2013 IEEE International Conference on Computer Vision

Robust Dictionary Learning by Error Source Decomposition

Zhuoyuan Chen Ying Wu
Northwestern University
2145 Sheridan Road, Evanston, IL 60208
zhuoyuanchen2014@u.northwestern.edu,yingwu@eecs.northwestern.edu

Abstract

Sparsity models have recently shown great promise in many vision tasks. Using a learned dictionary in sparsity models can in general outperform predefined bases in clean data. In practice, both training and testing data may be corrupted and contain noises and outliers. Although recent studies attempted to cope with corrupted data and achieved encouraging results in testing phase, how to handle corruption in training phase still remains a very difficult problem. In contrast to most existing methods that learn the dictionary from clean data, this paper is targeted at handling corruptions and outliers in training data for dictionary learning. We propose a general method to decompose the reconstructive residual into two components: a non-sparse component for small universal noises and a sparse component for large outliers, respectively. In addition, further analysis reveals the connection between our approach and the "partial" dictionary learning approach, updating only part of the prototypes (or informative codewords) with remaining (or noisy codewords) fixed. Experiments on synthetic data as well as real applications have shown satisfactory performance of this new robust dictionary learning approach.

1. Introduction

With the development of harmonic analysis [4, 3], sparse models have received a lot of attention in recent years. The universal sparsity in real applications enables us to achieve good performance in many areas such as compressive sensing [3], image recovery [6] and classification [29]. We refer readers to [28] for a detailed summary.

Specifically, learning a sparse prototype model (or "dictionary") [15, 21, 6] to represent training data set is often applied as a first step. The advantages of dictionary learning over pre-defined fixed bases, such as DCT and FFT, have been shown in many applications [8, 23, 6]. Recent studies [26] also provided theoretical support for exact recovery of all codewords under that condition of sufficient sparsity and noise-free observations.

Most sparse coding methods [27, 15, 6, 17] make a basic assumption that the observed signals consist of a sparse linear combination of codewords plus dense Gaussian noises of small variation. However, though working well generally, this assumption does not hold in case of large corruptions and outliers, which is common in practice. For example, in face recognition, a sample face image can be considered as corrupted if the person accidentally wears sunglasses. As shown in [29], if the training data is clean, corrupted testing data can be handled by using sparse residual. This robust method demonstrated very encouraging face recognition results [29, 31, 12].

In practice, it may be inevitable to include corrupted sample and outliers in addition to dense Gaussian noises in the training data. Suppose we need to recognize faces for two people A and B, with a training set $\mathbf{T} = \{x_A^1, x_A^2, \ldots, x_B^1, x_B^2, \ldots\}$, where x_A^k and x_B^k are samples from A and B, respectively. If $\mathbf{T}$ is clean, we may be able to recognize the target under certain noise and corruption as shown in [29, 12]. However, if $\mathbf{T}$ itself is corrupted, e.g., x_A^k is person A accidentally wearing sunglasses, then it can be very ambiguous to recognize a corrupted input, e.g. B with sunglasses. It is clear that noisy and corrupted training data will largely result in low quality dictionary if learned by existing methods. As the data noise come multiple sources with different characteristics, we call this issue the *residual modality problem*. This also emerges in many other vision tasks, such as removing salt and pepper noises, and handling artificially added texts and other outliers in images.

In order to address this issue, we propose a robust dictionary learning approach based on the decomposition of the reconstructive residual into two modalities: one for dense small Gaussian noises an the other for large sparse outliers. We can have different residual penalty for different modalities. This paper provides a coordinate descent solution for robust dictionary learning, an online acceleration method, and its convergence property. This new approach allows us to learn a robust dictionary and identify outlier training data. In addition, our further study reveals a very interesting con-

1550-5499/13 $31.00 © 2013 IEEE
DOI 10.1109/ICCV.2013.276

nection between this source decomposition approach and the "partial dictionary update" approach. This residual decomposition method is an explicit way to handle corrupted data in dictionary learning. Moreover, we also propose an alternative that uses robust functions on reconstructive residual, which is an implicit means for corrupted data. We show these two methods are closely related, and they become equivalent in certain situations. Experiments on synthetic dataset, texture synthesis, and image denoising show that our model is able to achieve quite satisfactory results without using much heuristics.

2. Robust Dictionary Learning

The following notation is used throughout the paper: we denote a collection of observed data as $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ where $\mathbf{x}_i \in R^n$. We aim to learn a dictionary $D_{n\times m} = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_m\}$ to efficiently represent X as $\mathbf{x}_i = D\alpha_i$, where $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_N\}$ are sparse coefficients. As usual, the Frobenius norm is defined as $||X||_F \triangleq (\sum_{i=1}^{N}\sum_{j=1}^{n} X_{i,j}^2)^{1/2}$.

