IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 25, 2020, accepted September 7, 2020, date of publication September 18, 2020,
date of current version September 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024809

Data Restoration by Linear Estimation of the Principal Components From Lossy Data

YONGGEOL LEE[1] AND SANG-IL CHOI[2], (Member, IEEE)
[1]Police Science Institute, Korean National Police University, Asan 31539, South Korea
[2]Department of Computer Science and Engineering, Dankook University, Yongin 16890, South Korea

Corresponding author: Sang-Il Choi (choisi@dankook.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) under Grant 2018R1A2B6001400, and in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program, supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP), under Grant IITP-2020-2015-0-00363.

ABSTRACT In this article, we propose a method based on principal component analysis (PCA) to restore data after the occurrence of data loss due to sensor defects or environmental factors. In the L2-PCA feature space, the feature vector, which consists of principal components of the data, converges to a point known as the ‘‘convergence point’’ as the extent of data loss increases. Using these characteristics, we approximately linearly estimated the principal components of the original data from the feature vectors of the lossy data. The estimated principal components are used as coefficients in the linear combination of the projection vectors of the PCA feature space for data restoration. The restoration performance of the proposed method is not only superior; the method is also computationally more efficient than other data restoration methods. Experimental results for gas measurement data and facial image data confirm the excellent data restoration performance of the proposed method.

INDEX TERMS Data restoration, principal components, lossy data, approximately linear estimation, feature space, convergence point.

I. INTRODUCTION

A variety of methods for the analysis of sensor data [1]–[4] and the extraction of meaningful patterns from these data have been proposed in recent decades [5]. Data collected by various sensors such as image, voice, electromyography (EMG) and chemical sensors are used for different applications such as image recognition [6]–[8], speech recognition [9], [10], gesture recognition [11]–[14] and gas classification [15]–[20].

The performance of classification techniques using sensor data varies greatly depending not only on the amount of data collected but also on the quality of the data. In particular, when data are collected in a real environment rather than in a well-controlled laboratory environment, some of the data values may be lost owing to defects in the sensor itself or because of environmental variables. For example, data acquisition may either be temporarily interrupted as a result of unstable power supply, or dead pixels, in which a specific pixel value becomes 0 because of a defective image sensor, may be generated. These data, which are regarded as outliers compared to normal data, cause performance degradation when data are represented or recognized [18].

To overcome this problem, statistical methods have been proposed to restore the lossy data. For example, methods that identify the feature space that best represents the given data and project lossy data into the space to obtain the feature values were proposed [21], [22]. Then, the lossy data values are restored by employing a linear combination of the projection vectors using the feature values as the weights. Specifically, the projection vectors and features are extracted [21] to minimize the L2-Norm-based error between the original data and the data sample that was reconstructed using conventional principal component analysis (L2-PCA). However, this data reconstruction method is sensitive to outliers [23], although it minimizes the reconstruction error from the viewpoint of the mean squared error. This is because when the covariance matrix is calculated in the process of obtaining the projection vectors of L2-PCA, the outliers are squared, which excessively affects the covariance values. Other studies [22], [24] conducted PCA based on the L1-norm (L1-PCA) instead of the L2-norm. Although data

The associate editor coordinating the review of this manuscript and approving it for publication was Nadeem Iqbal.

reconstruction methods using L1-PCA are less susceptible to outliers than L2-PCA and show reliable restoration performance, they are computationally costly. To reduce the computational burden, PCA-L1 [22] that maximizes the L1-norm was proposed. The robust PCA [26] presented a method to recover the low-rank of the data matrix using convex optimization when a data matrix corrupted by noise or loss is provided. In other words, robust PCA aims to obtain the true low rank, i.e., projection matrices $\mathbf{W}$ and $\mathbf{V}$ from contaminated data. Another study [25] led to the development of an iteratively re-weighted fitting (IRF) strategy to repeatedly update feature values in the PCA feature space. The advantage of this approach is that the features are updated such that the reconstruction error of the data is reduced. Consequently, IRF improves the restoration performance rather than using the PCA feature values as they are; however, this also results in the iterative process becoming more computationally intensive.

In this article, we propose a new data restoration method that approximately linearly estimates the principal components of the original lossless data from the principal components of the lossy data. When a lossy data sample is projected onto the L2-PCA feature space, the data samples gradually approaches to one point of the feature space as the loss increases. Because all of the components in the data become "0" when loss occurs for all times or intervals, each data is converged into a single point in feature space regardless of the feature space. Regardless of the unique properties or information of the data in the event of loss, it all becomes the same zero-vector. Therefore all data samples converge to one point, so called "convergence point (CP)", in the feature space.

Therefore, if the loss rate of a given sample of data is known, the feature vector of lossless data can be approximately linearly estimated from the point at which data loss occurred, based on the straight line between the convergence point and the feature vector of the lossless data. The estimated feature values of the original data are used to restore the data as weights to the linear combination of the projection vectors that are the basis of the L2-PCA feature space (Fig. 1).

The restoration performance of the proposed method is not only high compared to that of existing methods, it also performs data restoration in a computationally simple manner. The experiments were conducted on three datasets to evaluate the restoration performance of the proposed method: the first comprised a collection of facial images [27], the second comprised gas data captured by an electronic nose [15], and the third comprised the occupancy rates [28] collected from car lanes on freeways. Our experimental results showed that the proposed method restores data more efficiently and accurately than other PCA-based reconstruction methods.

This article is organized as follows. Section II describes the PCA methods used for data restoration. Section III explains the relationship between the principal components of the original data and the lossy data, and presents the data restoration by estimating the principal components of the original data from the lossy data. The experimental results on data restoration and data classification are described in Section IV. The discussions and conclusions follow in Section V.

