

Robust Subspace Clustering With Complex Noise

Ran He, *Senior Member, IEEE*, Yingya Zhang, Zhenan Sun, *Member, IEEE*, and Qiyue Yin

Abstract—Subspace clustering has important and wide applications in computer vision and pattern recognition. It is a challenging task to learn low-dimensional subspace structures due to complex noise existing in high-dimensional data. Complex noise has much more complex statistical structures, and is neither Gaussian nor Laplacian noise. Recent subspace clustering methods usually assume a sparse representation of the errors incurred by noise and correct these errors iteratively. However, large corruptions incurred by complex noise cannot be well addressed by these methods. A novel optimization model for robust subspace clustering is proposed in this paper. Its objective function mainly includes two parts. The first part aims to achieve a sparse representation of each high-dimensional data point with other data points. The second part aims to maximize the correntropy between a given data point and its low-dimensional representation with other points. Correntropy is a robust measure so that the influence of large corruptions on subspace clustering can be greatly suppressed. An extension of pairwise link constraints is also proposed as prior information to deal with complex noise. Half-quadratic minimization is provided as an efficient solution to the proposed robust subspace clustering formulations. Experimental results on three commonly used data sets show that our method outperforms state-of-the-art subspace clustering methods.

Index Terms—Subspace clustering, subspace segmentation, correntropy, half-quadratic minimization.

I. INTRODUCTION

SUBSPACE clustering is a commonly used technique to segment high-dimensional data that are drawn from a union of multiple subspaces. It refers to the task of finding a low-dimensional subspace into which each group of data can fit its own subspace structure simultaneously [1]. Recently, subspace clustering has attracted a great attention due to its promising applications in computer vision and machine learning. A large number of subspace clustering methods have been proposed in the literature, which can be almost classified into four categories: iterative methods, algebraic methods, statistical methods, and spectral clustering-based methods [1].

Iterative methods, such as K-subspaces [2], first assign data to pre-defined multiple subspaces, and then update the

subspaces and reassign each data point to the nearest subspace. Iteratively repeating these two steps to convergence, one can obtain the segmentation result. One disadvantage of these methods is that they need to know the number of the subspaces and their dimensions in advance. Generalized Principal Component Analysis (GPCA) [3] uses an algebraic way to model and segment data. It fits the data with a polynomial, and the gradient of the polynomial at a point gives a normal vector that the point belongs to. But its computational complexity grows exponentially as the data dimension increases. Statistical approaches, such as Mixture of Probabilistic PCA (MPPCA) [4] and Multi-Stage Learning (MSL) [5], assume that sample points are drawn from a mixture of probabilistic distributions, and resort to the Expectation Maximization (EM) technique to estimate the subspaces and cluster the data iteratively. Although these methods have achieved good performance under certain conditions, they are often sensitive to noise and outliers in real-world applications.

Recent advances in subspace clustering are spectral clustering-based methods [6], which assume that each data point has a linear representation on all other data points in the same subspace. The representational coefficients are used to construct the affinity matrix, on which spectral clustering algorithms are applied to obtain correct segmentation. Elhamifar and Vidal [7] proposed the Sparse Subspace Clustering (SSC) algorithm, which aims to find the sparsest representation of each data point by imposing ℓ_1 norm on the coefficient matrix. Low-Rank Representation (LRR) [8]–[10], in another way, seeks to find the lowest-rank representation of all data points by using nuclear norm, which can capture the global structures of the data. Based on ℓ_2 norm regularization, Lu et al. [11] proposed the Least Square Regression (LSR) method for subspace clustering under Enforced Block Diagonal (EBD) conditions. Then, based on block-diagonal prior (or grouping effect), Feng et al. [12] studied block-diagonal subspace clustering methods, and Hu et al. [13] studied smooth representation subspace clustering model. Bayesian method [14], quadratic programming [15], least squares regression [11], [16], manifold regularization [17] and Markov random walks [18] were also used respectively to improve clustering accuracy. In addition, Patel et al. [19] proposed a latent space sparse subspace clustering method to simultaneously reduce dimension and segment data. Dictionary learning methods [16], [20] were adopted to learn a clean dictionary for subspace clustering. Constraint-based subspace clustering [21], distributed low-rank subspace segmentation [22] and fast low-rank subspace segmentation [23] were developed to save the computational costs of subspace clustering.

Manuscript received October 10, 2014; revised June 2, 2015; accepted July 9, 2015. Date of publication July 15, 2015; date of current version July 31, 2015. This work was supported by the National Basic Research Program of China under Grant 2012CB316300 and in part by the National Natural Science Foundation of China under Grant 61473289 and Grant 61135002. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joseph P. Havlicek.

The authors are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: rhe@nlpr.ia.ac.cn; yzyzhang@nlpr.ia.ac.cn; znsun@nlpr.ia.ac.cn; qyyin@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TIP.2015.2456504

The main differences among these methods include the regularization of the coefficient matrix and preprocessing (or post-processing) methods. However, the main challenge of subspace clustering is to handle complex noise. Although ℓ_2 norm and ℓ_1 norm can well deal with small Gaussian or Laplacian noise, complex noise is neither Gaussian nor Laplacian. Complex noise has much more complex statistical structures [24], and is from different kinds of corruptions or occlusions. For example, the face images occluded by sunglasses (Fig. 3) contain two types of noise incurred by sunglasses occlusion and highlight occlusion respectively. The errors incurred by noise often dominate the loss functions of subspace clustering and may lead to poor clustering results [25]. Despite the variety of the regularization, most of these methods assume that complex errors have a sparse representation, and correct these errors iteratively. However, the errors in real-world problems are often complex [26] and can not be well addressed by error correction methods [27].

This paper aims to propose a novel subspace clustering method that is robust against complex noise. Our basic idea is to minimize the influence of error data points on subspace clustering via a robust measure of the similarity (correntropy [28]) between data points and their subspace representations. The optimization model of the proposed robust subspace clustering method aims to achieve sparse representation coefficients and minimal reconstruction errors simultaneously. The proposed method can be naturally extended into a constrained clustering framework, in which must-link and cannot-link constraints [29] can be further used to improve clustering accuracy. An efficient iterative solution to the proposed problem based on half-quadratic (HQ) optimization is provided. In each iteration, the complex optimization problems are simplified to quadratic problems that have a closed form solution in half-quadratic optimization. The performance of our method is evaluated and compared with state-of-the-art subspace clustering methods on motion segmentation and face clustering problems. The main contributions of this work are summarized as follows:

- 1) The proposed method resorts to non-convex correntropy to deal with complex noise, which has a close relationship to M-estimation. It is one of the early solutions for the robust subspace clustering problem in terms of robust measures. Our M-estimation analysis in Sec. III-D also provides a convex Huber M-estimation formulation of the efficient dense subspace clustering in [16].
- 2) The proposed method elegantly integrates the pairwise link constraint information, which results in a novel semi-supervised clustering algorithm. To the best of our knowledge, the issue of must-link and cannot-link is less addressed in subspace clustering.
- 3) The introduced solution leads to considerable improvement in motion segmentation and face clustering without using any pre-processing or post-processing steps. It is robust to complex noise and can also handle real-world clustering problems where noise is often unpredictable.

This paper is an extension of our preliminary conference paper reported in [25]. In addition to provide more in-depth analysis of convex Huber M-estimation for subspace clustering

and more extensive experiments on five types of noise, the major difference between this paper and [25] is the introduction of pairwise link constraints. In many applications, since it is cheap to obtain some pairwise link information between data, pairwise link constraints can be used as prior information for all robust subspace clustering algorithms to improve clustering accuracy. Compared with other pre-processing or post-processing methods [10] that are proposed for a specific clustering task, the link information is a common strategy for any tasks. Our experiments suggest that for the complex noise in real-world problems, a robust subspace representation can be better learned by using a robust non-convex measure, leading to further improvement of clustering accuracy.

The rest of the paper is organized as follows. In Section II, we revisit sparse subspace clustering. In Section III, we derive new robust subspace clustering models from the viewpoint of correntropy, together with efficient half-quadratic solution in Section IV. In Section V, we compare the proposed approaches with related methods on two benchmark subspace clustering databases. Finally, we draw the conclusion in Section VI.

II. SPARSE SUBSPACE CLUSTERING

To better illustrate the main idea of our method in the context of sparse subspace clustering (SSC) [7], we briefly review the basic idea of SSC in this section.

Given a $D \times N$ data matrix X consisting of N vectors $\{x_i \in \mathbb{R}^D\}_{i=1}^N$, which are drawn from a union of linear or affine subspaces $\{S_i\}_{i=1}^K$ of unknown dimensions $d_k = \dim(S_k)$, $0 < d_k < D$, the task of subspace clustering is to find the number of subspaces K , their dimensions $\{d_k\}_{k=1}^K$ and segment the data vectors x_i into these subspaces. $X_{\hat{i}} \in \mathbb{R}^{D \times N}$ stands for the matrix obtained from X by replacing its i -th column x_i with the vector $\mathbf{0} \in \mathbb{R}^D$ of all zeros. Given a vector $z \in \mathbb{R}^p$, let $\text{diag}(z)$ be the diagonal $p \times p$ matrix whose i -th main diagonal element is the i -th entry of z . The matrix I stands for the identity matrix and $\mathbf{1}$ for a vector of all 1s. The j -th entry of the data point x_i is denoted by $(x_i)_j$.

