Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 5376087, 13 pages
http://dx.doi.org/10.1155/2016/5376087

Research Article

Customized Dictionary Learning for Subdatasets with Fine Granularity

Lei Ye,[1] Can Wang,[2] Xin Xu,[1] and Hui Qian[2]

[1]*College of Mechatronic Engineering and Automation, National University of Defense Technology, Changsha 410073, China*
[2]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*

Correspondence should be addressed to Lei Ye; yelei@nudt.edu.cn

Received 21 June 2016; Accepted 18 October 2016

Academic Editor: Simone Bianco

Sparse models have a wide range of applications in machine learning and computer vision. Using a learned dictionary instead of an "off-the-shelf" one can dramatically improve performance on a particular dataset. However, learning a new one for each subdataset (subject) with fine granularity may be unwarranted or impractical, due to restricted availability subdataset samples and tremendous numbers of subjects. To remedy this, we consider the dictionary customization problem, that is, specializing an existing global dictionary corresponding to the total dataset, with the aid of auxiliary samples obtained from the target subdataset. Inspired by observation and then deduced from theoretical analysis, a regularizer is employed penalizing the difference between the global and the customized dictionary. By minimizing the sum of reconstruction errors of the above regularizer under sparsity constraints, we exploit the characteristics of the target subdataset contained in the auxiliary samples while maintaining the basic sketches stored in the global dictionary. An efficient algorithm is presented and validated with experiments on real-world data.

1. Introduction

Sparse models are of great interest in machine learning and computer vision, owing to their applications for image denoising [1], face recognition [2–4], traffic sign recognition [5], visual-tactile fusion [6, 7], and so forth. In sparse coding, samples or signals are represented as sparse linear combinations of the column vectors (called atoms) of a redundant dictionary. This dictionary can be a predefined one, such as the DCT bases and wavelets [8], or a learned one based on a specific task or dataset of interest.

With sufficient samples, learning a specialized dictionary instead of using the "off-the-shelf" one has been shown to dramatically improve the performance. Generally, the dictionary and the coefficients are estimated by minimizing the sum of least squared errors under the sparsity constraint. Batch algorithms such as MOD [9] and K-SVD [10] and nonparametric Bayesian methods [11] have shown state-of-the-art performance. Further, Mairal et al. [12] developed an online approach to handle large amounts of samples.

Recently, theoretical analysis of sparse dictionary learning has attracted much attention. Schnass [13] presented theoretical results of the dictionary identification problem. Sample complexity has been estimated in [14, 15]. Gribonval et al. [16] analyzed the local minima of dictionary learning. Moreover, to extend the capacity, dictionary learning with specific motivations [17–19] has also attracted lots of interests. For instance, robust face recognition [3] is dedicated to particular applications, and Hawe et al. [20] require the dictionary to have a separable structure. While the learned dictionary has significant effects on a given dataset, attaining further specialized dictionaries for subdatasets with fine granularity is an interesting and useful concept as well. For instance, with a dictionary corresponding to facial images of all humans, we want to gain a customized dictionary for each particular individual. However, in this case, standard dictionary learning approaches may be unwarranted or impractical: on one hand, samples for a particular individual (subject) are restricted and insufficient in most cases; on the other hand, even with enough data, learning so many

dictionaries becomes inefficient for computation and storage. We demonstrate further examples, such as customizing handwritings to different styles, matching images of flower to various species, or matching paper corpora to specific proceedings.

In terms of classification tasks, approaches such as Yang et al. [18] and Ma et al. [2] learn a structured dictionary which consists of N subdictionaries on behalf of N different subjects. However they are often unfeasible: firstly, as a part of the global dictionary, the coding performance of the subdictionary is always worse than the global one. Secondly, the subdictionaries for N subjects must be learned together, which becomes inflexible and exacting for a huge N. Thirdly, once the global dictionary is obtained, specialization for a new $(N+1)$th subdataset would be impossible.

In this paper, we are looking for an effective, economic, and flexible dictionary customization approach, which are supposed to have the following characteristics:

(i) We specialize an existing global dictionary by utilizing auxiliary samples obtained from the target subdataset, valid for finer granularity and a small quantity of examples (hence less computations).

(ii) Compared with the global one, the customized dictionary has the same size but smaller reconstruction errors and better representation of the target subdataset.

(iii) The customization for each subdataset is independent; thus we can customize an arbitrary number of subdatasets or attain a particular one alone.

As depicted in Figure 1, we first observed that the corresponding dictionary atoms of the global and the particular subjects often look "similar." This is reasonable, as the dictionary atoms describe the sketches of the object and the basic shapes of all the subjects are consistent. For a more rigorous theoretical analysis, we further considered dictionary identifiability [13] for mixed bounded signal models, that is, signals that are generated from more than one source (reference dictionary). And we proved that if reference dictionaries were close in the sense of the Frobenius norm, the global dictionary learned from mixed signals would be close to each of them. In fact, the global dictionary grabs the common basic shapes of all the subdatasets, regarding characteristics of the subjects as noise and discarding them.

Thus, formulating the dictionary customization problem, we introduced a regularizer penalizing the difference between the global and the customized dictionary. By minimizing the sum of the reconstruction error and above regularizer under sparsity constraints, we exploit the characteristic of the target subdataset contained in the auxiliary samples while maintaining the basic shapes stored in the global dictionary. As a result, a better dictionary, closer to the global one, is obtained. The solution is an asymptotic unbiased estimation of the underlying dictionary and can be seen as a trade-off between learning a new one from data and using an existing one.

To minimize the object function, we considered a general strategy the same as dictionary learning, that is, coding the samples and updating the atoms alternately in each iteration. Further, we present an algorithm that shares the idea with K-SCVD [10], which we call C-Ksvd. The flow chart of our methods is demonstrated in Figure 2. Experiments on tasks such as denoising and superresolution illustrate that our approach can handle the customization problem effectively and efficiently, outperforming both the global one and the normal dictionary learning approach. In addition, our model is also promising for more tasks such as enhancing an insufficient learned dictionary.

2. Notations

Throughout this paper, we write matrices as uppercase letters and vectors as lowercase letters. Given $p > 0$, the l_p-norm of the vector $v \in \mathbb{R}^n$ is defined as $\|v\|_p \triangleq (\sum_{i=1}^n |v_i|^p)^{1/p}$. In particular, the l_0-norm $\|v\|_0$ counts the nonzero entries of v. Let $\operatorname{sign}(v)$ denote the vector such that its jth entry $[\operatorname{sign}(v)]_j$ is equal to zero if $v_j = 0$ and to one (resp., minus one) if $v_j > 0$ (resp., $v_j < 0$).

The *Frobenius norm* of the matrix M is denoted as $\|M\|_F \triangleq [\sum_{i=1}^m \sum_{j=1}^n |m_{ij}|^2]^{1/2}$ and *matrix* l_1-norm as $\|M\|_1 \triangleq \sum_{i=1}^m \sum_{j=1}^n |m_{ij}|^1$. Define the operator norm

$$\|M\|_{p,q} \triangleq \sup_{\|x\|_p \le 1} \|Mx\|_q, \tag{1}$$

where m_i denotes the ith column vector of M.

3. Dictionary Learning with Mixed Signals

Dictionary identifiability [13], that is, recovering a reference dictionary that is assumed to generate the observed signals, is important for the interpretation of the learned atoms. In particular, Gribonval et al. [16] proved that the loss function of dictionary learning admits a local minimum in the neighborhood of the dictionary generating the signals.

In this section, we consider that there are multiple reference dictionaries and that the signals generated from them are mixed. Further, we prove that if reference dictionaries are close to each other in the sense of the Frobenius norm, dictionary learning with mixed signals admits a local minimum near both reference dictionaries simultaneously.

Without loss of generality, we analyze the case of two signal sources S_1, S_2. In particular, for the signal source S_i ($i = 1$ or 2), assume its signals $x^i \in \mathbb{R}^m$ are generated by model

$$S_i\colon\ x^i = D^i \alpha^i + \varepsilon^i, \tag{2}$$

where $D^i \in \mathbb{R}^{m \times p}$ is the reference dictionary of S_i, $\alpha^i \in \mathbb{R}^p$ is the coefficient, and $\varepsilon^i \in \mathbb{R}^p$ is the noise.

