,\mathbf{x}_k} \|\mathbf{e}\|_1 \quad \text{s.t.} \quad \mathbf{y}' \circ \tau_k = \mathbf{A}_k\mathbf{x}_k + \mathbf{e}, \tag{2}$$

where $\mathbf{y}'$ is sequentially aligned to each subject $\mathbf{A}_k$ instead of the whole training set $\mathbf{A}$, mainly due to the difficulty of optimization associated with the later case, as discussed in [13]. [13] demonstrates the state-of-the-art face recognition performance in a practical access control setting. The success of SRC methods has also inspired many following works [14, 15].

In the context of statistical signal processing, it is well known that when using ℓ_1-norm to promote the sparsity in the errors $\mathbf{e}$, it assumes that each pixel is independently corrupted. However, for many practical face variations such as occlusion, disguise, or shadow caused by illumination change, errors due to these variations are typically spatially contiguous. It becomes inappropriate to model these variations using ℓ_1-norm minimization, as did in [3, 13, 14].

The theory of compressed sensing suggests that given contiguous structures, it is possible to recover sparse signals with fewer measurements [12]. This means that from a fixed number of measurements (pixels), we should expect to correct a larger fraction of errors and subsequently obtain improved recognition performance if the structural prior knowledge of the corruption can be properly harnessed. In particular, [11] has used a Markov Random Fields (MRF) model to estimate a contiguous error support from the obtained $\mathbf{e}$, and has demonstrated significantly improved performance over [3] for contiguous occlusion. However, the performance of the MRF model [11] drops drastically when test images are subject to slight misalignment. To handle misalignment [13] still resorts to promoting the sparsity on $\mathbf{e}$ with ℓ_1-norm.

In this paper we introduce a new class of norms that can promote error sparsity patterns with the properties of contiguity and spatial locality. Our motivation follows the recent development of new sparsity-inducing norms that are capable of encoding prior knowledge about the expected structured sparsity patterns. While ℓ_1-norm can only promote independent sparsity [16], one can partition variables into disjoint groups and promote group sparsity using the so called group Lasso regularization [17]. To induce more sophisticated structured sparsity patterns, it becomes essential to use structured sparsity-inducing norms built on overlapping groups of variables [20, 19]. In this paper, we consider to use a hierarchical tree-structured sparsity-inducing norm [20, 22] on the error $\mathbf{e}$ of a test face, as shown in Figure 1, where overlapping groups of pixels are from local patches of varying size and each group corresponds to a node of the tree. As shown in our experiments in Section 4, without knowing explicitly the number, locations, sizes, and shapes of contiguous errors caused by various face variations, our method performs better than [3] in terms of handling spatially contiguous errors. When test images are not well aligned with training images, unlike the MRF based method, we can effectively bring the images in alignment via minimizing the structured sparsity norm, by simply replacing the ℓ_1-norm in equation (2). In fact, experiments show that our method performs better than using the ℓ_1-norm for alignment and recognition [13], especially in cases when only partial face is visible due to occlusion or disguise.

To solve the corresponding optimization problems, we develop efficient algorithms based on the Augmented Lagrange Multiplier (ALM) method [23], in which a proximal problem associated with structured sparsity norm regularization can be efficiently solved using techniques given in [21, 22]. The better error correction capability of structured sparsity translates readily into improved face recognition performance. Experiments on benchmark face databases show that our methods achieve the state-of-the-art recognition results, and outperform other SRC-based methods in simultaneously handling illumination change, occlusion, and misalignment in the test face image.

2 Modeling using structured sparsity-inducing norms

In this section, we discuss how we could systematically develop sparsity-inducing norms that can incorporate prior structures on the support of the errors such as spatial continuity. We hope that such structures can better model corruptions in practical face images due to shadows, occlusion or disguise, and misalignment.

In this broader context, the work of [3] essentially considers a special case to the following problem

$$\min_{\mathbf{x},\mathbf{e}} \|\mathbf{x}\|_1 + \psi(\mathbf{e}) \quad s.t. \quad \mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} \tag{3}$$

with the regularizer $\psi(\cdot)$ on $\mathbf{e}$ chosen to be $\|\mathbf{e}\|_1$. The geometry of how ℓ_1-norm penalizing sparse errors is illustrated in Figure 2-(a), i.e., the unit ball of ℓ_1-norm. Clearly, the ℓ_1-norm regularization treats each entry (pixel) in $\mathbf{e}$

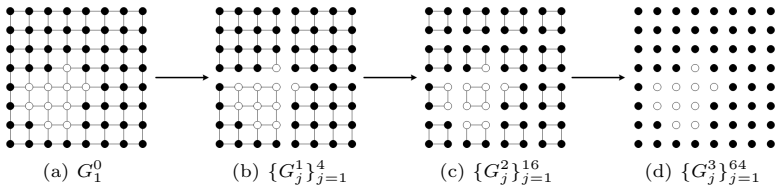

Fig. 1. Illustration of a four-level hierarchical tree group structure defined on the error image. Each circle represents a pixel, and connected circles represent a node/group in the tree. The 8×8 image in (a) is divided into 4 sub-images in (b) according to spatial locality, and each sub-image can be viewed as a child node of (a). The similar relation goes from (b) to (c), and from (c) to (d). Each group of connected black circles represents a node forced to zero, and white circles show the induced sparsity pattern by the tree-structured norm (4).

