

PhysioCover: Recovering the Missing Values in Physiological Data of Intensive Care Units

Sun-Hee Kim

Department of Computer Science
Chonnam National University, Gwangju 500-757, South Korea

Hyung-Jeong Yang*

Department of Computer Science
Chonnam National University, Gwangju 500-757, South Korea

Soo-Hyung Kim

Department of Computer Science
Chonnam National University, Gwangju 500-757, South Korea

Guee-Sang Lee

Department of Computer Science
Chonnam National University, Gwangju 500-757, South Korea

ABSTRACT

Physiological signals provide important clues in the diagnosis and prediction of disease. Analyzing these signals is important in health and medicine. In particular, data preprocessing for physiological signal analysis is a vital issue because missing values, noise, and outliers may degrade the analysis performance. In this paper, we propose PhysioCover, a system that can recover missing values of physiological signals that were monitored in real time. PhysioCover integrates a gradual method and EM-based Principle Component Analysis (PCA). This approach can (1) more readily recover long- and short-term missing data than existing methods, such as traditional EM-based PCA, linear interpolation, 5-average and Missing Value Singular Value Decomposition (MSVD), (2) more effectively detect hidden variables than PCA and Independent component analysis (ICA), and (3) offer fast computation time through real-time processing. Experimental results with the physiological data of an intensive care unit show that the proposed method assigns more accurate missing values than previous methods.

Key words: *Intensive Care Unit, Missing Values, Hidden variable, Real Time Processing and EM-Principle Component Analysis.*

1. INTRODUCTION

Physiological data are observations of physiological activities from neurons, cardiac rhythms, tissues and organs. They are measured by non-invasive methods such as surface sensors on the skin of a user, or invasive methods such as measurement method of Arterial Blood Pressure (ABP). They may provide a robust and accurate means of detecting and predicting diseases, because the signals correspond to internal physiology. Many physiological signals, such as Electroencephalography (EEG), Electrocardiography (ECG),

ABP, and Heart rate are recorded in digital format. Analyzing these digital signals to extract useful health information is an emerging research area in biomedical engineering.

Physiological signal analysis is performed for many reasons. These signals are mostly collected by multi sensors per patient over time. They are often sampled as a matrix to correctly analyze data. Recently, physiological signals have been measured by small, wearable, and wireless sensors, while the patient is moving in real time. Many studies using physiological signals focus on the prediction or classification of events, such as epileptic seizure, acute hypotensive episodes, emotion recognition, and so on [1]-[3]. To correctly analyze data, the collected data are required to be precise and reliable.

However, the dataset of most Physiological signals includes a lot of short- or long-term missing values, and noise

* Corresponding author, Email: hjyang@jnu.ac.kr
Manuscript received Dec. 23, 2013; revised Apr. 24, 2014;
accepted May. 02, 2014

(outliers). Most studies treat missing values by removing a specific signal, or by using a simple method, such as averaging of observed data, normalization, and linear interpolation [4]-[7]. These methods may cause a biased model because of the loss of information. Furthermore, they underestimate standard deviation, since they do not consider the uncertainty in missing values. Consequently, this problem may result in irreversible health damage and death through faulty analysis due to the characteristics of the physiological data.

In this paper, we propose *PhysioCover*, which recovers missing values of physiological signals by gradually updating the weight values of Principle Component Analysis (PCA) based on an Expectation Maximization approach. It also summarizes large data by detecting hidden variables in real-time. The proposed method can solve the problem of a biased model, which is inherited from missing data, and reduce the loss of information by recovering missing values. Experimental results with physiological data of an intensive care unit show that the proposed method replaces more accurate missing values than previous methods such as the traditional EM-based PCA (EM-PCA), linear interpolation, 5-average and Missing Value Singular Value Decomposition (MSVD) with respect to classification accuracy. Our contributions are as follows;

- *Robust missing value recovery*: The proposed method which combines an EM-based PCA and gradual approach provides a good recovery result for long-term missing values of physiological signals.
- *Hidden variable detection*: our method detects a few hidden variables, which summarize the whole signals.
- *Scalability*: *PhysioCover* provides a scalable approach that needs computation time $O(r \cdot k)$, where r is the number of iterations and k is the number of hidden variables in the model to recover missing values, and summarizes physiological signals. Therefore, we expect it will scale well for various real time series data of multi-dimensions.

The remainder of this paper is organized as follows: section 2 presents the proposed method for recovering missing values. Section 3 describes the result of experiments of the proposed method with physiological data of an intensive care unit. In section 4, we discuss the existing methods in comparison with our method for recovering missing values. Finally, conclusions are drawn in section 5.

2. MATERIALS AND METHODS

The major goal of analyzing these physiological time series data is to forecast or to detect disease. Many mathematical tools, such as Linear Regression and Auto-Regression [8], assume completely observed data. However, missing observations often occur in many real applications, and thus, it presents a major challenge to model physiological time series in the presence of missing data.