The original work of sparse dictionary learning was first proposed by Olshausen and Field [21] based on human perceptional system. Generally, the learning is commonly viewed as an optimization problem:

$$\min_{D,\alpha} \phi(X - D\alpha) + \psi(\alpha) \qquad s.t.\ ||\mathbf{d}_i||_{L^2} \leq 1 \quad (1)$$

where D is the dictionary, ϕ and ψ are cost functions. In the equation, the first term measures the residual (typically $\phi(.) = ||.||_F^2$), while the second regularizes the linear representation α. In sparse coding, an L^1-norm is always applied for ψ [15, 28, 17].

Recently, a lot of work has been done to improve the traditional dictionary learning model in Eqn (1) for specific tasks. Various formulations and properties for α and D have been investigated, such as heavy-tailedness [21], differentiability [1], hierarchy [14] and discriminative ability [13, 18, 19, 30]. Many variations are compared in [17].

However, not much attention has been paid to the modality of the residual, where the squared loss model $\sum_i(\mathbf{x}_i - D\alpha_i)_2$ is generally applied. Recently in SPAMS toolbox [17], Mairal extends it to a weighted square loss $||\Lambda(X - D\alpha)||_F^2$ to penalize different dimensions differently with a diagonal matrix Λ; Zhao [32] and Lu [16] assume that the residual observes a Laplacian distribution and use a pure L^1-norm. Zhou [33] studies the influence of residual modality parameter settings and suggests that a good estimation of noise level can enhance the performance of sparse coding. In contrast to these methods, we propose to decompose the residual into two sources rather than one Gaussian or Laplacian.

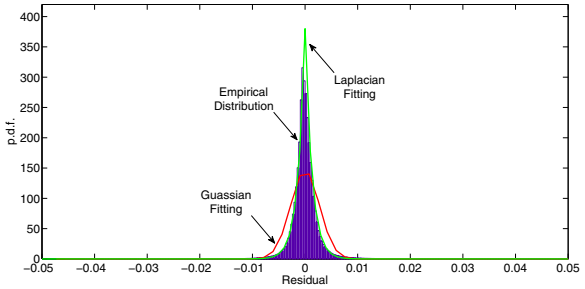


Figure 1. A statistical comparison for face recognition on extended Yale B [9]. The empirical residual distribution, its Gaussian and Laplacian fitting is shown in blue, red and green. We can see clearly that the true residual has ***smoother*** $p.d.f.$ near $Res = 0$ than Laplacian and ***heavier*** tails than Gaussian.

2.1. Over-smoothed or Over-sparsified Residual?

In Figure 1, we show a statistical comparison of the true residual with Gaussian and Laplacian fittings for a face recognition task on Extended Yale B dataset [9] by sparse coding [29]: we stack faces in columns as D and recognize query data by sparse coding:

$$\hat{\alpha} = \arg\min_{\alpha} ||X - D\alpha||_F^2 + \lambda||\alpha||_{L^1}$$

As we can see, it is obvious that the Gaussian fitting (red) tends to over-smooth the residual while the Laplacian (green) tends to over-sparsify. Similar results have also been observed in many other applications such as digit recognition and image recovery.

2.2. Sparse/Non-sparse Residual Decomposition

Rather than fitting one universal Gaussian or Laplacian model, we assume that the residual $Res = X - D\alpha$ contains two components:

$$Res \triangleq \begin{cases} N & \mathbf{x} \in D \setminus \Omega \\ \Xi & \mathbf{x} \in \Omega \end{cases} \quad (2)$$

where Ω denotes the corrupted region. Actually, this type of decomposition is also related in spirit to the Mumford-Shah model, or the membrane method [20, 11].

A simple illustration of our idea is given in Figure-2: we propose to learn a set of robust codewords $\{\mathbf{d}_1, \mathbf{d}_2\}$, to sparsely represent data points (diamonds and triangles) and ignore the outlier (the red diamond corrupted in z coordinate. A typical L^2-norm for residual penalty only obtains a compromised result $\{\mathbf{d}_1', \mathbf{d}_2\}$.