II. RELATED WORK

Several methods can be used to extract the principal components from a given data matrix [21], [22], [24]. Principal component analysis (PCA) [29], [30] is a statistical method based on multivariate analysis. PCA finds the projection vectors ($\mathbf{w}_t = [w_{t1}, w_{t2}, \ldots, w_{tn}]^T, t = 1, \ldots, n'$) to construct the feature space that best represents the data. Let us consider the set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of n-dimensional data samples $\mathbf{x}_k$s. Each data sample $\mathbf{x}_k = [x_{k1}, x_{k2}, \ldots, x_{kn}]^T$ can be represented by a linear combination of projection vectors $\mathbf{w}_t$s with principal components y_{kt}s in the PCA feature space. L2-PCA [21] and L1-PCA [24] define their objective function using the L2-norm and L1-norm, respectively, as follows.

$$\begin{aligned} \mathbf{W}_{PCA}^{L2} &= \arg\min_{\mathbf{W}} \sum_{k=1}^{N} ||\mathbf{x}_k - \sum_{t=1}^{n'} y_{kt}\mathbf{w}_t + \mathbf{m}||_2^2 \\ \mathbf{W}_{PCA}^{L1} &= \arg\min_{\mathbf{W}} \sum_{k=1}^{N} ||\mathbf{x}_k - \sum_{t=1}^{n'} y_{kt}\mathbf{w}_t + \mathbf{m}||_1, \end{aligned} \tag{1}$$

where $\mathbf{m} = [m_1, m_2, \ldots, m_n]^T$ is the mean of X. By solving the above objective functions, each method obtains the projection matrices $\mathbf{W}_{PCA}^{L2} = [\mathbf{w}_1^{L2}, \ldots, \mathbf{w}_{n'}^{L2}]$ and $\mathbf{W}_{PCA}^{L1} = [\mathbf{w}_1^{L1}, \ldots, \mathbf{w}_{n'}^{L1}]$, respectively, and the projection vectors constituting each projection matrix are the basis of their feature spaces. PCA-L1 [22] was used with the aim to maximize the L1 dispersion using the L1-norm in the feature space to obtain a subspace, which is not only robust to outliers and invariant to rotation, by using the following objective function.

$$\begin{aligned} \mathbf{W}_{PCA}^{L1'} = \ & \arg\max_{\mathbf{W}} \sum_{k=1}^{N}\sum_{t=1}^{n'}\sum_{i=1}^{n} |y_{kt}w_{ti}(x_{ki} - m_i)| \\ & \text{subject to } \ \mathbf{W}^T\mathbf{W} = I \end{aligned} \tag{2}$$

The process of finding the solution to the above equation is described in [22].

A feature vector $\mathbf{y}$ composed of principal components for a given data sample $\mathbf{x}$ can be represented by using $\mathbf{W}_{PCA}$, which is obtained by solving the objective function, as $\mathbf{y} = \mathbf{W}_{PCA}^T(\mathbf{x} - \mathbf{m})$.

III. PROPOSED METHOD

We first constructed the L2-PCA feature space from the original data unaffected by loss. The restoration was carried out by approximately linearly estimating the principal components of lossless data from the principal components of lossy data.

A. RELATIONSHIP BETWEEN LOSSY DATA AND THEIR PRINCIPAL COMPONENTS

Let $\mathbf{W}_{PCA}^{L2}$ and $\mathbf{V}_{PCA}^{L2}$ be the projection matrix obtained by L2-PCA from normal training data without loss and its transpose matrix, respectively. Then, the feature vector $\mathbf{y}$ for input

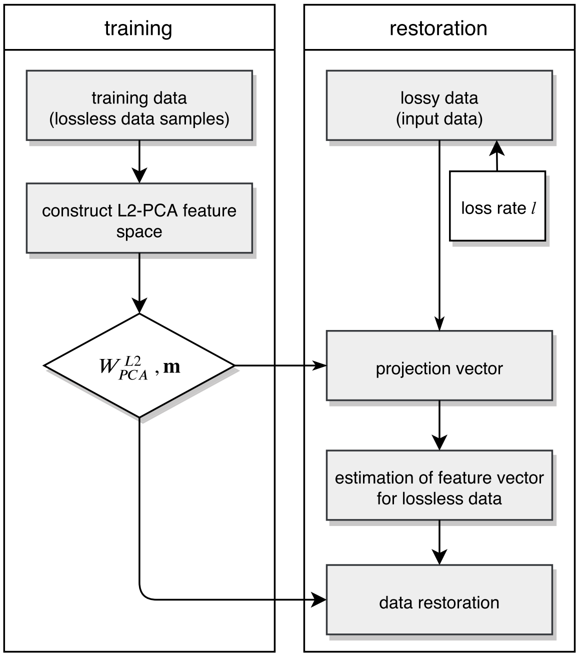
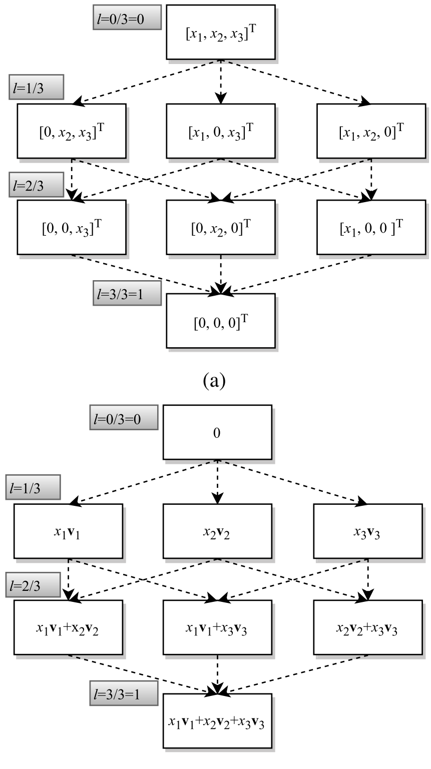


FIGURE 1. Overall procedure of the proposed method.



