

# Maximum Correntropy Criterion for Robust Face Recognition

Ran He, Wei-Shi Zheng, *Member, IEEE*, and Bao-Gang Hu, *Senior Member, IEEE*

**Abstract**—In this paper, we present a sparse correntropy framework for computing robust sparse representations of face images for recognition. Compared with the state-of-the-art  $l^1$ -norm based sparse representation classifier (SRC), which assumes that noise also has a sparse representation, our sparse algorithm is developed based on the maximum correntropy criterion, which is much more insensitive to outliers. In order to develop a more tractable and practical approach, we in particular impose non-negativity constraint on the variables in the maximum correntropy criterion, and develop a half-quadratic optimization technique to approximately maximize the objective function in an alternating way, so that the complex optimization problem is reduced to learning a sparse representation through a weighted linear least squares problem with non-negativity constraint at each iteration. Our extensive experiments demonstrate that the proposed method is more robust and efficient in dealing with the occlusion and corruption problems in face recognition, as compared to the related state-of-the-art methods. In particular, it shows that the proposed method can improve both recognition accuracy and receiver operator characteristic (ROC) curves, while the computational cost is much lower than the SRC algorithms.

**Index Terms**—Information Theoretical Learning, Correntropy, Linear Least Squares, Half-Quadratic Optimization, Sparse Representation, M-estimator, Face Recognition, Occlusion and Corruption.

## I. INTRODUCTION

Learning robust and discriminative face structure has been recognized as an important issue for real-world face recognition system. Well-known face recognition algorithms such as Eigenface [1], Fisherface [2] and Independent face [3] do not explicitly and effectively consider extracting robust facial features, although they are very useful for extracting descriptive or discriminative features.

To extract robust facial features, a number of algorithms such as modular Eigenspaces [4], modular linear regression based classification (LRC) [5], eigenimages [6] [7] [8], minimax probability machine [9], reconstructive and discriminative subspace method [10], and associative memories [11] [12] were developed in order to achieve more robust performance of face recognition. To address the alignment problem of face images, De la Torre and Black developed the robust parameterized component

analysis [13], which is also a modular Eigenspaces method, so that the performance such as face image reconstruction [13], Eigentracking [14] and active appearance models (AAM) [15] can be improved. Although notable improvement has been achieved by these methods, they can only deal with either the occlusion or corruption problem, or cannot deal with both of them very well. This is mostly because the effect of severe occlusion or corruption is not considered in these methods. In particular, some appearance models such as [14] [13] lack considering discriminant issues and thus may be less discriminative for robust recognition under tough condition [10]. To eliminate the effect of occlusions, occlusion masks are incorporated to learn a robust classifier in [16] [17]. However, it needs to first properly determine the occlusion masks [16] [17] [10], and the occlusion masks may discard useful redundant information [18] [19] and might not be suitable for dealing with corruptions. Hence, learning a robust representation for face recognition is still challenging, as any part of a face image could be corrupted in real-world scenarios, and sometimes the magnitude of noise may be arbitrarily large.

Recently, sparse signal representation has shown significant potential in solving computer vision problems [20]. Typically, many sparse techniques are casted into a  $l^1$  minimization problem, which is an approximately equal optimization problem as the  $l^0$  minimization problem under certain conditions [21] [22]. Learning theory shows that it is enough to obtain good generalization if the number of nonzero weights of training samples that are used to encode a test sample is small as compared to the size of training set [23]. Along this line, Wright et al. [24] recently proposed a sparse representation classifier (SRC) for robust face recognition against occlusions and corruptions, and impressive results were reported as compared to many well-known face recognition methods [1] [2] [25]. Although harnessing a ' $l^1$ -norm' will lead to a well-structured algorithm and can theoretically guarantee its strong performance [26], the SRC algorithm is still not robust enough to contiguous occlusion for face recognition as shown by Zhou et al. [18]. Furthermore, SRC would be hampered due to its expensive computational cost [24].

In this work, we propose to learn a robust sparse representation based on the correntropy [27] along with the use of  $l^1$ -norm penalty. The correntropy has been shown to obtain robust analysis [19] [28] in information theoretic learning (ITL) [29] [30] and efficiently handle non-Gaussian noise and large outliers [27]. Those works show the connection between entropy and Welsch M-estimator [31] [27] [28].

However, it is not easy to solve the optimization problem where the correntropy and  $l^1$  norm penalty are cooperated. To address this computational problem, we propose a new algorithm by first imposing non-negativity constraint on the variables in the correntropy. It not only greatly reduces the complexity of the model but also achieves significantly better performance as compared to the related state-of-the-art sparse algorithms for face

Manuscript received November 15, 2009

R. He is with the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China, and is also with the School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China. E-mail: rhe@nlpr.ia.ac.cn

W.-S. Zheng is with the Department of Computer Science, Queen Mary University of London, London, United Kingdom, and is also with Guangdong Province Key Laboratory of Information Security, Sun Yat-sen University, China. E-mail: wszheng@ieee.org

B.-G. Hu is with the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China. E-mail: hubg@nlpr.ia.ac.cn

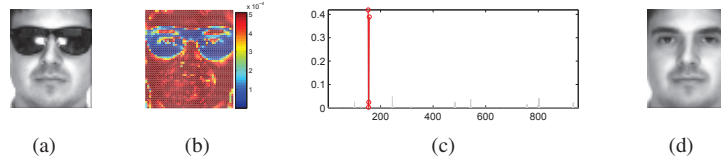


Fig. 1. **Overview of our approach for robust face recognition.** (a) A test face image with sunglasses occlusion. (b) The weight image learned by our approach. The entry with blue color has a small value while the entry with red color has a large value. The larger the value of the entry is, the more it contributes to the correntropy based objective function. Due to the occlusion caused by sunglasses, the pixels around two eyes are assigned small weights, which means that they are estimated as noise. (c) The sparse coefficients computed by our approach. The red coefficients correspond to the training images with the same class label of the test image. Our algorithm finds the true identity from 952 training images from 119 individuals in the standard AR face database. (d) The reconstructed image by a learned sparse linear combination of all the training images.

recognition. Second, the half-quadratic (HQ) optimization technique is adopted to optimize the maximum correntropy criterion in an approximate way, so that the optimization problem is further reduced to learning a sparse representation through a weighted linear least squares problem with non-negativity constraint. A new active-set technique is also developed for fulfilling such an efficient method.

Our study of robustness and sparsity in the ITL framework for face recognition makes our work novel in the following aspects:

1. **Robustness.** Different from SRC which assumes that the noise term has a sparse representation, our correntropy-based sparse representation (CESR) is based on correntropy, which can efficiently cope with non-Gaussian noise and outliers [27]. To our best knowledge, it has achieved much better results than what has been ever reported for the challenging scarf occlusion problem in face recognition.
2. **Sparsity.** CESR is a novel model to solve the problem of recovering a non-negative sparse signal. By imposing the non-negativity constraint in CESR, we further find that the solution of CESR is effective and sparse enough.
3. **Efficiency.** By developing a half-quadratic algorithm for optimization, CESR is presented to be a much more efficient algorithm for learning a sparse representation of a face image for recognition as compared to the SRC methods.

Fig. 1 presents an overview of the proposed robust algorithm for face recognition. Fig. 1 (b) gives a good illustration of the robustness of the proposed method. If a face is partially occluded, the pixels of occluded parts will be assigned small weights and thus less affect the maximum correntropy objective function. The learned sparse representation in Fig. 1 (c) is informative and the reconstructed image in Fig. 1 (d) is clear without occlusion.

The rest of this paper is structured as follows. We first briefly review existing linear representation and sparse representation methods for face recognition and discuss their limitations. In Section III, we propose the CESR model and develop a half-quadratic based active set algorithm for its optimization. In Section IV, we discuss the sparsity and robustness of the proposed approach. In Section V, the proposed approach is validated by conducting extensive experiments on the well-known and challenging face datasets along with the comparison with other related state-of-the-art techniques. Finally, we give the conclusion and discuss some future work in Section VI.

## II. RELATED WORKS

A fundamental task in pattern recognition is to correctly classify a new test sample  $y \in \mathbb{R}^{m \times 1}$  by using labeled training samples from  $k$  distinct classes. Let  $X_c \doteq [x_1^c, x_2^c, \dots, x_{n_c}^c] \in$

$\mathbb{R}^{m \times n_c}$  be a data matrix which consists of  $n_c$  training samples (columns of  $X_c$ ) from the  $c$ -th class. Let  $x_{ij}^c$  and  $y_j$  be the  $j$ -th entry of  $x_i^c$  and  $y$  respectively. Given a sufficiently expressive training set  $X_c$ , a test image  $y$  of subject  $c$  can be approximated by a linear combination of given training samples, i.e.  $y \approx X_c \beta^c$  for some coefficient vector  $\beta^c \in \mathbb{R}^{n_c}$ . By concatenating the training samples of all  $k$  classes, we get a new matrix  $X$  for the entire training set as:  $X \doteq [X_1, X_2, \dots, X_k] = [x_1^1, x_2^1, \dots, x_{n_k}^k] \in \mathbb{R}^{m \times n}$  where  $n = \sum_{c=1}^k n_c$ . Alternatively, a test sample  $y$  can also be expressed as a linear combination of all training samples:

$$y \approx X \beta \quad (1)$$

where the coefficient vector  $\beta \in \mathbb{R}^n$ . There are lots of work on learning this coefficient vector for face recognition. We in the following briefly review the major works.

### A. Non-sparse Linear Classifier

To seek the best representation for the test sample  $y$  in class  $c$ , the nearest neighbor (NN) classifier finds the nearest training sample in that class. That is it computes the minimal distance between  $y$  and all training samples of class  $c$ :

$$r_c^{NN}(y) = \min_{x_i^c} \|y - x_i^c\|_2 \quad (2)$$

where  $\|\cdot\|_2$  denotes the  $l^2$ -norm.

Compared with the nearest neighbor algorithm, linear representation methods seek the best representation by samples in each class. Among them, the simplest one is the nearest feature line (NFL) [32] that aims to extend the capacity of prototype features by computing a linear function to interpolate and extrapolate each sample pair belonging to the same class. In NFL, the distance is defined as follows:

$$r_c^{NFL}(y) = \min_{x_1^c, x_2^c, \beta_1} \|y - \beta_1 x_1^c + (1 - \beta_1) x_2^c\|_2 \quad (3)$$

where  $\beta_1 \in \mathbb{R}$ . We can also rewrite (3) as follows:

$$r_c^{NFL}(y) = \min_{x_1^c, x_2^c, \beta_1, \beta_2} \|y - (\beta_1 x_1^c + \beta_2 x_2^c)\|_2 \quad s.t. \quad \beta_1 + \beta_2 = 1 \quad (4)$$

If we extend (4) to a higher dimensional space, we get the following optimization problem:

$$r_c^{NCLC}(y) = \min_{\beta^c} \|y - X_c \beta^c\|_2 \quad s.t. \quad \sum_{i=1}^{n_c} \beta_i^c = 1 \quad (5)$$

Local subspace classifier [33] solves (5) by assuming  $X_c$  is orthonormal. The nearest constrained linear combination (NCLC) algorithm [34] and K-Local hyperplane (HKNN) algorithm [35]

solve (5) from a different geometric point of view. Note that by removing the equality constraint, we get

$$r_c^{NLC}(y) = \min_{\beta^c} \|y - X_c \beta^c\|_2 \quad (6)$$

The analytical solution of (6) can be directly computed by

$$\beta^c = (X_c^T X_c)^{-1} X_c^T y \quad (7)$$

The nearest linear combination (NLC) [34] solves (6) by a pseudo-inverse matrix and LRC [5] solves (6) by (7). LRC simply makes use of the downsampled images for the linear regression classification to achieve the state-of-the-art results as compared to the benchmark techniques [5]. Partial within-class match (PWCM) method [16] [17] solves (6) based on partial untermated pixels. The PWCM further utilizes a robust classification metric based on a M-estimator when all  $\beta^c$  have been computed, i.e.,

$$\arg \min_c \|y - X_c \beta^c\|_r \quad (8)$$

where  $\|\cdot\|_r$  is a robust norm based on M-estimators.