Under the assumption discussed above, we seek to estimate a dictionary, sparse coefficients and corruptions by minimizing the number of nonzero elements of α, Ξ as well as the negative log-likelihood of Gaussian residual N si-

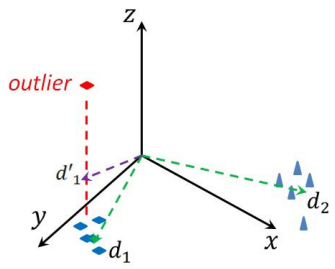


Figure 2. A demonstration of our idea: data points X are denoted by triangles and diamonds, with one outlier (marked in red). Ideally, two green codewords $\mathbf{d}_1, \mathbf{d}_2$ are desired, while the outlier brings $\mathbf{d}_1$ to $\mathbf{d}'_1$ using traditional dictionary learning [15].

multaneously:

$$\begin{aligned} & E(D, N, \Xi, \alpha) \triangleq ||N||_F^2 + \lambda_1 ||\Xi||_{L^0} + \lambda_2 ||\alpha||_{L^0} \\ & s.t. \qquad N + \Xi = X - D\alpha \qquad ||\mathbf{d}_i||_{L^2} \leq 1 \\ & \Leftrightarrow ||X - D\alpha - \Xi||_F^2 + \lambda_1 ||\Xi||_{L^0} + \lambda_2 ||\alpha||_{L^0} \\ & \Leftrightarrow ||X - [D\ I] \begin{bmatrix} \alpha \\ \Xi \end{bmatrix} ||_F^2 + \lambda_1 ||\Xi||_{L^0} + \lambda_2 ||\alpha||_{L^0} \end{aligned} \tag{3}$$

In practice, the optimization of Formula (3) is NP-hard. As customary, we relax it by minimizing its L^1 surrogate, such that

$$\begin{aligned} \{\hat{D}, \hat{\Xi}, \hat{\alpha}\} = \arg \min_{D,\Xi,\alpha} ||X - [D\ I] \begin{bmatrix} \alpha \\ \Xi \end{bmatrix} ||_F^2 + \\ \lambda_1 ||\Xi||_{L^1} + \lambda_2 ||\alpha||_{L^1} \qquad s.t.\ ||\mathbf{d}_i||_{L^2} \leq 1 \end{aligned} \tag{4}$$

For further details and related properties, we refer interested readers to [25], where the additive combination of i.i.d. Gaussian and Laplacian noises have been carefully studied and the analytical form of the $p.d.f.$ is deduced.

2.3. Robust Dictionary Learning– "Partial Codeword Updates"

Denoting the "augmented dictionary" by $\bar{D} := [D\ I]$, our model has an interesting interpretation in an EM based optimization process:
(1) *sparse coding step*: if we optimize Ξ and α with D fixed, our model becomes robust sparse coding [29];
(2) *dictionary update step*: if we update D with Ξ and α fixed, it is a "*partial*" dictionary learning: only the *informative* codewords $D_{Info} = \{\mathbf{d}_1, ..., \mathbf{d}_m\}$ are updated, while the *noisy* codewords with natural basis $D_{Noise} = \{\mathbf{e}_1, ..., \mathbf{e}_n\}$ are maintained.

A natural question is: what if we learn D_{Info} and D_{Noise} simultaneously? The non-convexity of dictionary learning method in Eqn (3) requires a good initialization; fixing D_{Noise} reasonably avoids local minima and enables us to obtain a better numerical solution.

3. Solution

As mentioned above, Eqn (4) is non-convex. We use a coordinate descent scheme to optimize D and Ξ, α alternatively:
(1) Fixing D, we optimize Ξ_i and α_i in Eqn (4):

$$\begin{aligned} \{\hat{\alpha}_i, \hat{\Xi}_i\} = \arg \min_{\alpha_i, \Xi_i} ||\mathbf{x}_i - [D\ I] \begin{bmatrix} \alpha_i \\ \Xi_i \end{bmatrix} ||_F^2 + \\ \lambda_1 ||\Xi_i||_{L^1} + \lambda_2 ||\alpha_i||_{L^1} \end{aligned}$$

This problem can be solved by shrinkage [10] efficiently and highly parallel in nature.
(2) Fixing sparse coefficients Ξ and α, we update D:

$$\hat{D} = \arg \min_{D} ||X - \Xi - D\alpha||_F^2 \qquad s.t.\ ||\mathbf{d}_i||_{L^2} \leq 1 \tag{5}$$

which is a constrained quadratic optimization problem and is solvable by Lagrange dual [15].