2.1 Background

Given a data matrix $X_{m \times n}$ that contains missing values, an improved PCA is proposed to use alongside the Expectation and Maximization (EM) algorithm [9], [10]. It recovers the values of missing data through the Expectation and the Maximization steps. As a first step for recovering missing values, the initial values of missing values are filled with the mean of the column vector, and the recovered data are projected by PCA [11]. The Expectation step can be easily derived with the projection data for learning. The Maximization step re-computes the principal components with the obtained value at the Expectation step, and the missing values are replaced with the updated unitary values. The optimal value to fill the missing values is predicted in the iterative process of EM.

$$\text{Expectation step: } W = (Y^T \cdot Y)^{-1} \cdot Y^T \cdot X, \quad (1)$$

$$\text{Maximization step: } Y = X \cdot W^T \cdot (W \cdot W^T)^{-1}, \quad (2)$$

where Y is a projected matrix of $m \times k$, and W is a $k \times n$ matrix of the unknown states. The columns of Y will span the space of the first k principal components. The data matrix X can be projected into this k dimensional subspace by computation of the corresponding eigenvectors and eigenvalues explicitly. The EM-based PCA projects the data matrix X using W_{new} , which is updated through the iterative process of Expectation and Maximization until convergence. The data matrix X can be reconstructed by Y and W_{new} as $\hat{X} = Y \cdot W_{new}$. The missing values of data matrix X are replaced by the reconstructed matrix \hat{X} . The EM-based PCA recovers the missing values well. However, it requires much execution time for a large multi-dimension dataset because of the batch process [10]. Therefore, we propose the novel method of a gradual approach that allows real-time processing for recovering missing values.

2.2 *PhysioCover*: A Gradual method with EM-based PCA

The proposed method is based on the gradual method and EM-based PCA. EM-based PCA demonstrated the advancement of recovering missing values [9]. Therefore, for the missing value recovery of the physiological time series data, we integrate a gradual model with the concept of EM-based PCA to update the weight vector in real time.

In the physiological time series data, $x_t \in \mathbb{R}$ is the n signal measurement column-vector at time tick t . That is, physiological data are represented as a $t \times n$ matrix. For real time processing, we apply the gradual method to update the weight vectors, w_i , at each time tick in the newly projected space. Each of the weight vectors, w_i is projected onto the input vector, x_t in the linear transformation of the data stream to obtain the hidden variables or components, y_t over time [8].

For the real time processing of the physiological data that include missing values, firstly, the number of hidden variables

is initialized by an arbitrary number, k . Given input data $x_t = x_{t,1}, x_{t,2}, \dots, x_{t,n}$ with n dimensions at time t , the i -th component, $y_{t,i}$, is obtained as follows:

$$y_{t,i} = \sum_{n=1}^N x_{t,n} \cdot w_{i,n}^T \quad (3)$$

where the input data x_t is computed by the previous weight vector, $w_{t-1,i} (1 \leq i \leq k)$. Second, we estimate the energy, $p_{t,i}$ and reconstruction error, $e_{t,i}$ by Eqs. (4) and (5), respectively, in order to adjust the number of hidden variables or components. The energy initializes with a small positive value.

$$p_{t,i} = \lambda p_{t,i} + y_{t,i}^2 \quad (4)$$

$$e_{t,i} = \hat{x}_{t,i} - x_{t,i} \quad (5)$$

This gradual approach uses the exponential forgetting factor, λ , to reflect more recent trends in the data stream. The exponential forgetting factor, λ , commonly uses values between 0.96 and 0.98 [8], [12]. It helps to reduce the huge memory usage, because no buffer space is required for the whole data. The magnitude of the estimates should also consider the past data captured by the participation weight vector $w_{t,i}$, because the update is inversely proportional to the current energy $E_{t,i}$ of the i -th hidden variable as follows:

$$E_{t,i} = (1/t) \sum_{\tau=1}^t y_{\tau,i}^2 \quad (6)$$

The participation weight vector is updated based on the following equation:

$$w_{t,i} = w_{t,i} + (y_{t,i} e_{t,i} / p_{t,i}) \quad (7)$$

Thirdly, we maximize the updated weight vector using Eqs. (1) and (2) by using the stopping criterion. In this study, we define the stopping criterion as follows: if the absolute value of the difference of a new weight vector and an old weight vector of w is smaller than δ (for example, $\delta = 0.001$), or if the absolute value of the sum of the new weight vector and old weight vector of w is smaller than δ , it is treated as stopping criterion. For recovering missing values, we compute the reconstruction data \hat{x}_t with the updated weight matrix w_{new} , and the new projected vector, y_t . The missing values of input data, x_t , are recovered by the reconstruction data, \hat{x}_t , but the observed values in x_t are not replaced by \hat{x}_t .