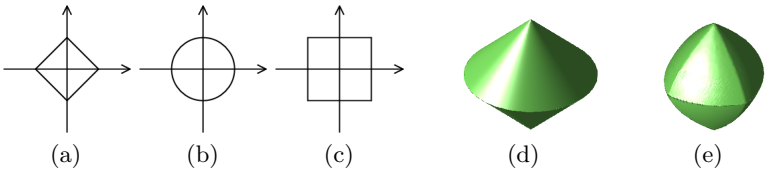

Fig. 2. Unit balls of different norms. (a), (b), and (c) are respectively for ℓ_1-norm, ℓ_2-norm, and ℓ_∞-norm in 2-dimensional space. (d) is for a non-overlapping group Lasso norm in 3-dimensional space: $\psi(\mathbf{e}) = \|\mathbf{e}_{\{1,2\}}\|_2 + |\mathbf{e}_3|$. (e) is for a structured sparsity norm with overlapping groups in 3-dimensional space: $\psi(\mathbf{e}) = \|\mathbf{e}_{\{1,2,3\}}\|_2 + \|\mathbf{e}_{\{1,2\}}\|_2 + |\mathbf{e}_1| + |\mathbf{e}_2| + |\mathbf{e}_3|$. Singular points appearing on these balls characterize the sparsity-inducing behavior of the underlying norms.

independently. It does not take into account any specific structures or possible relations among subsets of the entries. While in face recognition scenarios, shadows caused by illumination change, occlusion, misalignment, or even pose and expression changes normally have the structural properties of spatial contiguity and locality. Indeed, as reported in [3], SRC based on ℓ_1-norm performs better in case of random pixel corruption than contiguous occlusion. Unfortunately the later case is actually closer to practical situations in face recognition.

To encode prior knowledge, researchers have proposed to partition variables into disjoint groups, and use the so called group Lasso penalty [17] to promote sparsity on the group level. Given $\mathbf{e} \in \mathbb{R}^m$, the variables with indices $\{1, \ldots, m\}$ can be partitioned into a disjoint set of groups, denoted as $\mathcal{G}$, with each group $G \in \mathcal{G}$ containing a subset of these indices. A group Lasso norm used in [17] is defined as $\psi(\mathbf{e}) = \sum_{G \in \mathcal{G}} \|\mathbf{e}_G\|_2$. As expected, a regularized solution by this norm has the property that variables in the same group are prone to be zero or nonzero simultaneously. Figure 2-(d) shows a geometric interpretation. Applied to the face error image $\mathbf{e}$, it corresponds to divide $\mathbf{e}$ into non-overlapping local patches. However, the error patterns in $\mathbf{e}$ corresponding to various face variations could have arbitrary shapes, with unknown sizes and number. It is impossible to pre-design disjoint group structures in order to promote error patterns precisely matching corruptions in actual face images.

To induce more diverse and sophisticated sparse error patterns, we consider structured sparsity-inducing norms that involve overlapping groups of variables, motivated by recent advances in structured sparsity [20, 19]. Although it still assumes pre-defined group structures, the overlapping patterns of groups and the norms associated with the groups of variables allow to encode much richer classes of structured sparsity. Figure 2-(d) and -(e) give a geometric comparison between overlapping and non-overlapping group norms for a 3-dimensional vector. In this work, we consider a tree-structured sparsity-inducing norm. It involves a hierarchical partition of the m variables in $\mathbf{e}$ into groups, as shown in Figure 1. The tree is defined in a way that leaf nodes are singleton groups corresponding to individual pixels, and internal nodes/groups correspond to local patches of varying size. Thus each parent node contains a hierarchy of child nodes that are spatially adjacent to each other and constitute a local part in the face error image $\mathbf{e}$. As illustrated in Figure 1, when a parent node goes to zero all its descendents in the tree must go to zero. Consequently, the nonzero or support patterns are formed by removing those nodes forced to zero. This is exactly the desired effect of structured error patterns of spatial locality and contiguity.

To put formally, denote $\mathcal{G}$ as a set of groups from the power set of the index set $\{1, \ldots, m\}$, with each group $G \in \mathcal{G}$ containing a subset of these indices. The tree-structured groups used in this paper are defined as follows: A set of groups $\mathcal{G}$ is said to be *tree-structured* in $\{1, \ldots, m\}$ if $\mathcal{G} = \{\ldots, G_1^i, G_2^i, \ldots, G_{b_i}^i, \ldots\}$ where $i = 0, 1, 2, \ldots, d$, d is the depth of the tree, $b_0 = 1$ and $G_1^0 = \{1, 2, \ldots, m\}$, $b_d = m$ and correspondingly $\{G_j^d\}_{j=1}^m$ are singleton groups. Let G_j^i be the parent node of a node $G_{j'}^{i+1}$ in the tree, we have $G_{j'}^{i+1} \subseteq G_j^i$. For any $1 \leq j, k \leq b_i$, $j \neq k$, we also have $G_j^i \cap G_k^i = \emptyset$.