3.1. Online Acceleration

To accelerate, we set our algorithm in an online form. Assuming the training set is composed of $i.i.d.$ samples of a distribution $p(x)$, we add $\mathbf{x}_t$ sequentially into the system and minimize:

$$f_t(D_t) \triangleq \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} ||\mathbf{x}_i - D_t\alpha_i - \Xi_i||^2 + \lambda_1 ||\Xi_i||_{L^1} + \lambda_2 ||\alpha_i||_{L^1} \tag{6}$$

3.2. Convergence Analysis

We follow [17] to prove the convergence property of this new approach. Three reasonable assumptions have been made in [17]:
(A) compact support[1];
(B) strictly convex quadratic surrogate functions[2];
(C) unique sparse coding solution[3].
We keep (A)(B) unchanged and modify (C) slightly as:
(C') **Unique Sparse Solution**: the *informative* codewords $\{\mathbf{d}_1, \mathbf{d}_2, ...\}$ are sufficiently irrelevant to the *noisy* ones $\{\mathbf{e}_1, \mathbf{e}_2, ...\}$, i.e., $\exists \kappa'_2 > 0$, the smallest eigenvalue of $\bar{D}_\Lambda^T \bar{D}_\Lambda$ is larger than κ'_2.
Accordingly, with $f(D)$ strictly convex and the sparse solution α_i well defined, we have:

Proposition 1 (Convergence of D_t) *Under assumptions (A)(B)(C'), the distance between the informative D_t and the set of stationary points converges almost surely to 0 when $t \to \infty$ with probability 1.*

[1] The data admits a bounded probability density p with compact support K
[2] The smallest eigenvalue of matrix $A = E(\alpha\alpha^T)$ satisfies $eig(A) \geq \kappa_1$;
[3] $\exists \kappa_2 > 0$, $s.t.$, $\forall x \in K, D$, the smallest eigenvalue $D^T D \geq \kappa_2$

4. Dictionary Learning by Robust Penalty

The above residual decomposition approach model the residual explicitly. In this paper, we also propose an alternative that handles the residual implicitly. An interesting thing we observe is that these two treatments are closely related.

As mentioned in Section 2, we know that a good $p.d.f.$ of residual should: (1) be smoother around $Res = 0$ than Laplacian; (2) have heavier tails than Gaussian. Accordingly, we propose to take outliers into consideration implicitly:

$$\{\hat{D}, \hat{\alpha}\} = \arg\min_{D,\alpha} \sum_{i=1}^{N} \phi(\mathbf{x}_i - D\alpha_i) + \lambda||\alpha_i||_{L^1} \quad s.t.\ ||\mathbf{d}_j||_{L^2} \leq 1 \qquad j = \{1, 2, 3, ..., m\} \tag{7}$$

where $\phi(.)$ is a robust function for the residual.

In robust statistics [11], various forms of robust functions have been proposed, such as the Charbonnier penalty $\phi(s) = \sqrt{s^2 + \epsilon^2}$, Lorentzian, Geman-McClure function and so forth.

If we further regard the error source decomposition model as

$$\phi(s) = \inf_{\xi}(s - \xi)^2 + \lambda|\xi|$$

then the shape of $\phi(s)$ is very similar to the shape of the robust function. Especially, by varying λ, $\phi(s)$ is very close to the Charbonnier regularizer with different selection of ϵ.

Similar online optimization and convergence analysis can also be extended to the robust influence function models. We apply a stochastic gradient method for dictionary update as:

$$D_t = \Pi_C(D_{t-1} - \frac{\rho}{t}\sum_{i=1}^{t} \nabla_D \psi'(s_i)|_{s_i = X^i - D\alpha^i}) \tag{8}$$

where $0 < \rho < 1$ is a step-length and Π_C projects D_t to the unit ball. Empirically, we find it works well numerically, and the Charbonnier outperforms its highly non-convex alternatives. The convergence analysis in Section 3.2 still holds provided that $f_t(D_t)$ is strictly convex with lower bounded Hessian. However, most robust penalizers are non-convex except the Charbonnier. To enforce convexity, we can simply add an extra term $\frac{\kappa_1'}{2}||D||_F^2$, replacing Hessian matrix with $\frac{1}{t}\nabla^2 f_t(D_t) + \kappa_1' I$ so that the cost function remains convex to D_t.

Generally speaking, both the error source decomposition method and the robust penalty method perform well, but the former outperforms the latter in speed. Therefore, we use the former throughout the experiments.