Particularly, the coefficient α^i is drawn on index set $J \subset \{1, 2, \ldots, p\}$ such that $\alpha^i_{J^c}$ is a zero vector and α^i_J is a random vector. Assume α^i_J and ε^i satisfy the following assumptions similar to [15], where we denote $\xi^i = \operatorname{sign}(\alpha^i)$.

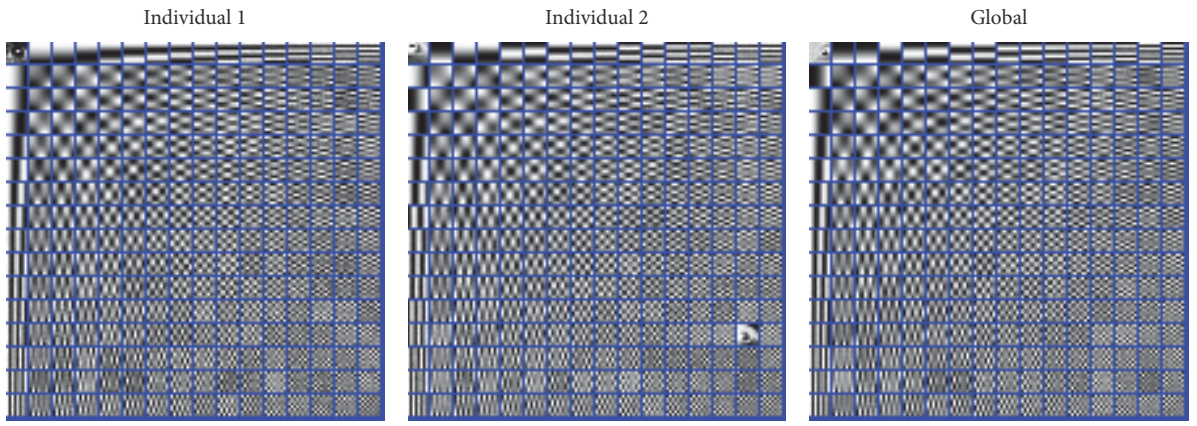

Figure 1: Three sorted dictionaries corresponding to two individual ones and the global one are demonstrated as images. Each one has a size of 64×256. The dictionaries for individuals 1 and 2 are trained from 40,000 patches picked from 24 of their corresponding facial images. The global one is trained from 80,000 patches sampled from 200 facial images belonging to 50 different individuals.

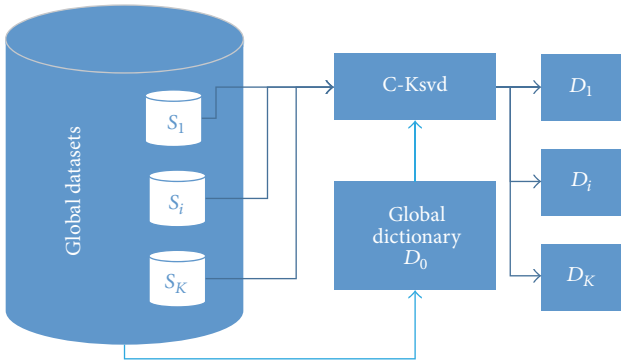

Figure 2: Flow chart of our methods.

Assumption 1 (basic and bounded signal assumption). There exist random variables α, ε, values $\underline{\alpha}$, M_α, and M_ε, such that

$$\begin{aligned}
&\mathbb{E}\left\{\alpha_J^i \left[\alpha_J^i\right]^T \mid J\right\} = \mathbb{E}\left\{\alpha^2\right\} \cdot I,\\
&\mathbb{E}\left\{\xi_J^i \left[\xi_J^i\right]^T \mid J\right\} = I,\\
&\mathbb{E}\left\{\alpha_J^i \left[\xi_J^i\right]^T \mid J\right\} = \mathbb{E}\left\{|\alpha|\right\} \cdot I,\\
&\mathbb{E}\left\{\varepsilon^i \left[\varepsilon^i\right]^T \mid J\right\} = \mathbb{E}\left\{\varepsilon^2\right\} \cdot I,\\
&\mathbb{E}\left\{\varepsilon^i \left[\alpha_J^i\right]^T \mid J\right\} = \mathbb{E}\left\{\varepsilon^i \left[\xi_J^i\right]^T \mid J\right\} = 0.\\
&\mathbb{P}\left(\min_{j\in J} \left|\alpha_j^i\right| < \underline{\alpha} \mid J\right) = 0,\\
&\mathbb{P}\left(\left\|\alpha^i\right\|_2 > M_\alpha\right) = 0,\\
&\mathbb{P}\left(\left\|\varepsilon^i\right\|_2 > M_\varepsilon\right) = 0.
\end{aligned} \tag{3}$$

$$(i = 1, 2).$$

Remark 2. Almost all sparse signal models such as k-sparse Gaussians and Laplacians satisfy the first five formulas, which can be seen as a kind of abstract and generalizing of the basic sparse signal model.

Further, the additional assumptions that the signal is upper-bounded and lower-bounded are standard and mainly used to make the analysis simple and clear [15]. In practice, as digital data is gathered with sensors with limited dynamics and stored in float format with limited precision, the boundedness assumption seems to be reasonably relevant.

The index set J is called the support of α^i and the sparsity s is defined as the number of elements in J. Thus the signal model is parameterized by the sparsity s, the expected coefficient energy $\mathbb{E}\alpha^2$, the minimum coefficient magnitude $\underline{\alpha}$, maximum norm M_α, and the flatness $\kappa_\alpha \triangleq \mathbb{E}|\alpha|/\sqrt{\mathbb{E}\alpha^2}$.

Note these assumptions can be generalized to multiple sources case easily, and thus we have the following definition.

Definition 3 (mixed bounded signal source). A mixed signal source S_O is defined as the union set of several signal sources $S_1, S_2, S_3, \ldots$; that is,

$$S_O = S_1 \cup S_2 \cup S_3 \cup \cdots, \tag{4}$$

where each source generates the signals by the way described in (2). Further, if $S_1, S_2, S_3, \ldots$ satisfy the basic and bounded signal assumptions (3) simultaneously, we say that S_O is a *mixed bounded signal source* or satisfies a *mixed bounded signal model*.

Further, for the two signal sources' case, assume D^1 and D^2 are close in the sense of Frobenius distance; that is, there is a small $\zeta \in \mathbb{R}$, s.t. $\|D^1 - D^2\|_F \leq \zeta$. (As discussed in [15], a dictionary is invariant by sign flips and permutations of the atoms, and we simply assume the atoms have been tuned to attain the minimum distance.) Denoting d_j the jth column

of D, the cumulative coherence of a dictionary D is defined as

$$\mu_k(D) \triangleq \sup_{|J|\le k}\sup_{j\notin J} \left\| D_J^T d_j \right\|_1. \tag{5}$$

The term $\mu_k(D)$ gives a measure of the level of correlation between columns of D. Moreover, the lower restricted isometry constant of a dictionary D, $\underline{\delta}(D)$, is the smallest number, for any $\alpha \in \mathbb{R}^m$, satisfying

$$\left(1-\underline{\delta}(D)\right)\|\alpha\|_2^2 \le \|D\alpha\|_2^2. \tag{6}$$

Recall that, for a set $X = [x_1,\ldots,x_n] \in \mathbb{R}^{m\times n}$, the loss function of dictionary learning is

$$F_X(D) \triangleq \frac{1}{n}\sum_{i=1}^{n}\inf_{\alpha_i\in\mathbb{R}^p}\frac{1}{2}\|x_i - D\alpha_i\|_2^2 + g(\alpha_i), \tag{7}$$

where $g(\alpha)$ is a penalty function promoting sparsity. Now consider a set of mixed signals $Y = [X_1, X_2] = [x_1^1,\ldots,x_n^1, x_{n+1}^2,\ldots,x_{2n}^2]$, where $X_1 \subset S_1$ and $X_2 \subset S_2$; the dictionary learning can be formulated as

$$\min_{D\in\mathscr{D}} F_Y(D) = \frac{1}{2}F_{X_1}(D) + \frac{1}{2}F_{X_2}(D), \tag{8}$$

where $\mathscr{D}$ denotes the set of dictionaries with unit l_2 norm atoms. Further, we have the following asymptotic result.