Sparse subspace clustering assumes that each data point sampled from a union of subspaces can be formulated as a sparse representation of other points in the dataset, which is called as self-expressiveness property in [7]. That is, for a data point x_i , one wants to find a coefficient vector c_i that satisfies,

$$x_i = X_{\hat{i}} c_i \quad (1)$$

In practice, this is often an ill-posed problem because the solution of (1) is not unique in general. A possible solution of this problem is to find the sparsest representation of x_i . That is, one can minimize the number of nonzero entries of the coefficient vector c_i w.r.t. x_i ,

$$\min_{c_i} \|c_i\|_1 \quad \text{s.t.} \quad x_i = X_{\hat{i}} c_i \quad (2)$$

where $\|c_i\|_1 = \sum_{j=1}^N |c_{ij}|$ is the relaxation of l_0 norm.

Since data points may contain small Gaussian noise, a natural way is to extend the original problem (2) to the following form by relaxing the equality constraint and adding

a penalty to cost,

$$\min_{\mathbf{c}_i} \|\mathbf{c}_i\|_1 + \gamma \|\mathbf{x}_i - X_{\hat{i}} \mathbf{c}_i\|_2^2 \quad (3)$$

The l_2 norm term $\|\mathbf{x}_i - X_{\hat{i}} \mathbf{c}_i\|_2^2$ measures the fidelity of sparse representation.

In more practical and general scenarios, the data points may be contaminated by both large corruptions and small dense noise. SSC can be extended as the following model using the technique in robust face recognition [30],

$$\min_{\mathbf{c}_i, \mathbf{e}_i} \|\mathbf{c}_i\|_1 + \lambda \|\mathbf{e}_i\|_1 + \gamma \|\mathbf{x}_i - X_{\hat{i}} \mathbf{c}_i - \mathbf{e}_i\|_2^2 \quad (4)$$

where the term $\mathbf{e}_i \in \mathbb{R}^D$ models large corruptions which have sparse nonzero entries.

SSC has become an important and popular method in subspace clustering. Wang and Xu [26] further studied the behavior of SSC under either adversarial or random noise. Soltanolkotabi et al. [31] resorted to the ideas from geometric functional analysis to show that SSC can accurately recover the underlying multiple subspaces. In robust sparse representation [24], the method to solve (4) can be categorized into error correction methods, which iteratively recover corrupted data from errors incurred by complex noise. Recent theoretical analysis and experimental results [24], [32] show that error correction methods may not deal with errors well when complex noise occurs.¹ Hence in the following sections, we will derive a robust subspace clustering method based on a robust measure.

III. ROBUST SUBSPACE CLUSTERING

In this section, we firstly introduce a robust measure (i.e., correntropy) for robust subspace clustering, and then formulate the robust subspace clustering problems as entropy minimization problems.

A. Correntropy

In order to deal with non-Gaussian noise and impulsive noise in signal processing, the concept of Correntropy was proposed in [28]. According to Renyi's quadratic entropy, the correntropy of two arbitrary variables $\mathbf{c}_u$ and $\mathbf{c}_v$ takes the form,

$$V_\sigma(\mathbf{c}_u, \mathbf{c}_v) = \mathbf{E}[k_\sigma(\mathbf{c}_u - \mathbf{c}_v)] \quad (5)$$

where $k_\sigma(\cdot)$ is a kernel function. And it is used to measure the generalized similarity between $\mathbf{c}_u$ and $\mathbf{c}_v$.

In practice, the joint probability density function may be unknown, and only a finite number of data $\{((\mathbf{c}_u)_i, (\mathbf{c}_v)_i)\}_{i=1}^N$ are available. Therefore, a sample estimator of correntropy is introduced,

$$V_\sigma(\mathbf{c}_u, \mathbf{c}_v) = \frac{1}{N} \sum_{i=1}^N k_\sigma((\mathbf{c}_u)_i - (\mathbf{c}_v)_i) \quad (6)$$

In this paper, we use the Gaussian kernel $g(t) = \exp(-t^2/\sigma^2)$ where σ is the kernel size.

¹Here complex noise indicates that there are several types of outliers.

Liu et al. [28] further proposed a metric for any two vectors in the sample space which is named as Correntropy Induced Metric (CIM) on the basis of (6). It is defined as follows,

$$\begin{aligned} CIM(\mathbf{c}_u, \mathbf{c}_v) &= \{k(0) - V(\mathbf{c}_u, \mathbf{c}_v)\}^{1/2} \\ &= (g(0) - \frac{1}{N} \sum_{i=1}^N g((\mathbf{c}_u)_i - (\mathbf{c}_v)_i))^{1/2} \end{aligned} \quad (7)$$

It should be noted that CIM is a decreasing function of correntropy, so the maximization of correntropy is equivalent to the minimization of CIM. Compared to the global metric mean square error (MSE), the correntropy is a local metric. That means, the value of correntropy is mainly decided by the kernel function along the line $\mathbf{x} = \mathbf{y}$ [28]. Moreover, it becomes practical to choose an appropriate kernel size for correntropy because of the close relationship between correntropy and M-estimators [28].

B. Proposed Robust Formulations

A main problem of existing subspace clustering methods is how to measure the reconstruction fidelity of outlier data points using a robust function. So correntropy is introduced as a robust measure in regularization terms of subspace clustering. A novel formulation of robust subspace clustering is proposed as follows:

$$\min_{\mathbf{c}_i} \sum_{j=1}^N \phi_r((\mathbf{c}_i)_j) + \gamma \sum_{j=1}^D \phi_c((\mathbf{x}_i - X_{\hat{i}} \mathbf{c}_i)_j) \quad (8)$$

where $\phi_c(x) = (1 - \exp(-x^2/\sigma^2))$ and $\phi_r(x) = \sqrt{x^2 + \alpha}$, which is called ℓ_1 - ℓ_2 loss function in M-estimation. γ is a positive scalar.

To be more specific, the first term in (8) controls the sparsity of the coefficient vector $\mathbf{c}_i$. Although ℓ_1 loss is convex, it is non-differentiable at zero-point. The optimization methods for ℓ_1 regularization often oscillate around the true optimum and slowly converge towards the optimum. Compared to ℓ_1 loss, ℓ_1 - ℓ_2 loss around the zero-point is differentiable and can be efficiently solved. Specially, when $\alpha \rightarrow 0$, the ℓ_1 - ℓ_2 loss turns to ℓ_1 loss. And the second term is the squared CIM, which is a robust measure of the reconstruction fidelity of high-dimensional data points with other data. The parameter γ is used to keep a balance between reconstruction fidelity and coefficient sparsity.

Because correntropy criterion is local along the bisector of the joint space and lies on the first quadrant of a sphere [28], the problem (8) treats the representation of each entry of $\mathbf{x}_i$ differently. For example, if there exist outliers in $\mathbf{x}_i$, those entries in $\mathbf{x}_i$ will be far away from the joint space between $\mathbf{x}_i$ and $X_{\hat{i}}$ so that they only have limited influence to the correntropy. Hence the data are weighted in a robust way depending on the residuals. Compared with most existing subspace clustering methods, our method can achieve a more robust solution because the subspace clustering optimization is mainly determined by the uncorrupted data entries and the errors caused by noise or corruptions in data are greatly suppressed.

In practice, the errors in data are usually unpredictable. An extension of our model by explicitly modeling error terms is proposed as follows:

$$\min_{c_i, e_i} \sum_{j=1}^N \phi_r((c_i)_j) + \lambda \sum_{j=1}^D \phi_r((e_i)_j) + \gamma \sum_{j=1}^D \phi_c((x_i - X_i c_i - e_i)_j) \quad (9)$$

Here the term e_i is used to model the errors existing in data. And the error term e_i is assumed to have a sparse representation, which is expressed by the ℓ_1 - ℓ_2 loss function. The third term of the objective function is a robust reconstruction fidelity measure of high-dimensional data using subspace representation and residual errors based on correntropy. Compared with the first formulation of robust subspace clustering defined in (8), the second formulation (9) is more suitable for the applications with large corruptions since its optimization model explicitly has error terms. The second and third items in (9) correspond to error correction and error detection respectively [24]. The usage of two types of error processing models makes our method more flexible to deal with complex noise.

Let $\omega_i = [c_i^T e_i^T]^T$ and $B = [X_i I]$ where I is the identity matrix, we can reformulate (9) as,

$$\min_{\omega_i} \sum_{j=1}^{N+D} \phi_r((\omega_i)_j) + \gamma \sum_{j=1}^D \phi_c((x_i - B\omega_i)_j) \quad (10)$$

Comparing (10) with (8), we learn that (10) and (8) are exactly the same minimization problem with different variables. Although c_i and e_i indicate linear coefficients and outliers respectively, both of them are assumed to have a sparse representation. Hence, in robust sparse representation [30], c_i and e_i are often solved simultaneously in terms of sparsity.