Similar group structures are also considered in [20, 22]. With the above notation, a general tree-structured sparsity-inducing norm can be written as

$$\psi(\mathbf{e}) = \sum_{i=0}^{d} \sum_{j=1}^{b_i} w_j^i \|\mathbf{e}_{G_j^i}\|_p, \tag{4}$$

where $\mathbf{e}_{G_j^i}$ is a vector with entries equal to those of $\mathbf{e}$ for the indices in G_j^i and 0 otherwise. w_j^i are positive weights for groups G_j^i. It is commonly chosen as $w_j^i = 1$. $\|\cdot\|_p$ denotes ℓ_p-norm with $p \geq 1$, and popular choices of p are $\{2, \infty\}$. Note that support patterns in the error image $\mathbf{e}$ corresponding to practical face variations are usually spatially localized and continuous, such as occlusion or shadow caused by illumination change. Pixels inside each of such error regions may have similarly large magnitude. When applying the sparsity-inducing norm $\|\cdot\|_p$ to $\mathbf{e}_{G_j^i}$, i.e., a group of pixels in a local patch, we expect similar errors in magnitude can be induced. For the ℓ_∞-norm, it is the maximum value of pixels in a group that decides if the group is set to nonzero or not, and it does encourage the rest of the pixels to take arbitrary (hence close to the maximum) values. Thus, in this paper we choose $p = \infty$ in the tree-structured norm (4). Figure 2-(b) and -(c) compares the unit balls of ℓ_∞ and ℓ_2 norms. The effectiveness of this choice is also corroborated with empirical evidences. The so defined norm

(4) promotes sparse error patterns more consistent to practical face variations than standard ℓ_1-norm. Figure 3 shows such an advantage by comparing with [3] on recovering a clean face from occlusion.

3 Robust face recognition via structured sparsity

In this section, we use the so defined structured sparsity-inducing norm to replace the ℓ_1-norm for modeling the error $\mathbf{e}$ in robust face recognition. Thus, the $(\ell_1\_\ell_1)$ objective function in the optimization program (1) is modified to the following

$$(\ell_1\_\ell_{struct}): \qquad \min_{\mathbf{x},\mathbf{e}} \|\mathbf{x}\|_1 + \lambda \sum_{i=0}^{d} \sum_{j=1}^{b_i} w_j^i \|\mathbf{e}_{G_j^i}\|_\infty \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{5}$$

where the sparse vector $\mathbf{x}$ induced by ℓ_1-norm is naturally discriminative and encodes the identity of the test sample $\mathbf{y}$. λ is a parameter controlling the trade-off between sparsity of $\mathbf{x}$ and structured sparsity of $\mathbf{e}$.

A drawback of formulation (5) is that $\mathbf{y}$ could be linearly represented by training samples of multiple subjects. As a consequence, the induced error $\mathbf{e}$ contains both within-class variation and between-class difference. On the other hand, identification of within-class variation is essential for face recognition since misclassification is mainly due to these variations. We thus propose another subject-wise face recognition method that involves solving

$$(\ell_{struct}): \qquad \min_{\mathbf{e}_k,\mathbf{x}_k} \sum_{i=0}^{d} \sum_{j=1}^{b_i} w_j^i \|\mathbf{e}_{k,G_j^i}\|_\infty \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}_k\mathbf{x}_k + \mathbf{e}_k, \tag{6}$$

w.r.t. each subject k of all the K subjects. If $\mathbf{y}$ belongs to subject k, solving (6) makes it possible to identify face regions of $\mathbf{y}$ that correspond to within-class variations. By discarding those regions a clean face image well-approximated by $\mathbf{A}_k$ can be recovered. The formulation (6) is thus a good approach to measure the capabilities of different methods for identifying within-class variations of test images. In this paper, we compare (6) with ℓ_1-norm variant of (6), which was considered in [11], in these settings. When optimizing (6) w.r.t. each subject, ideally the optimal $\mathbf{e}_k^*$ with the true subject would be smallest if based on some properly defined measure. (6) thus suggests new classification criteria which will be introduced shortly.

Both (5) and (6) are convex programs. To solve them we have developed algorithms based on Augmented Lagrange Multiplier (ALM) methods [23]. ALM has demonstrated its good balance between efficiency and accuracy in related sparse representation based face recognition methods [4, 13]. The notable difference here is that in our ALM framework, a subproblem concerns with a proximal problem associated with structured sparsity-inducing norm regularization. A few recently proposed techniques can be exploited to efficiently solve the proximal problems of such kind [21, 22, 20]. For the case of ℓ_∞-norm applied to overlapping groups considered in this paper, solutions can be found by solving a quadratic min-cost

flow problem [21]. Please refer to the supplemental material [1] for details of our developed algorithms for solving (5) and (6).

3.1 Alternative classification criteria

Given a test image $\mathbf{y}$, solving (5) enables us to obtain the optimal sparse vectors $\mathbf{x}^*$ and $\mathbf{e}^*$. When $\mathbf{y}$ is a face image from one of the K classes in the training set, we use the method in [3] for face classification. Denote $\delta_k(\mathbf{x})$ as a function to select coefficients from $\mathbf{x}$ corresponding to training samples of subject k, $\mathbf{y}$ can be classified as the class that minimizes the residuals

$$\text{identity}(\mathbf{y}) = \arg\min_k r_k(\mathbf{y}), \quad r_k(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}_k\delta_k(\mathbf{x}^*)\|_2. \tag{7}$$