Theorem 4. *S_O is a mixed bounded signal source described above which consists of two signal sources S_1 and S_2. Without loss of generality, let $\|D^1\|_{2,2}^2 \ge \|D^2\|_{2,2}^2$. And assume the cumulative coherence $\mu_s(D^i)$ and the sparsity level s satisfy*

$$\mu_s\left(D^i\right) \le \frac{1}{4},$$
$$s \le \frac{p}{16\left(\left\|D^i\right\|_{2,2}+1\right)^2}. \tag{9}$$
$$(i = 1, 2).$$

Further, we define

$$C_{\min}^i \triangleq 24\kappa_\alpha^2\frac{s}{p}\left(\left\|D^i\right\|_{2,2}+1\right)\left\|\left[D^i\right]^T D^i - I\right\|_F,$$
$$C_{\min} \triangleq \max\left(C_{\min}^1, C_{\min}^2\right),$$
$$C_{\max}^i \triangleq \frac{2}{7}\cdot\frac{E|\alpha|}{M_\alpha}\left(1-2\mu_s\left(D^i\right)\right), \tag{10}$$
$$C_{\max} \triangleq \min\left(C_{\max}^1, C_{\max}^2\right).$$

And assume $C_{\min} < C_{\max}$. Define $t = \mathbb{E}\|\varepsilon\|_2^2/\mathbb{E}\|\alpha^0\|_2^2$. Moreover, let $g(\alpha) = \lambda\|\alpha\|_1$ with a regularization parameter $\lambda \le (1/4)\underline{\alpha}$ and denote $\overline{\lambda} \triangleq \lambda/E|\alpha|$, $\pi = \pi(D^1 - D^2)$, $\underline{\delta} = \max(\underline{\delta}(D^1), \underline{\delta}(D^2))$. Then there exists a radius r which satisfies $C_{\min}\overline{\lambda} + \zeta < r < C_{\max}\overline{\lambda}$, $M_\varepsilon/M_\alpha < (7/2)(C_{\max}\overline{\lambda} - r)$, and

$$1 + t < \frac{\sqrt{(1-\underline{\delta})}}{8\pi\sqrt{s}\left\|D^2\right\|_{2,2}^2}\left(\frac{r}{\zeta}-1\right)\left(r-\zeta-C_{\min}\overline{\lambda}\right), \tag{11}$$

such that the expectation of the function $F_Y(D)$ admits a local minimum $\widehat{D}$ that $\|\widehat{D} - D^1\|_F < r$, $\|\widehat{D} - D^2\|_F < r$.

Let us consider in more detail the assumptions in Theorem 4.

(i) $\mu_s(D^i) \le 1/4$ and $s \le p/16(\|D^i\|_{2,2}+1)^2$ assume upper bounds of the correlation level between columns of D^i and the sparsity s. This is common in the analysis of sparse learning [21].

(ii) The condition $C_{\min} < C_{\max}$ would be satisfied with small s/p. The smaller s/p is, the larger $C_{\max} - C_{\min}$ would be.

(iii) $\lambda \le (1/4)\underline{\alpha}$ impose an upper limit on admissible regularization parameters. Note that limits on regularization parameters are also frequent [22].

(iv) $M_\varepsilon/M_\alpha < (7/2)(C_{\max}\overline{\lambda} - r)$ requires the level of noises. In particular, noiseless situation, that is, $M_\varepsilon = 0$, is a special case. Besides, t would be particularly small; for example, if the noise level is 30 dB, then $1 + t = 1.001$.

(v) Consider $C_{\min}\overline{\lambda} + \zeta < r < C_{\max}\overline{\lambda}$ and

$$1 + t < \frac{\sqrt{(1-\underline{\delta})}}{8\pi\sqrt{s}\left\|D^2\right\|_{2,2}^2}\left(\frac{r}{\zeta}-1\right)\left(r-\zeta-C_{\min}\overline{\lambda}\right), \tag{12}$$

and we can rewrite them as

$$\zeta \le \left(C_{\max} - C_{\min}\right)\overline{\lambda},$$
$$\zeta \lesssim \frac{\sqrt{(1-\underline{\delta})}}{8\pi\sqrt{s}\left\|D^2\right\|_{2,2}^2}, \tag{13}$$

so the conditions would be satisfied for small ζ, in line with that D^1 and D^2 are close.

To conclude, the assumptions will hold for small cumulative coherence $\mu_s(D^i)$, sparsity s, noise level M_ε, dictionary distance ζ, and chosen regularization parameter λ.

Remark 5. For the radius r, it is lower-bounded by $C_{\min}$, $\overline{\lambda}$, and ζ. While ζ is fixed, if the sparsity s is particularly small, $C_{\min}\overline{\lambda}$ will be very small as well and the lower bound of r will be close to ζ. While s is fixed, and ζ tends to zero, that is, the mixed signal model degenerated into one single source case, then

$$1 + t < \frac{\sqrt{(1-\underline{\delta})}}{8\pi\sqrt{s}\left\|D^2\right\|_{2,2}^2}\left(\frac{r}{\zeta}-1\right)\left(r-\zeta-C_{\min}\overline{\lambda}\right) \tag{14}$$

will be held forever and Theorem 4 degenerated into the case in [15], implying that the discussion in [15] can be seen as a special case of ours.

Moreover, the upper bound of r is implied to be less than 0.15, which can be concluded by a discussion similar to [15].

Remark 6. Theorem 4 can be generated to n ($n > 2$) sources case easily by considering a loss

$$\min_{D\in\mathscr{D}} nF_Y(D) = F_{X_1}(D) + F_{X_2}(D) + F_{X_3}(D) + \cdots + F_{X_n}(D), \tag{15}$$

and the proof is similar.

Proof. Define the closed ball of a dictionary D with radius r as

$$\mathscr{B}(D,r) = \left\{D' \in \mathscr{D} : \left\|D' - D\right\|_F \le r\right\}. \tag{16}$$

□

Now consider D^1 and D^2, as $\|D^1 - D^2\|_F \le \zeta$ and $\zeta < \zeta + C_{\min}\overline{\lambda} < r$, and the two balls $\mathscr{B}(D^1, r)$ and $\mathscr{B}(D^2, r)$ have intersection $U = \mathscr{B}(D^1, r) \cap \mathscr{B}(D^2, r)$ with D^1, D^2 contained. Denote $\mathscr{T}$ as the boundary of $\mathscr{U}$.

Further, for a set of samples X and two dictionaries D, D', define

$$\begin{aligned} f_X(D) &\triangleq \mathbb{E}F_X(D), \\ \Delta f_X\left(D, D'\right) &\triangleq f_X(D) - f_X\left(D'\right), \\ \Delta f_X\left(\mathscr{T}, D'\right) &\triangleq \inf_{D\in\mathscr{T}} \Delta f_X\left(D, D'\right). \end{aligned} \tag{17}$$

Note that $2f_Y(D) = f_{X_1}(D) + f_{X_2}(D)$; then we have

$$\begin{aligned} 2\Delta f_Y\left(\mathscr{T}, D^1\right) &= \Delta f_{X_1}\left(\mathscr{T}, D^1\right) + \Delta f_{X_2}\left(\mathscr{T}, D^1\right) \\ &= \Delta f_{X_1}\left(\mathscr{T}, D^1\right) + \Delta f_{X_2}\left(\mathscr{T}, D^2\right) \\ &\quad - \Delta f_{X_2}\left(D^1, D^2\right). \end{aligned} \tag{18}$$

When $g(\alpha) = \lambda\|\alpha\|_1$, the function $F_X(D)$ is Lipschitz continuous with respect to Frobenius metric on the compact constraint set $\mathscr{D} \subset \mathbb{R}^{m\times p}$ [16]. Thus by choosing a radius r such that $\Delta f_Y(\mathscr{T}, D^1) > 0$, the compactness of the closed set $\mathscr{U}$ will then imply the existence of a local minimum $\widehat{D}$ of $F_Y(D)$ such that $\|\widehat{D} - D^1\|_F < r$, $\|\widehat{D} - D^2\|_F < r$. Now let us bound each item of (18).