The minimal residual of (6) can also be taken as the distance from a query point to the space spanned by  $X_c$ . Then we get the nearest feature space (NFS) algorithm [36] that computes the distance between a query point and its projected point onto the feature space.

$$r_c^{NFS}(y) = \|y_p - y\|_2 \quad (9)$$

where  $y_p$  is the projection point of  $y$  onto the feature space spanned by  $X_c$ . According to (7), (6) can also be rewritten as

$$r_c^{NFS}(y) = \|y - X_c(X_c^T X_c)^{-1} X_c^T y\|_2 \quad (10)$$

If the columns of  $X_c$  are orthonormal, we have the nearest subspace algorithm [37] that solves the following optimization problem:

$$r_c^{NS}(y) = \|(I - X_c X_c^T)y\|_2 \quad (11)$$

After the residual or distance has been calculated, the label of the test sample  $y$  will be assigned  $\arg \min_c r_c(y)$ . Although those linear representation methods have indeed obtained promising improvement on performance, they still suffer two problems [24]:

1. The improvement of receiver operator characteristic (ROC) curve is quite limit.
2. They could not efficiently deal with the error incurred by severe occlusion and corruption.

### B. Sparse Linear Classifier

Different from traditional linear representation methods, the sparse representation based classifier (SRC) [24] makes use of the discriminative nature of sparse representation to perform classification. Also, all samples in the training set are used in SRC to represent the test sample, which is different from the other well-known face recognition methods in literatures. It aims to seek a sparse solution to (1) as follows,

$$\min \|\beta\|_0 \quad s.t. \quad y = X\beta \quad (12)$$

where the  $l^0$ -norm  $\|\cdot\|_0$  counts the number of nonzero entries in a vector. Originally inspired by the theoretical analysis [21] [22] that the solution of the  $l^0$  minimization problem is equal to the solution of the  $l^1$  minimization problem under some

conditions, SRC relaxes the above problem and seeks the sparse and informative vector  $\beta$  by solving the problem below:

$$\min \|\beta\|_1 \quad s.t. \quad y = X\beta \quad (13)$$

where  $\|\beta\|_1 = \sum_{i=1}^n |\beta_i|$ . Here, we denote the model (13) by SRC0. In real-world face recognition applications, occlusions and corruptions are always incurred. Hence, the above linear model in (1) is modified as:

$$y = X\beta + e \quad (14)$$

where  $e \in \mathbb{R}^m$  is the error (noise) vector. By assuming that noise vector  $e$  also has a sparse representation, SRC seeks an approximation of the sparsest solution to (14):

$$\min \|\beta\|_1 + \|e\|_1 \quad s.t. \quad y = X\beta + e \quad (15)$$

We denote the model (15) by SRC1.

Although SRC1 can effectively deal with the occlusion and corruption problems sometimes, it is still not robust enough to contiguous occlusion on facial images as shown in [18]. In addition, SRC1 would be hampered due to its expensive computational cost [24]. For example, it will take nearly hundreds of seconds for SRC1 to process a test image represented by a 700- $D$  vector.

## III. CORRENTROPY-BASED SPARSE REPRESENTATION

In this section, we first introduce correntropy and then we propose a robust correntropy-based sparse representation (CESR) for face recognition. A half-quadratic based active set algorithm is also developed for optimization.

### A. Correntropy

In real-world computer vision scenarios [7] [10] [24], face recognition against noise and outliers is critically challenging mainly due to the unpredictable nature of the errors (bias) caused by those noise and outliers. That means where these errors exist in an image can differ for different test samples and is hard for computer to predict in advance. The errors may be arbitrarily large in magnitude and therefore cannot be ignored or treated as small noise.

Recently, the concept of correntropy is proposed in ITL [27] to process non-Gaussian noise [27] and impulsive noise [38]. The correntropy is directly related to the Renyi's quadratic entropy [29] in which the Parzen windowing method is used to estimate the data distribution [39]. It is a local similarity measure between two arbitrary random variables  $A$  and  $B$ , defined by:

$$V_\sigma(A, B) = E[k_\sigma(A - B)] \quad (16)$$

where  $k_\sigma(\cdot)$  is a kernel function that satisfies Mercer theory [40] and  $E[\cdot]$  is the expectation operator. It takes advantage of kernel trick that nonlinearly maps the input space to a higher dimensional feature space. Different from conventional kernel methods, it works independently with pairwise samples. With a clear theoretic foundation, the correntropy is symmetric, positive, and bounded.

In practice, the joint probability density function of  $A$  and  $B$  is often unknown and only a finite number of data  $\{(A_j, B_j)\}_{j=1}^m$  is available. Hence, the sample estimator of correntropy is estimated by:

$$\hat{V}_{m,\sigma}(A, B) = \frac{1}{m} \sum_{j=1}^m k_\sigma(A_j - B_j), \quad (17)$$



TABLE I  
COMMONLY USED M-ESTIMATORS AND THEIR CORRESPONDING WEIGHT FUNCTIONS

name	$l^1$	$l^1 - l^2$	Fair	Cauchy	German-McClure	Welsch
$\rho(x)$	$ x $	$2(\sqrt{1+x^2/2} - 1)$	$c^2 \left[ \frac{ x }{c} - \log(1 + \frac{ x }{c}) \right]$	$\frac{c^2}{2} \log(1 + (x/c)^2)$	$\frac{x^2/2}{1+x^2}$	$\frac{c^2}{2} [1 - \exp(-(x/c)^2)]$
$w(x)$	$\frac{1}{ x }$	$\frac{1}{\sqrt{1+x^2/2}}$	$\frac{1}{1+ x /c}$	$\frac{1}{1+(x/c)^2}$	$\frac{1}{(1+x^2)^2}$	$\exp(-(x/c)^2)$
$\hat{w}(x)$	$\frac{1}{ x }$	$\frac{1}{\sqrt{1+x^2/2\sigma^2}}$	$\frac{1}{1+ x /\sigma}$	$\frac{1}{1+(x/\sigma)^2}$	$\frac{1}{(1+x^2/\sigma^2)^2}$	$\exp(-(x/\sigma)^2)$

where  $k_\sigma$  is Gaussian kernel  $g(x) \triangleq \exp(-\frac{x^2}{2\sigma^2})$ .

Based on (17), Liu et al. [27] [41] further extended the sample based correntropy criterion for a general similarity measurement between any two discrete vectors. That is they introduced the correntropy induced metric (CIM) [27] [41] for any two vectors  $A = (a_1, \dots, a_m)^T$  and  $B = (b_1, \dots, b_m)^T$  as follows:

$$CIM(A, B) = (g(0) - \frac{1}{m} \sum_{j=1}^m g(e_j))^{\frac{1}{2}} = (g(0) - \frac{1}{m} \sum_{j=1}^m g(a_j - b_j))^{\frac{1}{2}} \quad (18)$$

where the error  $e_j$  is defined as  $e_j = a_j - b_j$ . For adaptive systems, the below correntropy of error  $e_j$ :

$$\max_{\theta} \frac{1}{m} \sum_{j=1}^m g(e_j) \quad (19)$$

is called the maximum correntropy criterion (MCC) [27], where  $\theta$  is the parameter in the criterion to be specified later. MCC has a probabilistic meaning of maximizing the error probability density at the origin [41], and MCC adaptation is applicable in any noise environment when its distribution has the maximum at the origin [27]. Compared with the mean square error (MSE), a global metric, the correntropy is local. That means the correntropy value is mainly decided by the kernel function along the line  $A = B$  [27].

Correntropy has a close relationship with M-estimators [42]. If we define  $\rho(x) \triangleq 1 - \exp(-x)$ , (18) is a robust formulation of Welsch M-estimator [27]. Furthermore,  $\rho(x)$  satisfies  $\lim_{|x| \rightarrow \infty} \rho'(x) = 0$ , and thus it also belongs to the so called re-descending M-estimators [42], which has some special robustness properties [43]. A main merit of correntropy is that the kernel size controls all its properties. Due to the close relationship between m-estimation and methods of ITL, choose an appropriate kernel size [27] in the correntropy criterion becomes practical.

Considering the relationship between correntropy and M-estimators, we further list several commonly used M-estimators ( $\rho(x)$ ) and their corresponding weight functions  $w(x)$  in Table I for further analysis<sup>1</sup>. We also introduce the kernel size into other M-estimators by substituting the  $x$  with  $x/\sigma$  and setting  $c=1$ . We denote the modified weight function by  $\hat{w}(x)$ . Our experimental results show that the kernel size for  $\hat{w}(x)$  is also as effective as for correntropy.

### B. The Proposed Model

For face recognition, let vector  $A = (y_1, \dots, y_m)^T$  be a facial image vector, and vector  $B = (\sum_i x_{i1}\beta_i, \dots, \sum_i x_{im}\beta_i)^T$  be a linear representation of dataset  $X$ . We wish to find a sparse coding

<sup>1</sup>Although Huber's estimator can also be optimized by using the half-quadratic optimization, we don't discuss this estimator due to its threshold parameter. When there are noise and outliers, it is difficult to determine an appropriate value of the threshold.

vector  $\beta = (\beta_1, \dots, \beta_n)^T$  such that  $B$  becomes as correlated to  $A$  as possible under the maximum correntropy criterion (Eq. (19)). Then we have the following correntropy based sparse model:

$$J_{CESR} = \max_{\beta} \sum_{j=1}^m g(y_j - \sum_{i=1}^n x_{ij}\beta_i) - \lambda \|\beta\|_1. \quad (20)$$

where  $g(x) = \exp(-\frac{x^2}{2\sigma^2})$  is a Gaussian kernel function. Methods based on correntropy treat individual pixels of the representation differently and give more emphasis on those pixels corresponding to pixels of the same class as test sample  $y$ . This means if there are occlusions and corruptions in a test sample  $y$ , those pixels corresponding to outliers will have small contributions to the correntropy. Hence the noise can be handled uniformly within the correntropy framework.

Efficient minimization of the  $l^1$ -norm based and related convex functions is an active area of research [44]<sup>2</sup>. Generally, it is possible to optimize the correntropy criterion using gradient descend technique if the  $l^1$  penalty is not involved, albeit its significant usefulness for pursuing sparsity. In view of this difficulty, we relax the above  $l^1$  correntropy model by imposing nonnegativity constraint on the coding vector  $\beta$ . We then get the following maximum correntropy problem:

$$J_{CESR} = \max_{\beta} \sum_{j=1}^m g(y_j - \sum_{i=1}^n x_{ij}\beta_i) - \lambda \sum_{i=1}^n \beta_i \quad s.t. \beta_i \geq 0 \quad (21)$$

We note the optimization problem in (21) as *correntropy-based sparse representation* (CESR).

Note that in SRC1, both two  $l^1$ -norm terms are approximations of  $l^0$ -norm. The  $\|e\|_1 = \|y - X\beta\|_1$  is an approximation of  $\|y - X\beta\|_0$ . Hence, the SRC1 assumes that the noise also has a sparse representation and tries to estimate and correct the error incurred by noise [24] [26]. However, the first part in (21) of CESR is actually a robust M-estimator. As analyzed later and illustrated in Fig. 1, CESR aims to detect the noise and utilize the unterminated data to yield a robust sparse representation.

