LETTER **Communicated by José C. Principe**

A Regularized Correntropy Framework for Robust Pattern Recognition

Ran He
rhe@nlpr.ia.ac.cn
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China

Wei-Shi Zheng
wszheng@ieee.org
School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510275, China, and Department of Computer Science, Queen Mary University of London, London, U.K.

Bao-Gang Hu
bghu@nlpr.ia.ac.cn
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Xiang-Wei Kong
kongwx@dlut.edu.cn
School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China

This letter proposes a new multiple linear regression model using regularized correntropy for robust pattern recognition. First, we motivate the use of correntropy to improve the robustness of the classical mean square error (MSE) criterion that is sensitive to outliers. Then an l^1 regularization scheme is imposed on the correntropy to learn robust and sparse representations. Based on the half-quadratic optimization technique, we propose a novel algorithm to solve the nonlinear optimization problem. Second, we develop a new correntropy-based classifier based on the learned regularization scheme for robust object recognition. Extensive experiments over several applications confirm that the correntropy-based l^1 regularization can improve recognition accuracy and receiver operator characteristic curves under noise corruption and occlusion.

1 Introduction

Robust data classification or representation is a fundamental task with a long history in computer vision. Algorithmic robustness, which is derived

from the statistical definition of a breakdown point (Meer, Stewart, & Tyler, 2000; Sanja, Skocaj, & Leonardis, 2006), is the ability of an algorithm to tolerate a large number of outliers. Therefore, a robust method should be effective enough to reject outliers in images and perform classification on only uncorrupted pixels. A great deal of work has addressed subspace learning (Leonardis & Bischof, 2000; Torre & Black, 2003) and sparse signal representation (Mairal, Elad, & Sapiro, 2008; Wright et al., 2010) in order to develop more robust image-based object recognition. Despite significant improvement, performing robust classification is still challenging due to the nature of unpredictable outliers and noise in an image, since corruption may occupy any parts of an image and have arbitrarily large values in magnitude (Wright, Yang, Ganesh, Sastry, & Ma, 2009).

Recently information theoretic learning (ITL) (Principe, Xu, & Fisher, 2000; Xu, 1999) has shown its superiority in robust learning and classification. Yuan and Hu (2009), He, Hu, and Yuan (2009), Viola, Schraudolph, and Sejnowski (1995), and He, Hu, Yuan, and Zheng (2010), use Renyi entropy (Cover & Thomas, 2005) and correntropy (Santamaria, Pokharel, & Principe, 2006) as cost functions to learn robust subspaces under supervised and unsupervised learning. Liu, Pokharel, and Principe (2007) proved that correntropy is a robust function (in the sense of Huber) for linear and nonlinear regression. Xu (1999), Cover and Thomas (2005), Pokharel, Liu, and Principe (2009), and Yuan and Hu (2009) discussed the connection between entropy and robust functions. The correntropy MACE filter is a detector for automatic target detection and recognition (ATR) (Jeong, Liu, Han, Hasanbelliu, & Principe, 2009) and can cope with random subspace projections (Jeong & Principe, 2008). Principe, Xu, Zhao, and Fisher (2000), Torkkola (2003), and Yang, Zha, Zhou, and Hu (2009) achieve feature extraction by directly maximizing the quadratic mutual information between the class label and the features. Grandvalet and Bengio (2006) introduced the concept of entropy regularization in semisupervised learning to learn a robust structure. Numerical results on real-world machine learning tasks have demonstrated that the methods based on information-theoretic objectives can make the algorithm significantly robust to noise (Yuan & Hu, 2009).

In real-world applications, the classification problem is often not defined in a closed set and always has to deal with noisy and occluded data.[1] Crafting an appropriate cost function will improve results in these very difficult environments (Liu et al., 2007). Correntropy is a recently developed information-theoretic metric. Despite its advanced performance in terms of robustness, the solution is not necessarily sparse. Sparsity has been shown to be appropriate for alleviating the effect of outliers and noise due to

[1]Closed set domains assume that all classes of a domain have been known and can be used in training. In contrast, open set domains assume that new classes may be encountered. For example, face recognition and object recognition are open set problems.

selective processing (Wright et al., 2010). Hence, we wish to study robust image-based object recognition by regularizing the correntropy cost function. First, to overcome the weakness of mean square error (MSE) that is prone to the presence of outliers, a robust linear representation based on the correntropy-induced metric (CIM) (Liu et al., 2007) is studied. Then the l^1 regularization is further imposed on CIM to learn an informative and sparse model. The half-quadratic optimization technique is used to efficiently solve the nonlinear ITL optimization problem. Second, inspired by the nearest feature classifiers (Li, 1998; Chien & Wu, 2002; He, Ao, Xiang, & Li, 2008) and linear regression classification (Naseem, Togneri, & Bennamoun, 2010), which make use of the distance from a sample to a subspace to perform classification, we propose a regularized correntropy classifier. The proposed method is innovative in the field of robust recognition and makes the recognition of image-based objects possible under tough conditions.

Our study of regularized correntropy combines research in ITL (Principe, 2010), one-class linear classifiers (Li, 1998), and sparse signal representation (Wright et al., 2009) into a unified framework. This is significantly different from the related state-of-the-art sparse representation classifier (SRC) (Wright et al., 2009, 2010), which assumes that noisy items have a sparse representation. Our regularized correntropy method has no special assumptions about noise, so it can efficiently handle errors incurred by occlusion and corruption. Extensive experiments on several machine learning tasks demonstrate encouraging results in algorithmic robustness and efficacy as compared to SRC.

The remainder of this letter is organized as follows. In section 2, we briefly review previous robust methods and point out their main limitations. Then we present an l^1 regularized correntropy to learn multiple linear regression models for robust recognition in section 3. In section 4, the proposed approaches are verified by extensive experiments on machine learning tasks and comparisons with other state-of-the-art techniques.We summarize this work in section 5.

2 Related Work

Many machine learning techniques have been developed for robust automatic target detection and recognition (ATR) (Kumar, 1992). Here, we briefly review some principal work from two aspects: fixed-size and variable-length representations (Bengio, 2009).

2.1 Fixed-Size Representations. The earliest multivariate technique for multiclass pattern recognition is the synthetic discriminant function (SDF) proposed by Hester and Casasent (1980). The basic model builds a large linear network (or a template matcher) by using all the training images and can be computed analytically and effectively using frequency domain techniques (Jeong et al., 2009). This class of SDF models has been used for

both representation and recognition in, for example, ATR and face recognition (Kumar, Savvides, & Xie, 2006). One shortcoming of the conventional SDF is related to the noiseless model. Kumar (1986) then proposed the minimum variance SDF (MVSDF) filter by considering additive input noise, and Mahalanobis, Kumar, and Casasent (1987) developed the minimum average correlation energy (MACE) filter. The MACE minimizes the average correlation energy of the output over training samples to produce a sharp correlation peak. To further improve the generalization properties of MACE, Refregier and Figue (1991) proposed the optimal trade-off filters (OTSDFs) to combine the properties of various SDFs. However, both MVSDF and OTSDF are impractical since the additive noise properties cannot always be known in advance (Jeong et al., 2009). Hence, based on the correntropy cost function, Jeong et al. (2009) proposed a robust nonlinear extension to the MACE filter for ATR.

Another robust fixed-size representation is robust subspace learning (Torre & Black, 2003), which has been widely used for shape representation and tracking, for example, in computer vision applications. Different approaches have been explored in the literature to make the learning more robust. To alleviate the problem of occlusion, Pentland, Moghaddam, and Straner (1994) proposed modular eigenspaces. Modular linear regression-based classification (LRC) approaches (Naseem et al., 2010) have extended the concept of the modular representation for robust face recognition. Ohba and Ikeuchi (1997) presented the eigenwindow method to recognize partially occluded objects. These methods based on the eigenwindow partially alleviate the problems of occlusion but cannot solve them entirely (Leonardis & Bischof, 2000). Black and Jepson (1998) used a conventional robust M-estimator for computing the coefficients of subspace by substituting the MSE with a robust one. Leonardis and Bischof (2000) proposed a subsampling and hypothesize-and-test approach to reject outliers and learn the coefficients of subspace. Torre and Black (2003), Ding, Zhou, He, and Zha (2006), Kwak (2008), and He et al. (2010) developed robust principal component analysis methods to learn robust components and coefficients. These methods mainly focus on learning robust representation rather than performing classification, and therefore they are often noted as reconstructive methods.