First note that assuming $\mu_s(D^1) \le 1/4$, $s \le p/16(\|D^1\|_{2,2} + 1)^2$, $\lambda \le (1/4)\underline{\alpha}$, and $M_\varepsilon/M_\alpha < (7/2)(C_{\max}\overline{\lambda} - r)$, then, by the proof of theorem 1 in [15], for any radius $d \in (C_{\min}\overline{\lambda}, C_{\max}\overline{\lambda})$ and any dictionary D that $\|D - D^1\|_F = d$, we have

$$\Delta f_{X_1}\left(D, D^1\right) \ge \frac{\mathbb{E}\alpha^2}{8} \cdot \frac{s}{p} \cdot d\left(d - C^1_{\min}\overline{\lambda}\right) > 0. \tag{19}$$

Further, $\mathbb{E}\alpha^2/8 \cdot s/p \cdot d(d - C^1_{\min}\overline{\lambda})$ is monotonically increasing for $d > C_{\min}\overline{\lambda}$ and for $D \in \mathscr{T}$, we have $\|D - D^1\|_F \ge r - \zeta > C_{\min}\overline{\lambda}$. Thus

$$\Delta f_{X_1}\left(\mathscr{T}, D^1\right) \ge \frac{\mathbb{E}\alpha^2}{8} \cdot \frac{s}{p} \cdot (r - \zeta)\left(r - \zeta - C_{\min}\overline{\lambda}\right). \tag{20}$$

For the second item $\Delta f_{X_2}(\mathscr{T}, D^2)$, we have the same lower bound similarly.

Moreover, for the dictionary D^2 and any coefficient α with sparsity s, we have

$$\frac{1 - \underline{\delta}}{s}\|\alpha\|_1^2 \le (1 - \underline{\delta})\|\alpha\|_2^2 \le \left\|D^2\alpha\right\|_2^2. \tag{21}$$

Then by the theorem 2 and lemma 6 in [16], when $g(\alpha)$ is l_1 norm, we have

$$\left|F_{X_2}\left(D^1\right) - F_{X_2}\left(D^2\right)\right| \le L_{X_2}\left\|D^1 - D^2\right\|_{1,2}, \tag{22}$$

where $L_{X_2} = 2\sqrt{s}/(n\sqrt{1 - \underline{\delta}}) \cdot \|X_2\|_F^2$. Thus

$$\begin{aligned} &\left|F_{X_2}\left(D^1\right) - F_{X_2}\left(D^2\right)\right| \\ &\le \frac{2\sqrt{s}}{n\sqrt{1 - \underline{\delta}}} \cdot \frac{\pi}{p} \cdot \|X_2\|_F^2\left\|D^1 - D^2\right\|_F. \end{aligned} \tag{23}$$

Assume x is a sample in X_2 and its sparse coefficient and noise coefficient are α^0 and ε. As $x = D^2\alpha^0 + \varepsilon$, we have

$$x^2 = \left(D^2\alpha^0\right)^2 + \varepsilon^2 + 2\left(D^2\alpha^0\right)^T\varepsilon; \tag{24}$$

taking expectation on each side of it, by assumptions in (2), as $\mathbb{E}\|\alpha^0\|_2^2 = s\mathbb{E}\alpha^2$ and $\mathbb{E}(D^2\alpha^0)^T\varepsilon = 0$, then

$$\begin{aligned} \mathbb{E}\|x\|_2^2 &\le \left\|D^2\right\|_{2,2}^2 \mathbb{E}\left\|\alpha^0\right\|_2^2(1 + t) \\ &\le s\left\|D^2\right\|_{2,2}^2 \mathbb{E}\alpha^2(1 + t); \end{aligned} \tag{25}$$

thus $\mathbb{E}\|X_2\|_F^2 \le ns\|D^2\|_{2,2}^2\mathbb{E}\alpha^2(1 + t)$. Taking expectation on each side of (23), we have

$$\left|\Delta f_{X_2}\left(D^1, D^2\right)\right| < \frac{2s\sqrt{s}(1 + t)}{\sqrt{1 - \underline{\delta}}} \cdot \frac{\pi}{p} \cdot \mathbb{E}\alpha^2\left\|D^2\right\|_{2,2}^2 \cdot \zeta. \tag{26}$$

By (18), (20), and (26), as long as

$$1 + t < \frac{\sqrt{(1 - \underline{\delta})}}{8\pi\sqrt{s}\left\|D^2\right\|_{2,2}^2}\left(\frac{r}{\zeta} - 1\right)\left(r - \zeta - C_{\min}\overline{\lambda}\right), \tag{27}$$

we have

$$\begin{aligned} 2\Delta f_Y\left(\mathscr{T}, D^1\right) &\ge 2 \cdot \frac{s\mathbb{E}\alpha^2}{8p} \cdot (r - \zeta)\left(r - \zeta - C_{\min}\overline{\lambda}\right) \\ &\quad - \frac{2\pi s\sqrt{s}}{p\sqrt{1 - \underline{\delta}}} \cdot \mathbb{E}\alpha^2\left\|D^2\right\|_{2,2}^2 \cdot \zeta > 0, \end{aligned} \tag{28}$$

which means that $\mathbb{E}F_Y(D)$ admits a local minimum $\widehat{D}$ in $\mathscr{U}$; that is, $\|\widehat{D} - D^1\|_F < r$, $\|\widehat{D} - D^2\|_F < r$.

The result is reasonable, as when those reference dictionaries are similar, the dictionary learned from the mixed signals should be similar to each of them, in order to get less reconstruction errors for each subdataset and hence a lower total loss.

4. The Regularizer and Dictionary Customization Problem

Now we turn back to the dictionary customization problem. In particular, the dataset S_O consists of several separable subdatasets $S_A, S_B, S_C, \ldots$; that is, $S_O = S_A \cup S_B \cup S_C \cup \cdots$. Further, $D_0 \in \mathbb{R}^{m\times p}$ ($p \gg m$) is an existing global dictionary corresponding to S_O. This is common, as the dictionary for facial images would always be well trained, and the corresponding dataset can be divided by different individuals. Then we would like to customize D_0 with some auxiliary samples $X = \{x_i \mid x_i \in \mathbb{R}^m\}_{i=1}^n \subset S_A$, requiring that the customized dictionary D has the same size but behaves better on S_A.

Obviously, D should have sparse representations and small reconstruction errors on X, which corresponds to minimizing $\sum_{i=1}^n \|x_i - Dw_i\|$ under sparsity constraint $\|w_i\|_0 \le s$. Further, noting that $S_A, S_B, S_C, \ldots$ can be regarded as several signal sources and, hence, S_O as a mixed bounded model. Moreover, accounting for the fine granularity, the differences between those subdatasets $S_A, S_B, S_C, \ldots$ are small and the basic sketches of them are consistent, implying that the underlying dictionaries for all subdatasets are similar. Thus, D_0 should, according to Theorem 4, be close to our customized dictionary D as well, which is also in accordance with the practical observation. Considering the distance induced by the Frobenius norm, this leads directly to a regularizer $\|D - D_0\|_F$.

Denote $E = D - D_0$; then the customization model can be formulated as a sum of the reconstruction errors and the above regularizer; that is,

$$\begin{aligned} \arg\min_{E,W} \quad & \left\|X - (D_0 + E)W\right\|_F^2 + \gamma \|E\|_F^2, \\ \text{s.t.} \quad & \forall i,\ \|w_i\|_0 \le s, \end{aligned} \tag{29}$$

where w_i represents the ith column vector of W, $s \ll m$ is the sparsity number, and $\gamma \ge 0$ is the parameter balancing the prior knowledge of D_0 and the information in X.