Solving (6) w.r.t. each subject gives the optimal vectors $\{\mathbf{e}_k^*\}_{k=1}^K$ and $\{\mathbf{x}_k^*\}_{k=1}^K$. Since $\{\mathbf{x}_k^*\}_{k=1}^K$ are computed locally w.r.t. each subject, it is no longer available to use the criteria as above. Instead, it is natural to compare $\mathbf{e}_k^*$, $k = 1, \ldots, K$, to classify $\mathbf{y}$ if $\mathbf{y}$ is from one of the K training subjects. In this paper, we choose to classify $\mathbf{y}$ to the class that minimizes the structured group sparsity norms

$$\text{identity}(\mathbf{y}) = \arg\min_k \psi(\mathbf{e}_k^*),\ \psi(\mathbf{e}_k^*) = \sum_{i=0}^{d}\sum_{j=1}^{b_i} w_j^i \|\mathbf{e}_{k,G_j^i}^*\|_\infty. \tag{8}$$

This criteria outperforms the conventional ℓ_1-norm alternative, as reported in our experiments in Section 4.

The so obtained $\{\mathbf{e}_k^*\}_{k=1}^K$ provide information for identifying the regions of $\mathbf{y}$ that correspond to either within-class variation or between-class difference[2]. Intuitively, the size of support regions for within-class variation should be smaller than that for between-class difference. This suggects a new classification criteria based on support regions of $\mathbf{e}_k^*$ for $k = 1, \ldots, K$. To identify the support regions, [11] adopted a non-convex formulation based on a Markov random field model. Instead, we here consider a simple thresholding scheme in order to show the superiority of structured sparsity for identification of different face variations. In particular, we can normalize the range of entry values of each $\mathbf{e}_k^*$ to $[0, 1]$. Denote $0 < \tau < 1$ as a threshold parameter, and $\mathbf{s}_k \in \{0, 1\}^m$ as a support vector for each $\mathbf{e}_k^*$. $\mathbf{s}_k$ can be computed by setting $\mathbf{s}_k[i] = 0$ when $\mathbf{e}_k^*[i] \leq \tau$ and $\mathbf{s}_k[i] = 1$ otherwise. With the above notations the new classification criteria based on the sizes of support regions of $\{\mathbf{e}_k^*\}_{k=1}^K$ is defined as

$$\text{identity}(\mathbf{y}) = \arg\min_k \frac{\|\hat{\mathbf{e}}_k^*\|_1}{|\{i|\mathbf{s}_k[i] = 0\}|}\frac{1}{|\{i|\mathbf{s}_k[i] = 0\}|}, \tag{9}$$

where $\hat{\mathbf{e}}_k^*$ is a subvector of $\mathbf{e}_k^*$ with entries of indices corresponding to $\{i|\mathbf{s}_k[i] = 1\}$ removed. Thus the first part in (9) computes the averaged error value for each entry of $\hat{\mathbf{e}}_k^*$, and the introduction of the second part in (9) make this criteria favor $\mathbf{e}_k^*$ with smaller support regions.

[1] `http://web.adsc.com.sg/perception/publications.html`

[2] Usually entries of $\mathbf{e}_k^*$ will be very small in magnitude rather than exactly zero. And support regions of $\mathbf{e}_k^*$ cannot be directly obtained.

3.2 Robust face alignment via structured sparsity

So far we have assumed that the test image $\mathbf{y}$ is well aligned with the training images $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K]$. Precise alignment is crucial for success of sparse representation based face recognition methods – in fact, good alignment is important for any recognition tasks. However, practically observed test image $\mathbf{y}'$ could be subject to some pose change or misalignment, so that the above assumed linear model $\mathbf{y}' = \mathbf{A}_k\mathbf{x}_k + \mathbf{e}_k$ no longer holds for any k. In the context of practical face recognition, $\mathbf{y}'$ can be related to $\mathbf{y}$ by $\mathbf{y} = \mathbf{y}' \circ \tau$, where τ stands for some transformation in the image domain (e.g., 2D similarity transformation for correcting misalignment, or 2D projective transformation for handling some pose change). The objective thus becomes to find the correct τ so that after transformation the obtained $\mathbf{y}$ from $\mathbf{y}'$ can be represented linearly by the training images.

As suggested in [13], the assumption of sparsity itself provides a strong cue for finding the deformation τ. As an extension to the problem (6), based on our structured sparisty, we formulate the alignment problem as the following optimization objective

$$\tau_k^* = \arg\min_{\tau_k, \mathbf{e}_k, \mathbf{x}_k} \sum_{i=0}^{d} \sum_{j=1}^{b_i} w_j^i \|\mathbf{e}_{k,G_j^i}\|_\infty \quad \text{s.t.} \quad \mathbf{y}' \circ \tau_k = \mathbf{A}_k\mathbf{x}_k + \mathbf{e}_k, \tag{10}$$

for $k = 1, \ldots, K$. The problem (10) is a difficult, nonconvex optimization problem over the deformation τ_k, error $\mathbf{e}_k$ and coefficient vector $\mathbf{x}_k$. Fortunately, in practice a good initialization of τ_k can be obtained from the output of an automatic face detector [8]. To solve (10), we follow the strategy of [13] by repeatedly linearizing about the current estimate of τ_k, and seeking a deformation step $\Delta\tau_k$ via the following minimization problem