In robust discriminative models, MCD-estimators (Hawkins & McLachlan, 1997; Hubert & Driessen, 2003), S-estimators (He & Fung, 2000; Croux & Dehon, 2001), M-estimators (Mangasarian & Musicant, 2000), and information-theoretic estimators (Yuan & Hu, 2009; He et al., 2009) are used to replace the classical location of MSE or scatter matrix estimators. A main limitation of these methods is that they detect the whole example as an outlier and discard it from the learning process. Sanja et al. (2006) combined the discrimination power of discriminative methods and the reconstruction property of reconstructive methods into a framework. However, the reconstructive part in Sanja et al. (2006) depends on the heuristic coefficient

estimation procedure of Leonardis and Bischof (2000), which needs to generate a set of hypotheses and make use of the hypothesize-and-test to get an acceptable result.

Although previous methods improve the robustness, robust classification is still a challenge due to the nature of unpredictable outliers and occlusions (Wright et al., 2009). Moreover, the fixed-size representations compute a fixed-length representation for a test sample, so they may be less flexible in the design as compared with the variable-length representations (Bengio, 2009).

2.2 Variable-Length Representations. Let $y \in \mathbb{R}^{d\times 1}$ be a new test sample and $X_c \doteq [x_1^c, x_2^c, \ldots, x_{n_c}^c] \in \mathbb{R}^{d\times n_c}$ be a matrix whose columns are n_c training samples belonging to the cth class from k distinct classes. Let x_{ij}^c and y_j be the jth entry of x_i^c and y, respectively. Given a sufficiently expressive training set X_c, a test image y of class c can be approximated as a linear combination of the training set: $y \approx X_c\beta^c$ for the 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

$$X \doteq [X_1, X_2, \ldots, X_k] = [x_1^1, x_2^1, \ldots, x_{n_k}^k] \in \mathbb{R}^{d\times n}, \tag{2.1}$$

where $n = \sum_{c=1}^{k} n_c$. Alternatively, a test sample y can also be expressed as a linear combination of all training set,

$$y \approx X\beta, \tag{2.2}$$

where coefficient vector $\beta \in \mathbb{R}^n$. Equation 2.2 is also the SDF formulation of Hester and Casasent (1980).

The sparse representation classifier (SRC) (Wright et al., 2009, 2010) seeks the sparsest solution to equation 2.2,

$$\min \|\beta\|_0 \quad s.t. \quad y = X\beta, \tag{2.3}$$

where the l^0-norm $\|.\|_0$ counts the number of nonzero entries in a vector. Originally inspired by theoretical results (Candes & Tao, 2005; Donoho, 2006b) that the solution of the l^0 minimization problem is equal to the solution of the l^1 minimization problem if the solution is sparse enough, SRC seeks an approximate solution of β by solving the following convex relaxation,

$$\min \|\beta\|_1 \quad s.t. \quad y = X\beta, \tag{2.4}$$

where $\|\beta\|_1 = \sum_{i=}^{n} |\beta_i|$. Here, we denote the algorithm to solve equation 2.4 by SRC0.

Classically, the Lasso solution (Tibshirani, 1996) can be defined as an unconstrained minimizer of equation 2.4,

$$\min \|y - X\beta\|^2 + \lambda\|\beta\|_1, \tag{2.5}$$

where λ can be viewed as an inverse of the Lagrange multiplier in equation 2.4. To deal with occlusions and corruptions, Wright et al. (2009) further modified the linear model in equation 2.2 as

$$y = X\beta + e, \tag{2.6}$$

where $e \in \mathbb{R}^d$ is a vector of errors. Assuming that a fraction of e is nonzero (i.e., the noisy vector e has a sparse representation), SRC seeks the sparsest solution to the robust system of equation 2.6:

$$\min \|\beta\|_1 + \|e\|_1 \quad s.t. \quad y = X\beta + e. \tag{2.7}$$

We denote the algorithm to solve equation 2.7 by SRC1.

Although SRC1 can achieve impressive results against traditional methods in terms of recognition rate and receiver operator characteristic (ROC) curves, the sparse assumption of a noisy vector e can bring unsatisfactory results under severe occlusion and noise corruption. In addition, the computational cost of SRC1 is quite high. It will take nearly 100 seconds for SRC1 to process a test image stacked in a 700-D vector.

3 Regularized Correntropy Costs

In order to address the weakness in MSE-based linear model, this section develops a correntropy-based linear model for outlier detection. In order to learn a robust and sparse model, the l^1 regularization strategy is further adopted, and the optimization method is presented.

3.1 Correntropy. Recently the concept of correntropy (Liu et al., 2007) was proposed within ITL. It is defined as a generalized similarity measure between two arbitrary random variables A and B (Liu et al., 2007),

$$V_\sigma(A, B) = E[k_\sigma(A, B)], \tag{3.1}$$

where $k_\sigma(.)$ is the kernel function that satisfies Mercer's theory (Vapnik, 1995) and $E[.]$ denotes the mathematical expectation. It takes advantage of the kernel technique that nonlinearly maps the input space to a higher-dimensional space. It has a clear theoretical foundation and is symmetric, positive, and bounded.

In practice, the joint probability density function is often unknown, and only a finite number of data $\{(a_i, b_i)\}_{i=1}^{d}$ are available, which lead to the following sample estimator of correntropy:

$$\hat{V}_{d,\sigma}(A, B) = \frac{1}{d}\sum_{i=1}^{d} k_\sigma(a_i, b_i). \tag{3.2}$$

When k_σ is the gaussian kernel function $k_\sigma(a_i, b_i) \triangleq g(a_i - b_i) = \exp(-\frac{(a_i-b_i)^2}{2\sigma^2})$, we can rewrite equation 3.2 as

$$\hat{V}_{d,\sigma}(A, B) = \frac{1}{d}\sum_{i=1}^{d} g(a_i - b_i). \tag{3.3}$$

The maximum of correntropy of error in equation 3.2 is called the maximum correntropy criterion (MCC) (Liu et al., 2007). Given two vectors $A = (a_1, \ldots, a_d)$ and $B = (b_1, \ldots, b_d)$, the sample estimator of correntropy defines a metric in the sample space and is named as the correntropy-induced metric (CIM) (Liu et al., 2007).

Compared with the global measure mean square error (MSE), MCC is local along the bisector of the joint space and lies on the first quadrant of a sphere, which means that the value of correntropy is mainly decided by the kernel function along the line $A = B$ (Liu et al., 2007). Correntropy has a close relationship with redescending m-estimators (Huber, 1981). Liu et al. (2007) first imposed it and proved that correntropy is a robust function (in the sense of Huber) for linear and nonlinear regression. One of its main merits is that the kernel size controls all the properties of correntropy (Liu et al., 2007). It establishes a close relationship between the m-estimation and methods of ITL and provides a practical way to choose an appropriate kernel size (Liu et al., 2007).

3.2 Correntropy-Based Linear Model. In machine learning, the solution of the linear model in equation 2.2 is typically computed by the following least squares problem:

$$\min_{\beta} \|y - X\beta\|^2 = \sum_{j=1}^{d}\left(y_j - \sum_{i=1}^{n} x_{ij}\beta_i\right)^2. \tag{3.4}$$

This solution is based on MSE measurement that has been recognized to be sensitive to outliers (Martinez, 2002; Sanja et al., 2006). Here outliers are corrupted image pixels that are significantly different from uncorrupted image pixels. Outliers would seriously reduce the reliability of most machine learning models. They are significant challenges mainly due to the

unpredictable nature of the error—the error may be arbitrarily large in magnitude and therefore cannot be ignored or treated with methods developed for small noise (Wright et al., 2009).