It is worth noting that problem (29) is connected with the matrix version of total least squares (TLS) problems [23], which generalized the least squares by assuming noises in both dependent and independent variables. This is interpretable: as mentioned above, the atoms of the global dictionary only grab the main sketches. They regard the characteristics belonging to different subdatasets as noise and discard them. As a result, when considering a particular subdataset, the characteristic information is absent and thus the corresponding atoms of D_0 can be seen as noisy. Different from TLS, the tuning parameter γ is necessary, as noises in D_0 and X are different and should be balanced. We further depict model (29) with the following properties.

Theorem 7. *Consider customization problem (29), where $X = \{x_i\}_{i=1}^n$ is the auxiliary data, D_0 is the global dictionary, D_* is the true one corresponding to the target subdataset, and $\overline{D}$ is the customized one attained from (29); then*

(1) *denote $\mathscr{W} = \{w \in \mathbb{R}^p \mid \|w\|_0 \le s\}$; for any $\gamma \ge 0$,*

$$\inf_{w_i\in\mathscr{W}} \sum_{i=1}^n \left\|x_i - \overline{D}w_i\right\|_F \le \inf_{w_i\in\mathscr{W}} \sum_{i=1}^n \left\|x_i - D_0 w_i\right\|_F; \tag{30}$$

(2) *for a fixed γ, when n tends to infinity, $\overline{D}$ will converge to D_*; in other words, the minimizer of (8) is an asymptotically unbiased estimator of D_*;*

(3) *the tuning parameter γ reflects the confidence in D_0; in particular, if $\gamma \to \infty$, $\overline{D} = D_0$; if $\gamma = 0$, (29) will degrade into a common dictionary learning problem.*

Proof. For 1, as $\overline{D}$ is the optimal solution of problem (29), then, for any $\gamma \ge 0$ and $D \in \mathbb{R}^{m\times p}$, we have

$$\begin{aligned} &\inf_{w_i\in\mathscr{W}} \sum_{i=1}^n \left\|x_i - \overline{D}w_i\right\|_F + \gamma \left\|\overline{D} - D_0\right\|_F^2 \\ &\quad \le \inf_{w_i\in\mathscr{W}} \sum_{i=1}^n \left\|x_i - Dw_i\right\|_F + \gamma \left\|\overline{D} - D_0\right\|_F^2. \end{aligned} \tag{31}$$

Let $D = D_0$; then we have

$$\begin{aligned} \inf_{w_i\in\mathscr{W}} \sum_{i=1}^n \left\|x_i - \overline{D}w_i\right\|_F &\le \inf_{w_i\in\mathscr{W}} \sum_{i=1}^n \left\|x_i - \overline{D}w_i\right\|_F \\ &\quad + \gamma \left\|\overline{D} - D_0\right\|_F^2 \\ &\le \inf_{w_i\in\mathscr{W}} \sum_{i=1}^n \left\|x_i - D_0 w_i\right\|_F, \end{aligned} \tag{32}$$

and the equality holds only when $\overline{D} = D_0$.

For 2, reshape the loss function as

$$\arg\min_{E,\{w_i\}} \frac{1}{n}\sum_{i=1}^n \left\|x_i - (D_0 + E)w_i\right\|_2^2 + \frac{\gamma}{n}\|E\|_F^2. \tag{33}$$

When n tends to infinity, the penalty will tend to zero and thus the loss function will degenerate into the common dictionary learning form.

For 3, it is easy to see that γ reflects the weight of the penalty in the loss function and the conclusion is reasonable. □

According to the third property of Theorem 8, customization can be seen as a trade-off between learning a dictionary and using an existing one, which fills the void between them and implies a more flexible dictionary selection strategy. In particular, for datasets with coarse granularity, select dictionary learning with large amounts of samples. For subjects with fine granularity, customize the existing one with some auxiliary samples, and use a predefined dictionary if no sample is available.

We also emphasize that our model (29) will be valid as long as the assumption is satisfied (i.e., $\|D_0 - D_*\|_F \le \zeta$). As demonstrated in the experiments, there are more applications, such as improving an insufficient learned dictionary or correcting a contaminated one. In addition, for the

regularizer, more matrix norms can be selected as well. For example, consider the distance induced by *matrix* l_1-norm; then E will be sparse and the consumptions in storage and transmission will be reduced greatly.

5. Optimization

In this section, we first introduce a general optimization strategy and then devise a more straightforward dictionary updating strategy similar to K-SVD [10].

5.1. A General Strategy. A general optimization strategy, not necessarily leading to a global optimum, can be found by splitting the problem into two parts which are alternately solved within an iterative loop. The two parts are as follows.

5.1.1. Sparse Coding. Keeping E fixed, find W by

$$\begin{aligned} \min_{W} \quad & \left\| X - (D_0 + E) W \right\|_F^2 \\ \text{s.t.} \quad & \forall i, \ \left\| w_i \right\|_0 \le s. \end{aligned} \tag{34}$$

This can be solved by pursuit algorithms such as OMP [24], FOCUSS [25], or relaxed to Lasso [26].

5.1.2. Dictionary Updating. Keeping W fixed, find E by

$$\min_{E} \left\| X - (D_0 + E) W \right\|_F^2 + \gamma \left\| E \right\|_F^2. \tag{35}$$

This is a quadratic programming problem with a closed-form solution $E = (X - D_0 W) W^T (\gamma I + W W^T)^{-1}$.

5.2. C-Ksvd Algorithm. We now turn to a more involved dictionary updating strategy: rather than freezing the coefficient matrix W, we update $D = D_0 + E$ together with the nonzero coefficients (i.e., only the support is fixed).

In particular, assume that both W and E are fixed except for one column e_k in the correction matrix E and the coefficients that correspond to it, the kth row in W, denoted as w_T^k. Then the loss function can be rewritten as

$$\begin{aligned} & \left\| X - (D_0 + E) W \right\|_F^2 + \gamma \left\| E \right\|_F^2 \\ & = \left\| X - \sum_{i \ne k}^{p} \left(d_i^0 + e_i \right) w_T^i - \left(d_k^0 + e_k \right) w_T^k \right\|_F^2 \\ & \quad + \gamma \left\| e_k \right\|_2^2 + C \\ & = \left\| M_k - \left(d_k^0 + e_k \right) w_T^k \right\|_F^2 + \gamma \left\| e_k \right\|_2^2 + C, \end{aligned} \tag{36}$$

where d_k^0 is the kth column of D_0, $C = \gamma \sum_{i \ne k}^{p} \left\| e_i \right\|_2^2$ is a constant, and $M_k = X - \sum_{i \ne k}^{p} (d_i^0 + e_i) w_T^i$ represents the error when the kth dictionary atom is removed.

Now we shrink the loss function to the support of row vector w_T^k. Define δ_k as the group of indices pointing to samples $\{x_i\}$ that use the atom $d_k = d_k^0 + e_k$; that is, $\delta_k = \{i \mid 1 \le i \le p, w_T^k(i) \ne 0\}$. Further, define $\Omega_k \in \mathbb{R}^{m \times |\delta_k|}$ as ones on the $(\delta_k(i), i)$th entries and zeros elsewhere. Then problem (29) is transformed to

$$\min_{e_k, w_R^k} \left\| M_k^R - \left(d_k^0 + e_k \right) w_R^k \right\|_F^2 + \gamma \left\| e_k \right\|_2^2, \tag{37}$$

where $M_k^R = M_k \Omega_k$ and $w_R^k = w_T^k \Omega_k$. For this subproblem, we have the following results.

Theorem 8. *Suppose the largest singular value and the corresponding singular vectors of matrix* $[\sqrt{\gamma} d_k^0, M_k^R] \in \mathbb{R}^{m \times (n+1)}$ *are* σ_1, u_1, *and* v_1. v_{11} *is the first element of* v_1. *Then, the unique solution for problem (37) is*

$$\begin{aligned} e_k &= \frac{\sigma_1 v_{11}}{\sqrt{\gamma}} u_1 - d_k^0, \\ w_R^k &= \frac{d_k^T M_k^R}{\left\| d_k \right\|_2^2}, \end{aligned} \tag{38}$$

where $d_k = d_k^0 + e_k = (\sigma_1 v_{11} / \sqrt{\gamma}) u_1$.