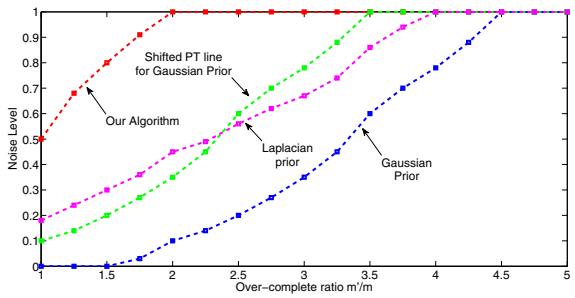


Figure 3. Phase Transition line comparison: the red, blue and magenta lines are boundaries of successful/failure regions of our model, traditional methods with Gaussian prior [15] and Laplacian prior [32] respectively. To make the comparison "fair", we shift the phase transition line of Gaussian prior to the left (green), since more bases are implicitly used in the other two methods.

5. Experimental Results

5.1. Phase Transition on Synthetic Data

We first demonstrate the validity of our algorithm on a synthetic dataset. Suppose we observe a number of N noisy data $Y = \widetilde{D}\alpha + n_1 + n_2$. The "true" dictionary $\widetilde{D} = \{\widetilde{\mathbf{d}_1}, \widetilde{\mathbf{d}_2}, ..., \widetilde{\mathbf{d}_m}\}$ is generated from $i.i.d.$ Gaussian; $\alpha = \{\alpha_1, ..., \alpha_N\}$ are N sparse vectors; $n_1 \sim N(0, \sigma_1^2)$ is an $n \times N$ residual matrix with Gaussian noises of small variance; n_2 is a sparse corruption matrix with large Gaussian noise for nonzero entries.

We train an over-complete dictionary $D_{m' \times p}$ with $m' > m$ bases for candidates. In our experiments, we use $x \in R^{50}$, $m = 30$. $N = 1000$. Similar to [26], we use a more direct criteria as "every codeword $\widetilde{d_i}$ is recovered exactly":

$$\min_i\{\max_j\{|\widetilde{\mathbf{d}_i} \cdot \mathbf{d}_j|\}\} \geq thr \tag{9}$$

Typically, we set the threshold as $thr = 0.97$.

In Figure-3, We compare the performance of traditional dictionary learning with Gaussian prior [15] and Laplacian prior [32] with our model. The horizontal axis is the over-complete ratio, (i.e., if we train $m' = 60$ potential codewords for a true dictionary of size $m = 30$, the ratio is $m'/m = 2$) [4]; the vertical axis is the variance of sparse noises n_2. The dashed line are transition boundaries of "successful" and "failure" regions obtained by logistic regression.

We can see clearly that our robust model (blue) has more tolerance to mixed heavy-tail noises than both [15] (green and red for with/out self-taught bases) and [32] (purple lines). Similar results have been observed with different parameter settings of m, n, N.

[4] the more potential codewords we train, the more likely we can recover all codewords $\widetilde{d_i}$

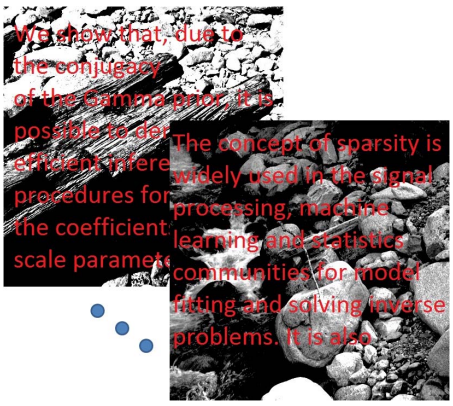


Figure 4. We use 10 images available at [21] for 13×13 basis learning. The red characters are added manually as outliers.

5.2. Robust Dictionary Learning on Contaminated Images

Our second experiment is to test the robustness of our algorithm on contaminated images.

As shown in Figure-4, we train a dictionary D on the SparseNet image dataset [21] with small Gaussian noises (5dB) and sparse large outliers (red characters) added. We randomly crop 13×13 patches as X and initialize D_0 with gray-scale DCT. A visual comparison of traditional dictionary learning [15] and our algorithm is shown in Figure-5(a)(b) respectively. In the experiment, we set $\lambda_1 = \lambda_2 = 0.2$ for the sparse regularization term.

We can see that both algorithms perform well to learn reasonable Gabor-like codewords, but our method is less likely influenced by outliers: 1.22% of our bases contain red patches, in comparison with 2.75% by the traditional [15]. Close scrutiny of Ξ coefficients reveals that a good initialization of D_{Noise} absorbs the corruptions and keeps D_{Info} away from sparse red outliers. We also tried Laplacian residual model [32]. The difference is less obvious and we omit them here. However, the advantages of our bases over the Laplacian model emerge when further applications are studied.

Next, we show two potential applications of our algorithm in robust image processing.