To tackle this issue, we introduce the correntropy in robust object recognition. By replacing MSE with CIM, we obtain the following correntropy-based model:

$$\max_{\beta} \sum_{j=1}^{d} g\left(y_j - \sum_{i=1}^{n} x_{ij}\beta_i\right) \tag{3.5}$$

Different from methods based on MSE that all entries of the representation will contribute equally to the value of the measurement, methods based on CIM treat individual entries of the representation differently and give more emphasis on the entry that is close to the entry y_j of test sample y. This means that if entry y_j is occluded or corrupted, the entry corresponding to outliers will provide small contributions to the objective in equation 3.5. As a result, the noise can be handled uniformly within the framework of correntropy. Although the models trained by correntropy are robust to outliers, they are not necessarily sparse.

3.3 l^1-Regularized Maximum Correntropy. Recent advances in sparse signal representation and compressive sensing (Candes & Tao, 2005; Donoho, 2006a; Wright et al., 2009) have shown that the sparse code computed by l^1 regularization could be informative as well as discriminative. We further impose the l^1 regularization on equation 3.5 in order to learn a robust and discriminative linear model, which merges sparse signal representation and information-theoretic learning. Then we get the following l^1-regularized maximum correntropy problem:

$$J_{l^1} = \max_{\beta} \sum_{j=1}^{d} g\left(y_j - \sum_{i=1}^{n} x_{ij}\beta_i\right) - \lambda\|\beta\|_1. \tag{3.6}$$

Since the objective of l^1-regularized maximum correntropy in equation 3.6 is also nonlinear, it is difficult to directly optimize the correntropy cost. Fortunately, we recognize that the half-quadratic technique (Yuan & Hu, 2009), the EM method (Yang et al., 2009), and the conjugate gradient algorithm (Grandvalet & Bengio, 2006) can be used to solve this ITL optimization problem. In this study, we follow the approach of Yuan and Hu (2009) and use the half-quadratic technique to solve the regularized entropy maximization problem. According to the property of the convex conjugated function (Boyd & Vandenberghe, 2004), we have

Proposition 1. *There exists a convex conjugated function* φ *of* $g(x) = exp(-x^2/2\sigma^2)$[2] *such that*

$$g(x) = \max_{p'} \left(p' \frac{\|x\|^2}{\sigma^2} - \varphi(p') \right), \tag{3.7}$$

and for a fixed x*, the maximum is reached at* $p' = -g(x)$ *(Yuan & Hu, 2009).*

Substituting equation 3.7 into equation 3.6, we have the augmented objective function in an enlarged parameter space,

$$\hat{J}_{l^1} = \max_{\beta, p} \sum_{j=1}^{d} \left(p_j \left(y_j - \sum_{i=1}^{n} x_{ij} \beta_i \right)^2 - \varphi(p_j) \right) - \lambda \|\beta\|_1, \tag{3.8}$$

where $p = [p_1, \ldots, p_d]^T$ stores the auxiliary variables introduced in the half-quadratic optimization. According to proposition 1, for a fixed β, the following equation holds:

$$\max_{\beta} J_{l^1}(\beta) = \max_{\beta, p} \hat{J}_{l^1}(\beta, p). \tag{3.9}$$

Then we can conclude that maximizing J_{l^1} is identical to maximizing the augmented $\hat{J}_{l^1}$. Obviously a local maximizer (β, p) can be calculated in an alternate way,

$$p_j^{t+1} = -g\left(y_j - \sum_{i=1}^{n} x_{ij} \beta_i^t \right), \tag{3.10}$$

$$\beta^{t+1} = \arg\min_{\beta} (y - X\beta)^T P (y - X\beta) + \lambda \|\beta\|_1, \tag{3.11}$$

where t means the tth iteration and P is a diagonal matrix whose diagonal element $P_{jj} = -p_j^{t+1}$. It is clear that the optimization in equation 3.11 is a robust formulation of the Lasso in equation 2.5. The auxiliary variables $-p$ are weights. The optimal problem in equation 3.11 can also be restated as the following l^1-regularized quadratic problem:

$$\min_{\beta} \frac{1}{2} \beta^T \hat{X}^T \hat{X} \beta - (\hat{X}^T \hat{y})^T \beta + \frac{\lambda}{2} \|\beta\|_1, \tag{3.12}$$

where $\hat{X} = \sqrt{P} X$ and $\hat{y} = \sqrt{P} y$.

[2]Strictly speaking, φ is the conjugate function of the exponential function $\exp(-x)$. Here we apply the variable substitution method (substitute x with x^2/σ^2), which is commonly used in HQ.

Let θ be a d-dimensional vector whose element $\theta_i \in \{-1, 0, 1\}$ denotes sign(β_i) and F and G be two subsets of $\{1, \ldots, n\}$ such that $F \cup G = \{1, \ldots, n\}$ and $F \cap G = \phi$. And let F and G be the working set and inactive set in the active set algorithm, respectively. Then we obtain the following partitions of $\hat{X}$, β, and θ:

$$\hat{X} = [\hat{X}_F, \hat{X}_G], \ \beta = [\beta_F, \beta_G], \ \theta = [\theta_F, \theta_G], \tag{3.13}$$

where $\hat{X}_F \in \mathbb{R}^{d \times |F|}$, $\hat{X}_G \in \mathbb{R}^{d \times |G|}$, $\beta_F \in \mathbb{R}^{|F| \times 1}$, $\beta_G \in \mathbb{R}^{|G| \times 1}$, $\theta_F \in \mathbb{R}^{|F| \times 1}$, $\theta_G \in \mathbb{R}^{|G| \times 1}$, and $|F|$, $|G|$ are the number of F and G, respectively. According to the feature-sign search algorithm (Lee, Battle, Raina, & Ng, 2006), when θ has been estimated, we can iteratively solve the l^1-regularized problem of equation 3.12 by the following quadratic program (QP):

$$\min \frac{1}{2}\beta_F^T \hat{X}_F{}^T \hat{X}_F \beta_F - (\hat{X}_F{}^T \hat{y})^T \beta_F + \frac{\lambda}{2}\theta_F^T \beta_F. \tag{3.14}$$

Instead of finding the optimal solution of equation 3.12, we can find only a local solution to increase the objective. The l^1-regularized maximum correntropy algorithm is shown in algorithm 1. It maintains the working set F of potentially nonzero coefficients in β, their corresponding signs, and the inactive set G of all other zero coefficients. In the feature-sign step, it systematically searches for the optimal working set and coefficient signs to reduce the objective in equation 3.11. It gives a current guess for the active set and the signs and computes the analytical solution $\hat{\beta}_F$ to the resulting unconstrained QP, and then it updates the solution by an efficient discrete line search between the current solution β_F and $\hat{\beta}_F$. In the half-quadratic step, it alternatively maximizes the correntropy objective according to current β. Proposition 2 shows that each alternative maximum step increases the correntropy and that the overall algorithm converges to the optimal solution:

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

Proof. According to lemmas 3.1 and 3.2 of Lee et al. (2006), we learn the objective is increased at each feature sign step. Combining with proposition 1, we have

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

Therefore, the cost function increases at each alternate maximization step. The sequence $\{\hat{J}_{l^1}(\beta^t, p^t), t = 1, 2, \ldots\}$ is nondecreasing. We can easily verify that $J_{l^1}(\beta)$ is bounded (Liu et al., 2007), and by equation 3.9, we get that $\hat{J}_{l^1}(\beta^t, p^t)$ is also bounded. Consequently we can conclude that algorithm 1 will converge.

Algorithm 1: l^1-Regularized Maximum Correntropy.

Input: data matrix X, test sample y, kernel size σ, $p^1 = -\vec{1}$, $F = \phi$, $\hat{X} = X\sqrt{P}$, $\hat{y} = \sqrt{P}y$, $\theta = \vec{0}$, and $\beta = \vec{0}$.