Proof. Denote $d_k = d_k^0 + e_k$; then

$$\begin{aligned} & \left\| M_k^R - \left(d_k^0 + e_k \right) w_R^k \right\|_F^2 + \gamma \left\| e_k \right\|_2^2 \\ & = \left\| M_k^R - d_k w_R^k \right\|_F^2 + \left\| \sqrt{\gamma} \left(d_k^0 - d_k \right) \right\|_2^2 \\ & = \left\| \left[\sqrt{\gamma} d_k^0, M_k^R \right] - \left[\sqrt{\gamma} d_k, d_k w_R^k \right] \right\|_F^2. \end{aligned} \tag{39}$$

□

As $[\sqrt{\gamma} d_k, d_k w_R^k] = d_k [\sqrt{\gamma}, w_R^k]$ is the product of two vectors, its rank is one. Then problem (37) can be rewritten as

$$\min_{d_k, w_R^k} \left\| \left[\sqrt{\gamma} d_k^0, M_k^R \right] - d_k \left[\sqrt{\gamma}, w_R^k \right] \right\|_F^2. \tag{40}$$

And thus $d_k [\sqrt{\gamma}, w_R^k]$ is the best rank-one approximation of $[\sqrt{\gamma} d_k^0, M_k^R]$. By Eckart-Young-Mirsky Theorem [27], we have $d_k [\sqrt{\gamma}, w_R^k] = \sigma_1 u_1 v_1$; thus $d_k = (\sigma_1 v_{11} / \sqrt{\gamma}) u_1$, and $e_k = d_k - d_k^0$. Taking e_k back into the original problem (37), it becomes a least squares problem and we have

$$w_R^k = \frac{d_k^T M_k^R}{\left\| d_k \right\|_2^2}. \tag{41}$$

Thus, problem (37) has a closed-form solution and the main computation is top SVD of $[\sqrt{\gamma} d_k^0, M_k^R]$. In the dictionary updating stage, we can suggest minimization with respect to each column d_k (for simplicity, omitting e_k, we directly use d_k in updating) and corresponding w_T^k in sequence, forcing the support of the coefficients fixed. The complete algorithm, named "C-Ksvd", is described as Algorithm 9. Noting that while K-SVD computes the top SVD of matrix $M_k^R \in \mathbb{R}^{n \times |\delta_k|}$ for the kth column, C-Ksvd computes that of $[\sqrt{\gamma} d_k^0, M_k^R] \in \mathbb{R}^{m \times (|\delta_k| + 1)}$.

Assuming that the sparse coding stage is performed perfectly, a local minimum is guaranteed, as the loss function is guaranteed to be nonincreasing at the update step for d_k and a series of such steps ensures a monotonic reduction. Compared with the general strategy, the updating for d_k is more straightforward as it allows tuning of the values of the corresponding coefficients. In addition, each atom can have a unique parameter γ_i as well, on behalf of the confidence level of the ith atom d_i.

Algorithm 9 (C-Ksvd algorithm).

Initialization: a global dictionary D_0, samples $\{x_i\}_{i=1}^n$.

Repeat:

(i) Sparse coding stage: use any sparse recovery algorithm to compute the coefficients w_i for each sample x_i by approximating the solution of

$$\begin{aligned} \min_{\{w_i\}} \quad & \left\|x_i - (D_0 + E)\, w_i\right\|_2^2 \\ \text{s.t.} \quad & \left\|w_i\right\|_0 \le s. \end{aligned} \tag{42}$$

(ii) Dictionary updating stage: for each column $k = 1, 2, \ldots, p$ in D, update it by

(a) Compute M_k by $M_k = X - \sum_{i \neq k} d_i w_T^i$.

(b) Define the group of samples that use this atom as δ_k. Restrict M_k and w_T^k by choosing the columns corresponding to δ_k, obtain M_k^R and w_R^k.

(c) Apply top SVD decomposition to $[\sqrt{\gamma} d_k^0, M_k^R]$, obtain σ_1, u_1, v_1. Update

$$\begin{aligned} d_k &= \frac{\sigma_1 v_{11}}{\sqrt{\gamma}} u_1, \\ w_R^k &= \frac{d_k^T M_k^R}{\left\|d_k\right\|_2^2}. \end{aligned} \tag{43}$$

Until convergence (stopping rule).

Output: a better dictionary D.

6. Experiments

We first showed the effectiveness of our approach on the denoising task, with analysis of the customized dictionary and the tuning parameter γ. Further, a novel superresolution experiment was illustrated, sharing the idea of transferring knowledge from a related auxiliary data source. In addition, we conducted an experiment that enhances an insufficient learned dictionary by C-Ksvd, illustrating that our model was also valid for more tasks.

6.1. Denoising. We demonstrated the customization results by denoising tasks on facial images drown from PIE Database [28]. The denoising process was similar to [1], which included sparse coding of each patch of the noisy image. As the coding performance relied heavily on the dictionary, we could assess the dictionary by the denoising results, which were evaluated by PSNR (Peak Signal Noise Ratio).

Table 1: Denoising results (PSNR, dB) on facial images of different individuals with the noise level $\sigma = 30$. For each image, three kinds of D_0 were considered.

Individual	Original	Type of D_0	D_0	K-SVD	Customized
1	18.77	Global I	26.62	26.63	27.34
		Global II	26.37	26.62	27.12
		DCT	25.42	26.34	26.61
2	18.52	Global I	26.73	26.24	27.39
		Global II	26.39	26.06	27.10
		DCT	25.72	25.85	25.82
3	18.63	Global I	26.83	27.16	27.78
		Global II	26.41	26.82	27.31
		DCT	25.84	26.49	26.57

In particular, the noisy images were produced by adding Gaussian noises with mean zero and different standard deviation σ. The patch size and the redundant factor were set as 16×16 and 4. (We chose them for the best visual effect while similar comparisons can be attained for different value.) OMP was used for coding and atoms were accumulated until the average error passed the threshold, chosen empirically to be $\varepsilon = 1.15 \cdot \sigma$. Results corresponding to three dictionaries were compared, that is, the global dictionary D_0, the one generated by K-SVD, and the one produced by our customization approach. In K-SVD and customization, D_0 was used as initialization and the iteration number was set to 10. Moreover, three kinds of D_0 were considered, denoted as "global I", "global II," and "DCT," respectively: (1) a dictionary learned by K-SVD, with 40,000 noiseless patches picked from 100 individuals; (2) similar to (1), but learned with noisy patches ($\sigma = 20$); (3) predefined DCT (discrete cosine transform).

Each experiment was repeated 5 times and results are depicted in Table 1 and Figure 3. It was seen that customization outperformed the global dictionary and K-SVD on both PSNR and visual effects, accounting for the fact that both the common sketches in D_0 and characteristics in X had been utilized. Particularly, note that denoising by D_0 tended to be too smooth, and results by K-SVD were likely to be too rough. Regarding DCT as a suboptimal global dictionary, the results also showed that our customization is valid for a wide range of D_0. Conducted on a i7-3770 CPU and processed with the same dataset X, the average running time for K-SVD and customization were 173.34 s and 48.21 s, respectively, showing that our approach is competitive. In particular, for K-SVD, 119.31 s were used for removing identical atoms. We also display the three dictionaries as images in Figure 4, showing that the customized one was similar to the global one, while the one corresponding to K-SVD was not.

In addition, we plotted the relations of the tuning parameter γ, the average number (AN) of coefficients for patches, and the PSNR after denoising in Figure 5. It was shown that γ could be chosen as the one attaining the minimum average number of coefficients by a quick one-dimensional search.

FIGURE 3: Examples of denoising different facial images. For the parameters of three rows, D_0 was chosen as global I, II, and DCT, and $\delta = 30, 20, 25$, respectively.