(a)

(b)

FIGURE 2. (a) Process of generating lossy data according to the loss rate. (b) error between lossy and lossless data in the feature space.

data $\mathbf{x}$ can be rewritten as follows.

$$\begin{aligned}\mathbf{y} &= (\mathbf{W}_{PCA}^{L2})^T(\mathbf{x}-\mathbf{m}) = \mathbf{V}_{PCA}^{L2}(\mathbf{x}-\mathbf{m})\\ &= \begin{bmatrix} v_{11}x_1+\cdots+v_{1n}x_n\\ v_{21}x_1+\cdots+v_{2n}x_n\\ \vdots\\ v_{n'1}x_1+\cdots+v_{n'n}x_n\end{bmatrix} - \mathbf{V}_{PCA}^{L2}\mathbf{m}\\ &= x_1\begin{bmatrix} v_{11}\\ v_{21}\\ \vdots\\ v_{n'1}\end{bmatrix}+\cdots+x_n\begin{bmatrix} v_{1n}\\ v_{2n}\\ \vdots\\ v_{n'n}\end{bmatrix} - \mathbf{V}_{PCA}^{L2}\mathbf{m}\\ &= x_1\mathbf{v}_1+\cdots+x_n\mathbf{v}_n - \mathbf{V}_{PCA}^{L2}\mathbf{m} \end{aligned} \tag{3}$$

In (3), each element x_i of the input data is a coefficient of the column vector $\mathbf{v}_i$. If the i-th element of the input data is lost, the contribution of $\mathbf{v}_i$ is 0 in determining $\mathbf{y}$, and thus the difference between the feature vectors ($\mathbf{y}$ and $\bar{\mathbf{y}}$) of the original data without loss ($\mathbf{x}$) and the lossy data $\bar{\mathbf{x}}$ is $x_i\mathbf{v}_i$. In the worst case, if all elements of the n-dimensional data sample are lost, the difference in the feature space is $\sum_{i=1}^{n} x_i\mathbf{v}_i$.

Let us define a flag vector $\mathcal{L}(k) \in R^{n\times 1}$, of which the element $\mathcal{L}(k)_i$ indicates whether the i-th element of the k-th data sample $\mathbf{x}_k$ is as follows.

$$\mathcal{L}(k)_i = \begin{cases} 1, & \text{if loss occurred at } x_{ki}\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

For a given n-dimensional data sample, if a loss occurs in r elements, the loss rate l for this data sample becomes $l = \frac{\sum_{i=1}^{n}\mathcal{L}(k)_i}{n} = \frac{r}{n}$. For a single data sample, because the number of cases of loss occurrence at the same loss rate is $N_l = {}_nC_r = n!/r!(n-r)!$, the average error ($\phi(\bar{e}|l)$) between the lossy data and the lossless data in the feature space for a given loss rate l is calculated as

$$\phi(\bar{e}|l) = 1/N_l\sum_{k=1}^{N_l}\sum_{i=1}^{n} x_{ki}\mathbf{v}_i\mathcal{L}(k)_i \tag{5}$$

On the other hand, the process of loss occurring from data is characterized by a relationship between data samples in which r elements are lost and those in which $r-1$ elements are lost. For example, let us consider three-dimensional lossless data $\mathbf{x}$ (Fig. 2). For loss rate $l = 1/3$, it is possible to generate three types of loss data ($[0, x_2, x_3]^T$, $[x_1, 0, x_3]^T$, $[x_1, x_2, 0]^T$), whereas another three types of loss data ($[x_1, 0, 0]^T$, $[0, x_2, 0]^T$, $[0, 0, x_3]^T$) can be generated for loss rate $l = 2/3$. In Fig. 2(a), $[0, 0, x_3]^T$ with a loss rate of 2/3 result from $[0, x_2, x_3]^T$ and $[x_1, 0, x_3]^T$ with a loss rate of 1/3, but loss cannot occur from $[x_1, x_2, 0]^T$. This means that the type of data sample with $l = 2/3$ is dependent upon the previous state of $l = 1/3$. Fig. 2(b) shows the error between the lossless data and the lossy data in the feature space according to the loss rate l. If the process in Fig. 2(b) is extended to the case of n-dimensional data samples, the average error for loss rate l in (5) can be expressed as follows.

$$\phi(\bar{e}|l) = l\sum_{i=1}^{n} x_i\mathbf{v}_i \tag{6}$$

In (6), as l increases, the average error increases proportionally, and when all the data values are lost ($l = 1$), the data samples always converge to one point (the convergence point, CP) in the feature space. In other words, in the PCA feature space, as the input data becomes increasingly lossy, the feature vector of the lossy data approximately linearly approaches the CP.

This is demonstrated by using the following toy example. We generated 10 samples of 1,000-dimensional data and plotted the samples in the L2-PCA feature space composed of two dominant projection vectors (Fig. 3). In Fig. 3, the color of a point represents the type of data sample, and the shape of a point represents the different loss rate of each lossy data value(square→ diamond→ circle). As shown in Fig. 3, regardless of the type of data, as the loss rate increases, the samples gradually converge to the CP.

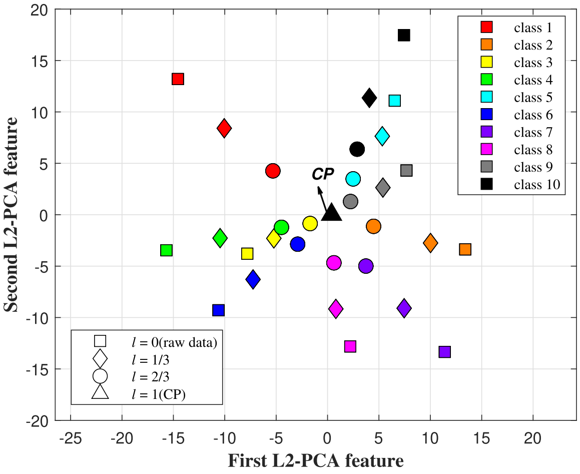


FIGURE 3. Toy example for convergence of feature vectors as the loss rate (*l*) increases.