Output: β

1. **Update the working set**: From zero coefficients of β, compute $r = \arg\max_r \left|\frac{\partial||\hat{y}-\hat{X}\beta||^2}{\partial\beta_r}\right|$.
   If $\left|\frac{\partial||\hat{y}-\hat{X}\beta||^2}{\partial\beta_r}\right| > \lambda$ Then set $\theta_r = -sign(\frac{\partial||\hat{y}-\hat{X}\beta||^2}{\partial\beta_r})$ and update $F = F \cup r$.
2. **Feature-sign step**:
   2.1: Compute the analytical solution to the resulting unconstrained quadratic problem of equation 3.14. $\hat{\beta}_F = (\hat{X}_F^T\hat{X}_F)^{-1}(\hat{X}_F^T\hat{y} - \lambda\theta_F/2)$.
   2.2: Perform a discrete line search on the closed-line segment from β_F to $\hat{\beta}_F$: Check the objective value at $\hat{\beta}_F$ and all points where any coefficient changes sign; update β_F to the point with the lowest objective value.
   2.3: Remove zero coefficients of β_F from F and update $\theta = sign(\beta)$.
3. **Half-quadratic step**: Update the auxiliary vector p^{t+1} according to equation 3.10.
4. **Check the optimality conditions**:
   a. Optimality condition for nonzero coefficients:
   $\frac{\partial||\hat{y}-\hat{X}\beta||^2}{\partial\beta_j} + \lambda sign(\beta_j) = 0, \forall\beta_j \neq 0$
   If condition a is not satisfied, update $\hat{y}$ and $\hat{X}$ according to p^{t+1} and go to step 2; else check condition b.
   b. Optimality condition for zero coefficients: $\left|\frac{\partial||\hat{y}-\hat{X}\beta||^2}{\partial\beta_j}\right| \leq \lambda, \forall\beta_j = 0$.
   If condition b is not satisfied, update $\hat{y}$ and $\hat{X}$ according to p^{t+1} and go to step 1; otherwise return β as the solution.

The computation cost of algorithm 1 mainly involves three steps: calculation of a working set, feature-sign, and half-quadratic. Let I_1 and C_F denote the number of iterations of algorithm 1 and the maximum number of the nonzero entries of β, respectively. Since algorithm 1 finds a sparse solution, $C_F \ll n$. The cost of calculating matrix $\hat{X}$, $\hat{y}$, and gradient in updating a working set is $o(d \times n + d)$. The calculation of the square matrix $\hat{X}_F^T\hat{X}_F$ and its inverse are $d \times C_F^2$ and C_F^3, respectively. Hence, the cost of the feature-sign step is $o(d \times C_F^2 + C_F^3)$. The cost of the half-quadratic step is $o(d \times n)$. As a result, the computation cost of algorithm 1 is $o(I_1 \times (d \times n + d \times C_F^2 + C_F^3))$. Since $C_F \ll n$, we also have $C_F^2 \ll n$. When n trends to be large, the computation cost of algorithm 1 trends to be $o(I_1 \times d \times n)$.

3.4 Robust Sparse Representation-Based Classification. A major issue in pattern classification is the use of the distance metric. The frequently used

Algorithm 2: Robust Sparse Representation-Based Classification (RSRC).

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

Output: $identity(y)$

1. Normalize the columns of X to have unit l^2-norm.
2. Solve the maximum correntropy problem defined in equation 3.6 and obtain the β^*.
3. Calculate the correntropy-induced metric CIM_c, for $c = 1, \ldots, k$
4. $identity(y) = \arg\min_c CIM_c(y, y_c)$

metrics include Euclidean distance, Mahalanobis distance, cosine distance, and Hamming distance. These distances are defined between two samples in a feature subspace. In many applications, multiple samples are available for a class. Nearest feature classifiers (NFC) (Li & Lu, 1999; Chien & Wu, 2002) and linear representation (LR) (Li, 1998; Naseem et al., 2010) utilize these information (or linear representation) of multiple samples to improve classification performance.

Inspired by NFC and LR, we first compute the β of l^1 regularized correntropy to obtain a sparse linear representation. Note that our criterion in equation 3.6 aims to learn a linear representation of sample y using existing training samples. Ideally, training samples from the same class of y should give the best linear representation (Li, 1998; Wright et al., 2009). For each class c, let $\delta_c : \mathbb{R}^n \to \mathbb{R}^{n_c}$ be a function that selects the coefficients belonging to the cth class. For $\beta \in \mathbb{R}^n$, $\delta_c(\beta) \in \mathbb{R}^{n_c}$ is a vector whose entries are the entries in β corresponding to class c. Utilizing only the coefficients associated with the cth class, one can compute a linear representation of the given sample y as $\hat{y}_c = X_c\delta_c(\beta)$ (Wright et al., 2009). Then y is classified by assigning it to the class that minimizes the correntropy-induced metric between y and $\hat{y}_c$:

$$\min_c CIM_c(y, y_c) = \left(1 - \sum_{j=1}^{d} g(y_j - \hat{y}_{cj})\right)^{\frac{1}{2}}. \tag{3.15}$$

Equation 3.15 actually defines a robust metric between the input sample y and a linear subspace of training samples. Algorithm 2 summarizes the complete classification procedure.

4 Experiments

In this section, the proposed methods are systematically compared with support vector machine (SVM), minimum average correlation energy

(MACE) filter (Kumar et al., 2006), linear regression classification (LRC) (Naseem et al., 2010), reconstructive and discriminative subspace method (LDAonK) (Sanja et al., 2006), and sparse representation-based classification (SRC) (Wright et al., 2009). Recognition rate, ROC curve, and computational costs of compared methods are reported. All methods are implemented in Matlab on an AMD Quad-Core 1.80 GHz Windows XP machines with 2 GB memory.

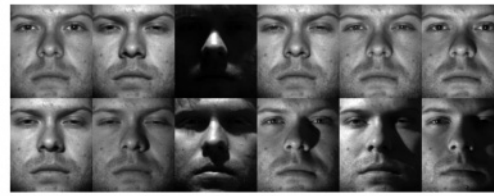
4.1 Experimental Setting and Databases

4.1.1 Databases. Four image databases were selected to evaluate different methods. All the images were converted to grayscale. All facial images were aligned by fixing the locations of two eyes. The four databases are the following:

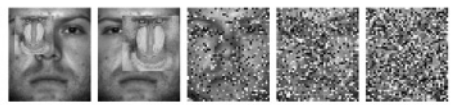
- *COIL database (Nene, Nayar, & Murase, 1996)*: Columbia Object Image Library (COIL-100) is a database of 7200 color images of 100 objects (72 images per object). The objects were placed on a motorized turntable against a black background. The turntable was rotated through 360 degrees to vary object pose with respect to a fixed color camera. Images of the objects were taken at pose intervals of 5 degrees. This corresponds to 72 poses per object. All images are resized to 32×32. Some of them are shown in the first row of Figure 1a.
- *MNIST database* (LeCun, Bottou, Bengio, & Haffner, 1998). The MNIST database of handwritten digits has a training set A of 60,000 examples and a testing set B of 10,000 examples. The digits were centered in a fixed size (28×28) and normalized. Here, we randomly took 100 images for each digit in the testing set A as our training set and test all methods on the whole testing set B.
- *AR database* (Martinez & Benavente, 1998). The AR database consists of over 4000 facial images of 126 subjects (70 men and 56 women). For each subject, 26 facial images were taken in two separate sessions. These images include more facial variations, including various facial expressions, illumination change, and occlusion modes (sunglass and scarf). In the experiment, we selected a subset of the data set consisting of 65 male subjects and 54 female subjects. The images were cropped to size 112×92 pixels. Figure 1b shows 8 cropped images used in the experiments.
- *Extended Yale B database* (Georghiades, Belhumeur, & Kriegman, 2001; Lee, Ho, & Kriegman, 2005). The Extended Yale B database consists of 2414 frontal-face images of 38 subjects (Georghiades et al., 2001) under various lighting conditions. The cropped 192×168 face images were captured under various laboratory-controlled lighting conditions (Lee et al., 2005). Figure 1c shows illustrative faces of the first subject in Yale B database. For each subject, half of the images were randomly

(a) Some of the objects in COIL database and occluded objects.

(b) Cropped facial images of the first subject in AR database.