5.2.1 Robust Image Recovery

First, we consider image denoising. To deal with outliers as well as Gaussian noises simultaneously, we propose a robust image denoising algorithm based on robust codewords as following:

(1) *Robust Dictionary Learning*: we train a dictionary D on noisy dataset with our model in Eqn (4);

(2) *Local Patch Denoising*: then for each patch $\mathbf{x}_i$, we do sparse coding as:

PSNR(dB)	House	Jetplane	Lake	Lena
(ours)	33.96	31.32	29.00	31.60
[6]	33.59	31.16	28.96	31.38
PSNR(dB)	Mandril	Peppers	Pirate	Cameraman
(ours)	27.83	31.10	29.32	32.47
[6]	27.43	31.06	29.09	32.10

Table 1. Performance comparison on standard image processing dataset with K-SVD [6].

PSNR(dB)	Our Method	[6]	[32]	[24]
$\sigma = 5$	37.41	37.36	36.13	36.77
$\sigma = 10$	33.36	33.16	31.95	31.27
$\sigma = 15$	31.17	30.85	28.42	28.73

Table 2. Performance comparison with K-SVD [6], Laplacian [32] and total-variation [24] on denoise benchmark [7] with random sparse corruptions added.

$$\{\alpha_i, \Xi_i\} = \arg\min_{\alpha_i, \Xi_i} ||\mathbf{x}_i - D\alpha_i - \Xi_i||^2 + \lambda_1|\alpha_i| + \lambda_2|\Xi_i|$$

then the denoised patch $\widetilde{\mathbf{x}}_i = D\alpha_i$ is obtained with both sources of residual removed;

(3) *Non-local Refinement*: finally, we process the overlapping regions with a weighted mean filtering: $\hat{\mathbf{x}}_i = \sum_{j \in N(\mathbf{x}_i)} w_j \widetilde{\mathbf{x}}_j$, where $N(\mathbf{x}_i)$ is the neighbor set of $\mathbf{x}_i$. Following [2], we use the weights w_j to achieve the best PSNR performance as:

$$w_j = \frac{1}{Z_i} e^{-\lambda(||\mathbf{x}_j - D\alpha_j - \Xi_j||^2 + \lambda|\Xi_j|)} \tag{10}$$

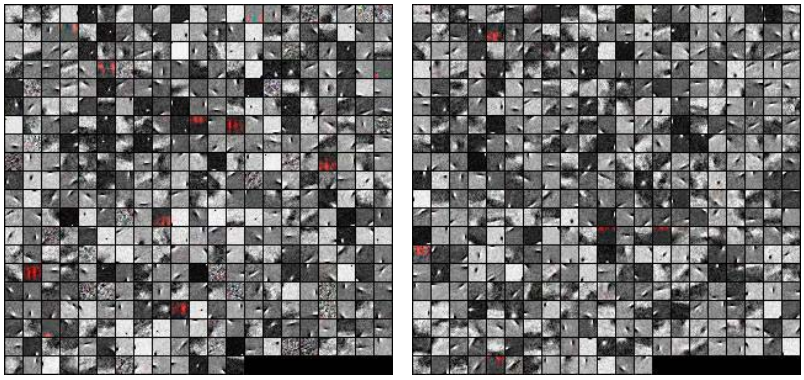
where $Z_i = \sum_{j \in N} w_j$ is a normalization constant.

We add synthetic Gaussian noises of $\sigma = 20$ and sparse outliers of $\sigma = 30$ (about 3% pixels are corrupted) to standard images. In Table-1, we compare PSNR performance of our algorithm with K-SVD denoising [6]. Some denoised results are shown in Figure-6, from which we can see that the "dotted" salt and pepper corruptions are eliminated successfully.

For an extensive study, we carry out a complete experiment of image denoising on the benchmark [7]. Besides Gaussian noises with $\sigma = \{5, 10, 15\}$, we corrupts 1% pixels with $\sigma = 25$. In Table-2, we compare average PSNR performance with classic K-SVD [6], Laplacian [32] and total-variation denoising [24]. This clearly demonstrate that the error source decomposition model outperforms others in case of heavy-tailed noise removal.

5.2.2 Robust Texture Synthesis

Another potential application of our model is robust texture synthesis. Sparse modeling of texture analysis has been studied [22] for exemplar-based synthesis. We exploit the



(a) Learned basis by [15] (b) Learned basis of RDL

Figure 5. (a) The training results by [15] and our robust model are shown in (a) and (b).