In particular, we consider a special case when  $\lambda$  is set zero. The rationale of this special model is that the positive vector  $\beta$  is actually playing as a clustering indicator, because each entry  $\beta_i$  reflects the importance of sample  $x_i$  in reconstructing pattern  $y$ . Hence it should be expected that more weights would be assigned to the samples of the same class label of  $y$ , while weights of the others should be small and zero in the optimal case. Therefore,  $\beta$  can also be sparse in this case. When  $\lambda > 0$ , CESR can yield a sparser solution and further improve the recognition accuracy. We will further verify this scenario in experiment V-G.

<sup>2</sup>In [44], the sparse solution of (13) is computed via solving the unconstrained optimization problem  $J(\beta) + h(X\beta - y)$  where  $J(\cdot)$  is convex and  $h(\cdot)$  has to be convex and differentiable.

### C. Algorithm of CESR

Since the objective function of CESR in (21) is nonlinear, CESR is difficult to be directly optimized. Fortunately, we recognize that the half-quadratic technique [28] and expectation maximization (EM) method [45] can be utilized to solve this ITL based optimization problem. According to the property of convex conjugate function [46], we have:

**Proposition 1:** There exists a convex conjugate function  $\varphi$  of  $g(x)$ <sup>3</sup>, such that,

$$g(x) = \max_{p'} \left( p' \frac{\|x\|^2}{\sigma^2} - \varphi(p') \right), \quad (22)$$

and for a fixed  $x$ , the maximum is reached at  $p' = -g(x)$ . [28]

Substituting (22) into (21), we have the augmented objective function in an enlarged parameter space

$$\hat{J}_{CESR} = \max_{\beta, p} \sum_{j=1}^m (p_j(y_j - \sum_{i=1}^n x_{ij}\beta_i)^2 - \varphi(p_j)) - \lambda \sum_{i=1}^n \beta_i \quad (23)$$

s.t.  $\beta_i \geq 0$

where  $p = [p_1, \dots, p_m]^T$  are the auxiliary variables introduced by half-quadratic optimization. According to Proposition 1, for a fixed  $\beta$ , the following equation holds

$$J_{CESR}(\beta) = \max_p \hat{J}_{CESR}(\beta, p) \quad (24)$$

It follows that

$$\max_{\beta} J_{CESR}(\beta) = \max_{\beta, p} \hat{J}_{CESR}(\beta, p) \quad (25)$$

Then we can conclude that maximizing  $J_{CESR}(\beta)$  is identical to maximizing the augmented function  $\hat{J}_{CESR}(\beta, p)$ . Obviously, a local maximizer  $(\beta, p)$  can be calculated in an alternating maximization way

$$p_j^{t+1} = -g(y_j - \sum_{i=1}^n x_{ij}\beta_i^t) \quad (26)$$

$$\beta^{t+1} = \arg \max_{\beta} (y - X\beta)^T \text{diag}(p)(y - X\beta) - \lambda \sum_i \beta_i \quad (27)$$

s.t.  $\beta_i \geq 0$

where  $t$  means the  $t$ -th iteration and  $\text{diag}(\cdot)$  is an operator to convert the vector  $p$  to a diagonal matrix. It is clear that the optimization problem in (27) is a weighted linear least squares problem with non-negativity constraint.<sup>4</sup> The auxiliary variables  $-p$  can be viewed as weights in (27).

The optimal problem in (27) can be re-formulated as the following quadratic program:

$$\min_{\beta} \left( \frac{\lambda}{2} - \hat{X}^T \hat{y} \right)^T \beta + \frac{1}{2} \beta^T \hat{X}^T \hat{X} \beta \quad \text{s.t. } \beta_i \geq 0 \quad (28)$$

where  $\hat{X} = X \text{diag}(\sqrt{-p^{t+1}})$  and  $\hat{y} = \text{diag}(\sqrt{-p^{t+1}})y$ . Since  $\hat{X}^T \hat{X}$  is a positive semidefinite matrix, this quadratic program in (28) is convex. Based on the Karush-Kuhn-Tucker optimal conditions, the following monotone linear complementary problem (LCP) is derived [47]:

$$\alpha = \hat{X}^T \hat{X} \beta - \hat{X}^T \hat{y} + \frac{\lambda}{2}, \quad \alpha \geq 0, \beta \geq 0, \beta^T \alpha = 0 \quad (29)$$

<sup>3</sup>Strictly speaking,  $\varphi$  is the conjugate function of the exponential function. Here we harness the variable substitution method (substitute  $x$  with  $x^2/\sigma^2$ ) which is commonly used in HQ.

<sup>4</sup>According to Proposition 1, we can learn that  $p \leq 0$ . By replacing the  $p$  with  $-p$ , we can get the equivalent minimal problem.

If the matrix  $\hat{X}$  has full column rank ( $\text{rank}(\hat{X}) = n$ ), the convex program in (28) and the LCP in (29) have unique solutions for each vector  $\hat{y}$  [47].

Let  $F$  and  $G$  be two subsets of  $\{1, \dots, n\}$  such that  $F \cup G = \{1, \dots, n\}$  and  $F \cap G = \emptyset$ . And let  $F$  and  $G$  be the working set and inactive set in the active set algorithm respectively. Consider the following column partition of the matrix  $\hat{X}$ :

$$\hat{X} = [\hat{X}_F, \hat{X}_G] \quad (30)$$

where  $\hat{X}_F \in \mathbb{R}^{m \times |F|}$ ,  $\hat{X}_G \in \mathbb{R}^{m \times |G|}$ , and  $|F|$ ,  $|G|$  are the number of  $F$  and  $G$  respectively. We then can rewrite (29) as:

$$\begin{bmatrix} \alpha_F \\ \alpha_G \end{bmatrix} = \begin{bmatrix} \hat{X}_F^T \hat{X}_F & \hat{X}_F^T \hat{X}_G \\ \hat{X}_G^T \hat{X}_F & \hat{X}_G^T \hat{X}_G \end{bmatrix} \begin{bmatrix} \beta_F \\ \beta_G \end{bmatrix} - \begin{bmatrix} \hat{X}_F^T \hat{y} \\ \hat{X}_G^T \hat{y} \end{bmatrix} + \frac{\lambda}{2}$$

where  $\beta_F, \alpha_F \in \mathbb{R}^{|F|}$ ,  $\beta_G, \alpha_G \in \mathbb{R}^{|G|}$ ,  $\beta = (\beta_F, \beta_G)$  and  $\alpha = (\alpha_F, \alpha_G)$ . Then we can compute the values of variables  $\beta_F$  and  $\alpha_G$  by the following iterative procedure [47]:

$$\min_{\beta_F \in \mathbb{R}^{|F|}} \|\hat{X}_F \beta_F - \hat{y}\|^2 + \lambda \sum_{i \in F} \beta_i \quad (31)$$

$$\alpha_G = \hat{X}_G^T (\hat{X}_F \beta_F - \hat{y}) + \frac{\lambda}{2} \quad (32)$$

And the optimal solution is given by  $\beta = (\beta_F, 0)$  and  $\alpha = (0, \alpha_G)$ . At each alternating maximum step in (27), instead of finding the optimal solution of the optimization problem, we simply learn a  $\beta$  in the feasible region to increase the objective. Combining (26), (31) and (32), we have the Half-Quadratic based active-set algorithm for the proposed CESR model.

**Input:** data matrix  $X$ , test sample  $y$ ,  $p^1 = -\mathbf{1}$ ,  $F = \emptyset$ ,  
 $G = \{1, \dots, n\}$ ,  $\beta = \mathbf{0}$ , and  $\alpha = -X^T y$ .

**Output:**  $\beta$

- 1: Compute  $\hat{X} = X \text{diag}(\sqrt{-p^t})$  and  $\hat{y} = \text{diag}(\sqrt{-p^t})y$ .
- 2: Compute  $r = \arg \min\{\alpha_i : i \in G\}$ . If  $\alpha_r < 0$ , set  $F = F \cup r$ ,  $G = G - r$ .  
Otherwise stop:  $\beta^* = \beta$  is the optimal solution.
- 3: Compute  $\bar{\beta}_F$  by solving (31). If  $\bar{\beta}_F \geq 0$ , set  $\beta^t = (\bar{\beta}_F, 0)$  and go to step 4.  
Otherwise let  $r$  be such that:  
$$\theta = \frac{-\beta_r}{\bar{\beta}_r - \beta_r} = \min\left\{ \frac{-\beta_i}{\bar{\beta}_i - \beta_i} : i \in F \text{ and } \bar{\beta}_i < 0 \right\}$$
  
and set  $\beta^t = ((1 - \theta)\beta_F + \theta\bar{\beta}_F, 0)$ ,  $F = F - r$ ,  
 $G = G \cup r$ . Return to step 2.
- 4: Compute  $\alpha$  according to (32).
- 5: Update the auxiliary vector  $p^{t+1}$  and kernel size  $\sigma$  according to (26) and (33) respectively. Return to step 1.

**Algorithm 1:** Algorithm of Correntropy-based Sparse Representation (CESR)

Algorithm 1 summarizes the optimization procedure. From Step 1 to Step 4, Algorithm 1 finds a feasible  $\beta$  to maximize the objective function for current  $p^t$ . In Step 5, Algorithm 1 computes the auxiliary variables of the Half-Quadratic optimization. It alternately maximizes the augmented objective function in (23) until it converges (Proposition 2). Note that Algorithm 1 may not reach the maximum value of the correntropy objective in (21) because it will return to Step 2 when a feasible solution occurs. Experimental results illustrate that this solution is good enough to achieve significant improvements over the related state-of-the-art methods.

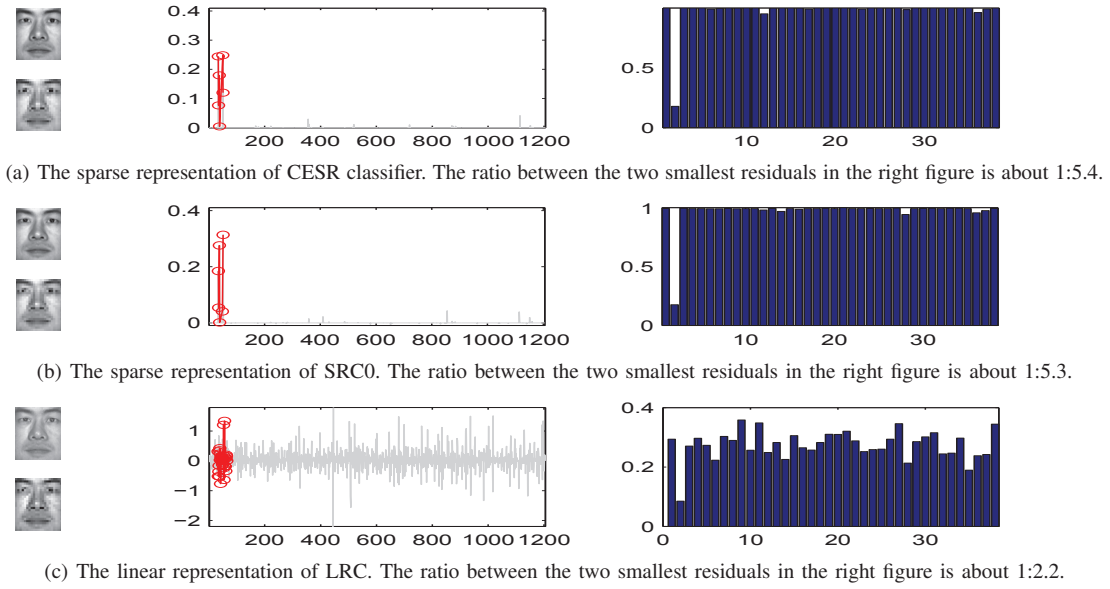


Fig. 2. **Recognition using  $24 \times 23$  downsampled image as features.** The image  $y$  belongs to the subject 2 of Yale B. Left: the reconstructed images in original  $192 \times 168$ -D feature space and downsampled  $24 \times 23$ -D feature space respectively; Middle: the coefficients computed by different algorithms. Nonzero coefficients associated to the subject 2 are highlighted by red color. Right: the residual ( $r'_c(y) = \|y - X_c \delta_c(\beta)\|_2$ ) of the input image from the subject 2.

