ROBUST SPARSE LEARNING BASED ON KERNEL NON-SECOND ORDER MINIMIZATION

Miaohua Zhang[1]*, Yongsheng Gao*[1]*, Changming Sun*[2]*, and Michael Blumenstein*[3]

[1] School of Engineering, Griffith University, QLD, Australia.
[2] CSIRO Data61, Marsfield, NSW, Australia.
[3] Faculty of Engineering & Information Technology, University of Technology Sydney, NSW, Australia.

ABSTRACT

Partial occlusions in face images pose a great problem for most face recognition algorithms due to the fact that most of these algorithms mainly focus on solving a second order loss function, e.g., mean square error (MSE), which will magnify the effect from occlusion parts. In this paper, we proposed a kernel non-second order loss function for sparse representation (KNS-SR) to recognize or restore partially occluded facial images, which both take the advantages of the correntropy and the non-second order statistics measurement. The resulted framework is more accurate than the MSE-based ones in locating and eliminating outliers information. Experimental results from image reconstruction and recognition tasks on publicly available databases show that the proposed method achieves better performances compared with existing methods.

Index Terms— Sparse representation, kernel non-second order measurement, correntropy, sparse recovery.

1. INTRODUCTION

Face recognition has become an important research direction in the pattern recognition field in recent years due to their wide uses in real-world applications, such as security systems, video surveillance, and pedestrian tracking [1] [2] [3]. The effectiveness of a face recognition technology largely depends on how to learn discriminative features from given data samples, which plays an important role in enhancing recognition precision [4] [5]. Another challenge is to solve occlusion and corruption problems, namely weakening the influence from these artificial defects when extracting features [6] [7]. Obtaining discriminative features under these difficult environments makes face recognition even more challenging.

In past years, sparse representation has shown great potential in tackling computer vision problems, such as image denoising, image restoration, and image classification [8] [9]. Wright et al. [10] proposed a sparse representation classifier (SRC) and boosted the research of sparse representation of face recognition. It can better solve the occlusion and corruption problems in comparison with existing face recognition methods. However, SRC still cannot solve the contiguous occlusion for face recognition [11]. The computational cost of SRC is also expensive. Recently, information theoretic learning (ITL) [12] has shown its superiority in robust learning and classification. For example, He et al. [11] proposed a robust face recognition method based on maximum correntropy criterion (CESR) which is much less sensitive to outliers.

However, the loss function in [11] is based on the second order statistical measure in the kernel space, which is still sensitive to outliers. The non-second order loss function [13] [14] [15] has been frequently used in the learning system to learn good features due to the fact that it is more efficient than the second order statistics based methods in handling non-Gaussian noise or large outliers in the training data. In this work, we proposed to learn a robust sparse representation based on a kernel non-second order minimization (KNS-SR). The KNS-SR algorithm uses the non-second order correntropy loss function and can efficiently cope with contiguous occlusions in real world applications. To optimize the objective function, we transform the loss function into a reweighted least squared problem and effectively solve it by the active set algorithm [16]. A new classifier based on the non-second order kernel measurement is also developed to minimize the effect from outliers for classification.

2. RELATED WORKS

Given training data $D = [D_1, \dots D_c] \in \mathcal{R}^{m \times n}$ from c different classes, and the new test samples $Y = [y_1, \dots, y_N] \in \mathcal{R}^{m \times N}$. Normally any test sample $y \in \mathcal{R}^m$ can be approximately represented by a linear combination of given training samples from the same class or all the training samples:

$$y \approx D_i \alpha_i = D\alpha \tag{1}$$

where α_i and α correspond to the sparse coefficients of class i and all the classes.

He et al. [11] (CESR) took the advantage of the correntropy in handling non-Gaussian noise and large outliers and proposed a maximum correntropy criterion based robust

sparse face recognition method for face recognition:

$$J_{\text{CESR}} = \max_{\alpha} \; g\left(y_j - \sum_{i=1}^{n} D_{i,j}\alpha_i\right) - \lambda||\alpha||_1, \quad (2)$$

where $g(x) = \exp(-\frac{x^2}{2\sigma^2})$. The correntropy based classifier is given by: $\text{label}_y = \underset{c_i}{\text{argmax}}\, g(y - D\delta_{c_i}(\alpha))$.