Finally, we obtain the actual hidden variables, $y_t = x_{t,replace} \cdot w_{t,i}^T$ at time t , which is computed by newly

recovered missing values of the input data x_t and the weight matrix $w_{t,i}$ by Eq. (8), which is whitened to maximize the weight vector.

$$w_{t,i}^{new} = w_{t,i} / \text{norm}(w_{t,i}) \quad (8)$$

To automatically determine the number of hidden variables, we compute the energy E_{hv} based on the values of the hidden variables. In practice, we do not know the number of hidden variables k . Therefore, we use an energy threshold to determine the number of hidden variables. The energy threshold corresponds to a bound, which contains the upper bounds FE_x and lower bounds fE_x of the energy [8], [13]. The energy of the hidden variable E_{hv} is compared with the predefined upper and lower bounds. If $E_{hv} < fE_x$, the number of hidden variables, k , increases. On the other hand, if $E_{hv} > FE_x$, k decreases. We keep the number of hidden variables within the range FE_x to fE_x . If the lower bound of energy is too low, the useful information of the data may be lost. In *PhysioCover*, we use the upper and lower energy thresholds of 0.98 and 0.95, respectively. This means that the energy of input data x_t is retained between 95% and 98%.

When new data x_{t+1} that include missing values arrives, missing values are recovered with the updated weight vector through the iterative process of the Expectation and Maximization step. The number of hidden variables will be adjusted automatically, while maintaining the predefined bounds. The algorithm of *PhysioCover* is shown in Table 1.

Table 1. *PhysioCover* algorithm

Input: New input $x_t \in \mathbb{R}^n$

Output: Recovered data $x_{t,replace}$ and Projected data y_t

Algorithm

if input vector x_t does not include missing values

{for $i=1$ to k // k is the number of hidden variables

Eqs. (3)–(8) // carry out

end }

else input vector x_t include missing values

{ $x_{t,replace} = x_{t,miss} = x_{mean}$ // Initialize the missing values by x_{mean} , while the observed values of x_t remain unchanged

for $i=1$ to k // k is the number of hidden variables

while stopping criterion

Expectation and Maximization step: Eqs. (1) and (2) // maximize weight vector

end

$w_i = w_i / \text{norm}(w_i)$

end

$y_t = x_{replace} \cdot w_{new}^T$ // Compute hidden variables

```

 $\hat{x}_{t,replace} = y_t \cdot w_{new}$  // Reconstruction of input data  $x_t$ 
 $x_{t,replace} = \hat{x}_{t,replace}$  // Recover missing values
}
end
 $E_{hv} = \lambda E_{hv} + y_t^2$  // Compute energy of total hidden variables
 $E_x = \lambda E_x + x_{t,replace}^2$  // Compute energy of total input data
if  $E_{hv} < fE_x$ , then  $k = k + 1$  // Increase the number of hidden variables
else  $E_{hv} > fE_x$ , then  $k = k - 1$  // Decrease the number of hidden variables
end

```

3. EXPERIMENT RESULTS

3.1 Data descriptions

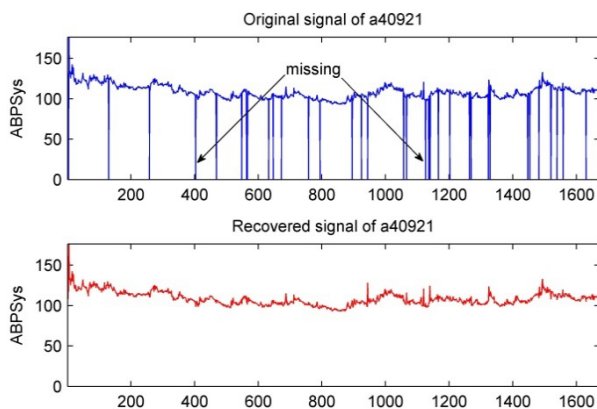
The dataset we used to verify the efficiency of the proposed method was obtained from a public access Intensive Care Unit (ICU) database [7], [14]. We used the dataset of 923 patients over 45 hours from the ICU database. The physiological signals of each patient were monitored using 18 sensors, such as heart rate (HR), arterial blood pressure (ABP), noninvasive indirect blood pressure (NBP), respiration, and saturation of oxygen measured by a pulse oximeter (SpO2), and so on. Each patient belongs to the Acute Hypotensive Episode (AHEs) or Non-AHEs class. The patients were separated into two groups; the group of Acute Hypotensive Episodes (AHEs) includes an AHE in the forecast window, and the other group does not include an AHE in the forecast window section. AHE group has 314 patients and the non-AHE group has 609 patients. In this dataset, several sensors were not recorded signal during the monitoring period from most of the patients (e.g. from 8 to 21th sensors). Therefore, we used only the monitored signals excluding all missing signals. We used the seven signals: HR, ABPSys, ABPDias, ABPMean, Pulse,

RESP, and SpO2. The normal HR beats 50-100 per one minute. ABPSys and ABPDias mean systolic and diastolic pressures, respectively. ABPMean is the mean arterial pressure, and Pulse is the heart beats that are strong enough to be felt, at the wrist/knee/ankle, etc. RESP means respiration (13-16/min), and SpO2 is saturation of oxygen measured by a pulse oximeter. These signals include many missing values. We evaluated the performance of the proposed method using both simulated sample data and real time data. The simulated sample data were generated from some of the real data, and it not includes missing values. To measure the accuracy of recovering short and long-term missing values, we artificially removed the values of area during 3 or 4 hours in the simulated sample data.