(c) Cropped facial images of the first subject in Extended Yale B database.



(d) Occluded and corrupted images from Extended Yale B. Images from left to right are: 20% contiguous occlusion, 40% contiguous occlusion, 30% corruption, 50% corruption, and 70% corruption.

Figure 1: Illustrative images used in the experiments.

selected for training (about 32 images per subject), and the other half were used for testing. Randomly choosing the testing set ensures that our experimental results will not depend on the selection of testing data.

4.1.2 Algorithm Setting. The details for the comparisons are as follows:

- *SRC*: We implemented its two models, which differ in robustness and computation strategy. For the first *SRC* algorithm (*SRC1*) the implementation minimizes the l^1-norm in equation 4.1 by a primal-dual algorithm for linear programming based on Boyd and Vandenberghe (2004) and Candes and Romberg (2005):[3]

$$\min \|\beta\|_1 + \|e\|_1 \quad s.t. \quad \|y - X\beta + e\|_2 \leq \varepsilon, \tag{4.1}$$

[3]Matlab source code: http://www.acm.caltech.edu/l1magic/.

where ε is a given nonnegative error tolerance. For the second SRC algorithm ($SRC2$), the implementation minimizes the l^1-norm in equation 4.2 via an active set algorithm based on Lee et al. (2006):[4]

$$\min \|y - X\beta + e\|_2 + \lambda(\|\beta\|_1 + \|e\|_1) \tag{4.2}$$

where λ is a given sparsity penalty, which can be viewed as an inverse of the Lagrange multiplier associated with the constraint in equation 4.1. The models in equation 4.1 and 4.2 are different when there are noise. The ε can be interpreted as a pixel noise level, whereas λ cannot (Wright et al., 2009).

- *RSRC*: We followed the approach of Torre and Black (2003) and Liu et al. (2007) and empirically set σ in algorithm 1 to 0.0005.
- *WLRC*: Since the model in equation 3.5 can be viewed as a weighted LRC from the viewpoint of iteratively reweighted least square (IRLS), we note the model in equation 3.5 as weighted LRC (WLRC). The classification method used in WLRC is the same as that used in LRC.
- *LDAonK*: The parameters of LDAonK follow the suggestion in Leonardis and Bischof (2000) and Sanja et al. (2006).
- *SVM*: The multiclass SVM was implemented by SVM Toolbox.[5] For each training set, the kernel parameter was selected by fivefold cross-validation on each dimension.[6]
- *MACE*: The peak sharpness in MACE was measured by the peak-to-side-lobe ratio (PSR).

In order to estimate the parameter ε of SRC1 and λ of SRC2 and RSRC automatically, fivefold cross-validation on each dimension of data for each training set was used, where the candidate value set for both ε and λ 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 possible. For the same reason, we can report the experimental results of SRCs and SVM only in the lower-dimensional feature space. More experimental results on parameter selection in the regularized correntropy method are shown in section 4.6.

4.2 Recognition under Sunglasses Occlusion. The occluded images came from a subset of the AR Face Database. For training, we used 952 images (about 8 per subject) of unoccluded frontal views with varying facial expression. Figure 1b shows an example of eight selected images of the first

[4]Matlab source code: http://redwood.berkeley.edu/bruno/sparsenet/.

[5]Matlab source code: http://asi.insa-rouen.fr/enseignants/~arakotom/toolbox/index.html

[6]Although SVM can achieve very high accuracy on a training set—for example, it achieves 100% recognition rate on the AR training set—its recognition rate is quite low on the testing set due to occlusions and corruptions.

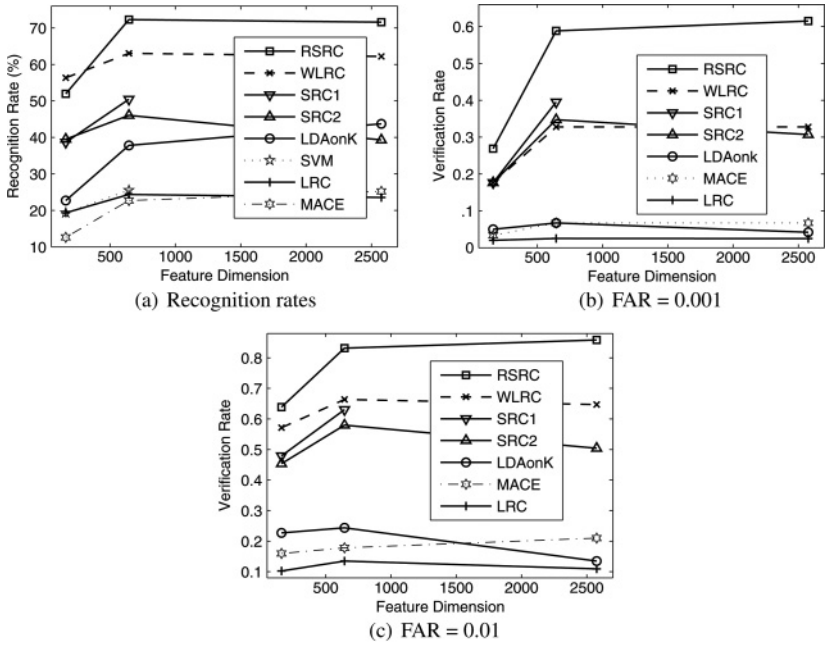


Figure 2: Recognition rates and verification rates for various feature spaces and classifiers under occlusion of sunglasses.

subject. For testing, we used images of the subjects wearing sunglasses, which occlude roughly 20% of facial images.

To quantitatively evaluate different methods, we compute the recognition rates respective to the feature (downsampled image) space dimensions 161, 644, and 2576, respectively.[7] Those numbers correspond to downsampling ratios of 1/8, 1/4, and 1/2, respectively. Figure 2a shows the recognition performance for the various features, in conjunction with eight classifiers: RSRC, WLRC, SRC1, SRC2, LRC, LDAonK, SVM, and MACE. Since the computational cost of SRC1 and SVM is very high in high-dimensional space, we report the experimental results only in lower-dimensional space.

On the training set, SVM and MACE achieve a 100% and 96.43% recognition rate, respectively, on dimension 644. However, when there is occlusion, it is unlikely that the test image will be very close to any single image or linear subspace in the training set, and thus LRC, SVM, and MACE perform

[7]In face recognition (Wright et al., 2009; Naseem et al., 2010), the experimental results show that using simply downsampled images can also achieve superior results compared to the benchmark techniques (Naseem et al., 2010).

Table 1: Recognition Rates (%) of Different Methods.

Database	SVM	MACE	LRC	LDAonK	SRC2	WLRC	RSRC
MNIST	**91.38**	41.50	90.80	-	90.60	82.03	88.20
COIL	37.82	20.58	22.45	35.60	61.05	**87.01**	78.62

Note: The numbers in bold are the highest recognition rates for each configuration.

poorly. Due to the robustness of correntropy, two correntropy methods achieve the best results. Correntropy-based methods can accurately model noise and then use uncorrupted pixels to achieve the highest recognition rates.

The ROC curve is also illustrated in Figure 2. This curve is an important tool to evaluate different algorithms. It can be represented equivalently by plotting the fraction of the false acceptance rate (FAR) versus. the verification rate (VR) and is often used to measure the accuracy of outlier rejection. A good algorithm should achieve high verification rates even at very low false acceptance rates. One often concerns verification rates when false acceptance rates are 0.001 and 0.01 (Yi, Liu, Chu, Lei, & Li, 2007).

Figures 2c and 2d show the verification rates when FAR $= 0.001$ and FAR $= 0.01$, respectively.[8] The verification rates of LDAonK, LRC, SVM, and MACE are quite low. Although the recognition rates of WLRC are higher than those of SRC1, verification rates of WLRC are lower than those of SRC1 when FAR $= 0.001$. It seems that sparse representation-based methods can achieve higher ROC curves. RSRC achieves the highest verification rates, performing significantly better than the two SRC methods.

4.3 Recognition under Contiguous Occlusion. In this section we simulate various types of contiguous occlusions by replacing a randomly selected local region in each testing image with an unrelated image (Wright et al., 2009) and a white square (Sanja et al., 2006).