3. PROPOSED FRAMEWORK

3.1. KNS-loss

Correntropy is a second order statistical measure in the kernel space, which is a local measurement between two random variables A and B [15], defined by $V(A,B) = E[\langle\varphi(A), \varphi(B)\rangle_{\mathcal{H}}]$, where $E[\cdot]$ denotes the expectation operator. $\varphi(a)$ is a nonlinear mapping function and transforms a from the original space to the Hilbert space, satisfying $\langle\varphi(a), \varphi(b)\rangle_{\mathcal{H}} = k(a,b)$. In this paper, we use the Gaussian function as the kernel function: $k(a,b) = \exp\left(-\frac{(a-b)^2}{2\sigma^2}\right)$, where σ is the kernel bandwidth. Motivated by the non-second order statistic measure having the advantages in improving the ability of eliminating the information from outliers, in this work we use the kernel non-second order correntropy loss (KNS-loss):

$$\begin{aligned} J_{\text{KNS-loss}}(A,B) &= 2^{-p/2}E[\| \varphi(A) - \varphi(B) \|_{\mathcal{H}}^{p}] \\ &= E[(1 - k_\sigma(A-B))^{p/2}], \end{aligned} \quad (3)$$

where $p > 0$ is the power parameter. Obviously, the above equation includes the case for the c-loss function when p is 2.

3.2. KNS-SR

Given a training data set $D = [D_1, \ldots, D_c] = [d_1, \ldots, d_n] \in \mathcal{R}^{m\times n}$ and a testing sample $A = (y_{i1}, y_{i2} \ldots, y_{im})^T \in \mathcal{R}^{m\times 1}$ transformed from the i-th facial image, the testing image then can be approximately represented by a linear combination of training samples in D. Let B be this representation, namely $B = (\sum_i d_{i1}\alpha_i, \ldots, d_{im}\alpha_i)$, the sparse representation under the KNS-loss is given by:

$$\begin{aligned} J_{\text{KNS-loss}}(\alpha) &= \frac{1}{m}\sum_{j=1}^{m}(1 - k_\sigma(y_j - \sum_{i=1}^{n} d_{ji}\alpha_i))^{\frac{p}{2}} + \lambda||\alpha||_1 \\ &= \frac{1}{m}\sum_{j=1}^{m}\rho(||e_j||_2) + \lambda||\alpha||_1, \end{aligned} \quad (4)$$

where $e_j = (y_j - \sum_{i=1}^{n} d_{ji}\alpha_i)$, and $\rho(||e_j||_2) = (1 - \exp(||e_j||_2^2/2\sigma^2))^{p/2}$ belongs to the M-estimation type of robust cost function. Based on the theory of M-estimation, the problem in (4) can be transformed into a weighted least square problem and can be solved by the iteratively reweighted least square algorithm which has already been successfully applied in computer vision and face recognition [17].

3.3. Optimization Algorithm for KNS-SR

According to the general framework of M-estimator, the first term $(\rho(||e_j||_2))$ in (4) will be equivalent to the following weighted least square problem [12]:

$$\min\sum_{j=1}^{m}\rho(||e_j||_2) = \min\sum_{j=1}^{m}\gamma(||e_j||_2)||e_j||_2^2, \quad (5)$$

where $e_j = y_j - \sum_{i=1}^{n} d_{ji}\alpha_i$ and the weight function $\gamma(||e_j||_2)$ is defined by:

$$\gamma(||e_j||_2) = \rho'(||e_j||_2)/||e_j||_2, \quad (6)$$

in which $\rho'(||e_j||_2)$ is the derivative of $\rho(||e_j||_2)$. Therefore we have

$$\gamma(||e_j||_2) = \frac{p}{2\sigma^2}\left[1 - \exp\left(-\frac{||e_j||_2^2}{2\sigma^2}\right)\right]^{\frac{p}{2}-1}\exp\left(-\frac{||e_j||_2^2}{2\sigma^2}\right), \quad (7)$$

which means that a large error gets larger attenuation and the learned subspace will have little influence from outliers. Based on (5)-(7), the loss function in (4) can be rewritten in a weighted least square problem as:

$$\underset{\alpha}{\text{argmin}} \quad (y - D\alpha)^T\Gamma_{jj}(y - D\alpha) + \lambda\sum_i \alpha_i, \text{ s.t. } \alpha_i > 0, \quad (8)$$

where Γ_{jj} is a diagonal matrix with each element being calculated from (6). After some algebraic operations, the optimization problem in (8) can be rewritten in a function of α:

$$J(\alpha) = \min\ (\frac{\lambda}{2} - \hat{y}^T\hat{D})\alpha + \alpha^T\hat{D}^T\hat{D}\alpha, \quad \text{s.t. } \alpha_i > 0, \quad (9)$$

where $\hat{D} = \text{diag}(\sqrt{-\gamma_j})D$ and $\hat{y} = \text{diag}(\sqrt{-\gamma_j})y$. Since $\hat{D}^T\hat{D}$ is a positive semidefinite matrix, this quadratic problem in (10) is convex and is actually a monotone linear complementary problem (LCP) [16]. Based on the Karush-Kuhn-Tucker optimal conditions, the standard LCP framework of (9) is derived as follows [18]:

$$\begin{aligned} &\text{LCP}(\hat{D}, \hat{y}): \ \text{find } (\omega, \alpha) \\ &\text{s.t. } \omega = \nabla J(\alpha) = \hat{D}^T\hat{D}\alpha - \hat{D}^T\hat{y} + \frac{\lambda}{2} \\ &\quad \alpha^T\omega = 0, \quad \omega \geq 0, \ \alpha \geq 0. \end{aligned} \quad (10)$$

We propose to use the active set algorithm to solve the LCP [11] [16]. The convex nonnegative least squares problem works with complementary solutions, namely those vectors $[\alpha; \omega]$ verifying $\omega = \hat{D}^T\hat{D}\alpha - \hat{D}^T\hat{y} + \frac{\lambda}{2}$ and $\alpha^T\omega = 0$. Let us denote the working set by $\mathcal{A} \subset \{1, \ldots, n\}$ and its complement by $\mathcal{B} = (1:n) \setminus \mathcal{A}$, partitioning the matrices, vectors accordingly as follows:

$$\begin{aligned} &\hat{D} = [\hat{D}_{\mathcal{A}} \in \mathcal{R}^{m\times|A|}, \hat{D}_{\mathcal{B}} \in \mathcal{R}^{m\times|B|}], \\ &y = [y_{\mathcal{A}} \in \mathcal{R}^{|A|}, y_{\mathcal{B}} \in \mathcal{R}^{|B|}], \\ &\omega = [\omega_{\mathcal{A}} \in \mathcal{R}^{|A|}, \omega_{\mathcal{B}} \in \mathcal{R}^{|B|}], \end{aligned} \quad (11)$$

where $|A|$ and $|B|$ are the number of elements in $\mathcal{A}$ and $\mathcal{B}$. Then ω in (11) can be rewritten in another form as

$$\begin{bmatrix}\omega_{\mathcal{A}}\\ \omega_{\mathcal{B}}\end{bmatrix} = \begin{bmatrix}\hat{D}_{\mathcal{A}}^T\hat{D}_{\mathcal{A}} & \hat{D}_{\mathcal{A}}^T\hat{D}_{\mathcal{B}}\\ \hat{D}_{\mathcal{B}}^T\hat{D}_{\mathcal{A}} & \hat{D}_{\mathcal{B}}^T\hat{D}_{\mathcal{B}}\end{bmatrix}\begin{bmatrix}\alpha_{\mathcal{A}}\\ \alpha_{\mathcal{B}}\end{bmatrix} - \begin{bmatrix}\hat{D}_{\mathcal{A}}^T\hat{y}\\ \hat{D}_{\mathcal{B}}^T\hat{y}\end{bmatrix} + \frac{\lambda}{2}. \quad (12)$$

The optimal solution $\alpha_{\mathcal{A}}$ and $\omega_{\mathcal{B}}$ can be obtained from the following iterative procedure [11] [16]:

$$\begin{aligned}\min_{\alpha_{\mathcal{A}}\in\Re^{|\mathcal{A}|}} ||\hat{D}_{\mathcal{A}}\alpha_{\mathcal{A}} - \hat{y}||_2^2 + \lambda\sum_{i\in\mathcal{A}}\alpha_i\\ \omega_{\mathcal{B}} = \hat{D}_{\mathcal{B}}^T(\hat{D}_{\mathcal{A}}\alpha_{\mathcal{A}} - \hat{y}) + \frac{\lambda}{2}.\end{aligned} \quad (13)$$

The optimal solution is given by $\alpha = (\alpha_{\mathcal{F}}, 0)$ and $\omega = (0, \omega_{\mathcal{B}})$. During the optimization of the proposed algorithm, the kernel size is an important factor. We calculate the kernel size σ empirically [11] by the following equation:

$$\sigma^2 = \frac{1}{2m}(D_{\mathcal{A}}\alpha_{\mathcal{A}} - y)^T(D_{\mathcal{A}}\alpha_{\mathcal{A}} - y). \quad (14)$$

Algorithm 1 summarizes the optimization procedures of the proposed approach.