B. DATA RESTORATION BY PRINCIPAL COMPONENT ESTIMATION OF LOSSLESS DATA

The solution ($\mathbf{W}_{PCA}^{L2}$) for the objective function of L2-PCA in (1) can be obtained by singular value decomposition (SVD) [31] on the covariance matrix of the training data samples $\mathbf{C}_{tr} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T$, ($\mathbf{C}_{tr} \in R^{n\times n}$), where $\mathbf{m}$ is the mean of the training samples. The projection matrix $\mathbf{W}_{PCA}^{L2}$ consists of projection vectors ($\mathbf{w}_t$), which are the eigenvectors of $\mathbf{C}_{tr}$. Then, the feature vector $\bar{\mathbf{y}}$ composed of the principal components y_t, $t = 1, \ldots, n'$ for the lossy sample $\bar{\mathbf{x}}$ is $(\mathbf{W}_{PCA}^{L2})^T(\bar{\mathbf{x}} - \mathbf{m})$.

On the other hand, as previously mentioned, the feature vector ($\bar{\mathbf{y}}$) for the data sample ($\bar{\mathbf{x}}$) with the loss rate l in the L2-PCA feature space is located on the straight line that connects the feature vector of the lossless data ($\mathbf{y}$) with the CP. Thus, when the loss rate for $\bar{\mathbf{x}}$ is known, the feature vector of the original data can be estimated approximately linearly from $\bar{\mathbf{y}}$ by using the proportional equation. As the loss rate l increases, $\bar{\mathbf{y}}$ gradually approaches CP, the estimation ($\mathbf{y}^+$) for $\mathbf{y}$ can be calculated by multiplying the reciprocal of $(1 - l)$ by $\bar{\mathbf{y}} - \text{CP}$ as follows.

$$\mathbf{y}^+ = \frac{(\bar{\mathbf{y}} - \text{CP})}{\epsilon(1-l)} + \text{CP} \tag{7}$$

Here, ϵ is a regularization term to prevent the L2-PCA feature space from over-fitting to the training data.

The restored data $\mathbf{x}^+$ for $\bar{\mathbf{x}}$ can be obtained by the following linear combination using the feature vector $\mathbf{y}^+$ estimated by (7) and the n' projection vectors ($\mathbf{w}_t$, $t = 1, \ldots, n'$) constituting the L2-PCA feature space.

$$\mathbf{x}^+ = \sum_{t=1}^{n'} \mathbf{y}_t^+ \mathbf{w}_t + \mathbf{m} \tag{8}$$

Here, we set the value of n' as $N - 1$.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

We demonstrated the effectiveness of the proposed method (LEPC) by performing data restoration experiments on the aforementioned three datasets containing facial image data [27], electronic nose (E-nose) data [15] and PEMS dataset [28] (Table 1). In addition, we compared the results of the proposed method ($\mathbf{x}_{LEPC}^+$) with those of other data restoration methods using L2-PCA ($\mathbf{x}_{L2PCA}^+$) [17] PCA-L1 ($\mathbf{x}_{PCAL1}^+$) [22], and PCAL1+IRF ($\mathbf{x}_{L1+IRF}^+$) [18].

TABLE 1. Characteristics of the datasets used in the experiments.

	AR	E-nose	PEMS
Dimension of data	6,400 (80 × 80)	32,000 (2, 000 × 16)	138,672 (24 × 6 × 963)
No. of subject	118	8	7
No. of training data	826	80	267
No. of test data	118	80	173

We evaluated the data restoration performance of each method by measuring the root mean squared (RMS) error [32], [33] and the peak signal-to-noise ratio (PSNR) from the reconstructed sample of lossy data and the original lossless data sample. In addition, we showed the effectiveness of the proposed method indirectly by conducting classification experiments on the reconstructed data samples. This was achieved by extracting discriminant features for classification using the discriminant common vector (DCV) [34] and the one nearest neighbor rule was used as a classifier with the $l2$ norm as the distance metric.

A. AR FACIAL IMAGE DATASET

The AR face database [27] consists of over 4,000 frontal images of 126 subjects differentiated by facial variations such as illumination, expression, and occlusion. We chose a subset of the database consisting of facial images of 64 male subjects and 54 female subjects. Among these, images without partial occlusion were used in the experiments. The center of each eye was manually located and the eyes were rotated to be aligned horizontally as in [35], [36]. Each facial image was cropped and rescaled such that the center of each eye was placed at its fixed point in an image of 80 × 80 (pixels) [35].