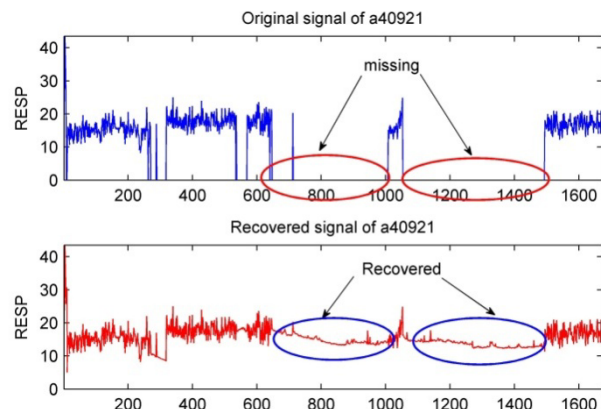
3.2 Recovery of Missing Values

For detection or prediction of AHEs, the signals that are monitored at least one hour before an AHEs event occurs are important [15]. However, if signal of this important point is missing, the detection or prediction of an AHEs event may be difficult. The purpose of the proposed method is to impute the missing values to prevent failures of detection and prediction because of the incomplete data. Given multiple physiological time series data with missing values, we propose *PhysioCover*, which recovers the missing values, finds the latent variables, and summarizes data.

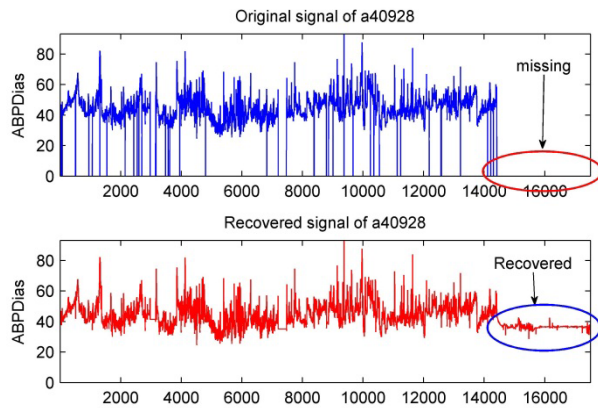
Fig. 1 shows the original input signals of patient a40921 and the recovered signals by the proposed method. In Figure 1, if the value of a signal is zero, it means a missing value. Also, in the figures, the x-axis means the time point and y-axis means the real values that were recorded from the sensors. In Figure 1 (a), the ABPSys signal of patient a40921 has much short-term missing data. That is, the dropped down point to zero is a missing value. The missing signal is recovered without failure, which is the red color signal at the bottom of Figure 1(a). The RESP signal of a40921 has long-term missing values (in Fig. 1. (b)).



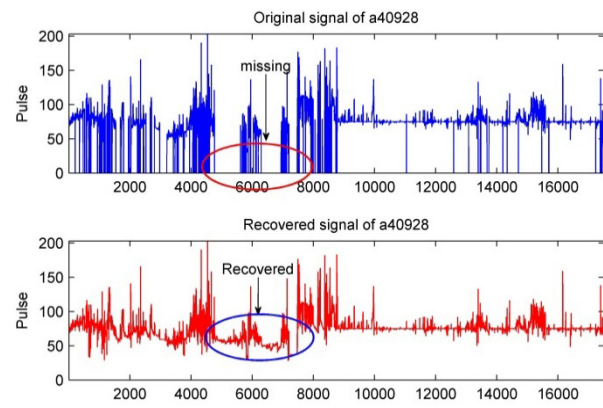
(a) Original and recovered signals of ABPSys of a40921



(b) Original and recovered signals of RESP of a40921



(c) Original and recovered signals of ABPDias of a40928



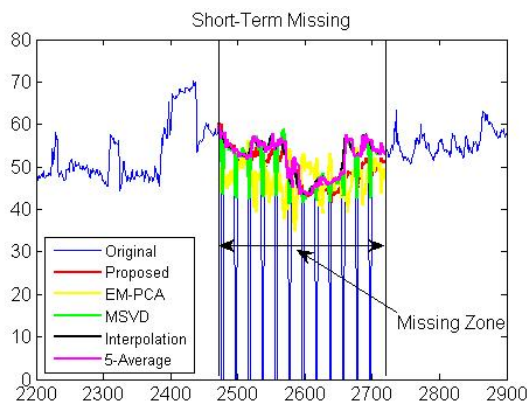
(d) Original and recovered signals of Pulse of a40928