Figure 6. 1st and 3rd columns: noisy images; 2nd and 4th: results of our robust image recovery method.

self-similarity of textures with outlier removal by integrating our model into image quilting [5]:

(1) *Robust Dictionary Learning*: given an textured image, we first learn D:

$$\{D, \alpha\} = \arg\min_{D,\alpha} ||X - D\alpha - \Xi||_F^2 + \lambda||\Xi||_{L^1} \quad s.t. \quad ||\alpha_i||_{L^0} \leq 1 \tag{11}$$

We apply a typical block coordinate descent optimization scheme to update D and $\{\Xi, \alpha\}$ alternatively.

(2) *Robust Patch Processing*: for a new patch y to be added "agreeing" with the neighbors based on the criteria in [5], we decide whether it is also consistent with learned codewords D by:

$$f(y) = \min_{\alpha,\xi} ||y - D\alpha - \xi||^2 + \lambda|\xi| \; s.t. \; ||\alpha||_{L^0} \leq 1 \tag{12}$$

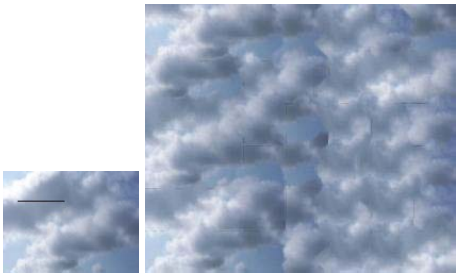
if $f(y)$ is within a threshold $f(y) < e$, we directly add y; otherwise, we add $D\alpha$ instead.

Figure 7. Texture patches with sparse corruptions.

(3) *Minimum Inconsistent Boundary-cut [5]*: we use the dynamic programming method to smooth the overlapping regions for each added patch.

In Figure-7, we randomly add some outliers to original patches and the synthesized textures are shown in Figure-8. As we can see, our model achieve visually pleasant results. A heuristic explanation is: if we choose $\lambda << 1$ in Eqn (11), the cost function is very close to an L^1- norm. Then, for a codeword $\mathbf{d}^i$ and its examples $X^{d^i} := \{X^{i,1}, X^{i,2}, ...\}$, we have $\mathbf{d}^i \approx Med(X^{d^i})$, which is actually an exemplar-based dimension-wise median filter.

Figure 8. 1st and 3rd columns: direct image quilting [5]; 2nd and 4th columns: robust texture synthesis.



(a) (b)

Figure 9. A typical failure case of our algorithm. To remove the artificially added outliers (the black line), we eliminate some infrequent patterns in the input. The result turns to be over-repetitive on stochastic textures.

We have also carried out a complete evaluation on the CMU-NRT Database[5] with sparse noises added. The experiment shows that our method performs well on more regular patterns rather than stochastic ones. We show a failure case in Figure-9: the internal patterns need to be more frequent than outliers to be synthesized, and our algorithm sometimes achieve over-uniform textures during step(2).

5.3. Robust Discriminative Dictionary Learning

Finally, we propose to learn a robust dictionary for classification. There have been some work on discriminative models [13, 18, 23], relying either on the reconstructive residual, or on the discriminative ability of sparse coding coefficients.

Following [30], we considering a k-class classification $c_i = \{1, 2, ..., k\}$. We aim to infer a set of dictionaries $D = \{D^1, ..., D^k\}$ and related sparse coefficients $\alpha = \{\alpha^1, ..., \alpha^k\}$ for each class satisfying following two conditions:
(1) Given $\mathbf{x}_i \in c_j$ we have $\mathbf{x}_i = D\alpha_i \approx D^j \alpha_i^j$;
(2) the within-class scatter is small, while the between-class scatter is large.

Accordingly, we have:

$$\{D, \alpha\} = \arg\min_{D,\alpha} \sum_{c_i=1}^{k} r(X, D, \alpha) + \lambda_1 ||\alpha||_{L^1} + \lambda_2 (tr(S_W(\alpha) - S_B(\alpha))) + \eta ||\alpha||_F^2 \quad (13)$$

In the equation, inter-class scatter and between-class are defined as:

$$S_W(\alpha) = \sum_{c_i=1}^{k} \sum_{x_j \in c_i} (\alpha_j - \mathbf{m}_{c_i})(\alpha_j - \mathbf{m}_{c_i})^T$$

$$S_B(\alpha) = \sum_{c_i=1}^{k} (\mathbf{m}_{c_i} - \mathbf{m})(\mathbf{m}_{c_i} - \mathbf{m})^T$$

where $\mathbf{m}_{c_i}$ and $\mathbf{m}$ are the mean of X^{c_i} and X.