Like any kernel method, the selection of kernel size will affect the performance of the proposed technique, and kernel size is often determined empirically. In this study, the kernel size (bandwidth) is computed by:

$$\sigma^2 = \frac{\theta}{m} (X_F \beta_F - y)^T (X_F \beta_F - y) \quad (33)$$

where  $\theta$  is a constant to control the noise. We set  $\theta$  to 1 throughout the paper.

**Proposition 2:** The sequence  $\{\hat{J}_{CESR}(\beta^t, p^t), t = 1, 2, \dots\}$  generated by CESR converges.

*Proof:* According to (27) and Proposition 1, we have

$$\hat{J}_{CESR}(\beta^t, p^t) \leq \hat{J}_{CESR}(\beta^t, p^{t+1}) \leq \hat{J}_{CESR}(\beta^{t+1}, p^{t+1})$$

The cost function increases at each alternating maximization step. Therefore, the sequence  $\{\hat{J}_{CESR}(\beta^t, p^t), t = 1, 2, \dots\}$  is non-decreasing. We can verify that  $J_{CESR}(\beta)$  is bounded (property of correntropy [27]), and by (25)  $\hat{J}_{CESR}(\beta^t, p^t)$  is also bounded. Consequently CESR converges. ■

#### D. Classifier of CESR

We aim to classify an unseen sample  $y$ , which comes from existing classes. Note that our criterion in Eq. (21) aims to reconstruct a test sample  $y$  using existing training samples as well as possible. Ideally, training samples from the same class of  $y$  should give the best reconstruction performance. This gives us the motivation to design a class specific reconstruction classifier similar to the sparse classifier proposed by [24], but the major difference is that the classifier introduced here is based on the correntropy criterion.

Therefore, we classify  $y$  as follows. For each class  $c$ , let  $\delta_c : \mathbb{R}^n \rightarrow \mathbb{R}^{n_c}$  be a function which selects the coefficients belonging to class  $c$ , i.e.  $\delta_c(\beta) \in \mathbb{R}^{n_c}$  is a vector whose entries are the entries in  $\beta$  corresponding to class  $c$ . Utilizing only the coefficients associated to class  $c$ , the given sample  $y$  is reconstructed as  $\hat{y}_c = X_c \delta_c(\beta)$ . Then  $y$  can be classified by assigning it to the class

$c$  corresponding to the maximal nonlinear difference between  $y$  and  $\hat{y}_c$ , i.e.,

$$\max_c r_c(y) \doteq g(y - X_c \delta_c(\beta)) \quad (34)$$

where the kernel size  $\sigma$  in  $g(x)$  is calculated by:

$$\sigma^2 = \frac{\theta_r}{k} \sum_{c=1}^k \|y - X_c \delta_c(\beta)\|_2^2 \quad (35)$$

We set  $\theta_r$  to 1 throughout the paper. Algorithm 2 summarizes the classification procedure.

**Input:** data matrix  $X = [X_1, X_2, \dots, X_k] \in \mathbb{R}^{m \times n}$  for  $k$  classes, a test sample  $y \in \mathbb{R}^{m \times 1}$

**Output:**  $\text{identity}(y)$

1: Solve the maximum correntropy problem ( $\lambda = 0$ ):

$$\beta^* = \arg \max_{\beta} \sum_{j=1}^m g(y_j - \sum_{i=1}^n x_{ij} \beta_i) \quad \text{s.t. } \beta_i \geq 0$$

2: Calculate the residuals  $r_c(y) = g(y - X_c \delta_c(\beta^*))$ , for  $c = 1, \dots, k$

3:  $\text{identity}(y) = \arg \max_c r_c(y)$

**Algorithm 2:** Classifier of Correntropy-based Sparse Representation

#### IV. SPARSITY AND ROBUSTNESS OF CESR

In this section, we demonstrate the visual results of CESR for robust face recognition along with the comparison with SRC [24]. More results will be reported in the experiment section.

##### A. Sparsity of CESR

It has been shown that sparse representation is more efficient than the non-sparse one in many challenging cases for pattern analysis, as it is able to select the most representative training samples for representation of each test sample [23]. To show the

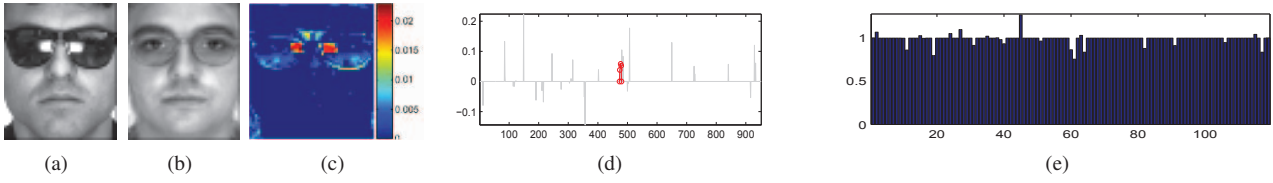


Fig. 3. **Sparse representation computed by SRC1.** (a) A test face image from the 60th subject in AR database with sunglasses occlusion. (b) Reconstructed image as a sparse linear combination of all the training images. (c) The error image (by reshaping error vector  $e$ ). The entry with red color has a large value, which means that this entry is estimated as noise by SRC1. (d) The sparse coefficients learned by SRC1. The red coefficients correspond to the training images with the same class label of the test image. (e) The residuals  $r'_c(y)$  with respect to the coefficients for different classes. The minimal residual does not correspond to the same class of this test subject.

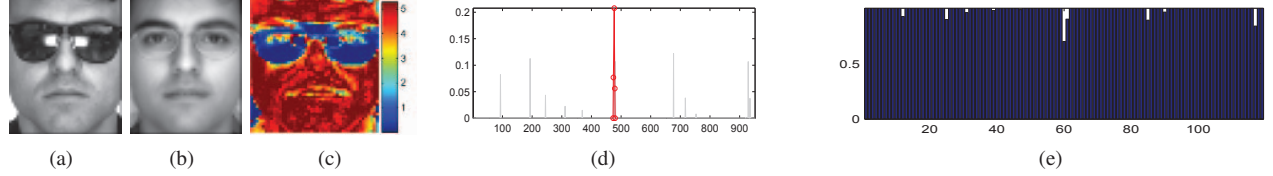


Fig. 4. **Sparse representation computed by CESR.** (a) The same occluded image as the one in Fig. 3. (b) The reconstructed image via a sparse linear combination of all the training images. (c) The weight image (by reshaping auxiliary variable  $-p$ ). The entry with blue color receives a small weight, which means that this entry is estimated as noise by CESR. (d) The sparse coefficients computed by CESR. The red coefficients correspond to the training images with the same class label of the test image and are the largest coefficients. (e) The residuals  $r'_c(y)$  with respect to the coefficients for different classes.

strength of learning sparse features by CESR for face recognition, we compare it with SRC and show that the coefficients computed by CESR are more sparse and more effective. Results of LRC are also presented for reference.

We randomly selected half of the 2414 images in the Extended Yale B database as the training set and the rest for testing, where the database will be introduced in the experiment section. The images are downsampled from the original size  $192 \times 168$  to size  $24 \times 23$ <sup>5</sup>. As in [24], each image vector will have unit norm by normalization<sup>6</sup>.

Middle columns in Fig. 2 show the coefficients computed by different algorithms for a test image from the second subject in Yale database. The nonzero coefficients associated with subject 2 are highlighted by red color. The numbers of nonzero coefficients ( $l^0$  norm) computed by CESR, SRC0 and LRC are 28, 39 and 1205 respectively. For CESR and LRC, we directly compute the number of nonzero coefficients. Since the coefficients of SRC0 solved by primal-dual algorithm have many small nonzero entries, we compute the number of nonzero coefficients of SRC0 by:

$$\|\beta\|_0^{SRC0} = \sum_{i=1}^n I(\beta_i) \quad (36)$$

where  $I(x)$  is an indicator function:

$$I(x) = \begin{cases} 1 & |x| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \quad (37)$$

where  $|\cdot|$  is the absolute value operator. Here, we set  $\varepsilon = 0.001$  to fairly compare different sparse methods. If we reduce the value of  $\varepsilon$  to a smaller value, such as 0.0001, the number of nonzero entries computed by SRC0 will be larger than 100.

Fig. 5 further shows the average number of nonzero coefficients in the case of random corruption. (See experiment V-C.2 for details of the setting). We varied the percentage of corruption from 10% to 80% and computed the average numbers of nonzero

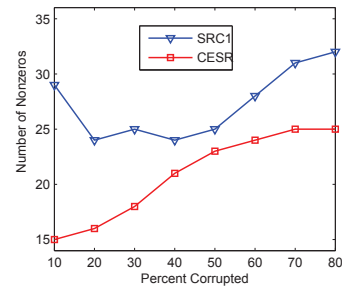


Fig. 5. **Average numbers of nonzero coefficients with respect to different levels of random corruptions.**

coefficients. We find that CESR can also obtain sparse representations in this case and achieve much sparser representation than the one obtained by the  $l^1$  norm based SRC1.

The right part of Fig. 2 shows the residuals  $r'_c(y) = \|y - X_c \delta_c(\beta)\|_2$  with respect to the corresponding projected coefficients  $\delta_c(\beta)$ ,  $c = 1, \dots, 38$ . Like SRC0, CESR can correctly classify the test sample, where the dominant coefficients correspond to the second subject which is of the same class as the test image. The ratios between the two smallest residuals of CESR, SRC0 and LRC are about 1:5.4, 1:5.3 and 1:2.2 respectively. The higher the ratio value is, the more discriminative the algorithm can be. As we will show later in the experiment section that higher ratio value also always implies higher ROC curve.

The left part of Fig. 2 shows the reconstructed images in the original image space (top) and downsampled image space (bottom) using the coefficients learned by three different algorithms respectively. For LRC, the coefficients associated to subject 2 are used to reconstruct the test image. Since the dominant coefficients of CESR and SRC0 are associated with subject 2, the reconstructed images are quite similar to the original images.

### B. Robustness of CESR

In real-world applications, facial images always incur occlusions and corruptions. SRC1 presents a novel approach to deal

<sup>5</sup>In [24] [5], face recognition methods make use of downsampled images as features to achieve impressive results.

<sup>6</sup>In SRC [24], normalization is a necessary preprocessing step. Hence, we normalized the facial images to fairly compare different methods.



with the errors caused by occlusions and corruptions by assuming that they also have sparse representations [24]. Although 'sparse' may not mean 'very few' [24], this assumption may make SRC1 fail when there are severe occlusions and corruptions. Fig. 3 shows such an example. As demonstrated in Fig. 3 (b), obvious shadows around the eyes can be found in the image reconstructed by SRC1. This is due to the inaccurate estimation of the noise. In Fig. 3 (a), the occlusion of sunglasses is roughly 20%. Fig. 3 (c) further shows the noise item  $e$  in (15) when SRC1 converges (the  $e$  is reshaped to an image for visual illustration). Note that large entry value is marked in red color in the image. We can observe that only the two red regions in Fig. 3 (c) corresponding to the two highlighted regions of sunglasses in Fig. 3 (a) are estimated as noise. This inaccurate estimation makes the coefficients estimated by SRC1 less sparse and also less informative, so that the largest coefficients in Fig. 3 (d) still contain much noisy information. This makes SRC1 fail to identify the correct subject in some cases.