In the first experiment, we evaluate different methods on the MNIST data set without occlusions. Table 1 shows the recognition rates of different methods. In terms of recognition performance, all compared methods can be ordered as MACE, WLRC, RSRC, SRC2, LRC, and SVM, where MACE performs the worst and SVM performs the best. Since there are no occlusions, SVM and LRC can better use the linear structure of training samples (Li, 1998; Li & Lu, 1999; He et al., 2008; Chien & Wu, 2002) so that they outperform other methods. These experimental results are also coincidence with those reported in (Wright et al., 2009; Naseem et al., 2010). The two correntropy-based methods do not perform better than other methods, and

[8]Since the SVM Toolbox outputs only the label information instead of continuous scores, we cannot show the ROC curves of multiclass SVM.

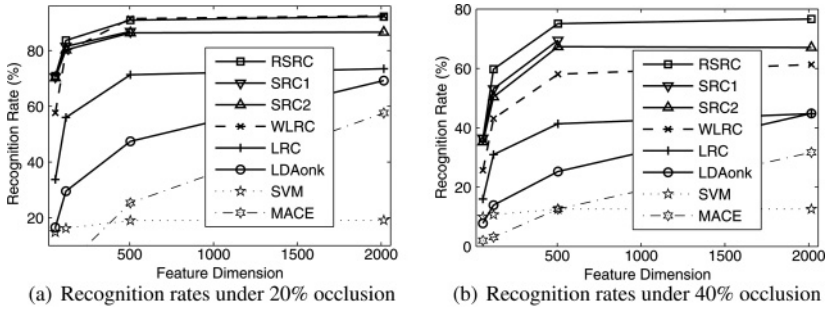


Figure 3: Recognition rates under different levels of contiguous occlusions.

it may be because the value of kennel size is still not optimally tuned for all experiments. For large kernel sizes, correntropy defaults to MSE (Liu et al., 2007).

In the second experiment, a subset of COIL database was used to evaluate different methods. In the testing set, each image was occluded with a white square (Sanja et al., 2006). Recognition rates in Table 1 are used to illustrate the discriminative ability. Both WLRC and RSRC significantly outperform other methods due to the accurate detection of the occluded region. WLRC seems to detect the occluded region more accurately than RSRC, so that WLRC achieves the highest recognition rate.

In the third experiment, we quantitatively evaluate different methods on the Extended Yale B Face Database (Wright et al., 2009). For each subject, half of the images were randomly selected for training (about 32 images per subject) and the other half for testing. Figure 1d shows two occluded images under 20% and 40% occlusions. The training set and testing set contain 1205 and 1209 images, respectively. We compute the recognition rates with the feature (downsampled image) space dimensions 56, 120, 504, and 2016. Those numbers correspond to downsampling ratios of 1/24, 1/16, 1/8, and 1/4, respectively.

Figure 3 shows the recognition rates under different levels of contiguous occlusions. When occlusion is 20%, both WLRC and RSRC achieve the highest recognition rates. When occlusion is 40%, the sparse representation-based methods can perform better than the three nonsparse representation methods. Still, RSRC obtains the best results. Since the pixel values of unrelated monkey images are close to those of facial images, LDAonK fails to detect the occluded region so that it performs worse than other methods.

4.4 Recognition under Random Pixel Corruption. In some practical scenarios, the test image y may be partially corrupted. We evaluate the robustness on the Extended Yale B Face Database. For each subject, half of the images are randomly selected for training (about 32 images per subject)

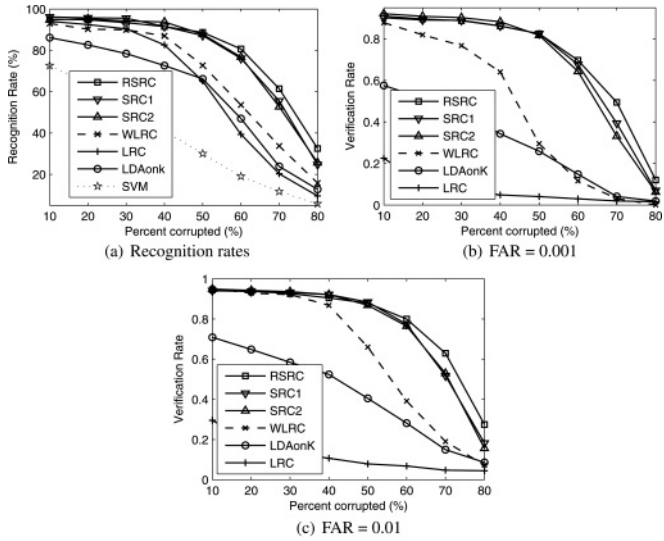


Figure 4: Recognition rates and verification rates for various feature spaces and classifiers under varying level of random corruption.

and the other half for testing. The training set and testing set contain 1205 and 1209 images, respectively. All the images are resized to 24×21, which is stacked in a 504-D vector. Each image in the testing set is corrupted by replacing a percentage of randomly chosen pixels with i.i.d. samples from a uniform distribution (uniform over [0, 255]). The corrupted pixels are randomly chosen for each test image, and the locations are unknown to the computer. We vary the percentage of corrupted pixels from 10% to 80%. Figure 1d shows three examples to those corruptions. To the human eye, beyond 50% corruption, the corrupted images are barely recognizable as face images.

Figure 4a shows the recognition accuracy of different methods, as a function of the level of corruption. [9] We see that three sparse representation-based methods perform better than other methods. In particular, the RSRC achieves the highest recognition rates when corruption is larger than 50%. Figures 4b and 4c further show the verification rates under 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%). RSRC, SRC1, and SRC2 can significantly improve the ROC curves due to the sparse representation. When

[9]Since the recognition rates of SVM and MACE are significantly lower than the other methods in this experiment, we do not report their results here

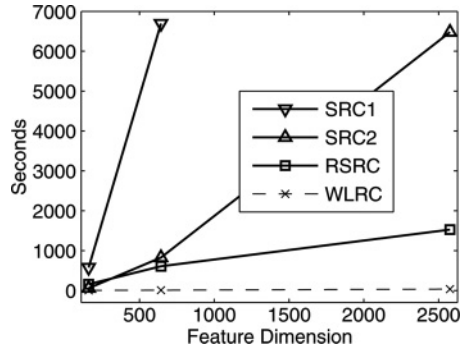


Figure 5: Computation time for various feature spaces and classifiers.

the corruption is larger than 50%, the verification rates of RSRC become higher than those of the other methods. Thus, RSRC could be an effective method to deal with severe corruption for robust recognition.

4.5 Comparison of the Computational Complexity. Computational complexity of an algorithm is often a major concern in computer vision. Figure 5 shows the overall computation time of using various features on the AR database, with the same experimental setting in section 4.2. The computation costs of two SRC methods become extremely high as the dimension is increased. When the feature dimension is 644-D, SRC1, SRC2, RSRC, and WLRC take 56, 6.9, 5.2, and 0.13 seconds per test image, respectively. (In Wright et al., 2009, SRC1 requires about 75 seconds per test image on a PowerMac G5.) The computational advantage of WLRC is clearly seen over these three sparse methods. When dimension is high, the computational costs of two SRC methods are much larger than RSRC, because RSRC estimates only n variables for sparse representation and d auxiliary variables can be optimized by HQ, whereas the SRC models have to estimate all $d + n$ variables. Hence, WLRC and RSRC are more suitable for robust and real-time pattern recognition tasks.

4.6 Parameter Selection. In real-world scenarios, corruptions are often unknown beforehand. Hence, a cross-validated parameter on an uncorrupted training set may be unsuitable for corrupted testing data. For this reason, we follow the suggestion in Liu et al. (2007) to investigate parameter selection for our method. The experimental setting is the same as that in section 4.4. All experiments are averaged over 20 random corruptions. In each random corruption, corrupted pixels are randomly chosen for each test image in the testing set.

Table 2: Average Recognition Rates (%) ± Standard Deviation under Different Values of λ in RSRC.