Fig. 1. Comparison between Original signal and Recovered signal by our method

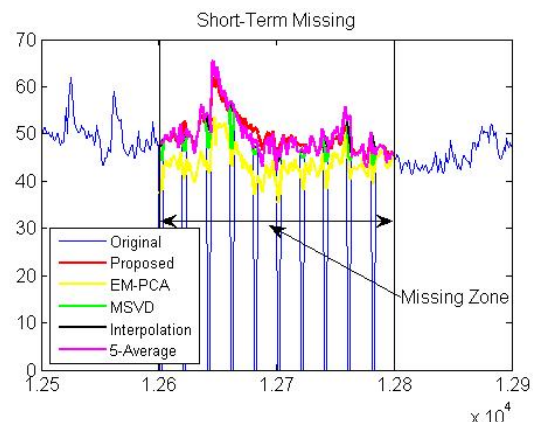
In this paper, we define long-term missing values as values that have been missing for more than an hour. Red ellipses indicate the missing area, and long-term missing lasted for 5 hours and 7 hours. In the observed signal, one time point is 1 minute. In the recovered signal graph of RESP, blue ellipses indicate the recovered signal by the proposed method. Figures 1 (c) and (d) show the original and recovered signals of the ABPDias and Pulse of patient a40928, respectively. The ABPDias signal of a40928 has a strong long-term missing period from the 14500 to 17500 time point. This means missing of more than 40 hours. In the case of the normal patient, ABPDias appears near 80mmHg. However, ABPDias of a40928 are between 40 and 60mmHg, and the recovered signal also appears nearby in the range of the original signal. The Pulse is mostly equal to the heart rate. The normal HR beat is between 50 and 100 per minute. Blue ellipses in Figures 1 (d) indicate the recovered signals by the proposed method, and they appear in a similar range to the original signal of Pulse. The similarity between the original and recovered signals may mean a suitable method for recovering of a missing value from a physiological signal.

To verify the effectiveness of *PhysioCover* in recovering short-term and long-term missing values, we generated simulated sample data that were extracted from a part without missing values in the original data, and we coercively created

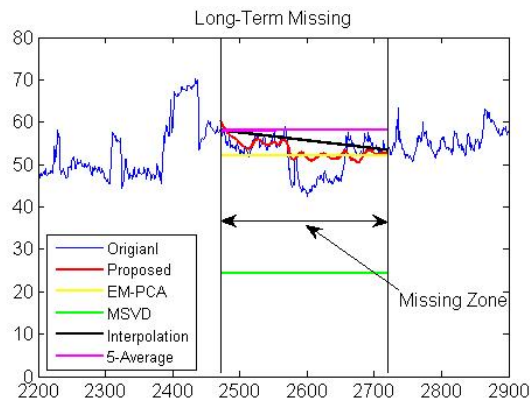
missing values in a part of the sample data. This experiment contemplated two cases: short-term missing and long-term missing. Short-term missing periodically arises for several minutes spread over 3 or 4hour. Long-term missing is complete missing for 3 or 4hour. To compare the imputation power of the missing values, we applied five methods to the simulated sample data: traditional EM-PCA [10], [16], MSVD [17], Linear Interpolation, 5-average, and our method. Fig. 2 shows the recovered result in the short and long-term missing alongside a comparison with the existing methods. Fig. 2(a) is the result of short-term missing recovery of the simulated data, which were extracted from ABPDias of patient a41770, and almost recovered signal appeared similar to original signal, but MSVD has the highest RMSE as Table 2. In case of the long-term missing over 4 hours, it shows surprising results in Fig. 2 (c). The existing methods have the unvaried values while missing values recovered, but our proposed method flexibly recovered missing values of long-term. In traditional EM-PCA, MSVD, and 5-average, the recovered signal appears as a nearly straight line in the long-term missing section. Linear Interpolation method is a traditional approach of missing value recovery in the research such as detection and prediction of AHE [5][7]. The recovered signal shows linear lines, because it draws a straight line between the starting point of missing and the end point of missing.