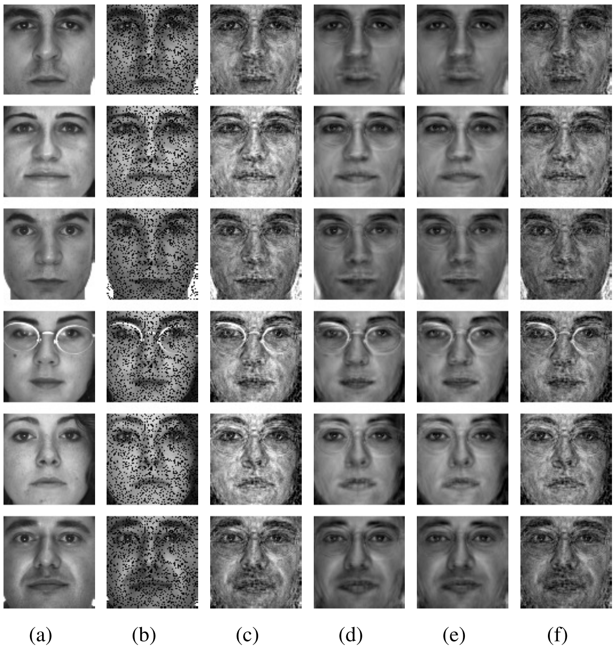


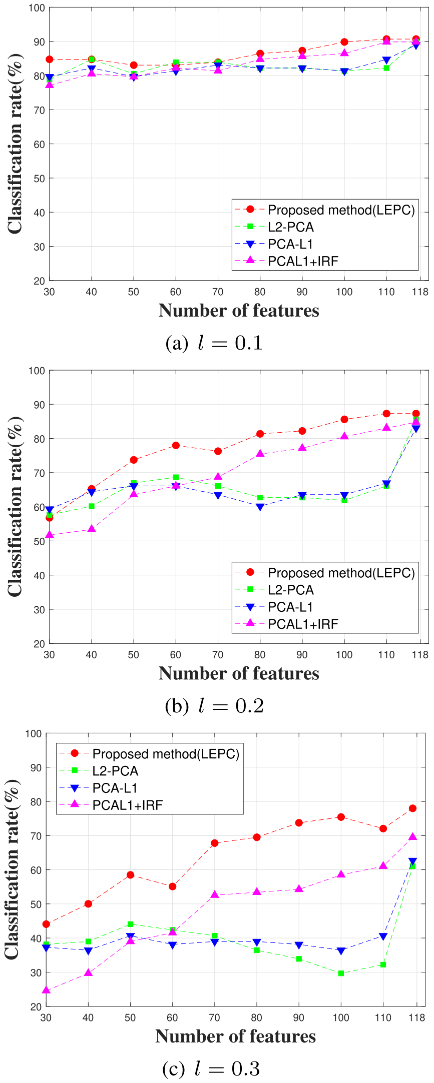
FIGURE 4. AR facial images. (a) original lossless data; (b) lossy data ($l = 0.2$); (c) restored data by using the proposed method; (d) restored data by using L2-PCA; (e) restored data by using PCA-L1; (f) restored data by using PCAL1+IRF.

TABLE 2. RMS error and PSNR between the restored data and original data (AR facial image).

l	measure	LEPC	L2-PCA	PCA-L1	PCAL1+IRF
0.1	RMS	5.67	5.57	**5.56**	6.55
	PSNR	22.58	22.75	**22.76**	21.32
0.2	RMS	**7.99**	9.09	9.07	10.39
	PSNR	**19.60**	18.47	18.49	17.31
0.3	RMS	**10.10**	12.79	12.80	14.12
	PSNR	**17.56**	15.51	15.51	14.65

The training set consisted of 826 images with variations in lighting and facial expression. The other 118 facial images that were not included in the training set were used for testing. We randomly selected 25% of all the pixels in each test image to create lossy data by setting the values of these pixels to 0, after which the proposed method was used to restore the data ($\epsilon = 0.7$).

Fig. 4 shows the original images, images with loss ($l = 0.2$), and the images restored by using several methods. In Fig. 4, the images reconstructed by other methods ($\mathbf{x}^{+}_{L2PCA}$, $\mathbf{x}^{+}_{PCAL1}$, and $\mathbf{x}^{+}_{L1+IRF}$) seem to have a cleaner skin texture than the images obtained by the proposed method ($\mathbf{x}^{+}_{LEPC}$). However, the shapes of the facial components such as the eyes, nose, and mouth that reflect the characteristics of individual faces are more clearly preserved when using $\mathbf{x}^{+}_{LEPC}$. Table 2 shows the RMS error and PSNR between the original and the restored data. In Table 2, when the loss rate is small ($l = 0.1$), all methods show similar performance in RMS error and PSNR. However, as the loss rate increases ($l = 0.2$ and $l = 0.3$), LEPC performed better than the other methods.



(a) $l = 0.1$

(b) $l = 0.2$

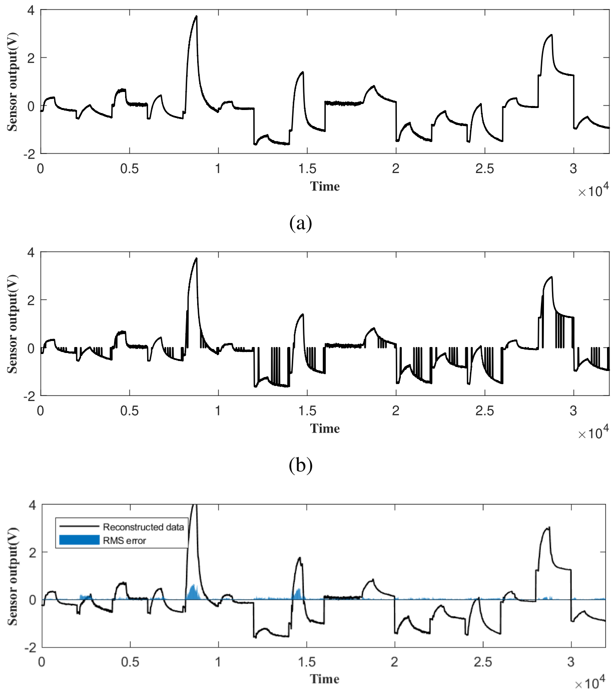
(c) $l = 0.3$

FIGURE 5. Comparison of classification rates between the proposed method and other methods (AR facial image).