C. Extensions to Constrained Clustering

In many web applications, it is cheap to obtain some pairwise link information that tells whether two data points are from the same cluster or not. In constrained clustering methods [33]–[36], these pairwise links involve data level must-link and cannot-link constraints. A must-link constraint enforces that two data points must be placed in the same cluster whereas a cannot-link constraint enforces that two data points must not be placed in the same cluster. The must-link and cannot-link constraints are often used to guide clustering and to promise the feasibility of clustering results [29]. Considering the problem of clustering data X , we can define the two types of constraints as [29] and [37]:

Must-Link Constraints: Each must-link constraint involves a pair of points x_i and x_j ($i \neq j$) and indicates that points x_i and x_j must be clustered to the same class in any feasible clustering.

Cannot-Link Constraints: Each cannot-link constraint involves a pair of distinct points x_i and x_j ($i \neq j$), and indicates that points x_i and x_j must not be clustered to the same class in any feasible clustering.

Denote O be the set of must-link and cannot-link constraint pairs. For subspace clustering problem, denote A is a partially observed link matrix defined as follows,

$$A_{ij} = \begin{cases} 1 & \text{must-link pair} \\ 0 & \text{cannot-link pair} \end{cases} \quad (11)$$

Given link pairs, the constrained subspace clustering problem can be formulated as the following minimization problem,

$$\min_{c_i} \sum_{j=1}^N \phi_r((c_i)_j) + \gamma \sum_{j=1}^D \phi_c((x_i - X_i c_i)_j) + \lambda \sum_{ij \in O} (c_{ij} - A_{ij})^2 \quad (12)$$

where λ is const. The last item in (12) aims to make the learned sparse representation c_i agree with the link constraints in A .

Note that both (9) and (12) are extensions of (8) to deal with complex noise. In particular, (12) can further apply prior information to improve clustering accuracy. Since (9) can be formulated as (10), we can also extend (12) to deal with link constraints problems for (9).

D. Connections to Huber M-Estimation

Since correntropy has a close relationship to Welsch M-estimator, we further discuss a convex case of (8) from the viewpoint of M-estimators [38]. In computer vision, M-estimators [38] are widely used to improve the robustness of learning algorithms. They are generalized maximum likelihood estimation to the minimization of the sums of functions of the data. Commonly used technique to minimize M-estimators is half-quadratic (HQ) minimization [39]. By applying the multiplicative and the additive HQ reformulation of M-estimators, the original M-estimation problem can be minimized (or maximized) in an alternating way of an augmented objective function. Some popular M-estimators include ℓ_p function, $\ell_1 - \ell_2$ function, Huber function, and Welsch function [24].

If we substitute the Welsch function in (8) with Huber function, (8) becomes,

$$\min_{c_i} \sum_{j=1}^N \phi_r((c_i)_j) + \lambda_1 \sum_{j=1}^D \phi_H((x_i - X_i c_i)_j) \quad (13)$$

where $\phi_H(\cdot)$ is Huber function defined as,

$$\phi_H(v) = \begin{cases} \frac{1}{2}v^2 & |v| \leq \lambda_2 \\ \lambda_2|v| - \frac{1}{2}\lambda_2^2 & |v| > \lambda_2 \end{cases} \quad (14)$$

In the additive form of HQ, the dual potential function of Huber function is the absolute function, i.e.,

$$\phi_H(v) = \min_u (v - u)^2 + \lambda_2 |u| \quad (15)$$

The optimal solution of u is determined by the following additive minimizer function,

$$\delta_A(v) = \begin{cases} 0 & |v| \leq \lambda_2 \\ v - \lambda_2 \text{sign}(v) & |v| > \lambda_2 \end{cases} \quad (16)$$

Considering ϕ_r is a square loss function, (13) becomes the following standard ℓ_2 regularized M-estimation problem,

$$\min_{\mathbf{c}_i} \frac{1}{2} \|\mathbf{c}_i\|_2^2 + \lambda_1 \sum_{j=1}^D \phi_H((\mathbf{x}_i - X_i \mathbf{c}_i)_j) \quad (17)$$

By applying the additive form in (15), the augmented objective function of (17) takes the following form,

$$\min_{\mathbf{c}_i, \mathbf{e}_i} \frac{1}{2} \|\mathbf{c}_i\|_2^2 + \lambda_1 \sum_{j=1}^D \left\{ ((x_i - X_i \mathbf{c}_i - \mathbf{e}_i)_j)^2 + \lambda_2 |(e_i)_j| \right\} \quad (18)$$

where $\lambda_2 |e_{ij}|$ is the dual potential function of Huber function [39]. (18) can be further reformulated in a matrix form, i.e.,

$$\min_{C, E} \frac{1}{2} \|C\|_F^2 + \frac{\lambda_1}{2} \|X - XC - E\|_F^2 + \lambda_2 \|E\|_1 \quad (19)$$

By introducing a new variable $D = XC$, we have the following constrained problem,

$$\begin{aligned} \min_{D, C, E} \quad & \frac{1}{2} \|C\|_F^2 + \frac{\lambda_1}{2} \|X - D - E\|_F^2 + \lambda_2 \|E\|_1 \\ \text{s.t.} \quad & D = XC \end{aligned} \quad (20)$$

In subspace clustering, most of methods rely on the self-expressiveness of X . Since X may be corrupted by noise, Ji et al. [16] proposed to learn a dictionary to perform self-expressiveness. By modifying the equality constraint in (20) as $D = DC$, we have the optimization problem in [16] as follows,

$$\begin{aligned} \min_{D, C, E} \quad & \frac{1}{2} \|C\|_F^2 + \frac{\lambda_1}{2} \|X - D - E\|_F^2 + \lambda_2 \|E\|_1 \\ \text{s.t.} \quad & D = DC \end{aligned} \quad (21)$$

Fixing D and C , we know that the solution of $\min_E \|X - D + E\|_F^2 + \lambda_2 \|E\|_1$ is determined by the scalar soft-thresholding operator in (16) (i.e., the additive minimizer function in HQ). If the scalar soft-thresholding operator in (16) is adopted to minimize (21), minimizing (21) is equal to minimizing the following Huber M-estimation problem,

$$\begin{aligned} \min_{D, C, E} \quad & \frac{1}{2} \|C\|_F^2 + \frac{\lambda_1}{2} \sum_{i=1}^N \sum_{j=1}^D \phi_H(X_{ij} - D_{ij}) \\ \text{s.t.} \quad & D = DC \end{aligned} \quad (22)$$

Hence the subspace clustering method proposed in [16] has a potential relationship to Huber M-estimation. In robust M-estimation, the parameter λ_2 in (21) corresponding to Huber function affects the robustness of (21).

IV. SOLUTION OF THE PROPOSED FORMULATIONS

An iterative regularization method based on half-quadratic optimization is firstly proposed to solve the proposed formulations in (8), (9) and (12). And then the convergence of the solution is analyzed. Finally the whole procedure of our subspace clustering algorithm is given. Since the optimization problem in (8) is a special case of (9), only the solutions to (9) and (12) are provided.

TABLE I
LOSS FUNCTIONS AND THEIR MINIMIZER FUNCTIONS [27]

	Functions $\phi(\cdot)$	Minimizer functions $\delta(\cdot)$
ℓ_1 - ℓ_2	$\sqrt{\alpha + x^2}$	$1/\sqrt{\alpha + x^2}$
Welsch	$1 - \exp(-\frac{x^2}{\sigma^2})$	$\exp(-\frac{x^2}{\sigma^2})$

A. The Half-Quadratic Approach

Since the problems (8) and (9) are not convex, it is difficult to optimize them directly. Fortunately, the half-quadratic technique [40] can be utilized to optimize the non-convex function by minimizing its augmented function alternately.

According to the conjugate function theory [41] and HQ theory [40], [42], we have:

Lemma 1: Suppose that $\phi(x)$ is a function that satisfies some conditions listed in [40], then for a fixed x , there exists a dual potential function $\psi(\cdot)$, such that

$$\phi(x) = \inf_{s \in \mathbb{R}} \{sx^2 + \psi(s)\} \quad (23)$$

where s is an auxiliary variable which is determined by the minimizer function $\delta(\cdot)$ with respect to $\phi(x)$ (two specific functions and their minimizer functions are listed in Table I).

According to Lemma 1, the function $\phi_r(\cdot)$ in (9) takes the following half-quadratic form,

$$\begin{aligned} \sum_{j=1}^N \phi_r((c_i)_j) &= \min_{\mathbf{r}} \sum_{j=1}^N (\mathbf{r}_j (c_i)_j)^2 + \psi_r(\mathbf{r}_j) \\ &= \min_{\mathbf{r}} (\mathbf{c}_i^T R \mathbf{c}_i + \sum_{j=1}^N \psi_r(\mathbf{r}_j)) \end{aligned} \quad (24)$$

where $\mathbf{r} \in \mathbb{R}^N$ is an auxiliary vector and $R = \text{diag}(\mathbf{r})$. We can also reformulate other HQ functions in (9) in a half-quadratic way. Then, the augmented cost-function $\mathcal{J}$ of (9) reads

$$\begin{aligned} \mathcal{J}(\mathbf{c}_i, \mathbf{e}_i, \mathbf{r}, s, \mathbf{q}) &= \mathbf{c}_i^T R \mathbf{c}_i + \lambda \mathbf{e}_i^T S \mathbf{e}_i \\ &\quad + \gamma (\mathbf{x}_i - X_i \mathbf{c}_i - \mathbf{e}_i)^T Q (\mathbf{x}_i - X_i \mathbf{c}_i - \mathbf{e}_i) \\ &\quad + \sum_{j=1}^N \psi_r(\mathbf{r}_j) + \sum_{j=1}^D \psi_r(s_j) + \sum_{j=1}^D \psi_c(\mathbf{q}_j) \end{aligned} \quad (25)$$

where $\mathbf{s} \in \mathbb{R}^D$, $\mathbf{q} \in \mathbb{R}^D$ are auxiliary vectors, and $S = \text{diag}(\mathbf{s})$, $Q = \text{diag}(\mathbf{q})$.