However, CESR can correctly detect the occlusion in this scenario. As shown in Fig. 4 (b), the whole facial image is reconstructed much better by using CESR. The occluded regions around two eyes are also much more clear. Fig. 4 (c) further shows the auxiliary variables  $-p$  (We also reshape it to an image) in the Half-Quadratic optimization when CESR algorithm converges. We can find that the weights near two eyes have the lowest values whereas others have large values, which makes CESR learn a robust sparse representation mainly from the non-occluded region of facial images. Fig. 4 (d) and (e) show the coefficients and residuals computed by CESR respectively. Although the sunglasses occlude the area of two eyes, CESR classifier can still find the true identities (the 60th individual) from 952 training images of 119 individuals.

## V. EXPERIMENT

To evaluate the proposed correntropy-based sparse representation algorithm, we compared it with two related state-of-the-art methods for robust face recognition: linear regression classification (LRC) and sparse representation-based classification (SRC). Experiments were performed on three public face recognition databases, namely the AR [48], CMU [49] and FRGC [50] databases. The recognition rate, ROC curves, and computational cost of the compared methods will be reported. All algorithms were implemented in MATLAB on an AMD Quad-Core 1.80GHz Windows XP machine with 2GB memory.

### A. Experimental Setting and Face Databases

**Database.** All grayscale images of the three public face databases are aligned by eyes' locations which are manually located. Descriptions of the three databases are as follows.

- 1) *AR Database* [48]: The AR database consists of over 4,000 facial images from 126 subjects (70 men and 56 women). For each subject, 26 facial images are 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 occlusion by sunglasses or scarf. This database is always used for evaluating robust face recognition algorithm. In the experiment, we selected a subset of the dataset that consists of 65 male subjects and



Fig. 6. **Cropped facial images of the first subject in AR database.** The images in the first row are from the first session and the images in the second row are from the second session.

54 female subjects. The grayscale images were resized to resolution  $112 \times 92$ .

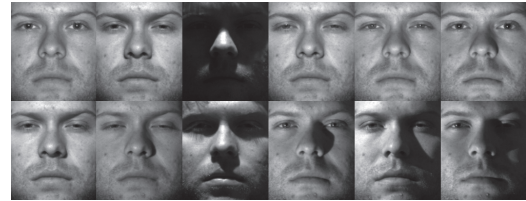


Fig. 7. **Cropped facial images of one subject in YALE B database.**

- 2) *Extended Yale B Database* [49] [51]: The Extended Yale B database consists of 2,414 frontal face images from 38 subjects [49] under various lighting conditions. The cropped and normalized  $192 \times 168$  face images were captured under various controlled lighting conditions [51] [24]. Fig. 7 shows some face images of the first subject in the Yale B database. For each subject, half of the images were randomly selected for training (i.e., about 32 images for each subject), and the rest were for testing.



Fig. 8. **Cropped facial images of one subject in FRGC database.**

- 3) *FRGC Database* [50]: The FRGC version 2 face database is a challenging benchmark face recognition database. There are 8014 images from 466 subjects in the query set for FRGC experiment 4. These uncontrolled still images suffer the variations of illumination, expression, time, and blurring. In our experiment, we only selected the objects which has over 10 facial image in the database. Then got 3160 facial images from 316 subjects. Each facial image was in grey scales and cropped into resolution 32 in pixels by fixing the positions of two eyes. For each selected subject, 8 images were randomly selected for training and the rest were for testing.

**Algorithm Setting.** For SRC, we conducted its three models, which are different in the aspects of formulation and computational strategy. As suggested by [24], we implemented models of SRC by normalizing the facial image with unit norm. In our experiments, the details of these three models are:



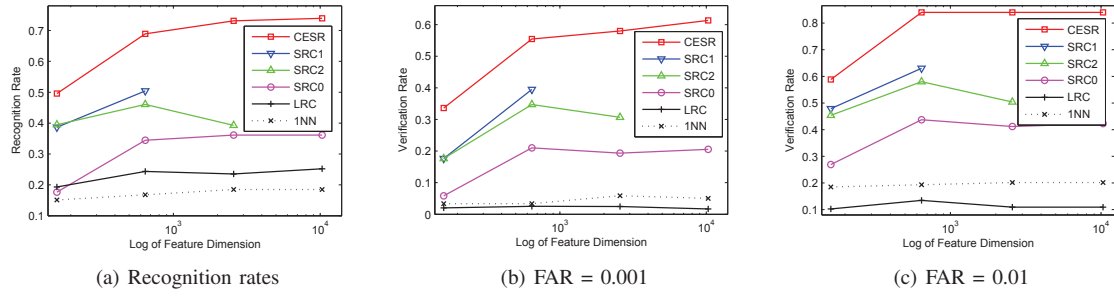


Fig. 9. Recognition rates and verification rates against sunglasses occlusion in AR database.

- 1) *SRC0*: It minimizes the standard SRC in (12) via an active set algorithm based on [52].<sup>7</sup>
- 2) *SRC1*: It minimizes the  $l^1$ -norm in (38) via a primal-dual algorithm for linear programming based on [46], [53].<sup>8</sup>

$$\min \|\beta\|_1 + \|e\|_1 \quad s.t. \quad \|y - X\beta + e\|_2 \leq \varepsilon \quad (38)$$

where  $\varepsilon$  is a given non-negative error tolerance.

- 3) *SRC2*: In many cases, since the noise level  $\varepsilon$  in SRC1 is unknown beforehand [54], we can use the Lasso optimization algorithm to recover the sparse solution by<sup>9</sup>

$$\min \|y - X\beta + e\|_2^2 + \lambda(\|\beta\|_1 + \|e\|_1) \quad (39)$$

where  $\lambda$  is a given regularization parameter, which can be viewed as an inverse of the Lagrange multiplier associated with the constraint in (38). The  $\varepsilon$  can be interpreted as a pixel noise level, whereas  $\lambda$  cannot [24]. Since the noise level is also unknown in the testing set, the experimental results of SRC1 and SRC2 are not always the same due to different values of  $\varepsilon$  and  $\lambda$ .

In order to estimate the parameters  $\varepsilon$  and  $\lambda$  automatically, five-fold cross-validation on each dimension for each training set was used, where the candidate value set for both  $\varepsilon$  and  $\lambda$  is  $\{1, 0.5, 0.25, 0.1, 0.075, 0.05, 0.025, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$ . Note that due to the large computational cost of SRC1 and SRC2, exhaustive search of the parameter value is not doable. Also for the same reason, we can only report the experimental results of SRCs in the lower dimensional feature space.

For CESR, we mainly evaluated the proposed algorithm using the non-negativity constraint with  $\lambda = 0$ . Discussion of  $\lambda$  will be given at the end of experiment section. More experimental results on parameter selection in CESR will be shown in Section V-G later.

### B. Real-world Malicious Occlusion

We applied CESR to real face recognition scenarios against malicious occlusion.

1) **Sunglasses Occlusion**: For training, we used 952 non-occluded frontal view images (about 8 for each subject) with varying facial expression. Fig. 6 shows an example of 8 selected images of the first subject. For testing, we used images occluded by sunglasses. Fig. 3 (a) shows a facial image of the 60th individual from this testing set.

Figure 9 (a) shows the recognition performance of different methods using different downsampled images of dimensions 161, 644, 2576, and 10304. Those numbers correspond to downsampling ratios of 1/8, 1/4, 1/2, and 1, respectively. We see that if occlusions exist, it is unlikely that the test image will be very close to any single training image of the same class, so that the nearest-neighbor classifier performs poorly. Although SRC0 and LRC can improve the recognition rates compared to 1NN, their improvements are limited because they are based on the MSE criterion which is sensitive to outliers. As illustrated in Fig. 4, CESR can model the noise more accurately and therefore perform significantly better.

The ROC curve is also illustrated in Fig. 9. ROC is important for evaluating different methods for face recognition in order to measure the accuracy of outlier rejection. A good algorithm should achieve high verification rate (VR) at a very low false acceptance rate (FAR). In a face recognition system, users often care about the verification rates when the false acceptance rates are 0.001 and 0.01 [55]. Fig. 9 (b) and (c) show the verification rates when FAR = 0.001 and FAR = 0.01 respectively. The verification rates of 1NN and LRC are quite low. Although LRC has better recognition performance than 1NN, LRC cannot get better ROC curves and the verification rate of LRC is even lower than that of 1NN. Still, the verification rate of CESR is the highest among the compared methods as shown.

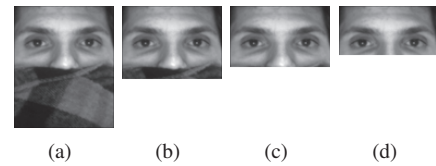


Fig. 10. Samples of cropped faces with scarf occlusion from AR database. (a) Cropped images of the first subject (b) 40% pixels (white region) are treated as missing pixels. (c) 50% pixels are treated as missing pixels. (d) 60% pixels are treated as missing pixels.

2) **Scarf Occlusion**: For training, we used 952 non-occluded frontal view images (about 8 for each subject) with varying facial expression. Fig. 6 shows an example of 8 selected images of the first subject. For testing, we used images with scarf (one image for each person), which makes facial image roughly 45% occluded. Fig. 10 (a) shows a facial image of the first subject from this testing set.

In the case of scarf occlusion, we can treat the region occluded by scarf as missing pixels. Fig.10 (b), (c) and (d) show different levels of missing pixels. In real world face recognition systems, we often need to manually crop the facial image when there is

<sup>7</sup>The MATLAB source code: <http://redwood.berkeley.edu/bruno/sparsenet/>

<sup>8</sup>The MATLAB source code: <http://www.acm.caltech.edu/1lmagic/>

<sup>9</sup>The MATLAB source code: <http://redwood.berkeley.edu/bruno/sparsenet/>

severe and continuous occlusion. For example, the mouth in a face image is occluded by scarf [5]. It is convenient for users to simply locate the continuous occlusion as a rectangle region during manual alignment. In view of this, during the optimization of CESR, the entries of the auxiliary variable  $p$  corresponding to the missing pixels would be set zero.

TABLE II  
 RECOGNITION RATES (%) FOR VARIOUS FEATURE SPACES AND  
 DIFFERENT LEVEL OF MISSING VALUES.

Missing pixels	$14 \times 11$	$28 \times 23$	$54 \times 46$
60%	85.71%	93.28%	93.28%
50%	90.76%	97.48%	<b>98.32%</b>
40%	91.60%	95.80%	95.80%

Table II tabulates the recognition rates of our approach using various feature spaces (downsampled images) at different levels of missing values. To the best of our knowledge, the recognition accuracy 98.3% obtained by the proposed approach is probably the best result that has ever been reported against scarf occlusion. Note that for 100 subjects the previous best results were 95.5% and 93.5% obtained by the SRC approach [24] and modular LRC approach in [5] respectively, and for 50 subjects it was 93% reported in [10].