Fig. 5 shows the classification rates corresponding to each dimension of DCV feature space for $\mathbf{x}^{+}_{L2PCA}$, $\mathbf{x}^{+}_{PCAL1}$, $\mathbf{x}^{+}_{L1+IRF}$, and $\mathbf{x}^{+}_{LEPC}$. As shown in Fig. 5, the classification rate of the data restored by the proposed method was higher than that of the other methods. Similar to the results in Table 2, as the loss rate increases, LEPC showed better classification performance than the other methods.

B. ELECTRONIC NOSE DATASET

The E-nose dataset [15] contains measurements of eight different gases. More specifically, the dataset consists of a total of 160 gas data samples, i.e., 20 samples for each type of gas (Table 1). Each sample comprises measurements of 2,000 points collected within 200 seconds at a sampling frequency of 10 Hz. The measurements collected from all 16 channels were stored in the form of a $2,000 \times 16$ matrix and were then transformed into a vector in 32,000-dimensional spaces by using a lexicographic ordering operator [17]. All data samples used in the experiments were



(a)

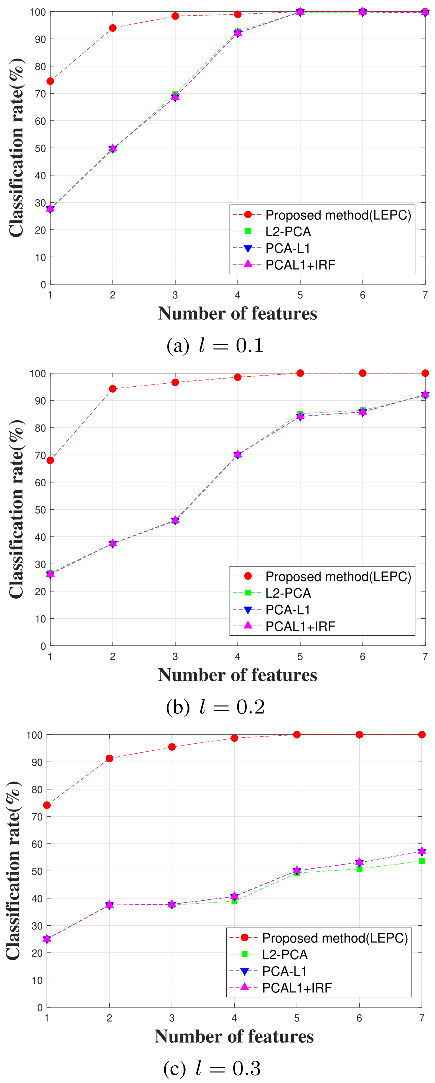
(b)

(c)

FIGURE 6. Representation of electronic nose data sample in vector form. (a) original lossless data; (b) lossy data ($l = 0.3$); (c) data restored by using the proposed method.

TABLE 3. RMS error and PSNR between the restored data and original data (electronic nose).

l	measure	LEPC	L2-PCA	PCA-L1	PCAL1+IRF
0.1	RMS	**14.97**	24.81	24.91	25.13
	PSNR	**35.41**	29.96	29.86	29.68
0.2	RMS	**18.84**	43.09	43.36	45.26
	PSNR	**32.92**	24.71	24.67	24.16
0.3	RMS	**24.38**	61.70	60.00	64.90
	PSNR	**30.23**	21.46	21.80	20.91



(a) $l = 0.1$

(b) $l = 0.2$

(c) $l = 0.3$

FIGURE 7. Comparison of classification rates between the proposed method and other methods (electronic nose).

normalized to the mean and standard deviation of the training data samples [15]. Of the 160 samples in the dataset, 80 data samples were randomly selected for the training set and the remaining 80 samples were tested. We lost some elements at a loss rate l from each test sample, and then restored the lost elements by using the proposed method ($\epsilon = 1.001$). To increase the statistical confidence, the above procedure was repeated ten times and the average results were reported.

Fig. 6 shows the original lossless data sample expressed as a 32,000-dimensional vector ($\mathbf{x}$), lossy data samples with loss ($\bar{\mathbf{x}}$) with $l = 0.3$, and data samples restored by the proposed method. The blue solid line in Fig. 6(c) represents the RMS error between the original data and the restored data. As shown in Fig. 6(c), the proposed method was able to reconstruct the shapes of the data samples such that they closely resembled the respective shapes of the original data. Table 3 lists the RMS error and PSNR between the original data and the restored data. These results show that the RMS error of the proposed method is $0.38 \sim 0.60$ times lower than that of L2-PCA, L1-PCA, and PCAL1+IRF for l between 0.1 and 0.3. In the case of the PSNR, the proposed method outperformed the other methods by $1.18 \sim 1.45$ times.

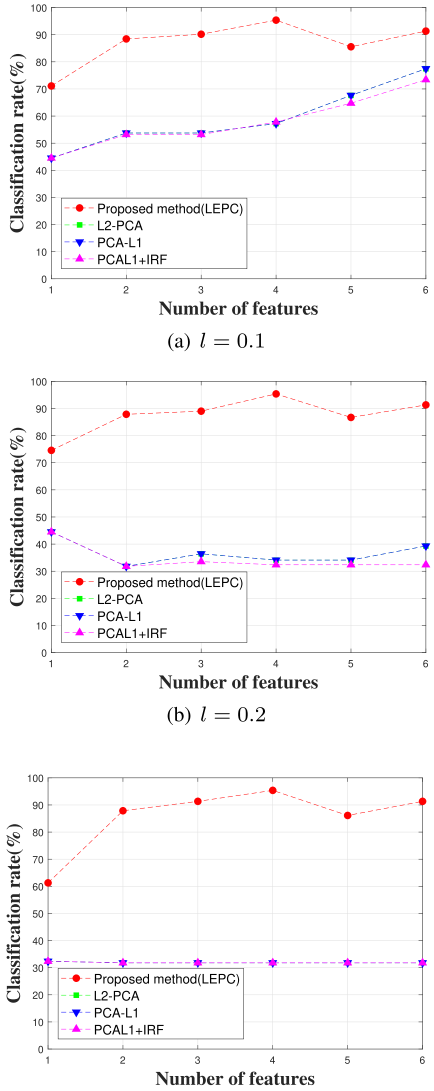
The efficacy of the proposed method was also verified by assessing the gas classification performance. Fig. 7 shows the classification rates corresponding to each dimension of DCV feature space for $\mathbf{x}^+_{L2PCA}$, $\mathbf{x}^+_{PCAL1}$, $\mathbf{x}^+_{L1+IRF}$, and $\mathbf{x}^+_{LEPC}$ with different loss rates. As shown in Fig. 7, the classification rate of the data reconstructed by the proposed method was approximately 0.85% and 38.98% higher than that of the other methods. In Fig. 7, as the loss rate gradually increases from $l = 0.1$ to $l = 0.3$, the classification performance of the other methods is greatly reduced, whereas that of the proposed method ($\mathbf{x}^+_{LEPC}$) remains at a high level of 100.00% even when $l = 0.3$.