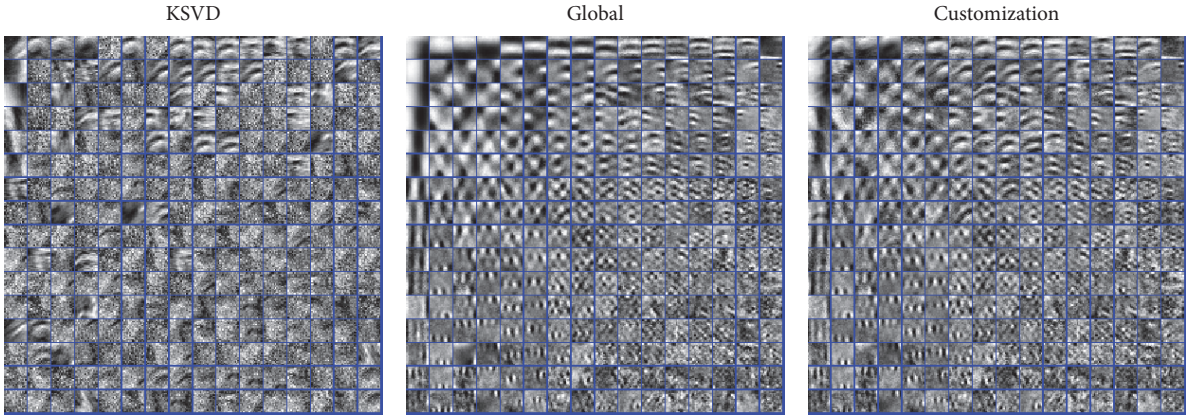

FIGURE 4: Three dictionaries were sorted and quarters at the top left corner were demonstrated while denoising a noisy image with $\sigma = 30$. For the convenience of comparison, the global dictionary learned by noisy patches is placed in the center.

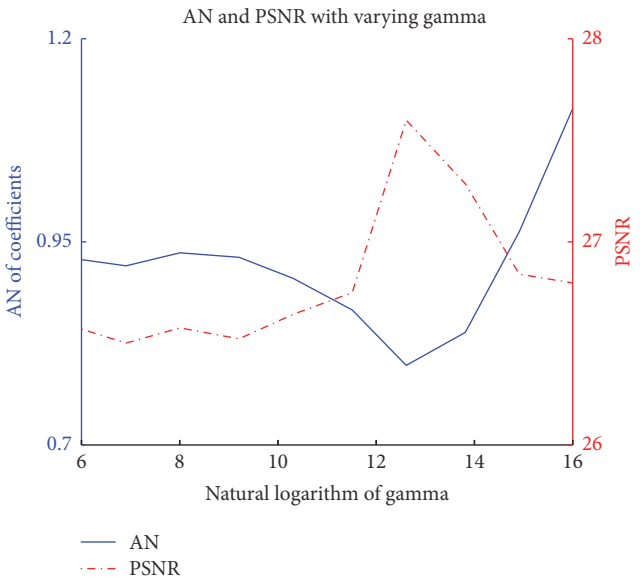

Figure 5: γ with highest PSNR was attained by the minimum average number of coefficients.

What is more, experimentally, for a fixed D_0, the best γ for different individuals was the same, which implies we only need to tune it once while customizing.

6.2. Superresolution. Yang et al. [29] proposed a scale-up algorithm via sparse signal representation, which contains two steps: dictionary learning and patch-pairs construction. To reduce the dimension and speed-up processing, Elad [30] applied PCA on the samples and used K-SVD for training. However, this learned dictionary was still a global dictionary, which means that we can further improve the performance of superresolution by customization.

Consider a global dictionary D_0 and patches $X = \{x_i\}_{i=1}^n$ sampled from related high-resolution images. We can customize this dictionary to a finer granularity. In particular, by substituting K-SVD, the low-resolution dictionary D_l and the coefficient W was customized by C-Ksvd, and the corresponding high-resolution dictionary was attained by

$$D_h = D_{h_0} + \left(X - D_{l_0} W\right) W^T \left(\gamma I + W W^T\right)^{-1}, \quad (44)$$

where D_{h_0} and D_{l_0} denoted the initial high-resolution and low-resolution dictionaries, respectively.

In this experiment, similar to the settings in [31], we evaluated the proposed approach on the Yale Face Database [32], which contains 11 different 100×100 facial images for each of the 15 individuals. A downscaled image was taken as the low-resolution object, and the down-scale factor was set to 3. Further, other images of the same individual were considered as high-resolution auxiliary data. The patch size was set to 3×3. The global dictionary D_0 was trained by 34,650 patches sampled from the 80% downscaled total dataset. (This is to highlight the relevance of auxiliary data and simulate real conditions, as the total training set is relatively small and clean in our experiment.) Results produced by D_0, K-SVD (i.e., the original version with D_0 as initialization and X as training data), and customization were compared. For customization, 225 patches were taken from each auxiliary image. For K-SVD, the total number of sampled patches was fixed to 6,000, to gain the best results.

Table 2: PSNR for superresolution on test images.

Task	Bicubic	D_0	DL3	Cus3	DL6	Cus6	DL9	Cus9
1	32.5	34.4	33.1	**35.0**	34.1	**35.4**	35.6	**35.7**
2	33.8	36.2	36.1	**36.9**	35.0	**36.8**	37.0	**37.2**
3	32.9	35.8	32.4	**36.6**	31.8	**37.0**	36.7	**37.3**
4	32.0	34.0	34.0	**34.7**	32.1	**34.9**	33.8	**35.1**
5	36.2	38.1	37.6	**38.7**	36.7	**38.9**	38.7	**38.9**

Varying the number of the auxiliary images and repeating the experiments on different individuals, the performance was evaluated by PSNR. Some of the results are summarized in Table 2. "DL" and "Cus" represent K-SVD and customization, respectively, and "3," "6," or "9" denote the number of auxiliary images. "Bicubic," that is, simple Bicubic interpolation, is shown as a baseline method.

It is seen that if the number of auxiliaries is small, results produced by K-SVD are worse than the common dictionary, implying that the learning is meaningless. However, even when the auxiliary data is small (675 patches from 3 images), superresolutions by customization have significant improvements. Further, customization still outperforms or is no worse than the learning approach when new data is added. Remember that the number of patches which K-SVD needs is much larger than that needed for customization, meaning more computations and time are required. Also note that once the dictionary has been customized, it is valid for all the images of the person.

6.3. Enhancing. As was mentioned above, model (29) can be applied to more tasks, as long as the assumption that D_0 and the reference dictionary D_* are close. In this subsection, we consider the case of enhancing an existing dictionary by C-Ksvd and evaluate the performance on classification.

In particular, LC-KSVD [33], one of the state-of-the-art methods for image classification, introduced a triple model (D, A, W), where D represents the dictionary, A stands for parameters of the label consistent term, and W denotes the linear classifier. Regarding $(D^T, \sqrt{\alpha}A^T, \sqrt{\beta}W^T)^T$ as a new dictionary, the following object function can be solved by K-SVD:

$$\begin{aligned} \arg\min_{D,W,A,X} \quad & \|Y - DX\|_2^2 + \alpha \|Q - AX\|_2^2 \\ & + \beta \|H - WX\|_2^2, \\ \text{s.t.} \quad & \forall i, \ \|x_i\|_0 \le T. \end{aligned} \quad (45)$$

Sometimes the model $M_0 = (D_0, A_0, W_0)$ learned is likely not good enough, due to the fact that the training data may be insufficient or too noisy. Moreover, over time the past training information often becomes unavailable. In this case, we can further enhance it by our customization model, simply replacing the K-SVD procedure with C-Ksvd.