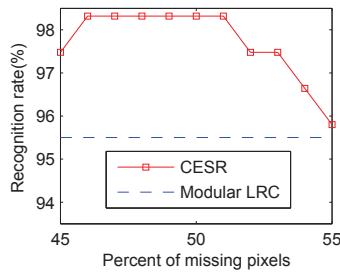


Fig. 11. Recognition rates against scarf occlusion under different levels of missing pixels. The size of image is  $54 \times 46$ .

Fig. 11 further shows the recognition rates with respect to different levels of missing pixels when the size of image is  $54 \times 46$ . Our proposed maximum correntropy framework shows a quite stable performance against scarf occlusion and achieves the highest performance as well.

### C. Contiguous Occlusion and Corruption

1) **Contiguous Occlusion of Random Block:** In this subsection we simulated various types of contiguous occlusion by replacing a randomly selected local region in each test image with an unrelated image [24]. Two images (top and bottom) in Fig. 14 (a) show two occluded images with 20% and 40% occlusions by monkey images respectively. Since the pixels of the unrelated monkey image are similar to the pixels of human face image, this contiguous occlusion is more challenging than the occlusion caused by random black or white dots.

We extensively evaluated different methods on the Extended Yale B Face Database against such a kind of contiguous occlusion. For each subject, half of the images were randomly selected for training (i.e., about 32 images for each subject), and the rest half were for testing, and the training set and the testing set contained 1205 and 1209 images respectively. We computed recognition

rates with respect to five feature (downsampled image) spaces of dimensions 56, 120, 504, 2016, and 8064. Those numbers correspond to the downsampling rates of 1/24, 1/16, 1/8, 1/4, and 1/2 respectively.

Fig. 12 (a) and Fig. 13 (a) show the recognition accuracies with respect to different feature spaces. CESR achieves the highest recognition rates (88.6% and 74.6%) on the dimension of 8064 where face images are with 20% and 40% occlusions. We note that SRC1 can slightly outperform CESR on the 56D space. This may be due to the non-optimality of the kernel size in the correntropy. As will be demonstrated in subsection V-G, a well tuned kernel size can further improve the recognition rate. Though SRC0 and LRC can also improve the recognition rates, their improvements are not good enough as compared with the other three robust methods. The verification results are also shown in Fig. 12 and Fig. 13. CESR is still overall better than the compared methods, especially when the occlusion is 40%.

Moreover, CESR can obtain much sparser coding. Fig. 14 shows two representative results of CESR and SRC1 against contiguous occlusions. In both examples (with 20% and 40% occlusions respectively), the estimated coefficients of CESR are sparser than the ones estimated by SRC1. Also, CESR is more efficient than SRC1. When the dimension is 2016-D ( $48 \times 42$ ), SRC1 takes nearly 800 seconds for computation of each test image on Matlab, while CESR only takes about 11 seconds.

2) **Corruption:** In practical face recognition scenarios, the test image  $y$  could be partially corrupted. We tested the robustness of CESR on the Extended Yale B Face Database. For each subject, half of the images were randomly selected for training, and the rest half were for testing. The training and testing set contained 1205 and 1209 images respectively. Since the images in the testing set have large variations due to different lighting conditions or facial expressions, it is a difficult recognition task. For CESR, SRC2, SRC0, LRC and 1NN, all the images were resized to  $48 \times 42$ ; for SRC1, we had to resize each image to  $24 \times 21$  and stacked it into a 504-D vector, because SRC1 is extremely expensive (if not impossible) to run SRC1 on high dimensional space. Each test image was corrupted by replacing a set of randomly selected pixels with random pixel value which follows a uniform distribution over  $[0, 255]$ . We varied the percentage of entire image pixels that will suffer corruption from 10% to 80%.

Fig. 15 (a) shows the recognition accuracies of CESR and its five competitors with respect to different levels of corruption. CESR dramatically outperforms the others when the corruption is beyond 50%. With 80% corruption, the recognition rate of CESR is still 67.7% whereas none of the other compared methods achieves higher than 25% recognition rate. If the parameter of CESR is well tuned, its recognition rate can reach 80.3% (see Fig. 19 (c)) against 80% corruption. Note that compared with the two robust SRCs, the improvement in terms of recognition rate by CESR is nearly 45%.

Fig. 15 (b) and (c) further show the verification rates with respect to various corruptions when FAR=0.001 and FAR=0.01. Although LRC can obtain 90% recognition rate when the corruption is smaller than 30%, its verification rates are extremely low (smaller than 30%). In contrast, CESR, SRC1, and SRC2 can significantly improve the ROC curves as they all seek sparse representation. When the corruption is larger than 50%, the verification rates of CESR are significantly higher than the others. Thus, CESR could be an effective method to deal with severe

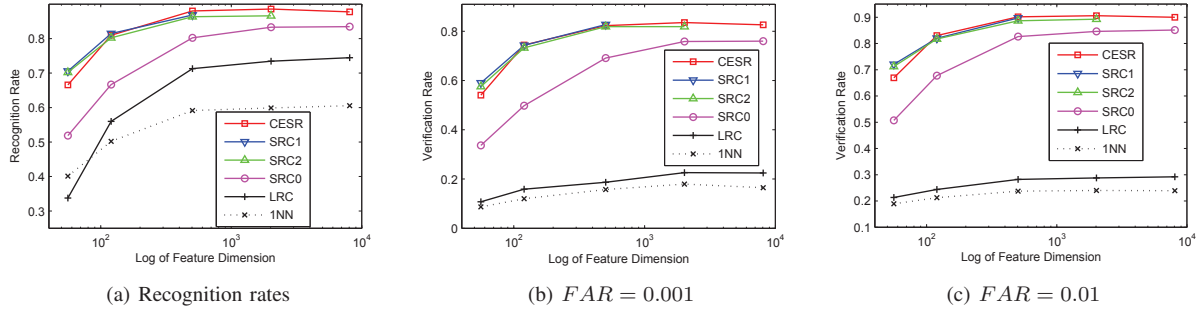


Fig. 12. Recognition rates and verification rates using various feature spaces and classifiers against 20% of contiguous occlusions.

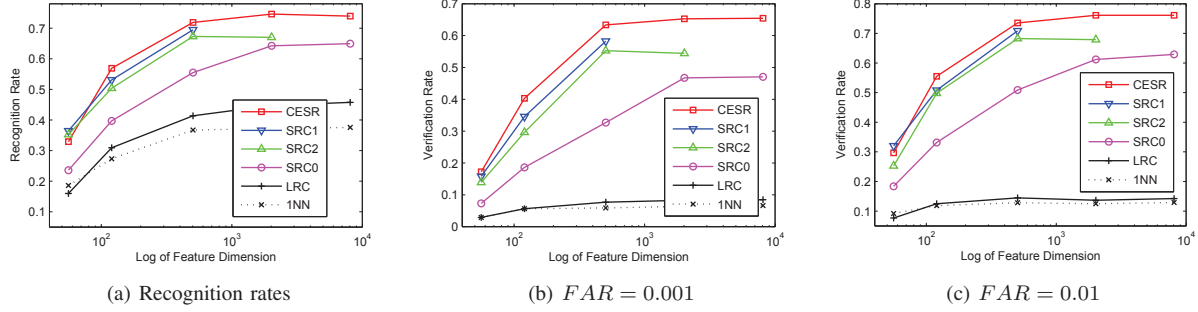


Fig. 13. Recognition rates and verification rates using various feature spaces and classifiers against 40% contiguous occlusions.

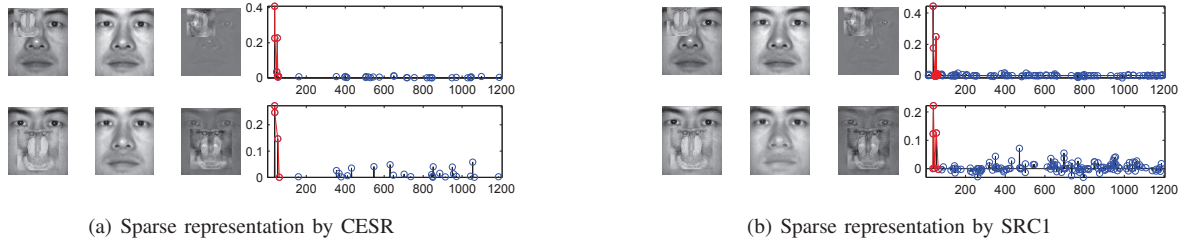


Fig. 14. Sparse representations by CESR and SRC against contiguous occlusion. The top row is with 20% occlusion and bottom row is with 40% occlusion. The red entries of coefficients correspond to training images of the same person as the test image. (a) Pictures from left to right are occluded images, the reconstructed images by using CESR, the difference image between original image and the reconstructed image, and the coefficients computed by CESR respectively. (b) Pictures from left to right are occluded images, the reconstructed images by using SRC1, the difference image between original image and the reconstructed image, and the coefficients computed by SRC1 respectively.

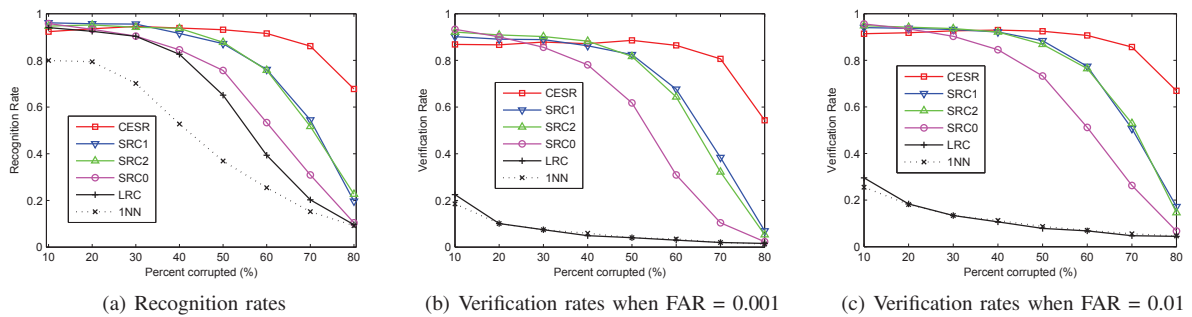


Fig. 15. Recognition rates and verification rates against random corruption using various classifiers. CESR significantly outperforms the others when there are severe corruptions.

corruption for face recognition.

#### D. Results on FRGC Database

In this subsection, we evaluated different methods on real facial images collected from uncontrolled conditions. We made use of a subset of the query set for FRGC experiment 4. The

FRGC version 2 face database is a challenging benchmark face recognition database. Two scenarios have been considered: 1) In the first scenario, for each selected subject, 8 images were randomly selected for training and the rest were for testing. 2) In the second scenario, we considered a real occlusion problem incurred in automatic face recognition system, especially in the



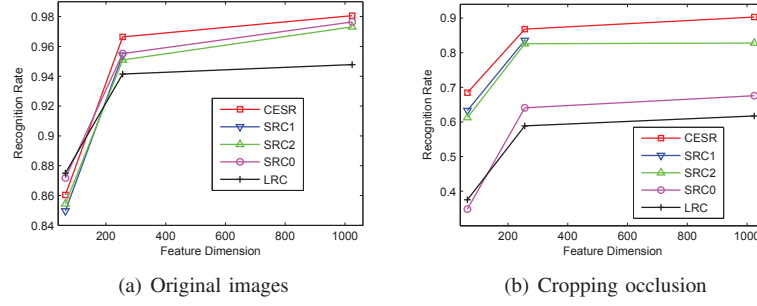


Fig. 16. Recognition rates of various classifiers on FRGC database.