C. PEMS TRAFFIC DATASET

The PEMS dataset [28] consists of an occupancy rate (between 0 and 1) collected from car lanes on the freeways

TABLE 4. RMS error and PSNR between the restored data and original data (PEMS traffic data).

l	measure	LEPC	L2-PCA	PCA-L1	PCAL1+IRF
0.1	RMS	**4.36**	5.75	5.75	6.23
	PSNR	**41.64**	34.13	34.13	33.13
0.2	RMS	**4.50**	7.82	7.82	8.19
	PSNR	**39.71**	30.68	30.68	30.24
0.3	RMS	**4.60**	10.16	10.16	10.46
	PSNR	**38.83**	28.23	28.23	27.98



(a) $l = 0.1$

(b) $l = 0.2$

(c) $l = 0.3$

FIGURE 8. Comparison of classification rates between the proposed method and other methods (PEMS traffic data).

in California, USA. A total of 440 data samples were collected from Monday to Sunday (indexed 1, 2, 3, 4, 5, 6, 7) except holidays. Each data sample was measured at intervals of 10 minutes for 24 hours using 963 sensors and recorded as 138, 672 (24 × 6 × 963)-dimensional vector. As provided in the PEMS dataset, we used 267 data samples as training data and the remaining 173 samples as test set ($\epsilon = 1.0$).

In Table 4, the proposed method outperformed the other methods in both RMS and PSNR.

In the classification experiments of Fig. 8, contrary to other methods that revealed the rapid degradation of classification performance with the increase of the loss rate, the proposed method appeared stable as the classification performance was maintained over 95%.

V. DISCUSSIONS AND CONCLUSION

In this article, we proposed a method based on principal component analysis to effectively reconstruct data from which some data values were lost. A data sample was represented by a feature vector, of which the elements are the principal components, in the L2-PCA feature space. Because each principal component value is calculated by projecting the data sample onto the projection vectors of L2-PCA, the feature vector approaches the convergence point as the amount of lossy data increases. This characteristic enabled the proposed method to approximately linearly estimate the feature vector of the original lossless data from the feature vector of the lossy data. The estimated feature values were then used as coefficients in the linear combination of projection vectors to restore the data. The proposed method is highly effective when the data loss rate is already known as a result of an analysis of the physical defects of a sensor or the occurrence of environmental instability such as temporary power interruption. We confirmed the effectiveness of the proposed method by performing data restoration and classification experiments on several datasets. The experimental results confirmed that the proposed method restores data efficiently and accurately compared to other methods based on principal component analysis.

The proposed method is motivated from the observation that the feature vector converges to CP in the principal component space as the amount of data loss increases. Therefore, this article does not directly deal with the case where data is corrupted by non-zero value noise rather than loss. However, when the noise value is small, data restoration methods using conventional principal component analysis can be effectively used, noise components of unusual values outside the normal range of data values can be treated with zero and then the proposed method can be applied. In the future, we plan to investigate ways to effectively restore data in a variety of situations by developing a method to accurately measure the extent of data loss from the data itself.

REFERENCES

[1] J. Davila, A.-M. Cretu, and M. Zaremba, "Wearable sensor data classification for human activity recognition based on an iterative learning framework," *Sensors*, vol. 17, no. 6, p. 1287, Jun. 2017.

[2] S. Alonso, D. Pérez, A. Morán, J. J. Fuertes, I. Díaz, and M. Domínguez, "A deep learning approach for fusing sensor data from screw compressors," *Sensors*, vol. 19, no. 13, p. 2868, Jun. 2019.

[3] M. G. Miranda, R. A. Salinas, U. Raff, and O. Magna, "Wavelet design for automatic real-time eye blink detection and recognition in EEG signals," *Int. J. Comput. Commun. Control*, vol. 14, no. 3, pp. 375–387, May 2019.

[4] L. Sun, J. Du, Z. Xie, and Y. Xu, "Auxiliary features from laser-Doppler vibrometer sensor for deep neural network based robust speech recognition," *J. Signal Process. Syst.*, vol. 90, no. 7, pp. 975–983, Jul. 2018.

[5] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Francisco, CA, USA: Academic, 2013.

[6] H. Zhang, A. Jolfaei, and M. Alazab, "A face emotion recognition method using convolutional neural network and image edge computing," *IEEE Access*, vol. 7, pp. 159081–159089, 2019.

[7] V.-H. Duong, Y.-S. Lee, J.-J. Ding, B.-T. Pham, M.-Q. Bui, P. T. Bao, and J.-C. Wang, "Projective complex matrix factorization for facial expression recognition," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 1, p. 10, Dec. 2018.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[9] M. J. Alam, V. Gupta, P. Kenny, and P. Dumouchel, "Speech recognition in reverberant and noisy environments employing multiple feature extractors and i-vector speaker adaptation," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 1, p. 50, Dec. 2015.