In HQ, $\mathcal{J}$ can be minimized using alternate minimization. The auxiliary vectors in $\mathcal{J}$ are only determined by their minimizer functions. When the values of the auxiliary vectors are given, $\sum \psi_r(\mathbf{r}_j)$, $\sum \psi_r(s_j)$, and $\sum \psi_c(\mathbf{q}_j)$ become constant and then can be eliminated. λ is a positive constant scalar, so we can reformulate the function (25) as follows,

$$\begin{aligned} \mathcal{J}(\mathbf{c}_i, \mathbf{e}_i) &= \mathbf{c}_i^T R \mathbf{c}_i + \mathbf{e}_i^T (\lambda S) \mathbf{e}_i \\ &\quad + \gamma (\mathbf{x}_i - X_i \mathbf{c}_i - \mathbf{e}_i)^T Q (\mathbf{x}_i - X_i \mathbf{c}_i - \mathbf{e}_i) \end{aligned} \quad (26)$$

or

$$\mathcal{J}(\mathbf{w}_i) = \mathbf{w}_i^T P \mathbf{w}_i + \gamma (\mathbf{x}_i - Y \mathbf{w}_i)^T Q (\mathbf{x}_i - Y \mathbf{w}_i) \quad (27)$$

Algorithm 1 Solving Problem (9) via HQ Minimization**Input:** A data point $\mathbf{x}_i \in \mathbb{R}^D$, the matrix $X \in \mathbb{R}^{D \times N}$.**Output:** $\mathbf{c}_i \in \mathbb{R}^N$ and $\mathbf{e}_i \in \mathbb{R}^D$.

- 1: $\mathbf{w}_i \leftarrow 0$, $Y \leftarrow [X_i, I]$ and $t \leftarrow 1$.
- 2: **repeat**
- 3: $p_j^t = \begin{cases} 1/\sqrt{(\mathbf{w}_i)_j^2 + \alpha} & j \leq N \\ \lambda/\sqrt{(\mathbf{w}_i)_j^2 + \alpha} & \text{otherwise} \end{cases}$;
- 4: $q_k^t = \exp(-(\mathbf{x}_i - Y\mathbf{w}_i)_k^2/\sigma^2)$;
- 5: $\mathbf{w}_i^t = \gamma(P^t + \gamma Y^T Q^t Y)^{-1} Y^T Q^t \mathbf{x}_i$;
- 6: $\sigma^2 = \frac{1}{2D} \|\mathbf{x}_i - Y\mathbf{w}_i^t\|_2^2$;
- 7: $t = t + 1$;
- 8: **until** Converges
- 9: $\mathbf{c}_i = \mathbf{w}_i(1:N)$ and $\mathbf{e}_i = \mathbf{w}_i(N+1:N+D)$.

where $\mathbf{w}_i = [\mathbf{c}_i^T, \mathbf{e}_i^T]^T \in \mathbb{R}^{(N+D)}$, $Y = [X_i, I] \in \mathbb{R}^{D \times (N+D)}$, $\mathbf{p} = [\mathbf{r}^T, \lambda \mathbf{s}^T]^T$ and $P = \text{diag}(\mathbf{p})$.

Based on the HQ optimization theory, the function $\mathcal{J}(\mathbf{w}_i, \mathbf{p}, \mathbf{q})$ can be alternately minimized as follows,

$$p_j^t = \begin{cases} 1/\sqrt{(\mathbf{w}_i)_j^2 + \alpha} & j \leq N \\ \lambda/\sqrt{(\mathbf{w}_i)_j^2 + \alpha} & \text{otherwise} \end{cases} \quad (28)$$

$$q_k^t = \exp(-(\mathbf{x}_i - Y\mathbf{w}_i)_k^2/\sigma^2) \quad (29)$$

$$\mathbf{w}_i^t = \arg \min_{\mathbf{w}_i} \mathcal{J}(\mathbf{w}_i, \mathbf{p}^t, \mathbf{q}^t) \quad (30)$$

where t is the iteration number.

The partial derivative of $\mathcal{J}(\mathbf{w}_i, \mathbf{p}^t, \mathbf{q}^t)$ with respect to $\mathbf{w}_i$ is

$$\frac{\partial \mathcal{J}(\mathbf{w}_i, \mathbf{p}^t, \mathbf{q}^t)}{\partial \mathbf{w}_i} = 2P^t \mathbf{w}_i + 2\gamma Y^T Q^t Y \mathbf{w}_i - 2\gamma Y^T Q^t \mathbf{x}_i \quad (31)$$

By setting the derivative to zero, we obtain the closed solution of (30)

$$\mathbf{w}_i^* = \gamma (P^t + \gamma Y^T Q^t Y)^{-1} Y^T Q^t \mathbf{x}_i \quad (32)$$

If link constraints are further considered, it is easy to derive that the optimal solution $\mathbf{w}_i^*$ for (12) becomes,

$$\mathbf{w}_i^* = (P^t + \gamma Y^T Q^t Y + \lambda \text{diag}(\mathbf{r}))^{-1} (\gamma Y^T Q^t \mathbf{x}_i + \lambda A_i) \quad (33)$$

where A_i is the i -th row of A in (12) corresponding to $\mathbf{c}_i$ and $\mathbf{r} \in \mathbb{R}^D$ takes the form,

$$\begin{cases} r_j = 1 & \{i, j\} \in O \\ r_j = 0 & \{i, j\} \notin O \end{cases}$$

Like any other robust M-estimation methods, the setting of parameter σ will affect the performance of the proposed method, therefore σ should be carefully determined to guarantee a non-increasing function for the objective function. The kernel size σ of this paper is computed as suggested in [27]

$$\sigma^2 = \frac{1}{2D} \|\mathbf{x}_i - Y\mathbf{w}_i\|_2^2 \quad (34)$$

Algorithm 1 outlines the complete algorithm.

Algorithm 2 Subspace Clustering via Half-Quadratic minimization (SCHQ)**Input:** Data matrix $X = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$.**Output:** Segmentation of the data.

- 1: For each data point $\mathbf{x}_i$, solve the problem (8) or (9) by the Algorithm 1. Obtain the coefficients matrix C .
- 2: Define the affinity matrix $W = |C| + |C|^T$.
- 3: Apply the spectral clustering algorithm [43] to the affinity matrix.

B. Convergence Analysis

According to the properties of HQ [27], [40], $J(\mathbf{w}_i^{t+1}, \mathbf{p}^{t+1}, \mathbf{q}^{t+1}) \leq J(\mathbf{w}_i^t, \mathbf{p}^{t+1}, \mathbf{q}^{t+1}) \leq J(\mathbf{w}_i^t, \mathbf{p}^t, \mathbf{q}^t)$. The cost function in (9) is non-increasing at each alternating minimization step. And according to the property of correntropy [28], the cost function is bounded. Hence the sequence $J(\mathbf{w}_i^t, \mathbf{p}^t, \mathbf{q}^t)$ in Algorithm 1 will converge. The inverse in (30) can be solved by a linear equations system. That is the solution of $\mathbf{x} = \text{inv}(\mathbf{A}) * \mathbf{y}$ can be obtained by solving the linear equations system $\mathbf{A}\mathbf{x} = \mathbf{y}$. If a rapid solver is adopted, the cost of a linear equations system can be further reduced. In addition, Algorithm 1 can be efficiently implemented in a parallel computing way because the representation of each data point can be computed independently.

C. Subspace Clustering

For each data point $\mathbf{x}_i$, we solve the optimization problem (8) or (9) by the Algorithm 1. Then we obtain the coefficient matrix $C = [\mathbf{c}_1, \dots, \mathbf{c}_N]$, and the affinity matrix is defined as $W = |C| + |C|^T$. Similar to SSC, we apply the spectral clustering algorithm of Ng et al. [43] to the affinity matrix and get the ultimate clustering results. The whole procedure of our subspace clustering method can be summarized in Algorithm 2.

V. EXPERIMENTS

Two real-world applications of subspace clustering, motion segmentation and face clustering, are used to evaluate the proposed SCHQ (Subspace Clustering based on Half-Quadratic Minimization) method. We compare SCHQ with state-of-the-art subspace clustering methods, e.g., Local Subspace Analysis (LSA) [44], Spectral curvature clustering (SCC) [45], LRR [9], block-diagonal LRR (BD-LRR) [12], LSR [11], Low-Rank Subspace Clustering (LRSC) [46], SSC [7], Latent Space SSC (LS3C) [19] and Sparse Additive Subspace Clustering (SASC) [47]. We denote the semi-supervised extension of SCHQ to deal with link constraints as S-SCHQ. Clustering error is used to evaluate different clustering methods. It is measured by counting the number of incorrectly assigned samples and dividing by the total number of samples.