λ	0.0005	0.001	0.005	0.01	0.05
10% corruption	96.48 ± 0.4	**96.76 ± 0.3**	95.28 ± 0.4	92.94 ± 0.5	83.28 ± 0.9
80% corruption	8.45 ± 1.1	10.76 ± 0.7	30.64 ± 0.6	**35.04 ± 0.8**	6.56 ± 1.2

Note: The numbers in bold are the highest recognition rates for each configuration.

Table 3: Average Recognition Rates (%) ± Standard Deviation for Different Kernel Sizes in RSRC.

θ	0.00010	0.00025	0.00050	0.00075	0.001	Silverman
10%	90.50 ± 0.9	93.89 ± 0.6	93.89 ± 0.8	**94.80** ± 0.4	94.63 ± 0.6	91.26 ± 0.7
80%	15.63 ± 1.8	**45.08** ± 1.4	28.37 ± 1.6	15.47 ± 1.3	11.66 ± 1.9	22.52 ± 1.7

Note: The numbers in bold are the highest recognition rates for each configuration.

4.6.1 Regularization Parameter. The regularization parameter λ in equation 3.6 is an important parameter to control the sparseness of the representation. To our best knowledge, it is still hard to find a method in the literature to determine an approximate value of λ for arbitrary corruptions. Alternatively, this section studies how λ affects recognition rates with respect to different levels of corruption.

Table 2 tabulates recognition rates of RSRC when different values of λ are set. Since the value of λ is related to sparse representation, its selection will affect the recognition performance of RSRC. When corruption is 10% and λ equals 0.001, RSRC achieves the highest recognition rate: 96.76%. When corruption is 80% and λ equals 0.01, RSRC achieves the highest recognition rate: 35.04%. These results further demonstrate that the robust sparse code induced by l^1 regularization is a powerful tool for robust recognition. It is interesting to observe that a small value of λ may be more suitable for small corruption and a large value of λ may make RSRC control the large corruption better.

4.6.2 Kernel Size σ. The kernel size σ is an important parameter that controls all robust properties of correntropy (Liu et al., 2007). An appropriate value of kernel size can effectively eliminate the effect of outliers and noise. Its selection is still an open issue in ITL (Liu, Pokharel, & Principe, 2006; Liu et al., 2007) and kernel density estimation. Considering the robust parameter estimation, we further investigate Silvermans' rule (Silverman, 1986) as suggested in correntropy (Liu et al., 2007).

The simulations in Table 3 were run to show how kernel size affects the performance of regularized correntropy methods under various corruptions. The experimental setting is the same as that in section 4.4. On 10%

corruption, the larger the θ is, the higher the recognition rate the algorithm will obtain. On 80% corruption, RSRC achieves the highest recognition rate, 45.08%, when $\theta = 0.00025$. The robust parameter estimation method (Silverman's rule) could not improve the recognition rates of RSRC, especially when the corruption is large.

When the percentage of corruption or occlusion is smaller than 50%, θ can be set to a larger value to give most of the uncorrupted pixels larger weight; when the corruption or occlusion is larger than 50%, θ can be set to a smaller value to dampen outliers significantly. If the error incurred by noise is unknown, we can simply choose a conservative way to set θ to a median value (e.g., 0.0005). Experimental results validate that this conservative choice can make regularized correntropy methods outperform other methods in most scenarios.

4.7 Discussion. From these experimental results, we have these observations:

- Correntropy-based methods versus other robust algorithms: Correntropy-based methods obtain encouraging performance improvements as compared to other robust algorithms. The RSRC always outperforms other robust methods under occlusion and noisy corruption. Correntropy-based methods can detect outliers more accurately so that they can utilize the uncorrupted subset of image pixels to perform classification. Experimental results further validate that correntropy is a powerful tool for handling nongaussian noise and large outliers.
- Nonsparse method (WLRC) versus sparse method (RSRC) based on MCC: Experimental results show that the nonsparse method WLRC can obtain better or similar classification results as compared to the sparse method RSRC. However, the recognition performance of WLRC drops significantly when noise increases (see Figures 3 and 4). The sparse methods (including SRC1, SRC2, and RSRC) seem to tackle large noise better. WLRC is advantageous due to its efficient computation. Hence, we recommend using WLRC for real-time recognition tasks under moderate occlusions and corruptions.
- ROC curves: Compared with LRC and LDAonK, WLRC can improve ROC curves. But WLRC's improvement is limited compared to the three sparse methods. The scores (residuals) with respect to the coefficients computed by the l^1 norm are still sparse, which seems to favor higher ROC curves.
- Nonlinear template matcher: Recent experimental results show that a nonlinear template matcher can further improve classification accuracy in both the real domain (Jia & Martinez, 2008, 2009) and the frequency domain (Jeong & Principe, 2008; Jeong et al., 2009). Taking advantage of the linear structure of the reproducing kernel

Hilbert space, the correntropy MACE (CMACE) (Jeong et al., 2009) can potentially improve MACE performance while preserving the shift-invariant property (Jeong et al., 2009). Algorithm 2 also develops a correntropy-based nonlinear classifier for classification.

- Parameter selection: Experimental results show that correntropy is an efficient tool to control large corruption (70%), while the kernel size controls its robustness. If the parameters are well tuned as in Table 3, recognition rate of RSRC can reach 45.08% at 80% corruption. Traditional parameter selection methods for Huber M-estimators, such as Silverman's rule (Silverman, 1986; Liu et al., 2007), are often based on a median so that they can efficiently deal only with small corruptions (smaller than 50%). How to automatically determine an appropriate kernel parameter of correntropy is still an open problem for large corruptions and occlusions.

5 Conclusion

We have presented a novel regularized correntropy framework for robust pattern recognition. The l^1 regularization scheme is imposed on correntropy. The half-quadratic optimization technique is used to solve the nonlinear correntropy optimization problem. Finally, a novel correntropy-based classification method is developed to favor efficient detection of outliers and robust recognition. Unlike the state-of-the-art SRC, which assumes that the noisy item has a sparse representation, our method is based on correntropy, which is much more insensitive to outliers. The experimental results validate that the l^1 regularization can further improve recognition accuracy and receiver operator characteristic curves under corruption and occlusion.

Acknowledgments

Thanks to Christian Ocier for proofreading this manuscript. We also greatly thank the associate editor and the reviewers for their valuable comments and advice. This work was supported in part by the Research Foundation for the Doctoral Program of the Ministry of Education of China (no. 20100041120009), the Natural Science of Foundation of China (nos. 61075051 60971095), the NSFC-GuangDong (no. U0835005), and 985 Project in Sun Yat-sen University (no. 35000-3181305).

References

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning, 2*(1), 1–127.

Black, M., & Jepson, A. (1998). Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision, 26*(1), 63–84.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Candes, E., & Romberg, J. (2005). l1-magic: Recovery of sparse signals via convex programming. http://www.acm.caltech.edu/l1magic/.

Candes, E., & Tao, T. (2005). Decoding by linear programming. *IEEE Trans. on Information Theory, 51*(12), 4203–4215.

Chien, J., & Wu, C. (2002). Discriminant waveletfaces and nearest feature classifiers for face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 24*(12), 1644–1649.

Cover, T., & Thomas, J. (2005). *Elements of information theory* (2nd ed.). Hoboken, NJ: Wiley.

Croux, C., & Dehon, C. (2001). Robust linear discrimination analysis using s-estimators. *Canadian J. Statistics, 29*, 473–492.

Ding, C., Zhou, D., He, X., & Zha, H. (2006). R1-PCA: Rotational invariant l1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd International Conference on Machine Learning*. New York: ACM.

Donoho, D. (2006a). Compressed sensing. *IEEE Trans. on Information Theory, 52*(4), 1289–1306.

Donoho, D. (2006b). For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics, 59*(6), 797–829.

Georghiades, A., Belhumeur, P., & Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 23*(6), 643–660.

Grandvalet, Y., & Bengio, Y. (2006). *Entropy regularization*. http://www.iro.umontreal.ca/∼lisa/pointeurs/entropy_regularization_2006.pdf.