[10] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Silent speech recognition as an alternative communication device for persons with laryngectomy," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2386–2398, Dec. 2017.

[11] J. Qi, G. Jiang, G. Li, Y. Sun, and B. Tao, "Intelligent human-computer interaction based on surface EMG gesture recognition," *IEEE Access*, vol. 7, pp. 61378–61387, 2019.

[12] S. Benatti, F. Casamassima, B. Milosevic, E. Farella, P. Schonle, S. Fateh, T. Burger, Q. Huang, and L. Benini, "A versatile embedded platform for EMG acquisition and gesture recognition," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 5, pp. 620–630, Oct. 2015.

[13] N. Boughnim, J. Marot, C. Fossati, and S. Bourennane, "Hand posture recognition using jointly optical flow and dimensionality reduction," *EURASIP J. Adv. Signal Process.*, vol. 2013, no. 1, p. 167, Dec. 2013.

[14] P. Nakjai and T. Katanyukul, "Hand sign recognition for thai finger spelling: An application of convolution neural network," *J. Signal Process. Syst.*, vol. 91, no. 2, pp. 131–146, Feb. 2019.

[15] S.-I. Choi, G.-M. Jeong, and C. Kim, "Classification of odorants in the vapor phase using composite features for a portable E-Nose system," *Sensors*, vol. 12, no. 12, pp. 16182–16193, Nov. 2012.

[16] X. Zhai, A. A. S. Ali, A. Amira, and F. Bensaali, "MLP neural network based gas classification system on zynq SoC," *IEEE Access*, vol. 4, pp. 8138–8146, 2016.

[17] S.-I. Choi, H.-M. Jeon, and G.-M. Jeong, "Data reconstruction using subspace analysis for gas classification," *IEEE Sensors J.*, vol. 17, no. 18, pp. 5954–5962, Sep. 2017.

[18] H.-M. Jeon, J.-Y. Lee, G.-M. Jeong, and S.-I. Choi, "Data reconstruction using iteratively reweighted L1-principal component analysis for an electronic nose system," *PLoS ONE*, vol. 13, no. 7, Jul. 2018, Art. no. e0200605.

[19] L. Han, C. Yu, K. Xiao, and X. Zhao, "A new method of mixed gas identification based on a convolutional neural network for time series classification," *Sensors*, vol. 19, no. 9, p. 1960, Apr. 2019.

[20] P. Peng, X. Zhao, X. Pan, and W. Ye, "Gas classification using deep convolutional neural networks," *Sensors*, vol. 18, no. 2, p. 157, Jan. 2018.

[21] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[22] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1680, Sep. 2008.

[23] J.-X. Mi, Y.-N. Zhang, Z. Lai, W. Li, L. Zhou, and F. Zhong, "Principal component analysis based on nuclear norm minimization," *Neural Netw.*, vol. 118, pp. 1–16, Oct. 2019.

[24] Q. Ke and T. Kanade, "Robust L_1 Norm factorization in the presence of outliers and missing databy alternative convex programming," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 739–746.

[25] W. Zuo, K. Wang, and D. Zhang, "Robust recognition of noisy and partially occluded faces using iteratively reweighted fitting of eigenfaces," in *Proc. Pacific-Rim Conf. Multimedia*, 2006, pp. 844–851.

[26] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 1–37, 2011.

[27] A. Martınez and R. Benavente, *The AR face database(Computer Vision Center)*. Barcelona, Spain: Univ. Autonoma Barcelona, 1998.

[28] D. Dua and C. Graff, "UCI machine learning repository," School Inf. Comput. Sci. Univ. California, Irvine, CA, USA, 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[29] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Phil. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, Nov. 1901.

[30] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, p. 417, 1933.

[31] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Linear Algebra*. Berlin, Germany: Springer, 1971, pp. 134–151.

[32] G. W. Brier and R. A. Allen, *Verification of Weather Forecasts*. Boston, MA, USA: AMS, 1951, pp. 841–848.

[33] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," *Geosci. Model Develop.*, vol. 7, no. 3, pp. 1247–1250, 2014.

[34] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 1, pp. 4–13, Jan. 2005.

[35] S.-I. Choi, C.-H. Choi, G.-M. Jeong, and N. Kwak, "Pixel selection based on discriminant features with application to face recognition," *Pattern Recognit. Lett.*, vol. 33, no. 9, pp. 1083–1092, Jul. 2012.

[36] Y. Lee, M. Lee, and S.-I. Choi, "Image generation using bidirectional integral features for face recognition with a single sample per person," *PLoS ONE*, vol. 10, no. 9, Sep. 2015, Art. no. e0138859.

YONGGEOL LEE received the B.S. degree from the Department of Applied Computer Engineering, Dankook University, in 2012, and the M.S. and Ph.D. degrees from the Department of Computer Science and Engineering, Dankook University, in 2014 and 2019, respectively. He is currently a Researcher with the Police Science Institute, Korean National Police University, Asan, South Korea. His research interests include computer vision, machine learning, and pattern recognition.

SANG-IL CHOI (Member, IEEE) received the B.S. degree from the Division of Electronic Engineering, Sogang University, in 2005, and the M.S. and Ph.D. degrees from the School of Electrical Engineering and Computer Science, Seoul National University, in 2007 and 2010, respectively. He was a Postdoctoral Researcher with the BK21 Information Technology, Seoul National University, in 2010, and the Computer Science Department, Institute for Robotics and Intelligent Systems, University of Southern California, Los Angeles, until August 2011. He is currently an Associate Professor with the Department of Computer Science and Engineering, Dankook University, South Korea. His research interests include pattern recognition, feature extraction and selection, machine learning, computer vision, and their applications.