Fig. 1. Some samples with the extracted features from the Hopkins 155 dataset [48].



Fig. 2. Some face samples of one subject from the YALE B database [49].

A. Datasets

The Hopkins 155 dataset [48], which is available online at <http://www.vision.jhu.edu/data/hopkins155/>, often serves as a benchmark database to evaluate subspace clustering algorithms. It consists of 120 2-motion and 35 3-motion video sequences² and each motion corresponds to a single subspace. The feature trajectories in the video sequences are automatically extracted with a tracker, and outliers in the dataset have been removed manually. Cornerness features are extracted and tracked across the video frames. Fig. 1 shows three examples with the extracted features from this dataset. Many clustering methods are tested and compared to the state-of-the-art methods on this dataset.

The Extended Yale Dataset B [49] is adopted as the benchmark for the face clustering problem. The dataset consists of frontal face images of 38 subjects taken under varying light conditions, and each image is cropped into 192×148 pixels. The face clustering experiments only use the facial images of the first 10 subjects for testing following other subspace clustering experiments in the literature. And each face image is resized to 48×42 pixels and stacked into a 2016D-vector. Fig. 2 shows some samples from this dataset.

The AR database [50] consists of over 4,000 face images from 126 subjects (70 men and 56 women). For each subject, 26 facial images were taken in two separate sessions. These images suffer different facial variations including various facial expressions (neutral, smile, anger, and scream), illumination variations (left light on, right light on, and all side lights on), and occlusions by sunglasses or scarf. To simulate different kinds of noise, we also replace a randomly selected local region in an uncorrupted image with an unrelated image [30]. Fig. 3 shows some samples from this database. It also shows two occluded images with 20 percent and 40 percent occlusions by monkey images, respectively.

²Actually, there are total 156 video sequences in the dataset with one sequence of 5 motions. However, only the results of 155 video sequences are reported following the experimental settings of other methods.

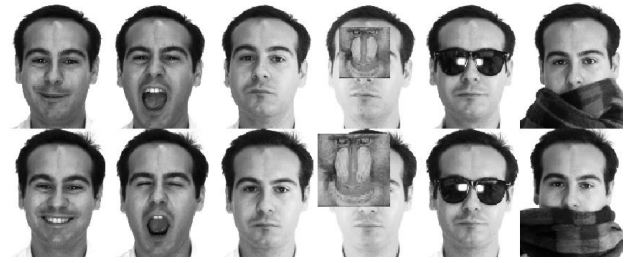


Fig. 3. Some face samples of the first subject from the AR database [50].

In our experiments, we used the first 20 male subjects to perform clustering experiments. Facial images were cropped and the resolution is 36×30 .

In the three databases, five types of noise are considered. In the Hopkins 155 dataset, there are only some small corruptions because outliers have been removed manually. In the Extended Yale Dataset B, complex noise is simulated by various occlusions incurred by different light conditions. In the AR database, sunglasses and scarf occlusions are used as two types of real world noise. In addition, we also simulate complex noise by randomly occluding an uncorrupted image with an unrelated image.

B. Settings

The proposed subspace clustering method (SCHQ) is implemented based on the Algorithm 2. Since the motion segmentation experiment is a problem of clustering slightly corrupted data points lying in a union of affine subspaces, the formulation of (8) is used following the similar technique in [7]. The sum of the coefficients is enforced to be 1, e.g., $\mathbf{1}^T \mathbf{c}_i = 1$. The optimization procedure is presented in Appendix. The formulation (9) is only used for the extended Yale B dataset because there are sharp intensity variations among intra-class face images. For the semi-supervised extension of SCHQ (S-SCHQ), we randomly label 1% link pairs from the overall data to simulate must-link and cannot-link constraints. The parameter λ for S-SCHQ is always fixed to 1.

The source codes of compared subspace clustering methods are downloaded or provided by the authors. We have tried our best to use the same algorithm settings described in the publications of compared methods. More specifically, the same setting of SSC [7] is used in the experiments, i.e., the noisy variation is used for motion segmentation and sparse outlying entries variation is used for face clustering. Both two versions of LSR described in [11], LSR1 and LSR2, are used in the experiments. Since a post-processing step is used in [9] to the coefficients matrix obtained by LRR, we report both the results of LRR and LRR-H (with post-processing). It should be noted the latest release of the algorithm [9] is implemented in this paper, which is different to the version in submission stage. The method in [46] is used for LRSC, i.e., Lemma 1 for motion segmentation and an ALM variant for face clustering. Because the authors do not provide the clustering algorithm and it is also based on low-rank representation, we use the same clustering algorithm as LRR-H to achieve the best

TABLE II
CLUSTERING ERRORS (%) ON HOPKINS 155 DATASET WITH THE ORIGINAL $2F$ -DIMENSIONAL DATA.
BEST RESULTS ARE HIGHLIGHTED IN BOLD. LRR-H INDICATES LRR WITH POST-PROCESSING

Algorithm	LSA	SCC	LRR	LRR-H	BD-LRR	LSR1	LSR2	LRSC	LS3C	SSC	SASC	SCHQ
<i>2 Motions</i>												
Mean	3.36	2.24	2.13	1.33	-	1.80	2.08	2.46	1.62	1.52	0.90	1.08
Median	0.50	0.00	0.00	0.00	-	0.11	0.10	0.00	0.00	0.00	0.00	0.00
<i>3 Motions</i>												
Mean	7.02	6.69	4.03	2.51	-	4.14	4.85	6.03	4.38	4.40	3.33	2.73
Median	1.45	0.40	1.43	0.00	-	1.60	1.83	2.20	0.56	0.56	0.60	0.43
<i>All</i>												
Mean	4.52	3.25	2.56	1.60	3.35	2.33	2.72	3.27	2.31	2.18	1.45	1.45
Median	0.57	0.00	0.00	0.00	0.37	0.31	0.31	0.00	0.00	0.00	0.00	0.00

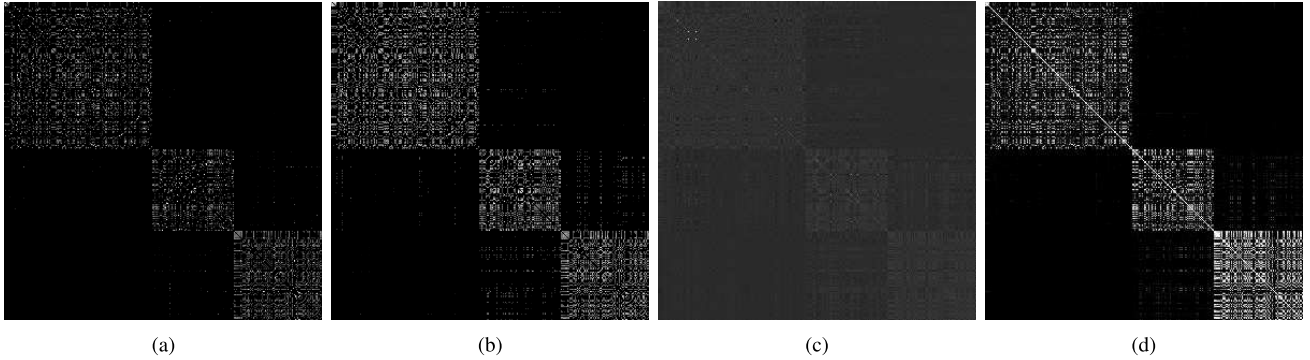


Fig. 4. Sparse coefficients for the 3-motion sequence cars10. (a) SCHQ. (b) SSC. (c) LRR. (d) LRR-H.

performance in comparison. The results of block-diagonal LRR (BD-LRR) are directly copied from [12].

C. Results on the Hopkins 155 Dataset

Motion segmentation is an important step in video sequences analysis [7]. Given a set of feature points tracked through the video sequences, the task of motion segmentation is to separate the trajectories of those feature points according to the motions belong to. As pointed out in [7], the set of all feature trajectories lie in a union of affine subspaces, which means the problem of motion segmentation can be reduced to segmenting the set of data points into a union of subspaces.

Preprocessing steps like PCA are usually used to reduce the dimension of the data [7], [11], [51]. Because the purpose of this paper is to study the robustness of subspace clustering, all tested methods are directly applied to the original $2F$ -dimensional data. The result of LSA on the 3 Motions is from [10] and [16]. All results of LRR are from [16] and [47].

The results of the clustering using different methods are shown in Table II. We observe that SCHQ is just slightly worse than SASC and LRR-H in the 2-motion case. SCHQ achieves lower clustering errors for all 155 video sequences. This may be because SCHQ has a relationship to M-estimation. In order to demonstrate the superiority of our method in constructing sparse coefficients, the coefficients matrix C of the 3-motion sequence cars10 obtained by the proposed SCHQ, SSC, LRR and LRR-H are shown in Fig. 4. And the clustering errors of these methods are 0.67%, 4.04%, 10.44%, and 5.39%, respectively. Note that the smaller number

of non-zeros entries lying outside the diagonal blocks, the better result in subspace clustering. Obviously, the matrix obtained by our method is much more clear than others outside the diagonal blocks. This may be the reason why our method works.