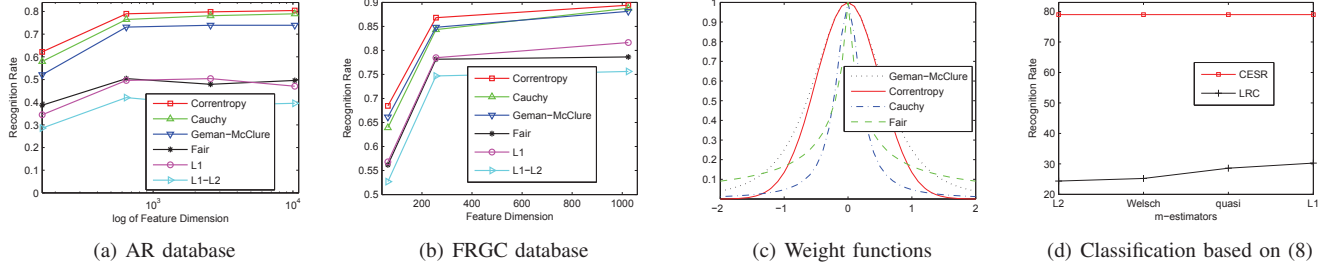


Fig. 17. Recognition rates of algorithm 1 based on different M-estimators. (a) The recognition rates on AR database. The experimental setting is the same as that in section V-B.1. (b) The recognition rates on FRGC database. The experimental setting is the same as that of the cropping occlusion in section V-D. (c) Weight functions of different M-estimators using the best tuned parameter. (d) The recognition rates of CESR and LRC based on (8) on AR database.

surveillance or infrared face recognition system. Note that when a person's face is close to the border of the camera [56], a face alignment program may only crop part of the face, and therefore the missing part needs to be fixed. To simulate this kind of cropping error, we randomly blocked each testing image by a rectangle near the image border.

For the first scenario, the recognition rates of the five compared methods are shown in Fig. 16 (a) as the functions of the number of feature dimensions. We see that the compared methods can be ranked in terms of ascending recognition rates by LRC, SRC1, SRC2, SRC0 and CESR. Although the improvement of CESR is limited, it can achieve a higher recognition rate with a small computation cost as compared to the SRC methods.

For the second scenario, we compared recognition rates against different levels of cropping error. Results are shown in Fig. 16 (b). We see that the recognition rates of both LRC and SRC0 drop fast as the level of cropping error increases. However, the three robust sparse methods can still obtain high recognition rates. As expected, CESR can well detect the cropping occlusion so that it achieves the highest recognition rates.

### E. M-estimators

Considering that the weight functions of  $l^1 - l^2$  and 'Fair' in Table I are exactly the minimizer functions relevant to the multiplicative form of the HQ [57], we can directly substitute the objective in (21) with  $l^1 - l^2$  and 'Fair' M-estimators. The simple modification of algorithm 1 is to substitute the minimizer function in (26) with other minimizer functions. Note that not all M-estimators can be optimized by half-quadratic technique. An estimator should satisfy some conditions as detailed in [57] and then can be optimized by the multiplicative form of the HQ. In order to evaluate different M-estimators, we still substituted

the minimizer function in (26) with other weight functions of  $l^1$ , 'Cauchy' and 'German-McClure' M-estimators. For each M-estimator, the results reported here are with the best tuned  $\theta$  in set  $\{2, 1.5, 1, 1/2, 1/5, 1/10, 1/20, 1/30, 1/40, 1/50\}$ . The best values of  $\theta$  for correntropy, 'Cauchy', 'German-McClure', 'Fair', and  $l^1 - l^2$  are 0.5, 0.05, 1, 0.2, and 0.025 respectively.

Fig. 17 (a) and (b) show the recognition rates on AR database and FRGC database respectively. We see that the recognition rates of correntropy, 'Cauchy' and 'German-McClure' are higher than those of 'Fair',  $l^1$ , and  $l^1 - l^2$ , on the two databases. Furthermore, the recognition rates of correntropy, 'Cauchy' and 'German-McClure' are very close. To further investigate the relationship among different M-estimators, we show the weight functions (or minimizer functions) of different M-estimators using the best tuned parameter in Fig. 17 (c). We see that the weight function of 'German-McClure' is similar to that of correntropy. The reason that correntropy can outperform 'German-McClure' may be because the weight function of 'German-McClure' assigns outliers larger weights. By comparing with 'Cauchy' and 'Fair', we see that the weight function of 'Fair' also give outliers larger weights so that the algorithm based on 'Fair' could not achieve a higher recognition rate. Therefore, according to the recognition rate and parameter selection, the correntropy is preferred at least for our problem compared to other M-estimators.

In the partial within-class match (PWCM) method [16] [17], the  $l^1$ -norm and 0.5-quasi-norm were used in (8) as robust estimator of classifier to further improve the classification performance after the coefficient has been computed. For further investigation, we implemented the classifiers of LRC and CESR by utilizing the 'Welsch' M-estimator,  $l^1$ -norm and 0.5-quasi-norm as the robust norm in (8). Fig. 17 (d) shows comparison results of the recognition rates between LRC and CESR. We see that the recognition rates of LRC increase when the robust estimators are

used in the classifier. However, the recognition rates of CESR are nearly the same. This may be due to the fact that the robust estimator of classifiers do not change the sparse coefficients computed by CESR dramatically and also the non-negative sparse representation is informative enough for classification.

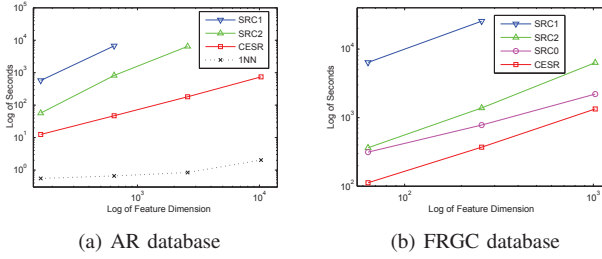


Fig. 18. **Computational cost on various feature spaces and classifiers.** (a) There are 952 images (8 images per subject) in the training set. (b) There are 2528 images (8 images per subject) in the training set.

### F. Comparison of the Computation Expense

Computational complexity of an algorithm is an important issue for face recognition. Fig. 18 (a) shows the overall computation time using various features on AR database, with the same experiment setting in Section V-B.1. And Fig. 18 (b) shows the overall computation time using various features on FRGC database, with the same experiment setting in Section V-D. On the AR database, SRC1, SRC2, and CESR take 56, 6.9 and 0.40 seconds for each test image respectively when the feature dimension is 644- $D$ . (In [24], SRC1 requires about 75 seconds per test image on a PowerMac G5.) Hence, the computation time of two SRC methods is extremely larger as compared to CESR. Although there are  $n$  linear variables and  $m$  auxiliary variables in CESR, CESR can efficiently estimate the  $m$  auxiliary variables in half-quadratic optimization. At each iteration, all the  $m$  auxiliary variables are updated by (26) according to Proposition 1. However, the SRC models treat  $m + n$  variables equally. When the dimension  $m$  is large, the computation cost of SRCs will increase rapidly and become extremely expensive<sup>10</sup>. Hence CESR is more suitable for robust and real-time pattern recognition tasks.

### G. Parameter Selection

1) **Kernel size  $\sigma$** : The kernel size  $\sigma$  is an important parameter which controls all robust properties of correntropy [27]. A well tuned kernel size value can eliminate the effect of outliers and noise much more effectively.

In this paper, we set the Gaussian kernel size as a single function of the average reconstruction error (defined in (33)). When the half-quadratic optimization is used to solve the correntropy, the correntropy will give outliers much smaller weights during iterations. The values of auxiliary variables (weights) are also computed by the Gaussian kernel function. A graphical description of this mechanism is given by plotting the Gaussian function as a function of the  $\theta$  (in Fig. 19 (a)). When the value of  $\theta$  is small, such as 0.5, the outliers will receive much smaller weights. When the value of  $\theta$  is large, such as 3, the outliers will receive relatively larger weights.

<sup>10</sup>Although SRC methods may be used on down-sampled images to reduce its computation cost, both its recognition and verification performances will decrease significantly if the resolution of down-sampled images is too small.

The simulations (in Fig. 19 (b) and (c)) were run to show how the kernel size affects the performance of CESR under various corruptions. The experimental setting is the same as that in Section V-C.2. On 10% corruption, the larger the  $\theta$  is, the higher the recognition rate CESR obtains. The margin between the maximum recognition rate and the minimum is 6%. On 80% corruption, CESR achieves the highest recognition rate when  $\theta = 0.5$ . The margin between the maximum recognition rate and the minimum is 46%. When the percentage of corruption or occlusion is smaller than 50%, the average reconstruction error in (33) is mainly dominated by uncorrupted pixels, and therefore  $\theta$  can be set to a larger value to give most of the uncorrupted pixels larger weights; when the corruption or occlusion is larger than 50%, the average reconstruction error is mainly dominated by corrupted pixels, and therefore  $\theta$  can be set to a smaller value to punish outliers significantly.

Fig. 19 (d) shows the ROC curves with different values of  $\theta_r$  in (35). It seems that the verification performance of CESR is robust to different  $\theta_r$ . Furthermore, the nonlinear scores learned by (34) can significantly improve the ROC curves compared with the scores (residuals) based on  $L_2$  distance.

2) **Regularization parameter  $\lambda$** : The regularization parameter  $\lambda$  in (21) is an important parameter to control the sparseness of sparse representation. We study how the  $\lambda$  affects recognition rates in the case of sunglasses occlusion on the AR database and cropping occlusion on the FRGC database. Specially, we normalized the columns of  $X$  to have unit  $l^2$ -norm so that we can easily tune the regularization parameter  $\lambda$  of CESR.

TABLE III  
 RECOGNITION RATES (%) AND AVERAGE NUMBER OF NONZEROS ( $l^0$  norm) UNDER DIFFERENT VALUES OF  $\lambda$ .

Value of $\lambda$	0	$1e^{-5}$	$1e^{-4}$	0.001	0.01	0.05
Image size $28 \times 23$ (AR database)						
Recognition rates	68.91	<b>69.75</b>	<b>69.75</b>	<b>69.75</b>	68.07	26.05
$l^0$ norm	10.55	10.55	10.54	10.50	9.92	5.88
Image size $56 \times 46$ (AR database)						
Recognition rates	<b>73.11</b>	<b>73.11</b>	<b>73.11</b>	72.27	68.07	25.21
$l^0$ norm	10.94	10.94	10.88	10.77	9.66	5.91
Image size $16 \times 16$ (FRGC database)						
Recognition rates	86.81	86.81	<b>88.55</b>	47.15	0.3	0
$l^0$ norm	15.31	15.04	12.54	4.28	0	0

Table III reports the recognition rates (%) and average number of nonzero entries of the learned sparse representation (in terms of  $l^0$  norm) when different values of  $\lambda$  are set. When we vary the value of  $\lambda$  from 0 to 0.05 on the AR database, the average number of nonzero entries decreases on both downsampling rates. This means that a larger value of  $\lambda$  will lead a more sparse solution. However, the sparsest solution may not yield the highest recognition rate. When  $\lambda$  is 0.05, the recognition rates under the two downsampling rates are only 26.05% and 25.21% on the AR database respectively.

On the FRGC database, we also observe that recognition rates can be further improved if  $\lambda$  in CESR is well tuned. This is because  $\lambda$  controls the sparsity of coefficient  $\beta$ . When the data is redundant, an appropriate value of  $\lambda$  makes the sparse representation more informative.