Hawkins, D., & McLachlan, G. (1997). High-breakdown linear discriminant analysis. *J. Am. Statistical Assoc., 92*, 136–143.

He, R., Ao, M., Xiang, S., & Li, S. (2008). Nearest feature line: A tangent approximation. In *Proceedings of the Chinese Conference on Pattern Recognition*. San Mateo, CA: IEEE Computer Society.

He, R., Hu, B., & Yuan, X. (2009). Robust discriminant analysis based on nonparametric maximum entropy. In *Proceedings of the 1st Asian Conference on Machine Learning*. Berlin: Springer.

He, R., Hu, B., Yuan, X., & Zheng, W. (2010). Principal component analysis based on nonparametric maximum entropy. *Neurocomputing, 73*, 1840–1852.

He, X., & Fung, W. (2000). High breakdown estimation for multiple populations with applications to discriminant analysis. *Multivariate Analysis, 72*(2), 151–162.

Hester, C., & Casasent, D. (1980). Multivariant technique for multiclass pattern recognition. *Appl. Opt., 19*, 1758–1761.

Huber, P. (1981). *Robust statistics*. Hoboken, NJ: Wiley.

Hubert, M., & Driessen, K. V. (2003). Fast and robust discriminant analysis. *Computational Statistics and Data Analysis, 45*, 301–320.

Jeong, K., Liu, W., Han, S., Hasanbelliu, E., & Principe, J. (2009). The correntropy MACE filter. *Pattern Recognition, 42*(5), 871–885.

Jeong, K., & Principe, J. (2008). Enhancing the correntropy MACE filter with random projections. *Neurocomputing, 72*, 102–111.

Jia, H., & Martinez, A. (2008). Face recognition with occlusions in the training and testing sets. In *Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition.* Piscataway, NJ: IEEE.

Jia, H., & Martinez, A. (2009). Support vector machines in face recognition with occlusions. In *Proceedings of the 8th 2009 IEEE Conference on Computer Vision and Pattern Recognition.* San Mateo, CA: IEEE Computer Society.

Kumar, B. (1986). Minimum variance synthetic discriminant functions. *J. Opt. Soc. Am., A3*(10), 1579–1584.

Kumar, B. V. (1992). Tutorial survey of composite filter designs for optical correlators. *Appl. Opt., 31*, 4773–4801.

Kumar, B. V., Savvides, M., & Xie, C. (2006). Correlation pattern recognition for face recognition. *Proc. IEEE, 94*(11), 1963–1976.

Kwak, N. (2008). Principal component analysis based on l1-norm maximization. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 30*(9), 1672–1677.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.

Lee, H., Battle, A., Raina, R., & Ng, A. (2006). Efficient sparse coding algorithms. In B. Schólkopf, J. Platt, & T. Hoffman (Eds.), *Neural information processing systems, 19* (pp. 801–808). Cambridge, MA: MIT Press.

Lee, K. C., Ho, J., & Kriegman, D. (2005). Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 27*(5), 684–698.

Leonardis, A., & Bischof, H. (2000). Robust recognition using eigenimages. *Computer Vision and Image Understanding, 78*(1), 99–118.

Li, S. (1998). Face recognition based on nearest linear combinations. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, (pp. 839–844). San Mateo, CA: IEEE Computer Society Press.

Li, S., & Lu, J. (1999). Face recognition using nearest feature line method. *IEEE Trans. Neural Network, 10*(2), 439–443.

Liu, W., Pokharel, P. P., & Principe, J. C. (2006). Correntropy: A localized similarity measure. In *Proceedings of the International Joint Conference on Neural Networks*. San Mateo, CA: IEEE, Computer Society Press.

Liu, W., Pokharel, P., & Principe, J. (2007). Correntropy: Properties and applications in nongaussian signal processing. *IEEE Trans. on Signal Processing, 55*(11), 5286–5298.

Mahalanobis, A., Kumar, B., & Casasent, D. (1987). Minimum average correlation energy filters. *Appl. Opt., 26*(17), 3633–3640.

Mairal, J., Elad, M., & Sapiro, G. (2008). Sparse representation for color image restoration. *IEEE Trans. on Image Processing, 17*(1), 53–69.

Mangasarian, O., & Musicant, D. (2000). Robust linear and support vector regression. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 22*(9), 950–955.

Martinez, A. (2002). Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 24*(6), 748–763.

Martinez, A., & Benavente, R. (1998). *The AR face database* (Tech. Rep.). Barcelona: Computer Vision Center, Autonomous University of Barcelona.

Meer, P., Stewart, C., & Tyler, D. (2000) Robust computer vision: An interdisciplinary challenge. *Computer Vision and Image Understanding, 78*, 1–7.

Naseem, I., Togneri, R., & Bennamoun, M. (2010). Linear regression for face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 32*, 2106–2112.

Nene, S., Nayar, S., & Murase, H. (1996). *Columbia Object Image Library (COIL-20)* (Tech. Rep. CUCS-005-96). New York: Columbia University.

Ohba, K., & Ikeuchi, K. (1997). Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 9*, 1043–1047.

Pentland, A., Moghaddam, B., & Straner, T. (1994). *View-based and modular eigenspaces for face recognition* (Tech. Rep. 245). Cambridge, MA: MIT Media Lab.

Pokharel, P., Liu, W., & Principe, J. (2009). A low complexity robust detector in impulsive noise. *Signal Processing, 89*(10), 1902–1909.

Principe, J. C. (2010). *Information theoretic learning: Renyi's entropy and kernel perspectives*. New York: Springer.

Principe, J. C., Xu, D. X., & Fisher, J. W. (2000). Information-theoretic learning. In S. Haykin (Ed.), *Unsupervised adaptive filtering* (pp. 265–319). Hoboken, NJ: Wiley.

Principe, J., Xu, D., Zhao, Q., & Fisher, J. (2000). Learning from examples with information theoretic criteria. *VLSI Signal Processing Systems, 26*, 61–77.

Refregier, P., & Figue, J. (1991). Optimal trade-off filter for pattern recognition and their comparison with Weiner approach. *Opt. Comput. Process., 1*, 3–10.

Sanja, F., Skocaj, D., & Leonardis, A. (2006). Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 28*(3), 337–350.

Santamaria, I., Pokharel, P. P., & Principe, J. C. (2006). Generalized correlation function: Definition, properties, and application to blind equalization. *IEEE Trans. on Signal Processing, 54*(6), 2187–2197.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B, 58*(1), 267–288.

Torkkola, K. (2003). Feature extraction by nonparametric mutual information maximization. *Journal of Machine Learning Research, 3*, 1415–1438.

Torre, F., & Black, M. (2003). A framework for robust subspace learning. *International Journal of Computer Vision, 54*(1–3), 117–142.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Viola, P., Schraudolph, N., & Sejnowski, T. (1995). Empirical entropy manipulation for real-world problems. In M. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Neural information processing systems, 9* (pp. 851–857). Cambridge, MA: MIT Press.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T. S., & Yan, S. (2010). Sparse representation for computer vision and pattern recognition. *Proceedings of IEEE, 98*(6), 1031–1044.

Wright, J., Yang, A., Ganesh, A., Sastry, S., & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 31*(2), 210–227.

Xu, D. (1999). *Energy, entropy and information potential for neural computation*. Unpublished doctoral dissertation, University of Florida.

Yang, S., Zha, H., Zhou, S., & Hu, B. (2009). Variational graph embedding for globally and locally consistent feature extraction. In *Proceedings of the Europe Conference on Machine Learning* (pp. 538–553). Berlin: Springer.

Yi, D., Liu, R., Chu, R., Lei, Z., & Li, S. (2007). Face matching from near infrared to visual images. In S. Li & S. W. Lee (Eds.), *Proceedings of IAPR/IEEE International Conference on Biometrics* (pp. 523–530). Berlin: Springer.

Yuan, X., & Hu, B. (2009). Robust feature extraction via information theoretic learning. In *Proceedings of the International Conference on Machine Learning*. New York: ACM.

Received January 13, 2010; accepted January 20, 2011.