Specially, we can also see that the post-processing of the coefficient matrix makes the matrix more ‘bright’ and ‘clean’ for LRR, which helps improve the clustering performance (LRR-H). The mean clustering error of the method proposed in [52] (a variation of LRR) is 2.95%, while the mean error decreases significantly to 0.85% after the post-processing step. The results also demonstrate the importance of post-processing in LRR method.

D. Results on the Extended Yale Dataset

Face clustering aims to group face images with fixed pose taken under varying illumination according to their subjects [7]. It has been demonstrated in [53] that, under the Lambertian assumption, face images of one subject obtained with varying illumination lie close to a $9D$ linear subspace. Hence, the set of face images of different subjects with varying illumination can be approximated by a union of $9D$ linear subspaces. In this subsection, we apply the clustering methods to the original data from the Extended Yale Dataset B database as we did in the motion segmentation problem.

Table III tabulates the clustering results of different subspace clustering methods. As we can see, as the number of subject increases, the clustering errors of all methods increase in different degrees. It may be due to the expansion of the

TABLE III
CLUSTERING ERROR (%) ON THE EXTENDED YALE B DATASET WITHOUT PRE-PROCESSING.
BEST RESULTS ARE HIGHLIGHTED IN BOLD

Algorithm	LSA	SCC	LRR	LRR-H	BD-LRR	LSR1	LSR2	LRSC	SSC	SCHQ	S-SCHQ
<i>5 Subjects</i>											
Error	54.06	60.56	14.69	1.88	12.97	13.75	5.31	3.75	5.11	0.00	0.00
<i>8 Subjects</i>											
Error	61.91	65.63	21.09	2.34	27.70	36.91	17.38	8.79	4.88	1.37	0.87
<i>10 Subjects</i>											
Error	67.34	73.59	33.75	8.91	30.84	37.81	34.38	10.47	9.38	2.03	1.51

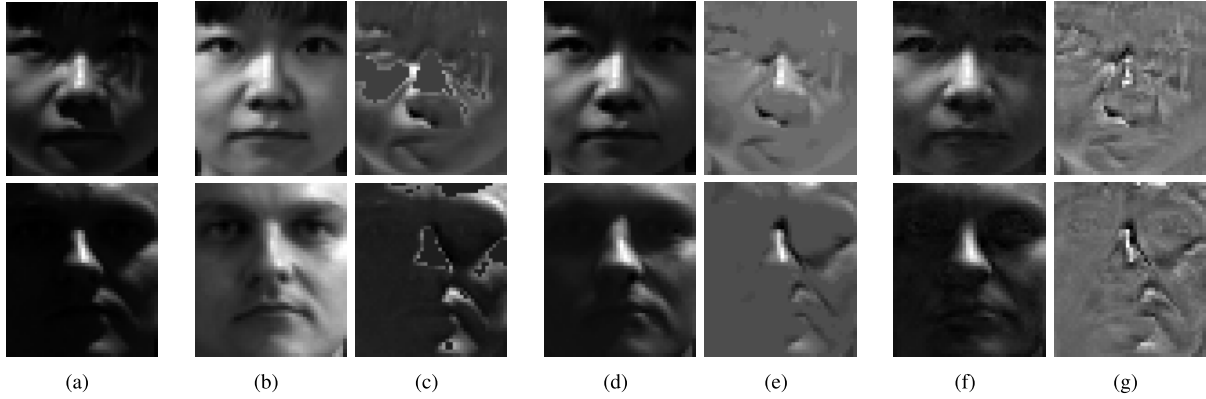


Fig. 5. Examples of reconstruction images and errors of different methods (a) Original images (b) Reconstruction images by our method (c) Reconstruction errors by our method (d) Reconstruction images by SSC. (e) Reconstruction errors by SSC. (f) Reconstruction images by LRR-H. (g) Reconstruction errors by LRR-H.

dictionary X . If there are more subjects in X , the corruptions and occlusions in face images will tend to be more complex. Obviously, our SCHQ method performs the best on 5, 8 and 10 subjects clustering problems. And the advantage of our method is much more significant on the 10-subject case, which indicates that our SCHQ method is more robust to the errors caused by complex noise. We also learn that the semi-supervised extension of SCHQ (S-SCHQ) can further reduce the clustering errors of SCHQ. This indicates that S-SCHQ can utilize must-link and cannot-link information that can be cheaply obtained in social network.

In Table III, we also observe that the clustering errors of LSA, SCC, LSR1, and LSR2 are larger than those of other subspace clustering methods. This may be because LSA, SCC, LSR1 and LSR2 are based on mean square errors (MSE, ℓ_2 norm). Since large errors will dominate the MSE, MSE based methods are sensitive to outliers that are significantly far away from the rest of the data points. Algorithmic robustness, which is derived from the statistical definition of a breakdown point, is the ability of an algorithm to tolerate a large number of outliers. From the viewpoint of robustness, these four methods may fail to deal with large outliers albeit they can work well on motion segmentation.

Some reconstructed face images and sparse errors of different methods are plotted in Fig. 5. We can see that even though the original face images are severely corrupted by shadow, our method can still recover the nearly perfect uncorrupted images while other methods all fail. It emphasizes the fact that our method can detect and correct the errors in data points simultaneously.

From Table III and Fig. 5, we also observe that although prior information (BD-LRR) and post-processing (LRR-H) can further improve clustering accuracy, these improvements are limited as compared with SCHQ. This may be because that previous state-of-the-art subspace methods mainly make use of convex ℓ_1 or $\ell_{2,1}$ norms to model errors. When there are complex errors in real world applications, these convex loss functions can not model errors well. Hence to improve the robustness of subspace clustering models may be more important to develop pre-processing or post-processing technique. The experimental results also suggest that correntropy can be used as a robust measure to deal with complex noise.

E. Results on the AR Dataset

In this experiment, we make use of the AR dataset [50] to further evaluate different subspace clustering methods. All the face images of the first 20 subjects (520 images) are used. Since different clustering methods depend on different parameter settings to achieve their best clustering accuracy, we split the 20 subjects into two equal subsets. The first 10 subjects are used as the training set to determine the parameters of different clustering methods, and the left 10 subjects are used as the testing set to report clustering errors. Since the source code of BD-LRR is not released, we do not compare these two algorithms on the AR dataset. To reduce clustering errors, PCA is used as a preprocessing step for LSA, SCC and LRR.

The tests are performed on three subsets. The first subset is composed of the 14 face images of each subject without occlusion. The second subset is composed of

TABLE IV
CLUSTERING ERRORS (%) ON THE AR DATASET WITHOUT OCCLUSION. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Algorithm	LSA	SCC	LRR	LRR-H	LSR1	LSR2	LRSC	LS3C	SSC	SCHQ	S-SCHQ
<i>5 Subjects</i>											
Error	42.86	25.71	8.57	8.57	8.57	8.57	18.57	8.57	8.57	4.29	2.86
<i>8 Subjects</i>											
Error	62.50	39.29	16.07	17.86	1.78	3.57	40.18	27.14	11.61	0.89	0.89
<i>10 Subjects</i>											
Error	67.14	40.71	22.86	26.43	15.00	15.71	40.71	30.71	18.57	12.86	11.43

TABLE V
CLUSTERING ERRORS (%) ON THE AR DATASET WITH OCCLUSION. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Algorithm	LSA	SCC	LRR	LRR-H	LSR1	LSR2	LRSC	LS3C	SSC	SCHQ	S-SCHQ
<i>5 Subjects</i>											
Error	46.92	35.39	39.23	34.62	15.38	23.08	39.23	29.23	33.85	10.00	6.92
<i>8 Subjects</i>											
Error	68.26	47.60	44.23	40.39	27.40	26.44	53.85	43.08	38.46	18.75	5.76
<i>10 Subjects</i>											
Error	73.84	53.85	43.08	42.31	31.92	30.77	57.69	47.12	48.46	22.69	13.46

all 26 face images of each subject with sunglasses and scarf occlusion. In the last subset, we randomly occlude 30% face images of the 10 subjects in the first subset with an unrelated image (as shown in Fig. 3). Table IV, Table V and Table VI tabulate the clustering errors of different subspace clustering methods. Although the AR database is not a difficult database for face recognition, it is challenging for clustering methods because each subject has the same facial variations, including various facial expressions, illumination and occlusions. More important, for each variation (e.g., anger expression), there are only two face samples in one subject. Different facial variations result in different subspaces so that it is difficult for a clustering algorithm to segment these subspaces under small number of samples.

Table IV shows clustering errors of different clustering methods without sunglasses and scarf occlusion. We observe that the clustering errors of different methods increase as the number of samples of each subject increases. When there are more subjects, a subspace clustering algorithm may potentially group the samples from the same facial variation into one class, which leads to a wrong clustering result. It seems that LSR1, LSR2, SSC and SCHQ outperform the four low-rank representation methods (LRR, LRR-H, LRSC, and LS3C). This may be due to that in the AR database each subject has many facial variations so that the low-rank representation methods can not determine an appropriate approximation of ranks.