$$\Delta\tau_k^* = \arg\min_{\Delta\tau_k, \mathbf{e}_k, \mathbf{x}_k} \sum_{i=0}^{d} \sum_{j=1}^{b_i} w_j^i \|\mathbf{e}_{k,G_j^i}\|_\infty \quad \text{s.t.} \quad \mathbf{y}' \circ \tau_k + J\Delta\tau_k = \mathbf{A}_k\mathbf{x}_k + \mathbf{e}_k, \tag{11}$$

where $J = \frac{\partial}{\partial \tau_k}\mathbf{y}' \circ \tau_k$ is the Jacobian of $\mathbf{y}' \circ \tau_k$ w.r.t. the transformation parameters τ_k. The notable difference of model (11) from that considered in [13] is the sparsity-inducing norm enforced on error $\mathbf{e}_k$: here we use structured group sparsity norm while ℓ_1-norm was used in [13]. We empirically observe that when $\mathbf{y}'$ contains large variations such as occlusion or disguise, our model is much better than that in [13] for face alignment and recognition, as reported in our experiments in Section 4. For solving (11), we have again developed an algorithm based on ALM. Please refer to the supplemental material for details of our algorithm. Similar to [13], it is important to normalize the warped image $\mathbf{y}' \circ \tau_k$ in optimization of (11), by replacing the linearization of $\mathbf{y}' \circ \tau_k$ with a linearization of the normalized version $\frac{\mathbf{y}' \circ \tau_k}{\|\mathbf{y}' \circ \tau_k\|_2}$.

After solving (10) w.r.t. all K subjects, the optimal $\{\tau_k^*\}_{k=1}^K$ and $\{\mathbf{e}_k^*\}_{k=1}^K$ can be obtained. The per-subject alignment residuals $\{\mathbf{e}_k^*\}_{k=1}^K$ can be naturally used

Algorithm 1: Robust face alignment and classification via structured sparsity

input : A test image $\mathbf{y}' \in \mathbb{R}^m$, initial transformations $\{\tau_k^0\}_{k=1}^K$, a matrix of well-aligned and normalized training samples of K subjects $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K] \in \mathbb{R}^{m \times n}$, a set of pre-defined tree-structured groups $\mathcal{G} = \{G_j^i\}$ with $i = 0, 1, \ldots, d$ and $j = 1, \ldots, b_i$, the weight $w_j^i \geq 0$ for each G_j^i, and a regularization parameter $\lambda > 0$.

for *each subject* k **do**
  let $\tau_k = \tau_k^0$,
  while *not converged* **do**
    compute an optimal step $\Delta\tau_k^*$ by solving (11): $\Delta\tau_k^* = \arg\min_{\Delta\tau_k, \mathbf{e}_k, \mathbf{x}_k} \sum_{i=0}^{d} \sum_{j=1}^{b_i} w_j^i \|\mathbf{e}_{k,G_j^i}\|_\infty \quad s.t. \quad \mathbf{y}' \circ \tau_k + J\Delta\tau_k = \mathbf{A}_k\mathbf{x}_k + \mathbf{e}_k$,
    update $\tau_k \leftarrow \tau_k + \Delta\tau_k^*$.
  end
end
keep the indices of top S candidates $c_1, \ldots, c_S$ among $\{1, \ldots, K\}$ with the smallest structured group sparsity norm $\psi(\mathbf{e}_k) = \sum_{i=0}^{d} \sum_{j=1}^{b_i} w_j^i \|\mathbf{e}_{k,G_j^i}\|_\infty$.
set $\tilde{\mathbf{A}} \leftarrow [\mathbf{A}_{c_1} \circ \tau_{c_1}^{*-1}, \ldots, \mathbf{A}_{c_S} \circ \tau_{c_S}^{*-1}]$.
compute an optimal $\tilde{\mathbf{x}}^*$ via solving $\tilde{\mathbf{x}}^* = \arg\min_{\tilde{\mathbf{x}}, \mathbf{e}} \|\tilde{\mathbf{x}}\|_1 + \lambda \sum_{i=0}^{d} \sum_{j=1}^{b_i} w_j^i \|\mathbf{e}_{G_j^i}\|_\infty \quad s.t. \quad \mathbf{y}' = \tilde{\mathbf{A}}\tilde{\mathbf{x}} + \mathbf{e}$.
compute the residuals $r_k(\mathbf{y}') = \|\mathbf{y}' - \tilde{\mathbf{A}}_k \delta_k(\tilde{\mathbf{x}}^*)\|_2$ for $k = c_1, \ldots, c_S$.
output : identity$(\mathbf{y}') = \arg\min_k r_k(\mathbf{y}')$.

for robust face recognition. For example, we can use (8) to classify the test image $\mathbf{y}'$ to one of the K subjects. To further improve the recognition performance, a global sparse representation problem (5) can be solved by aligning training samples of each $\mathbf{A}_k$ to $\mathbf{y}'$ using the computed τ_k^*. We thus get a discriminative representation $\mathbf{x}^*$ in terms of the entire training set, and (7) can be used as the criteria for face classification. The complete procedure of our robust face classification with automatic alignment is summarized as Algorithm 1, where the parameter S is used to reduce the number of subjects used in the global sparse representation problem (5), leaving a much smaller problem to solve.

4 Experiments

In this section, we conduct experiments to test the effectiveness of enforcing structured sparsity on the error $\mathbf{e}$ for robust and practical face recognition. We use three publicly available databases including the Extended Yale B [5, 7], AR [10] and Multi-Pie [9] databases. We compare our method with those closely related sparse representation based face recognition methods [3, 11, 13], and also with other baseline classifiers such as Nearest Neighbor (NN), Nearest Subspace (NS), and Support Vector Machine (SVM). We will first present how different methods perform when both training and test images are well aligned, and then present experiments of practical face recognition by automatic face alignment.