Algorithm 1 Algorithm of KNS-SR

Input:

Training data D (normalized), and define $\mathcal{A} = \phi$, $\mathcal{B} = 1, \ldots, n$. $\alpha = \mathbf{0}^{n\times 1}$, $\omega = -D^T y$, $\gamma = -\mathbf{1}^{m\times 1}$.

Output:

Step 1: let $k \leftarrow k+1$, compute $\hat{D} = diag(\sqrt{-\gamma^k})D$ and $\hat{y} = diag(\sqrt{-\gamma^k})y$.

Step 2: compute $\omega = \min\{\omega_i : i \in \mathcal{B}\}$. If $\omega_s < 0$, then add s to $\mathcal{A}$, and delete s from $\mathcal{B}$, otherwise stop with $\alpha^* = \alpha$.

Step 3: compute the temporal variables $\alpha_{\mathcal{A}}$ by solving (13). If $\bar{\alpha}_{\mathcal{A}} > 0$, then let $\alpha^{k+1} \leftarrow [\bar{\alpha}_{\mathcal{A}}, 0]$, and go to step 4 otherwise. Let s be a constraint such that:

$$\theta = \frac{-\alpha_s}{\bar{\alpha}_s - \alpha_s} = \min\left\{\frac{-\alpha_i}{\bar{\alpha}_i - \alpha_i} : i \in \mathcal{A}_k, \bar{\alpha}_i < 0\right\}$$

and let $\alpha^{k+1} \leftarrow [\alpha_{\mathcal{A}} + \theta(\bar{\alpha}_{\mathcal{A}} - \alpha_{\mathcal{A}}); 0]$. Delete from $\mathcal{A}_k$ all index j (s among them) such that $\alpha_j = 0$, then add them to $\mathcal{B}_k$ and got to step 3 with $k \leftarrow k+1$.

Step 4: compute $\omega_{\mathcal{B}}$ by (13) and go back to step 2 with $k \leftarrow k+1$.

Step 5: update the auxiliary vector γ^{k+1} and kernel size σ using (7) and (14), respectively. Return to step 1.

3.4. Learning Robust Classifier with KNS-SR

Unlike other works that design the classifier based on the Euclidean distance [19] [10] or correntropy [11] of the representation error, we design the classifier based on KNS-loss. Let $\delta_{c_i}(\alpha)$ be the sparse coefficients of class c_i, then we obtain the approximated representation for each class as $\hat{y}_{c_i} = D_{c_i}\delta_{c_i}(\alpha)$. Based on the KNS-loss, the test sample y can be classified by assigning it to the class corresponding to the minimal nonlinear difference between y and $\hat{y}_{c_i}$:

$$\underset{c_i}{\text{argmin}}\ r_{c_i}(y) = (1 - k_\sigma(y - D_{c_i}\delta_{c_i}(\alpha)))^{p/2}. \quad (15)$$

4. EXPERIMENTAL RESULTS

We carry out experiments on the public AR face database [20] and extended Yale B face database [21], which serves both to demonstrate the efficacy of the proposed KNS-SR algorithm and to validate the claims of the previous sections. Image reconstruction and classification are implemented successively to examine the learned sparse features using our framework, comparing performances across various evaluation measurements, and comparing to methods including LRC [19], SRC [10], and CESR [11].