Table V shows clustering errors of different clustering methods with sunglasses and scarf occlusion. Comparing the results in Table IV, we observe that algorithms' clustering errors increases significantly. This is because there are 12 face images occluded by sunglasses and scarf in this subset. From Fig. 3, we learn that the sunglasses/scarf occlusion occupies a large number of facial pixels. These occluded images are significantly different from uncorrupted ones. More important, compared with illumination and expression variations, there are 12 occluded facial images for each subject. These occluded images form entirely different subspaces, which makes

TABLE VI
CLUSTERING ERRORS (%) ON THE AR DATASET WITH DIFFERENT LEVELS OF OCCLUSION. BEST RESULTS ARE HIGHLIGHTED IN BOLD

	LRR-H	LSR1	LRSC	LS3C	SSC	SCHQ	S-SCHQ
20%	32.14	27.14	42.14	40.71	26.43	24.29	21.43
40%	39.29	38.57	47.86	45.71	40.71	33.57	31.43

this subset difficult for subspace clustering methods. More important, different facial variations result in different types of outliers. SCHQ achieves the lowest clustering errors. In addition, the improvements of the post-processing technique for LRR (LRR-H) seem limited on the AR dataset.

Table VI further shows clustering errors of some clustering methods under an unrelated image occlusion. Since the pixels of the unrelated monkey image are similar to the pixels of human face image, this contiguous occlusion is more complex than the noise caused by random black occlusion or white dots corruption [27], [30]. We observe that all compared clustering methods perform worse when face images are partly occluded an unrelated image. In addition, the errors of all methods increase when more face pixels are occluded. As expected, SCHQ outperforms its main competitors.

Experimental results also show the clustering errors of the semi-supervised extension of our SCHQ (S-SCHQ). By introducing some link information, S-SCHQ can further reduce the clustering errors of SCHQ. This indicates that the cannot-link and must-link constraints are useful for some challenging clustering tasks. If clustering results are not satisfied by a user, we can label some link constraints to further refine clustering results. S-SCHQ provides an efficient way for subspace clustering to deal with these constraints, which can also be easily extended to other subspace clustering methods.

F. Parameter Setting and Computational Costs

There are two important parameters in our basic correntropy modal in (8). The kernel size σ and the regularization parameter γ control robustness and sparsity respectively. The another

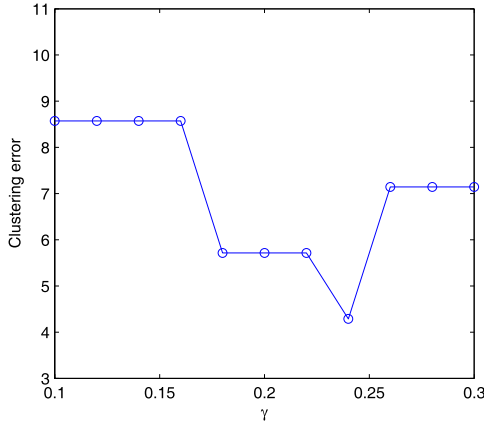
Fig. 6. Clustering errors as a function of parameter γ .

TABLE VII
CLUSTERING ERRORS (%) WITH RESPECTIVE TO DIFFERENT
M-ESTIMATORS ($\phi_r(\cdot)$ IN (8))

M-estimators	Welsch	ℓ_1 - ℓ_2	Huber	Fair	Cauchy
5 Subjects	10.00	14.62	19.23	14.62	13.85

parameter α is a constant, which makes the ℓ_1 - ℓ_2 loss function differentiable around the origin. α is often fixed to be a small value [39]. In this paper, we simply set α to be 0.000001 throughout all experiments. For the kernel size σ , we follow the suggestion in [24] and [27] to determine the value of σ . That is σ is computed by (34). Since the selection of σ has been well addressed in correntropy [24], [27], [28], the use of (34) to determine σ is suggested for convenience.

Simulations (see Fig. 6) were run on the AR dataset with occlusion (5 subjects) to show how the regularization parameter γ affects the clustering performance of SCHQ. The experimental setting is the same as that in Section V-E. We range the value of γ from 0.1 to 0.3. From Table IV, we learn that the lowest clustering error learned by the compared methods is 8.57%. We observe in Fig. 6 that there is a large range for γ to make SCHQ obtain lower errors than its competitors. More important, although there are variations of clustering when γ is set to different values, these variations are very small in adjacent values.

Table VII further shows clustering errors (%) with respect to different M-estimators ($\phi_r(\cdot)$ in (8)) under occlusions. All these five M-estimators can be minimized by the multiplicative form of HQ optimization [24], [27]. When $\phi_r(\cdot)$ in (8) is substituted by the M-estimators in Table VII, Algorithm 2 can be used to minimize (8). As demonstrated in [24] and [27], non-convex Welsch M-estimator and Cauchy M-estimator obtain the lowest clustering error, which suggests the use of non-convex M-estimators for complex noise in real world problems.

Since the representation of each data point in Algorithm 1 can be computed independently, Algorithm 1 can run in a parallel computing way to reduce computational costs. The major computational costs come from the computation of the linear equations system $Ax=y$. Simulations (see Fig. 7) were run on an Interl(R) Xeon (R) (2 processors, 2.67GHz, 2.66 GHz) Windows Server@Enterprise with 128 GB memory.

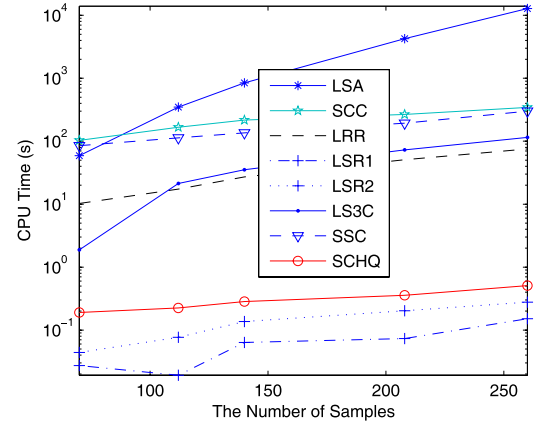


Fig. 7. CPU time as a function of the number of samples.

The AR dataset is used to evaluate the computational costs of SCHQ and its competitors. CPU time of different methods is plotted as a function of the number of samples. Since the difference between LRR and LRR-H lies on the post-processing step [9], the primal computational costs of LRR and LRR-H are similar. We only report the costs of LRR. We observe that the methods can be ordered in descending computational costs as LSA, SCC, SSC, LS3C, LRR, SCHQ, LSR2 and LSR1. The computational costs of SCHQ, LSR2 and LSR1 are significantly smaller than those of LSA, SCC, LS3C, SSC and LRR. Since SCHQ, LSR2 and LSR1 all involve the computation of a linear equations system, their computational costs are close. However, SCHQ has an inner loop to deal with outliers so that its costs are larger than the costs of LSR2 and LSR1.

VI. CONCLUSIONS

A robust subspace clustering method has been proposed to deal with complex noise. Correntropy has been demonstrated as a robust regularization term in subspace clustering. Must-link and cannot link constraints have been elegantly integrated into the proposed model, which results in a new semi-supervised subspace clustering algorithm. An efficient solution to the novel optimization problem is proposed based on half-quadratic minimization. Finally, experimental results on the Hopkins 155 dataset, the Extended Yale Database B and the AR database have shown that our method has achieved the state-of-the-art performance without any-preprocessing, and link information is helpful to further improve clustering accuracy.

One future work is to try other possible robust measures in subspace clustering because this paper has demonstrated the success of introduction of robust measures to the optimization problem of subspace clustering. Another future trend is to discuss rejection clustering problems [54] in subspace clustering. Since it is difficult to accurately cluster each data point into its corresponding group, self-expressiveness property of subspace clustering [7] can be used to assign some difficult data points into rejection class. Then clustering can be performed disregarding of those data in the rejection class.

APPENDIX

The following objective function is formulated if the constraint $\mathbf{1}^T \mathbf{c}_i = 1$ is integrated into problem (8):

$$\begin{aligned} \min_{\mathbf{c}_i} \quad & \sum_{j=1}^N \phi_r((\mathbf{c}_i)_j) + \gamma \sum_{j=1}^D \phi_c((\mathbf{x}_i - X_i \mathbf{c}_i)_j) \\ \text{s.t.} \quad & \mathbf{1}^T \mathbf{c}_i = 1 \end{aligned} \quad (35)$$

And the following Lagrangian function is obtained by introducing the Lagrangian multiplier λ to the function (35),

$$\begin{aligned} \mathcal{L}(\mathbf{c}_i) = \quad & \sum_{j=1}^N \phi_r((\mathbf{c}_i)_j) - \lambda(\mathbf{1}^T \mathbf{c}_i - 1) \\ & + \gamma \sum_{j=1}^D \phi_c((\mathbf{x}_i - X_i \mathbf{c}_i)_j) \end{aligned} \quad (36)$$

Note that λ is different from the penalty-term λ in (9).

The augmented cost-function $\mathcal{J}$ of (36) is as follows according to Section 4.1,

$$\begin{aligned} \mathcal{J}(\mathbf{c}_i, \mathbf{p}, \mathbf{q}) = \quad & \mathbf{c}_i^T P \mathbf{c}_i - \lambda(\mathbf{1}^T \mathbf{c}_i - 1) \\ & + \gamma (Y \mathbf{c}_i - X_i \mathbf{c}_i)^T Q (Y \mathbf{c}_i - X_i \mathbf{c}_i) \end{aligned} \quad (37)$$

where $Y = [\mathbf{x}_1, \dots, \mathbf{x}_i] \in \mathbb{R}^{D \times N}$ and $Y \mathbf{c}_i = \mathbf{x}_i$.