4.1 Robust face recognition with well aligned face images

Recognition with synthetic block occlusion. We use Extended Yale B database to test the robustness of our method against illumination change and

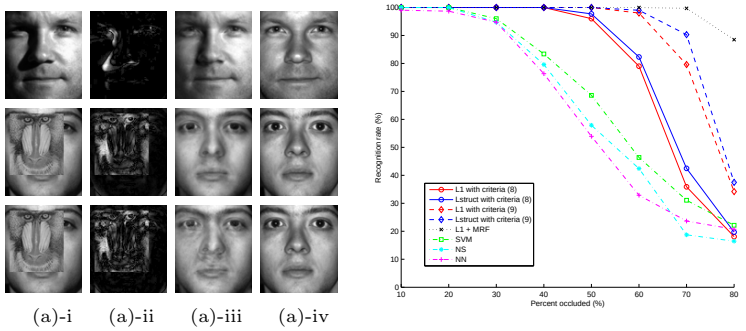

Fig. 3. Recognition on the Extended Yale B database (*better view the electronic version*). (a) shows example results for test images under extreme illumination condition or with large fraction of occlusion: (a)-i test images; (a)-ii estimated error images; (a)-iii recovered images; (a)-iv training images with frontal illumination. Top row in (a) is the result by our method ℓ_1_ℓ_{struct} on a test image under extreme illumination condition. Middle and bottom rows in (a) compare our method with the method ℓ_1_ℓ_1 [3] on a test image with 60% occlusion. (b) plots recognition results of our method ℓ_{struct} and its ℓ_1 variant under classification criteria (8) and (9), and compares with NN, NS, SVM, and the method $\ell_1 + \text{MRF}$ [11].

contiguous occlusion. There are 1238 frontal face images of 38 subjects captured under varying laboratory lighting conditions in Subsets 1, 2, and 3 of the Extended Yale B database. Subsets 1, 2, and 3 contain face images under mild, moderate, and extreme illumination conditions respectively. We choose four illuminations from Subset 1, two from Subset 2, and two from Subset 3 for testing, and the rest of the images are used for training. The total number of training and test images are respectively 935 and 303. All images are manually aligned and cropped to the size of 96×84. In our experiments we simulate various levels of contiguous block occlusion from 10% to 80%, by replacing a randomly located block of each test image with an unrelated image, where locations of the occlusion are unknown to the computer. We test both of our recognition methods, namely ℓ_1_ℓ_{struct} for equation (5) and ℓ_{struct} for equation (6). For ℓ_1_ℓ_{struct}, we set $\lambda = 1$, which is chosen to seek a balanced sparsity between $\mathbf{x}$ and $\mathbf{e}$. We compare our methods with NN, NS, SVM, and especially with related sparse representation based methods, dubbed ℓ_1_ℓ_1 for [3] and $\ell_1 + \text{MRF}$ for [11].

Figure 3-(a) shows example results using our method ℓ_1_ℓ_{struct}. For the case of no occlusion shown in the first row of Figure 3-(a), the obtained error image by our method compensates well for the shadow around nose, which is due to a violation of the assumed linear subspace model. Correspondingly a clean face without dark shadow is recovered. The second and third rows of Figure 3-(a) show results of our method and the method ℓ_1_ℓ_1 for an example test image with 60% occlusion. This is a difficult recognition task even for humans. Careful comparison between the second and third rows of Figure 3-(a) shows that our method performs better in terms of recovering the clean face with no occlusion.

Percent occluded	10%	20%	30%	40%	50%	60%	70%	80%
ℓ_1_ℓ_1 [3]	**100**%	**100**%	**100**%	99.7%	98.0%	68.4%	44.1%	22.4%
ℓ_1_ℓ_{struct}	**100**%	**100**%	**100**%	**100**%	**99.3**%	**73.7**%	**47.0**%	**24.1**%

Table 1. Recognition results of our method ℓ_1_ℓ_{struct} and the method ℓ_1_ℓ_1 [3] on the Extended Yale B database with varying levels of synthetic block occlusion.

We quantitatively compare the recognition performance of different methods in Table 1 and Figure 3-(b). We can see from Table 1 that up to 50% occlusion, our method ℓ_1_ℓ_{struct} performs almost perfectly, and it consistently outperforms the method ℓ_1_ℓ_1 up to 80% occlusion. For our method ℓ_{struct} (problem (6)), we report results in Figure 3-(b) by comparing with a variant of (6), dubbed "ℓ_1"[3], under classification criteria (8) and (9), where τ is set as 0.1 for criteria (9). Under criteria (8), enforcing structured sparsity by ℓ_{struct} gives better results than the ℓ_1 variant does. Under criteria (9), we also compare with NN, NS, SVM, and the method ℓ_1 + MRF [11]. ℓ_1 + MRF uses the ℓ_1 variant of (6) as initialization, and a complicated non-convex optimization method based on MRF to specifically address occlusion. Results by our method based on simple thresholding (cf. Section 3.1) are comparable with those from ℓ_1 + MRF up to 70% occlusion, and also consistently better than those from NN, NS, SVM, and the thresholding based ℓ_1 variant. It should be noted that ℓ_1 + MRF can only address the case that test images are well aligned, while our method is able to automatically align test images, as will be reported shortly. For the well aligned case, our method is also possible to be integrated with MRF to specifically address occlusion, as did by ℓ_1 + MRF [11]. Nevertheless, results in Table 1 and Figure 3 clearly demonstrate that structured sparsity-inducing norm is a better choice for robust face recognition.