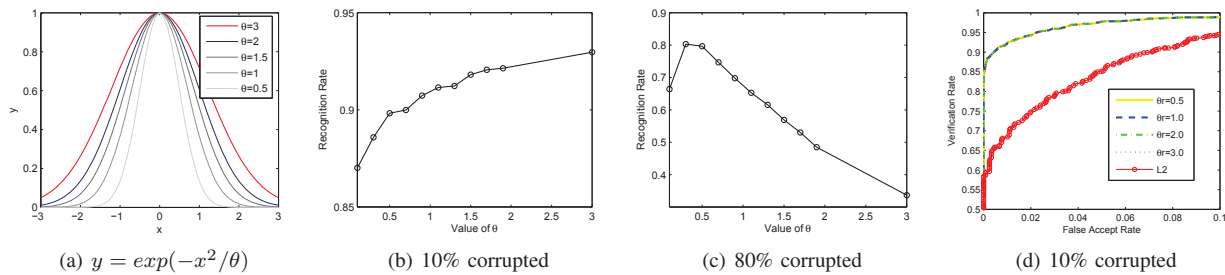


Fig. 19. **Recognition performance vs.  $\theta$  in Gaussian kernel size  $\sigma$ .** (a) Gaussian kernel as a function of  $\theta$ . (b) On 10% corruption, the recognition rate under various value of  $\theta$ . (c) On 80% corruption, the recognition rate under various value of  $\theta$ . (d) On 10% corruption, the ROC curves under various value of  $\theta_r$ . The  $L_2$  represents that the scores are computed by  $L_2$  distance.

## VI. CONCLUSION AND FUTURE WORKS

An effective sparse representation algorithm based on the maximum correntropy criterion is proposed for robust face recognition. The half-quadratic optimization technique is adopted to maximize the correntropy objective function, so that the difficult nonlinear optimization problem is reduced to learning a non-negative representation through a weighted linear least squares problem with non-negativity constraint at each iteration. Then a new active-set algorithm is developed to efficiently solve CESR. The sparse representation computed by CESR is robust to noise and can be computed more efficiently as compared to  $l^1$  norm based sparse algorithm. A classifier based on the sparse representation is proposed for robust face recognition. Experimental results show that CESR is able to provide new advanced ability to deal with errors caused by occlusions, corruptions and etc., and can achieve striking recognition performance in tough conditions.

The correntropy based sparse framework is presented to be a novel way to learn robust and informative presentations in this paper. This could suggest that improvement could be made if it is applied to learn an informative graph [28] for graph-based machine learning tasks such as data clustering, subspace learning, and semi-supervised learning, where the MSE criterion is still widely used in these algorithms. As robustness is an important issue for signal classification [58], image classification [59], and real-time object detection [24], we expect to investigate more use of the (maximum) correntropy criterion for these applications in future.

## ACKNOWLEDGMENT

We would like to greatly thank the associate editor and the reviewers for their valuable comments and advice. This work was supported in part by DUT R & D Start-up costs, the Natural Science of Foundation of China (#61075051), and the NSFC-GuangDong (U0835005).

## REFERENCES

- [1] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
- [2] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
- [3] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective representation using ica for face recognition robust to local distortion and partial occlusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1977–1981, 2005.
- [4] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," Tech. Rep. 245, 1994.
- [5] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2010.
- [6] A. Leonardis and H. Bischof, "Robust recognition using eigenimages," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 99–118, 2000.
- [7] A. M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 748–763, 2002.
- [8] S. Chen, T. Shan, and B. C. Lovell, "Robust face recognition in rotated eigenspaces," in *The Twenty-second International Image and Vision Computing New Zealand Conference*, 2007.
- [9] C. H. Hoi and M. R. Lyu, "Robust face recognition using minimax probability machine," in *ICME 04*, 2004, pp. 1175–1178.
- [10] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 337–350, 2006.
- [11] R. A. Vazquez, H. Sossa, and B. A. Garro, "Low frequency response and random feature selection applied to face recognition," in *International Conference on Image Analysis and Recognition (ICIAR)*, 2007.
- [12] R. A. Vazquez and H. Sossa, "Associative memories applied to pattern recognition," in *International Conference on Artificial Neural Networks*, 2008, pp. 111–120.
- [13] F. de la Torre and M. J. Black, "Robust parameterized component analysis: theory and applications to 2d facial appearance models," *Computer Vision and Image Understanding*, vol. 91, pp. 53–71, 2003.
- [14] M. Black and A. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *International Journal of Computer Vision*, vol. 26, no. 1, pp. 63–84, 1998.
- [15] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *European Conference on Computer Vision*, 1998, pp. 484–498.
- [16] H. Jia and A. M. Martinez, "Face recognition with occlusions in the training and testing sets," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2008.
- [17] —, "Support vector machines in face recognition with occlusions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [18] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma, "Face recognition with contiguous occlusion using markov random fields," in *IEEE International Conference on Computer Vision*, 2009.
- [19] K.-H. Jeong, W. Liu, S. Han, E. Hasanbelliu, and J. C. Principe, "The correntropy mace filter," *Pattern Recognition*, vol. 42, no. 5, pp. 871–885, 2009.
- [20] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.
- [21] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. on Information Theory*, vol. 51, no. 12, 2005.
- [22] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal  $l_1$ -norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.
- [23] Y. Bengio, "Learning deep architectures for ai," *Foundations and Trends in Machine Learning*, 2009.
- [24] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.



[25] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[26] J. Wright and Y. Ma, "Dense error correction via l1-minimization," preprint, submitted to *IEEE Trans. on Information Theory*, 2008.

[27] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," *IEEE Trans. on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, 2007.

[28] X. Yuan and B.-G. Hu, "Robust feature extraction via information theoretic learning," in *International Conference on Machine Learning (ICML)*, Montreal, Canada, 2009.

[29] J. C. Principe, D. Xu, and J. W. Fisher, "Information-theoretic learning," in S. Haykin, editor, *Unsupervised Adaptive Filtering, Volume 1: Blind-Source Separation*. Wiley, 2000.

[30] D. Xu, "Energy, entropy and information potential for neural computation," Ph.D. dissertation, University of Florida, 1999.

[31] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd edition. New York: John Wiley, 2005.

[32] S. Z. Li and J. Lu, "Face recognition using nearest feature line method," *IEEE Trans. Neural Network*, vol. 10, no. 2, pp. 439–443, 1999.

[33] J. Laaksonen, "Local subspace classifier," in *International Conference on Artificial Neural Networks*, 1997.

[34] S. Z. Li, "Face recognition based on nearest linear combinations," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 1998, pp. 839–844.

[35] P. Vincent and Y. Bengio, "K-local hyperplane and convex distance nearest neighbor algorithms," in *Advances in Neural Information Processing Systems*, vol. 14, 2001, pp. 985–992.

[36] J.-T. Chien and C.-C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1644–1649, 2002.

[37] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 11–18.

[38] P. P. Pokharel, W. Liu, and J. C. Principe, "A low complexity robust detector in impulsive noise," *Signal Processing*, vol. 89, no. 10, pp. 1902–1909, 2009.

[39] I. Santamaria, P. P. Pokharel, and J. C. Principe, "Generalized correlation function: Definition, properties, and application to blind equalization," *IEEE Trans. on Signal Processing*, vol. 54, no. 6, pp. 2187–2197, 2006.

[40] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[41] W. Liu, P. P. Pokharel, and J. Principe, "error entropy, correntropy and m-estimation," in *machine learning for signal processing*, 2006.

[42] P. Huber, *Robust statistics*. Wiley, 1981.

[43] I. Mizera and C. Muller, "Breakdown points of cauchy regression-scale estimators," *Statistics and Probability Letters*, vol. 57, no. 1, pp. 79–89, 2002.

[44] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for l1-minimization with applications to compressed sensing," *SIAM J. IMAGING SCIENCES*, vol. 1, no. 1, pp. 143–168, 2008.

[45] S. Yang, H. Zha, S. Zhou, and B.-G. Hu, "Variational graph embedding for globally and locally consistent feature extraction," in *Europe Conference on Machine Learning (ECML)*, 2009, pp. 538–553.

[46] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[47] L. F. Portugal, J. J. Judice, and L. N. Vicente, "A comparison of block pivoting and interior-point algorithms for linear least squares problems with nonnegative variables," *Mathematics of Computation*, vol. 63, no. 208, pp. 625–643, 1994.

[48] A. M. Martinez and R. Benavente, "The ar face database," *CVC, Tech. Rep.*, 1998.

[49] A. S. Georgiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[50] P. J. Phillips, P. J. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[51] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[52] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Neural Information Processing Systems NIPS*, 2006.

[53] E. Candes and J. Romberg, *l1-magic: recovery of sparse signals via convex programming*. <http://www.acm.caltech.edu/l1magic/>, 2005.

[54] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Computer Vision and Pattern Recognition*, 2009, pp. 2790–2797.

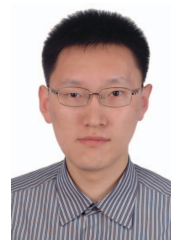
[55] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li, "Face matching from near infrared to visual images," in *Proceedings of IAPR/IEEE ICB*, 2007.

[56] R. He, B.-G. Hu, and X. Yuan, "Robust discriminant analysis based on nonparametric maximum entropy," in *Asian Conference on Machine Learning (ACML)*, NanJing, China, 2009.

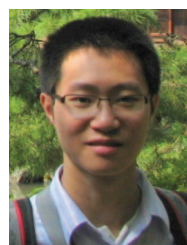
[57] M. Nikolova and M. K. NG, "Analysis of half-quadratic minimization methods for signal and image recovery," *Society for Industrial and Applied Mathematics*, vol. 27, no. 3, pp. 937–966, 2005.

[58] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Neural Information Processing Systems (NIPS)*, 2006.

[59] G. Peyre, "Sparse modeling of textures," *Journal of Mathematical Imaging and Vision*, vol. 34, no. 1, pp. 17–31, 2009.



**Ran He** received the BS degree in Computer Science from the Dalian University of Technology of China, and the Ph.D. degree in Pattern Recognition and Intelligent System from Institute of Automation, Chinese Academy of Sciences, in 2009. He is currently an assistant professor with NLPR (National Laboratory of Pattern Recognition), Institute of Automation, Chinese Academy of Science, Beijing, China. His research interests include information theoretic learning and computer vision.



**Wei-Shi Zheng** is a Postdoctoral Researcher at the Department of Computer Science, Queen Mary University of London, UK. He is now working on the European SAMURAI Research Project with Prof. Shaogang Gong and Dr. Tao Xiang. Prior to that, he received his Ph.D. degree in Applied Mathematics at Sun Yat-Sen University, China, 2008. He has been a visiting student working with Prof. Stan Z. Li at the Institute of Automation, Chinese Academy of Sciences, and an exchanged research student working with Prof. Pong C. Yuen at Hong Kong Baptist University. He was awarded the HP Chinese Excellent Student Scholarship 2008. His current research interests are in object association and categorization for visual surveillance. He is also interested in discriminant/sparse feature extraction, dimension reduction, kernel methods in machine learning, transfer learning, and face image analysis.



**Bao-Gang Hu** received his M.Sc. degree from the University of Science and Technology, Beijing, China in 1983, and his Ph.D. degree from McMaster University, Canada in 1993, all in Mechanical Engineering. From 1994 to 1997, Dr. Hu was a Research Engineer and Senior Research Engineer at C-CORE, Memorial University of Newfoundland, Canada. Currently, He is a Professor with NLPR (National Laboratory of Pattern Recognition), Institute of Automation, Chinese Academy of Science, Beijing, China. From 2000 to 2005, he was the Chinese Director of LIAMA (the Chinese-French Joint Laboratory for Computer Science, Control and Applied Mathematics). His main research interests include pattern recognition and plant growth modeling. He is a Senior Member of IEEE.