The function (37) is simplified as follows,

$$\mathcal{J}(\mathbf{c}_i, \mathbf{p}, \mathbf{q}) = \mathbf{c}_i^T G \mathbf{c}_i - \lambda(\mathbf{1}^T \mathbf{c}_i - 1) \quad (38)$$

where $G = P + \gamma(Y - X_i)^T Q(Y - X_i)$.

And the objective function can be solved alternately as follows,

$$p_j^t = 1/\sqrt{(\mathbf{c}_i)_j^2 + \alpha} \quad (39)$$

$$q_k^t = \exp(-(\mathbf{x}_i - X_i \mathbf{c}_i)_k^2 / \sigma^2) \quad (40)$$

$$\mathbf{c}_i^t = \arg \min_{\mathbf{c}_i} \mathcal{J}(\mathbf{c}_i, \mathbf{p}^t, \mathbf{q}^t) \quad (41)$$

The partial derivatives of $\mathcal{J}(\mathbf{c}_i, \mathbf{p}^t, \mathbf{q}^t)$ with respect to $\mathbf{c}_i$ and λ are

$$\frac{\partial \mathcal{J}(\mathbf{c}_i, \mathbf{p}^t, \mathbf{q}^t)}{\partial \mathbf{c}_i} = 2G \mathbf{c}_i - \lambda \mathbf{1} = 0 \quad (42)$$

$$\frac{\partial \mathcal{J}(\mathbf{c}_i, \mathbf{p}^t, \mathbf{q}^t)}{\partial \lambda} = \mathbf{1}^T \mathbf{c}_i - 1 = 0 \quad (43)$$

Then the closed solution of (41) is obtained,

$$\mathbf{c}_i^* = \frac{\lambda}{2} G^{-1} \mathbf{1} \quad (44)$$

The parameter λ can be adjusted to ensure that the sum of the entries of $\mathbf{c}_i$ is equal to 1. The rest steps are the same as the ones described in Algorithm 2.

ACKNOWLEDGMENT

The authors would like to greatly thank the associate editor and the reviewers for their valuable comments and advice.

REFERENCES

- [1] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, Mar. 2011.
- [2] P. Tseng, "Nearest q -flat to m points," *J. Optim. Theory Appl.*, vol. 105, no. 1, pp. 249–252, 2000.
- [3] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1945–1959, Dec. 2005.
- [4] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Comput.*, vol. 11, no. 2, pp. 443–482, 1999.
- [5] Y. Sugaya and K. Kanatani, "Geometric structure of degeneracy for multi-body motion segmentation," in *Statistical Methods in Video Processing*, Berlin, Germany: Springer-Verlag, 2004, pp. 13–25.
- [6] A. Aldroubi, "A review of subspace segmentation: Problem, nonlinear approximations, and applications to motion segmentation," *ISRN Signal Process.*, vol. 2013, 2013, Art. ID 417492.
- [7] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2765–2781, Nov. 2013.
- [8] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 663–670.
- [9] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
- [10] R. Vidal and P. Favaro, "Low rank subspace clustering (LRSC)," *Pattern Recognit. Lett.*, vol. 43, pp. 47–61, Jul. 2014.
- [11] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 347–360.
- [12] J. Feng, Z. Lin, H. Xu, and S. Yan, "Robust subspace segmentation with block-diagonal prior," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3818–3825.
- [13] H. Hu, Z. Lin, J. Feng, and J. Zhou, "Smooth representation clustering," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3834–3841.
- [14] S. D. Babacan and S. Nakajima, "Probabilistic low-rank subspace clustering," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, Inc., 2012.
- [15] S. Wang, X. Yuan, T. Yao, S. Yan, and J. Shen, "Efficient subspace segmentation via quadratic programming," in *Proc. AAAI*, 2011, pp. 519–524.
- [16] P. Ji, M. Salzmann, and H. Li, "Efficient dense subspace clustering," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Mar. 2014, pp. 461–468.
- [17] J. Liu, Y. Chen, J. Zhang, and Z. Xu, "Enhancing low-rank subspace clustering by manifold regularization," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4022–4030, Sep. 2014.
- [18] R. Liu, Z. Lin, and Z. Su, "Learning Markov random walks for robust subspace clustering and estimation," *Neural Netw.*, vol. 59, pp. 1–15, Nov. 2014.
- [19] V. M. Patel, H. Van Nguyen, and R. Vidal, "Latent space sparse subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 225–232.
- [20] L. Jing, M. K. Ng, and T. Zeng, "Dictionary learning-based subspace structure identification in spectral clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 8, pp. 1188–1199, Aug. 2013.
- [21] E. Fromont, A. Prado, and C. Robardet, "Constraint-based subspace clustering," in *Proc. Int. Conf. Data Mining*, 2009, pp. 26–37.
- [22] A. Talwalkar, L. Mackey, Y. Mu, S.-F. Chang, and M. I. Jordan, "Distributed low-rank subspace segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3543–3550.
- [23] X. Zhang, F. Sun, G. Liu, and Y. Ma, "Fast low-rank subspace segmentation," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1293–1297, May 2014.
- [24] R. He, W.-S. Zheng, T. Tan, and Z. Sun, "Half-quadratic-based iterative minimization for robust sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 261–275, Feb. 2014.
- [25] Y. Zhang, Z. Sun, R. He, and T. Tan, "Robust subspace clustering via half-quadratic minimization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3096–3103.
- [26] Y.-X. Wang and H. Xu, "Noisy sparse subspace clustering," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1–9.
- [27] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.

- [28] W. Liu, P. P. Pokharel, and J. C. Príncipe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007.
- [29] X.-T. Yuan, B.-G. Hu, and R. He, "Agglomerative mean-shift clustering via query set compression," in *Proc. Int. Conf. Data Mining*, 2009, pp. 223–234.
- [30] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
- [31] M. Soltanolkotabi, E. Elhamifar, and E. J. Candès, "Robust subspace clustering," *Ann. Statist.*, vol. 42, no. 2, pp. 669–699, 2014.
- [32] J. Chen and J. Yang, "Robust subspace segmentation via low-rank representation," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1432–1445, Aug. 2014.
- [33] D. Klein, S. D. Kamvar, and C. D. Manning, "From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering," in *Proc. Int. Conf. Mach. Learn.*, 2002, pp. 307–314.
- [34] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proc. Int. Conf. Mach. Learn.*, 2004, p. 11.
- [35] I. Davidson and S. S. Ravi, "Clustering with constraints: Feasibility issues and the k -means algorithm," in *Proc. Int. Conf. Data Mining*, 2005, pp. 138–149.
- [36] F. Wang, T. Li, and C. Zhang, "Semi-supervised clustering via matrix factorization," in *Proc. Int. Conf. Data Mining*, 2008, pp. 1–12.
- [37] X.-T. Yuan, B.-G. Hu, and R. He, "Agglomerative mean-shift clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 2, pp. 209–219, Feb. 2012.
- [38] P. Huber, *Robust Statistics*. New York, NY, USA: Wiley, 1981.
- [39] M. Nikolova and M. K. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM J. Sci. Comput.*, vol. 27, no. 3, pp. 937–966, 2005.
- [40] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Trans. Image Process.*, vol. 6, no. 2, pp. 298–311, Feb. 1997.
- [41] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [42] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 3, pp. 367–383, Mar. 1992.
- [43] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2001, pp. 849–856.
- [44] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 94–106.
- [45] G. Chen and G. Lerman, "Spectral curvature clustering (SCC)," *Int. J. Comput. Vis.*, vol. 81, no. 3, pp. 317–330, 2009.
- [46] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1801–1807.
- [47] X.-T. Yuan and P. Li, "Sparse additive subspace clustering," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 644–659.
- [48] R. Tron and R. Vidal, "A benchmark for the comparison of 3D motion segmentation algorithms," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
- [49] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.
- [50] A. M. Martinez and R. Benavente, "The AR face database," Centre de Visio per Computador, Univ. Autònoma Barcelona, Barcelona, Spain, Tech. Rep. 24, 1998.
- [51] R. Liu, S. Li, X. Yuan, and R. He, "Online determination of track loss using template inverse matching," in *Proc. Int. Workshop Vis. Surveill.*, 2008, pp. 1–8.
- [52] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1615–1622.
- [53] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, Feb. 2003.
- [54] X. Zhang and B.-G. Hu, "Learning in the class imbalance problem when costs are unknown for errors and rejects," in *Proc. IEEE 12th Int. Conf. Data Mining Workshops*, Dec. 2012, pp. 194–201.



research interests focus on information theoretic learning, pattern recognition, and computer vision.



Yingya Zhang received the B.E. degree in industrial automation from Sichuan University, Chengdu, China, in 2011, and the M.S. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2014. He is currently a Software Engineer with the Fujitsu Research and Development Center, Beijing. His research interests include subspace clustering and face recognition.



Recognition, CASIA, as a Faculty Member. His research focuses on biometrics, pattern recognition, and computer vision. He is a member of the IEEE Computer Society.



Qiyue Yin received the B.S. degree in automation control from Harbin Engineering University, Harbin, China, in 2012. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include clustering, recommender system, and computer vision.