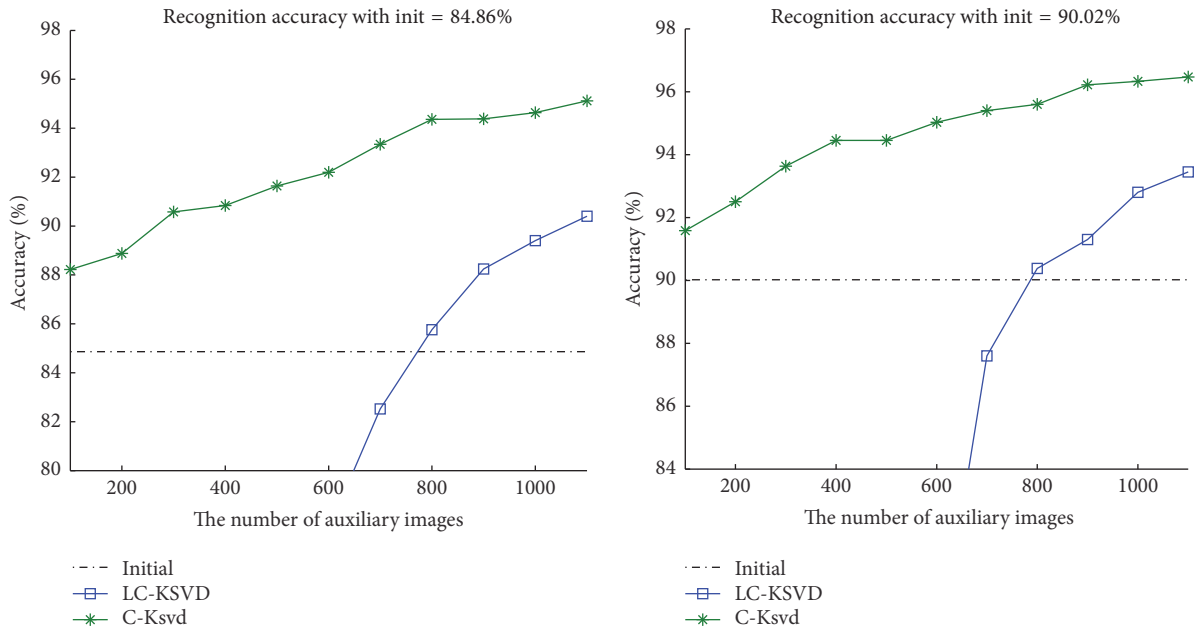

Figure 6: Varying the number of auxiliary images with fixed initial models.

Table 3: Accuracy (%) for different level initial and fixed number auxiliary images.

Init	76.76	79.32	84.86	88.04	90.41	93.30
LC-KSVD	85.32	85.36	85.76	88.69	90.23	90.83
C-Ksvd	**91.78**	**92.20**	**94.36**	**94.94**	**95.79**	**97.38**

Table 4: Accuracy (%) on enhanced classes, remainder classes, and all classes.

Classes	Init	LC-KSVD	C-Ksvd
Enhanced	83.90	92.05	**93.40**
Remainder	84.86	0.0	**82.83**
All	84.38	46.02	**88.11**

In accordance with [33], we used the "Extended Yale-B" dataset [34] to demonstrate that the performance and the data were divided into three parts: training data to obtain the initial model, auxiliary data for model enhancement, and test data to evaluate it. Parameters α and β were tuned while training the initial models and then kept fixed. The initial model M_0, LC-KSVD, and C-Ksvd were compared, where LC-KSVD used M_0 as initialization and X as training data. Results were analyzed in three ways.

(1) Initial models of different levels were obtained by tuning the number of training samples, and we tried to promote the model with 800 auxiliary images. After repeating the experiment 5 times for each level, the averaged recognition accuracies are summarized in Table 3.

It is seen that C-Ksvd is valid in a wide range of initial models and always significantly outperforms LC-KSVD. Besides, influences of the initial models on LC-KSVD are relatively small, in accordance with our previous analysis.

(2) For fixed initial models, varying the number of auxiliary images from 100 to 1100, we plotted the corresponding recognition results in Figure 6 and found that the accuracy had a significant increase, even when the auxiliary number was relatively small. To gain a competitive result for LC-KSVD, large amounts of images were required, which was unaffordable.

(3) In the previous discussion, the auxiliary images were uniformly sampled from all 38 individuals. Then we considered the nonuniform case where only images of several classes (named "enhanced classes") were available. After setting the number of enhanced classes as 19, and getting 31 images from each class, the results are reported in Table 4.

While C-Ksvd improved the accuracy on enhanced classes, the accuracy on the remainder was slightly reduced, owing to the similarity of the original and new dictionaries. LC-KSVD presented a sharp contrast.

7. Conclusion

In this paper, we considered the dictionary customization problem, which can be seen as a trade-off between learning a new dictionary from data and using an existing one. We investigated our hypothesis with theoretical analysis and formulated a model by raising a specific regularizer. An efficient algorithm was proposed, and experiments on real-world data demonstrate that our approach is promising.

Competing Interests

The authors declare that they have no competing interests.

References

[1] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[2] L. Ma, C. Wang, B. Xiao, and W. Zhou, "Sparse representation for face recognition based on discriminative low-rank dictionary learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2586–2593, Providence, RI, USA, June 2012.

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[4] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2691–2698, June 2010.

[5] H. Liu, Y. Liu, and F. Sun, "Robust exemplar extraction using structured sparse coding," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 26, no. 8, pp. 1816–1821, 2015.

[6] H. Liu, D. Guo, and F. Sun, "Object recognition using tactile measurements: kernel sparse coding methods," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 3, pp. 656–665, 2016.

[7] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual-tactile fusion for object recognition," *IEEE Transactions on Automation Science and Engineering*, 2016.

[8] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1999.

[9] K. Engan, S. O. Aase, and J. H. Husoy, "Method of Optimal Directions for frame design," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*, vol. 5, pp. 2443–2446, March 1999.

[10] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[11] M. Zhou, H. Chen, J. Paisley et al., "Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 130–144, 2012.

[12] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th International Conference on Machine Learning (ICML '09)*, pp. 689–696, ACM, Montreal, Canada, June 2009.

[13] K. Schnass, "Local identification of overcomplete dictionaries," https://arxiv.org/abs/1401.6354.

[14] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD," *Applied and Computational Harmonic Analysis*, vol. 37, no. 3, pp. 464–491, 2014.

[15] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert, "Sample complexity of dictionary learning and other matrix factorizations," https://arxiv.org/abs/1312.3790.

[16] R. Gribonval, R. Jenatton, and F. Bach, "Sparse and spurious: dictionary learning with noise and outliers," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6298–6319, 2015.

[17] C. Lu, J. Shi, and J. Jia, "Online robust dictionary learning," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 415–422, Portland, Ore, USA, June 2013.

[18] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 543–550, Barcelona, Spain, November 2011.

[19] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Proceedings of the Advances in Neural Information Processing Systems 22 (NIPS '09)*, pp. 1033–1040, 2009.

[20] S. Hawe, M. Seibert, and M. Kleinsteuber, "Separable dictionary learning," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 438–445, IEEE, Portland, Ore, USA, June 2013.

[21] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[22] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," in *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS '09)*, pp. 1348–1356, Vancouver, Canada, December 2009.

[23] I. Markovsky and S. V. Huffel, "Overview of total least-squares methods," *Signal Processing*, vol. 87, no. 10, pp. 2283–2302, 2007.

[24] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proceedings of the 27th Asilomar Conference on Signals, Systems & Computers*, pp. 40–44, IEEE, Pacific Grove, Calif, USA, November 1993.

[25] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 600–616, 1997.

[26] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 58, no. 1, pp. 267–288, 1996.

[27] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[28] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition (FGR '02)*, pp. 46–51, IEEE, 2002.

[29] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[30] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, New York, NY, USA, 2010.

[31] Y. Guo, "Robust transfer principal component analysis with rank constraints," in *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS '13)*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., pp. 1151–1159, 2013.

[32] A. Georghiades, *Yale Face Database*, Center for Computational Vision and Control at Yale University, 1997.

[33] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proceedings of the 2011 IEEE Conference on Computer Vision and*

Pattern Recognition (CVPR '11), pp. 1697–1704, IEEE, Colorado Springs, Colo, USA, June 2011.

[34] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.