4.1. Face Classification Against Real-World Occlusion

We first verify the face recognition ability of our algorithm on occluded images from the AR database which consists of 50 male and 50 female subjects. We choose 8 non-occluded images ($\{$1st-4th$\}$ and $\{$14th-17th$\}$) with varying expressions from each subject for training. All the images with sunglass occlusion, $\{$8th-10th$\}$ and $\{$21st-23rd$\}$, are selected for testing. First, we compare the sparsity and robustness of the proposed method against the CESR algorithm. Fig. 1 shows the results of both algorithms on a sunglass occlusion image from the 3rd male subject. The dark blue parts of the weight images in the 2nd column are the detected occlusion area. Although both algorithms can detect the sunglass parts, the weight distribution of the proposed algorithm with $p = 0.8$ is much more clear than that of CESR. This is because with a smaller p power acted on the kernel function, the weights corresponding to the outlier region or the occlusion region will be further reduced in comparison with the correntropy-based CESR. Subfigures in the 4th column show the sparse coefficients of the testing image. Although sparse coefficients from both algorithms give correct results, the coefficient vector from the proposed method is more sparse than that from CESR. Since the proposed method obtains more accurate sparse coefficients, the image reconstruction accuracy from the proposed algorithm is better than that from the CESR.

Then we carried out image classification on different downsampled images with downsampling ratio of $\{$1/16, 1/10, 1/4, 1/2, 1$\}$. To numerically evaluate the performance, we compare the recognition rates of the proposed algorithm and all the baseline algorithms in Fig. 2(a) which shows that the recognition rates of our algorithm with $p = 0.8$ are apparently higher than that of benchmarks because our method has the superiority in minimizing the effect of occlusion. These

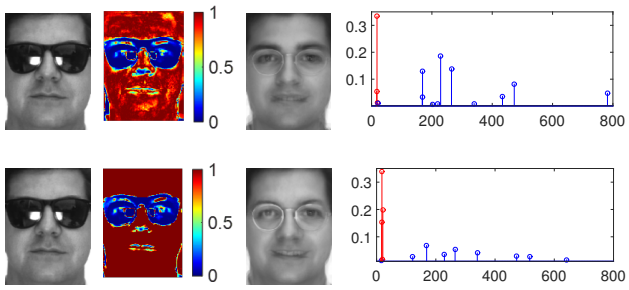


Fig. 1. Sparse representation of CESR (first row) and KNS-SR(second row) on an occluded image. Left to right: original images with sunglasses occlusion, weight images, reconstructed images, and sparse coefficients.

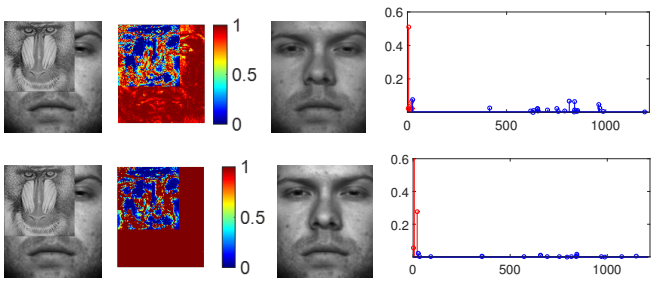
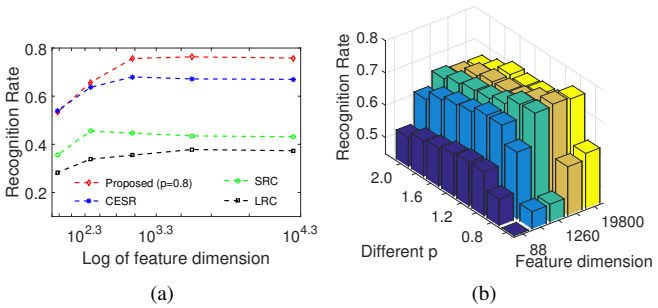


Fig. 3. Sparse representation by CESR (first row) and KNS-SR (second row) against contiguous occlusion. Left to right: original test images with contiguous occlusion, weight images, reconstructed images, sparse coefficients.

results also convey the information that the algorithms based on the mean square error loss function (e.g., LRC and SRC) and second-order statistics (CESR) are sensitive to outliers (occlusion part), and thus have difficulties in classifying occluded images. Fig. 2(b) plots the recognition rates of the proposed method with varying p values, and we know that the rates of the proposed method with $p < 2$ are apparently better than with $p = 2$.



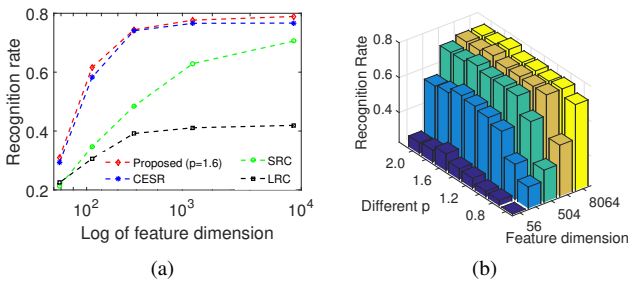
(a) (b)

Fig. 2. (a) The recognition rates of different methods under different sampling rates; (b) The recognition rates of the proposed method under different p.