Recognition with disguise. We test our method's ability to cope with real disguises using a subset of the AR database. The training set consists of 799 unoccluded face images of 100 subjects with different facial expressions [4]. We consider two separate test sets, each of which contains 200 face images. In the first test set are images of subjects wearing sunglasses, which occlude about 30% of each image. In the second test set are images of subjects wearing a scarf, which occludes roughly half of each image. All training and test images are resized to 83×60. Table 2-Left compares our method ℓ_1_ℓ_{struct} with NN, NS, SVM, and ℓ_1_ℓ_1 [3], where we again set $\lambda = 1$ for ℓ_1_ℓ_{struct}. Table 2-Right compares our method ℓ_{struct} with its ℓ_1 variant under the classification criteria (9) (τ is set as 0.1 for both ℓ_{struct} and its ℓ_1 variant), and also with the method ℓ_1 + MRF [11]. Table 2 shows that ℓ_1 + MRF achieves the best performance for the case of

[3] The ℓ_1 variant of (6) solves the problem: $\min_{\mathbf{e}_k, \mathbf{x}_k} \|\mathbf{e}_k\|_1 \, s.t. \, \mathbf{y} = \mathbf{A}_k \mathbf{x}_k + \mathbf{e}_k$, w.r.t. each subject k of all the K subjects.

[4] We use image IDs $\{1-4\}$ and $\{14-17\}$ for each subject in the AR database, except one corrupted image.

	NN	NS	SVM	ℓ_1_ℓ_1	ℓ_1_ℓ_{struct}	$\ell_1((9))$	$\ell_{struct}((9))$	ℓ_1+MRF
sunglasses	60.5%	59.0%	66.5%	91.0%	**92.5**%	99.0%	**99.5**%	**99.5**%
scarf	14.0%	15.0%	16.5%	64.0%	**69.0**%	84.0%	87.5%	**97.5**%

Table 2. Recognition results of different methods on the AR database with disguises.

occlusion by scarf. Since the scarf used in AR database [10] occludes half (the lower part) of each test image, and it happens to be with dark color and resembles some bearded men in the database, when pursuing sparse representation, there could be a degenerate solution that considers the scarf as the correct signal and the remainder of the face as error. In this case, the non-convex MRF approach in [11] is helpful in iteratively guiding the identification of error support into the scarf region, and hence getting improved performance. However, Table 2 also shows that our method ℓ_1_ℓ_{struct} outperforms ℓ_1_ℓ_1, and our method ℓ_{struct} outperforms its ℓ_1 variant, for both cases of sunglasses and scarf. It demonstrates that promoting structured sparsity on the error image is generally better than promoting standard sparsity using ℓ_1-norm in coping with real disguises.

4.2 Robust face recognition with automatic alignment

In this subsection, we test the effectiveness of our Algorithm 1 for automatic and robust face alignment and recognition, using the CMU Multi-Pie database. The CMU Multi-Pie database contains face images of 337 subjects captured in four sessions with simultaneous variations in illumination, pose, and expression. Of these 337 subjects, we use all the 249 subjects present in Session 1 as training subjects. For each of the 249 subjects we choose frontal images of 7 illuminations [5] with neutral facial expression as training images. As suggested in [13], these 7 extreme illuminations of frontal view are chosen in order to linearly represent other frontal illuminations well. We manually click outer eye corners in all the training images and crop them to the size of 80×60. The distance between the two outer eye corners is normalized to be 50 pixels. We start with experiments on region of attraction to verify the effectiveness of our alignment algorithm, and then present face recognition experiments with automatic alignment.

Experiments on region of attraction. In the CMU Multi-Pie database, we use frontal images of illumination 10 with neutral expression from Session 2 as our test images. We manually align these images in the same way as for training images, to provide ground truth for our region of attraction experiments. We introduce artificial deformation of translation, rotation, or scaling to these test images. To measure success of alignment, we use the structured sparsity norm on error $\mathbf{e}$, i.e., $\psi(\mathbf{e})$ defined in (4), as the alignment error. More specifically, let r_0 be the alignment error obtained by aligning a test image without any artificial perturbation, and r be the error for the case with perturbation. We consider

[5] They are illuminations $\{0, 1, 7, 13, 14, 16, 18\}$ of the total 20 illuminations.