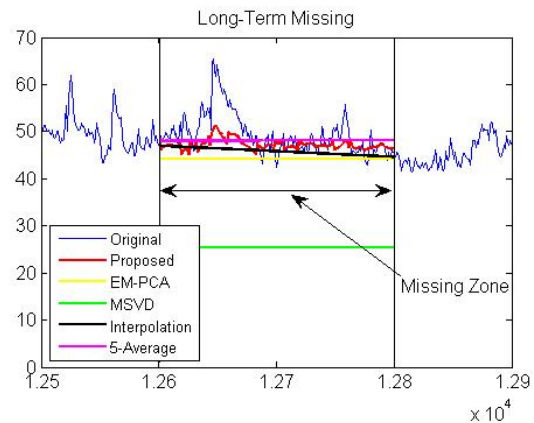
(a) Short-term missing of a41770(ABPDias)



(b) Short-term missing of a40384(ABPDias)



(c) Long-term missing of a41770(ABPDias)



(d) Long-term missing of a40384(ABPDias)

Fig. 2. Comparison between original and recovered signals by existing methods such as EM-PCA, MSVD, Linear Interpolation, 5-Average, and the proposed method using the sample data extracted from original data that do not include missing values. The extracted sample data contain missing values that were created coercively.

Table 2. RMSE of recovery methods for the sample data set

Method	Patient a41770		Patient a40384	
	Short-Term	Long-Term	Short-Term	Long-Term
Proposed Method	2.4919	3.6713	1.4711	1.8565
EM-PCA	6.6581	7.7201	4.4238	4.4238
MSVD	6.7617	28.0723	19.228	24.4166
Interpolation	0.8071	5.7714	0.7181	5.4757
5-Average	1.5299	7.5186	0.821	4.5837

RMSE of *PhysioCover* guaranteed the lowest value (see Table 2). Fig. 2 (b) and (d) show the results of simulated data from patient a40384 (ABPDias signal). The results validated that our proposed method is robust for long-term missing problem.

3.3 Detection of Hidden Variables

Our proposed method summarizes a large time series dataset by detecting a few hidden variables. To check the performance of hidden variables detection, we apply our proposed method, iPCA and PCA to a real dataset. iPCA is a remarkable method for real time processing [18]. This method can summarize or can detect a few hidden variables from a large multi-dimensional dataset. Therefore, it is useful for a time series health dataset. However, if features are independent, or signals have long-term missing values, it is impossible to recover missing values with iPCA. The experimental dataset is arbitrary extracted from real patient data that do not include long-term missing values. Patients are randomly selected from all of patients, and signals where the length of missing is not over 15 minutes are collected for 25 hours from the selected patients.

Fig. 3 (a) shows the signals of a patient. It is composed of 7 signals, and a few signals include short-term missing values.

Fig. 3 (b) shows the detected first hidden variables by the proposed method, iPCA, and PCA. iPCA and PCA were

applied on the recovered dataset by the proposed method to compare the patterns of the first hidden variable with the proposed method. In Fig. 3 (b), the blue line indicates the first hidden variable with the recovered missing values by the proposed method. The green line is the first hidden variable of iPCA, and the red line appears as the first hidden variable of PCA. These have the same patterns as the first hidden variable of the proposed method.

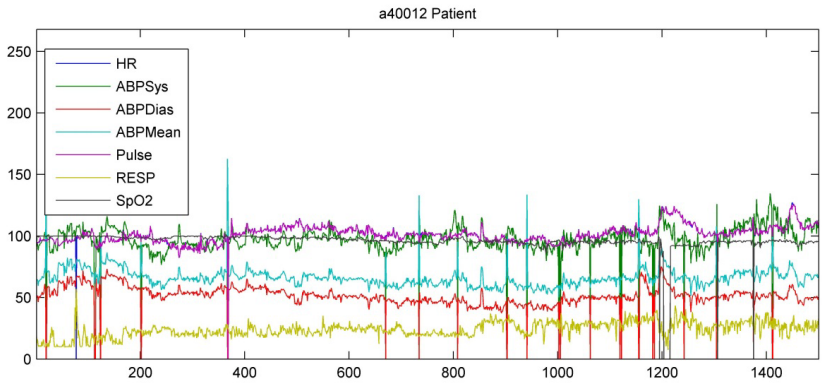
We compare hidden variables among the recovered signals of the proposed method and the existing methods to verify if the hidden variables are similar to the recovered data by each method. Traditional EM-PCA, MSVD, Linear Interpolation, and 5-average as the existing methods are applied to the real patients' dataset, which include short-term missing values as well as long-term missing values. After that, we projected the recovered dataset by PCA to compare the first hidden variable.

Fig. 4 shows the original signals of two patients. Fig. 4 (a) is the original signal of a40012, which includes short-term and long-term missing values. The a40802 of Fig. 4 (b) has mostly short-term missing values. Table 3 shows the first hidden variable of two patients (a40012 and a40802). The proposed method is able to recover missing values, and detect a hidden variable at the same time. Therefore, our method makes it unnecessary to use an extra dimension reduction or hidden detection method.