4.2. Face Classification Against Contiguous Occlusion

We now verify the image reconstruction and classification ability of the proposed algorithm on images with contiguous occlusions. The experiment are performed on the extended Yale B face database in which there are 30 subjects, and we randomly select 32 images per subject to construct the dictionary D. We simulated the contiguous occlusion by replacing a randomly selected local region in each testing image with an unrelated image. We compared the sparse representation and reconstruction on an occluded image (45% percent occluded) as shown in Fig. 3 where the weight image of the

KNS-SR are more clear than that from CESR, and the coefficients of KNS-SR are more sparse. We computed recognition rates with these occluded images under different downsampling rates $\{1/24, 1/16, 1/8, 1/4, 1/2\}$. Fig. 4(a) shows the recognition rates of different algorithms with various dimension of features. KNS-SR achieves the highest recognition rates under different feature spaces, and reaches the best result when sampling rates is $1/2$. Fig. 4(b) shows the recognition rates under different parameters for p. The best recognition accuracy is reached when p is 1.6.



(a) (b)

Fig. 4. (a) The recognition rates from different methods under different sampling rates; (b) The recognition rates of the proposed method under different p.

5. CONCLUSION

This paper presented a new framework based on the Information Theoretic Learning to improve the robustness of the representation ability on the occluded images. By introducing the non-second order statistic based loss function, more flexible constraints are imposed on the error loss, resulting in greater occlusion detection ability and improved performances in different image processing applications. Experimental results on the occluded images show that our algorithm outperforms existing methods in terms of the face reconstruction and recognition.

6. REFERENCES

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[2] E. G. Ortiz, A. Wright, and M. Shah, "Face recognition in movie trailers via mean sequence sparse representation-based classification," *CVPR*, pp. 3531–3538, 2013.

[3] O. Ocegueda, T. Fang, S. K. Shah, and I. A. Kakadiaris, "3D face discriminant analysis using Gauss-Markov posterior marginals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 728–739, 2013.

[4] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 337–350, 2006.

[5] S. Li and Y. Fu, "Learning robust and discriminative subspace with low-rank constraints," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 27, no. 11, pp. 2160–2173, 2016.

[6] H. Jia and A. M. Martinez, "Support vector machines in face recognition with occlusions," *CVPR*, pp. 136–141, 2009.

[7] ——, "Face recognition with occlusions in the training and testing sets," *FG*, pp. 1–6, 2008.

[8] M. Elad, M. A. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972–982, 2010.

[9] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[10] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[11] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, 2011.

[12] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer Science & Business Media, 2010.

[13] S.-C. Pei and C.-C. Tseng, "Least mean p-power error criterion for adaptive FIR filter," *IEEE J. Sel. Areas Commun.*, vol. 12, no. 9, pp. 1540–1547, 1994.

[14] B. Chen, L. Xing, Z. Wu, J. Liang, J. C. Príncipe, and N. Zheng, "Smoothed least mean p-power error criterion for adaptive filtering," *Digital Signal Process.*, vol. 40, pp. 154–163, 2015.

[15] B. Chen, L. Xing, X. Wang, J. Qin, and N. Zheng, "Robust learning with kernel mean p-power error loss," *IEEE Trans. Cybern.*, vol. 48, no. 7, pp. 2101 – 2113, 2017.

[16] 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," *Math. Comput.*, vol. 63, no. 208, pp. 625–643, 1994.

[17] C.-P. Wei and Y.-C. F. Wang, "Undersampled face recognition via robust auxiliary dictionary learning," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1722–1734, 2015.

[18] A. Santos-Palomo and P. Guerrero-Garcıa, "Solving a sequence of sparse least squares problems," in *Tech. Rep.* Department of Applied Mathematics, University of Malaga, 2001.

[19] S. Z. Li, "Face recognition based on nearest linear combinations," *CVPR*, pp. 839–844, 1998.

[20] A. Martinez and R. Benavente, "The AR face database, 1998," *Comput. Vis. Center, Tech. Rep.*, vol. 3, p. 5, 2007.

[21] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, 2001.