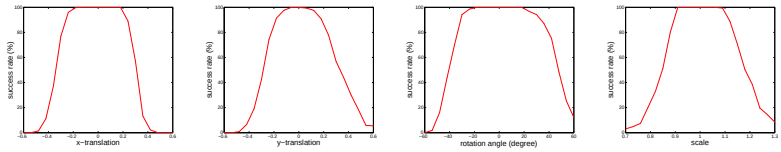

Fig. 4. Experiments on region of attraction. The amount of translation is defined as a fraction of the distance between the outer eye corners. From left to right: translation in x direction, translation in y direction, in-plane rotation, and scale change.

occlusion %	10%	20%	30%	40%	50%
[13],$S = 1$	99.2%	94.4%	76.7%	44.2%	18.5%
Alg.1,$S = 1$	**100**%	95.6%	81.1%	48.6%	20.9%
[13]	99.2%	95.2%	79.1%	48.2%	21.1%
Alg.1	**100**%	**96.8**%	**85.5**%	**52.6**%	**24.5**%

Session 2	Session 3	Session 4
90.7%	89.6%	87.5%
92.1%	90.6%	88.4%
93.9%	93.8%	92.3%
95.7%	**94.9**%	**93.7**%

Table 3. Accuracy of recognition with automatic alignment on the Multi-Pie database. Left table shows recognition results for test images from Session 1 under varying levels of synthetic block occlusion. Right table shows recognition results for test images from Sessions 2 - 4.

the alignment as successful if $|r - r_0| < 0.01r_0$. Region of attraction results for different kinds of deformation are plotted in Figure 4. Figure 4 shows that our algorithm works well when translation is below 20% of the eye corner distance (or 10 pixels) in both x- and y-directions, when in-plane rotation is below 30 degrees, or when change in scale is below 10%. As discussed in [13], outputs from Viola and Jones' face detector [8] fall safely inside this region of attraction.

Experiments on face alignment and recognition. We first test the robustness of our method against misalignment, illumination change, and contiguous occlusion. We use frontal images of illumination 10 from Session 1 (the same session used for training) of the Multi-Pie database as our test images. This choice is deliberate in order to remove other types of occlusion such as hair-style change across sessions. We simulate various levels of contiguous block occlusion from 10% to 50%, by replacing a randomly located block of each test image with an unrelated image. We compare our method with the closely related method [13], which is based on ℓ_1-norm minimization for alignment and recognition. For both methods, outputs from Viola and Jones' face detector [8] are used as initialization of the alignment process. Table 3-Left shows that our method performs reasonably well up to 30% of occlusion, and consistently outperforms [13] for both cases of $S = 1$ and $S = 10$ in Algorithm 1. These results show that enforcing structured sparsity on the error $\mathbf{e}$ is a better choice in simultaneously handling misalignment, illumination change, and contiguous occlusion.

We also test our method on frontal images of all the 20 illuminations from Sessions $2 - 4$ of the Multi-Pie database. Table 3-Right reports our results, and compares with those from [13]. Again, our method achieves better results.

Acknowledgments. This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapores Agency for Science, Technology and Research (A*STAR), and the funding of ONR N00014-09-1-0230, NSF CCF 09-64215, NSF IIS 11-16012, and DARPA KECoM 10036- 100471.

References

1. Turk, M., Pentland, A.: Eigenfaces for recognition. In: CVPR (1991)
2. Belhumeur, P., Hespanda, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. PAMI, vol. 19, no. 7, pp. 711-720 (1997)
3. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. PAMI, vol. 31, no. 2, pp. 210-227 (2009)
4. Yang, A. Y., Ganesh, A., Zhou, Z., Sastry, S., Ma, Y.: A review of fast ℓ_1-minimization algorithms for robust face recognition. Preprint (2010)
5. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. PAMI (2001)
6. Basri, R., Jacobs, D.: Lambertian reflectance and linear subspaces. PAMI (2003)
7. Lee, K., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. PAMI, vol. 27, no. 5, pp. 684-698 (2005)
8. Viola, P., Jones, M. J.: Robust real-time face detection. IJCV (2004)
9. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. In: FG (2008)
10. Martinez, A., Benavente, R.: The AR face database. CVC T.R., No. 24 (1998)
11. Zhou, Z., Wagner, A., Wright, J., Mobahi, H., Ma, Y.: Face recognition with contiguous occlusion using markov random fields. In: ICCV (2009)
12. Cevher, V., Duarte, M. F., Hegde, C., Baraniuk, R. G.: Sparse signal recovery using markov random fields. In: NIPS (2008)
13. Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mobahi, H., Ma, Y.: Towards a practical face recognition system: robust alignment and illumination by sparse representation. PAMI (2011)
14. Elhamifar, E., Vidal, R.: Robust classification using structured sparse representation. In: CVPR (2011)
15. Zhang, L., Yang, M., Feng, X. C.: Sparse representation or collaborative representation which helps face recognition?. In: ICCV (2011)
16. Tibshirani, R.: Regression shrinkage and selection via the Lasso. Journal of the Royal Stat. Soc., Series B, pages 267-288 (1996)
17. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Stat. Soc., Series B, 68(1):49-67 (2006)
18. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sci., 2(1):183-202 (2009)
19. Jenatton, R., Audibert, J.-Y., Bach, F.: Structured variable selection with sparsity-inducing morms. JMLR, 12(Oct):2777-2824 (2011)
20. Zhao, P., Rocha, G., Yu, B.: The composite absolute penalties family for grouped and hierarchical variable selection. Annals of Statistics, 37(6A):3468-3497 (2009)
21. Mairal, J., Jenatton, R., Obozinski, G., Bach, F.: Network flow algorithms for structured sparsity. In: NIPS (2010)
22. Liu, J., Ye, J.: Moreau-Yosida regularization for grouped tree structure learning. In: NIPS (2010)
23. Bertsekas, D.: Constrained optimization and Lagrange multiplier methods. Academic Press (1982)