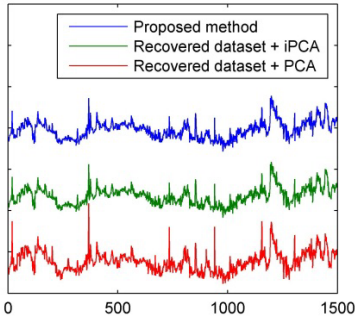
Table 3 (a) and (b) are the detected first hidden variables by the proposed method. The original dataset of Patient A40012 has long-term missing values, from 2462 to the end time point of ABPSys, ABPDias, ABPMean, and Pulse signal (over 4 hour) as in Fig. 4 (a). Table 3 (c) and (d) show the first hidden variable of each patient that was detected by traditional EM-PCA. This recovery method can recover and project the data at the same time. Therefore, the detection process is not needed for a hidden variable. (c) and (d) of Table 3 are the results that were detected in the recovering process autonomously. Table 3 (e) and (f) are the first hidden variable by PCA from the recovered missing values by MSVD. Table 3 (g) and (i) show the first hidden variables of data that recovered the missing values by Linear Interpolation and (i) and (j) are the results of 5-average, respectively. As a result, we can find a

singularity of the first hidden variables where the first hidden variable of the long-term missing section dropped down, when

MSVD and Linear Interpolation were used to recover missing values (see (e) and (g) of Table 3).

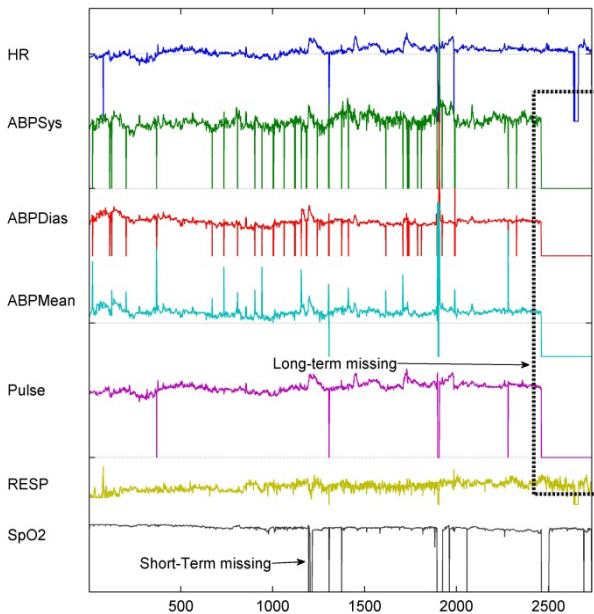


(a) Real signals of Patient a40012

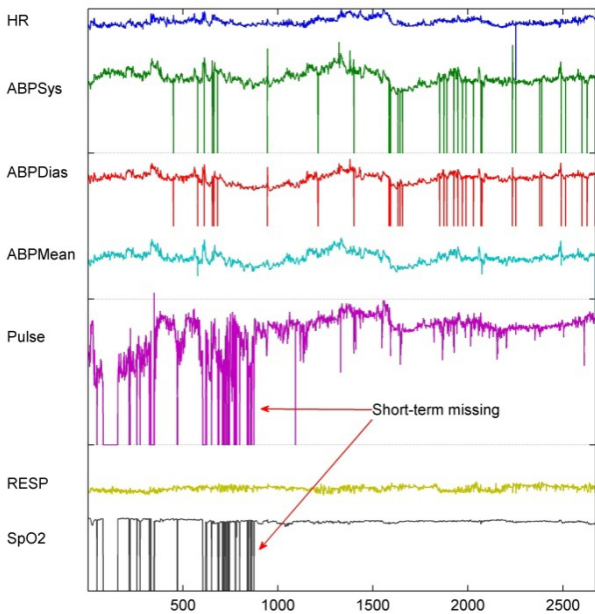


(b) The first hidden variable

Fig. 3. Detected first hidden variables by our proposed method, iPCA and PCA: (a) all signals of Patient a40012, and (b) the first hidden variable detected by the proposed method, iPCA, and PCA from the recovered signal by the proposed method.



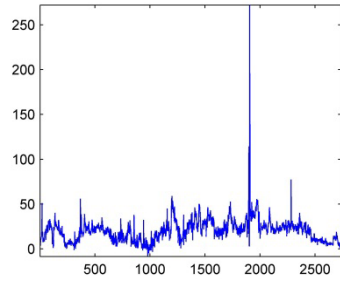
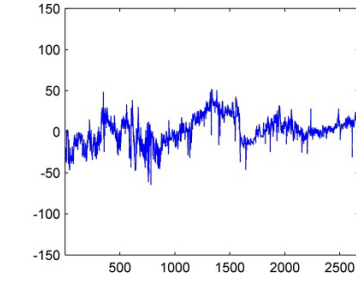
(a) Original signal of a40012