We apply the error source decomposition to the discriminative fidelity term as:

$$r(X, D, \alpha) = \sum_{c_j} \sum_{x_i \in c_j} (||\mathbf{x}_i - D\alpha_i - \Xi_{1,i}||_F^2 + ||\mathbf{x}_i - D^{c_j}\alpha_i^{c_j} - Xi_{2,i}||_F^2 + \lambda_3 ||\Xi_{1,i}||_{L^1} + \lambda_3 ||\Xi_{2,i}||_{L^1} + \sum_{l \neq c_j} ||D^l \alpha_i^l||_F^2)$$

[5] http://vivid.cse.psu.edu/texturedb/gallery/

Algorithms	Lasso	RSC [29]	Dirty [12]	SVM	FDDL[30]	ours
Error rate	7.9%	3.6%	3.5%	4.9%	1.6%	1.4%

Table 3. Performance comparison on face recognition benchmark [9].

For optimization, we initialize each D^{c_j} using a few iterations of K-SVD on each class separately as[18, 30]. Then, we iteratively update *sparse coding* for α^* and *dictionary update* for D. We omit further details due to lack of space and refer interested readers to [30].

We test our robust dictionary learning on Yale extended B benchmark [9], consisting of 2,414 frontal-face images from 38 individuals under different lighting condition. We randomly select half for training and the other half for testing. The comparison is shown in Table 3, which reveals that by adding robustness can enhance the performance of discriminative dictionary learning.

6. Conclusion

In this work, we introduce a novel generalized residual separation approach in robust dictionary learning to handle corruptions and outliers in training data. By exploiting the statistics on reconstructive residual, we observe that it comes from two sources: a large sparse corruption component and a small dense Gaussian component. Accordingly, we formulate a novel regularization to model the residual modality. Then, we propose an efficient online algorithm for optimization and analyze its convergence. Our experiments on the synthetic dataset as well as real image applications show that our approach can achieve satisfactory results.

7. Acknowledgement

This work was supported in part by National Science Foundation grant IIS-0916607, IIS-1217302, and DARPA Award FA 8650-11-1-7149.

References

[1] D. M. Bradley and J. A. Bagnell. Differentiable sparse coding. *NIPS*, 2008.

[2] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. *CVPR*, 2005.

[3] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *TIT*, 52(2):489–509, Feb. 2006.

[4] D. Donoho. Compressed sensing. *TIT*, 52:1289–1306, 2006.

[5] A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. *SIGGRAPH*, 2001.

[6] M. Elad and M. Aharon. Image denoising via learned dictionaries and sparse representation. *CVPR*, 2006.

[7] F. Estrada, D. Fleet, and A. Jepson. Stochastic image denoising. *BMVC*, 2009.

[8] P. J. Garrigues and B. A. Olshausen. Group sparse coding with a laplacian scale mixture prior. *NIPS*, 2010.

[9] 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. *PAMI*, 23(6):643–660, 2001.

[10] E. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for l1-minimization: Methodology and convergence. *SIAM: Journal on Optimization*, 19(3):1107–1130, 2008.

[11] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley and Sons Inc, 2009.

[12] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. *NIPS*, 2010.

[13] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. *CVPR*, 2011.

[14] A. B. Lee, B. Nadler, and L. Wasserman. Treelets-an adaptive multi-scale basis for sparse unordered data. *Annals of Applied Statistics*, 2(2):435–471, 2008.

[15] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. *NIPS*, 2006.

[16] C. Lu, J. Shi, and J. Jia. Online robust dictionary learning. *CVPR*, 2013.

[17] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *JMLR*, 11:19–60, 2010.

[18] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. *CVPR*, 2008.

[19] J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce. Discriminative sparse image models for class-specific edge detection and image interpretation. *ECCV*, 2008.

[20] D. Mumford and J. Shah. Optimal approximation of piecewise smooth functions and associated variational problems. *CPAM*, 42:577–685, 1989.

[21] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37:3311–3325, 1997.

[22] G. Peyrè. Sparse modeling of textures. *Journal of Mathematical Imaging and Vision*, 34(1):17–31, 2009.

[23] I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. *CVPR*, 2010.

[24] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[25] I. W. Selesnick. The estimation of laplace random vectors in additive white gaussian noise. *TSP*, 56(8):3482–3496, 2008.

[26] D. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. *COLT*, 2012.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267288, 1996.

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

[29] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 31(2):210 – 227, 2009.

[30] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. *ICCV*, 2011.

[31] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition. *CVPR*, 2011.

[32] C. Zhao, X. Wang, and W. kuen Cham. Background subtraction via robust dictionary learning. *EURASIP J. Image and Video Processing*, 2011.

[33] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin. Non-parametric bayesian dictionary learning for sparse image representations. *NIPS*, 2009.