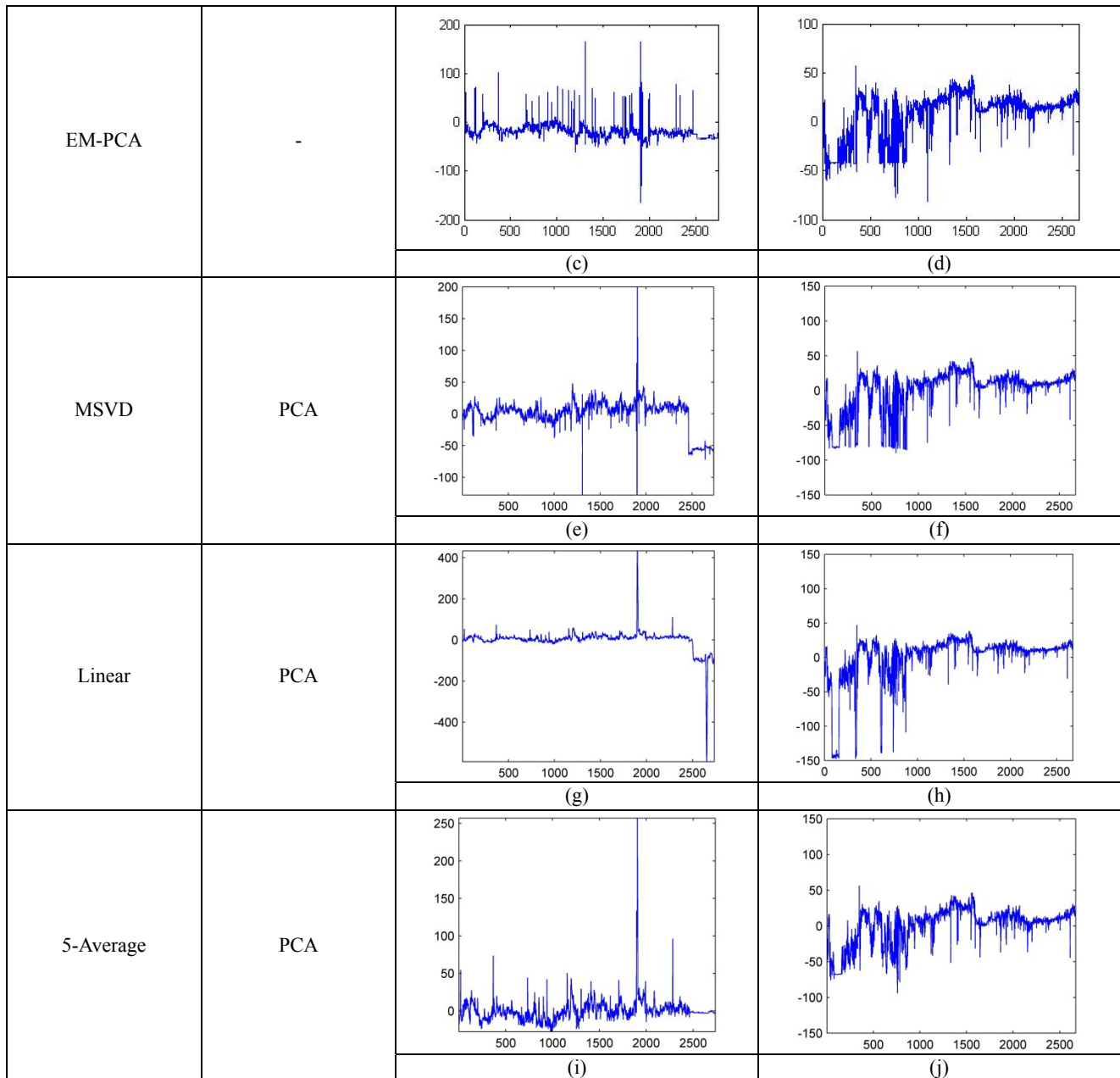


(b) Original signal of a40802

Fig. 4. Original signals of a40012 and a40802 that include short-term and long-term missing values

Table 3. Detected first hidden variables by the proposed method and PCA

Method	Hidden detection	a40012	a40802
Proposed Method	-		
		(a)	(b)



That is, the first hidden variable of the long-term missing section is relatively lower than the values of the range without long-term missing. Table 3 (f), (h), and (j) are the first principal components of Patient a40802. These data have many short-term missing values as in Fig. 4 (b). For the comparison, the short-term missing values of patient a40802 are recovered using traditional EM-PCA, MSVD, Linear Interpolation, and 5-average, and PCA is applied to detect hidden variables except in the traditional EM-PCA. In Table 3 (d), (f), (h), and (j), many drop down spikes appeared in the hidden variable before 1000 time points, because of the short-term missing values, but the patterns of first hidden variable among them are similar. Consequently, the proposed method clearly detects the high quality of the hidden variables from the recovered large dataset.

3.4 Real Time Processing

In this section, we compared the execution time of the proposed method with those of other approaches. The execution time was measured for both missing value recovery and feature extraction. In the case of existing methods to recover missing values, these require a method to extract features such as PCA or ICA.

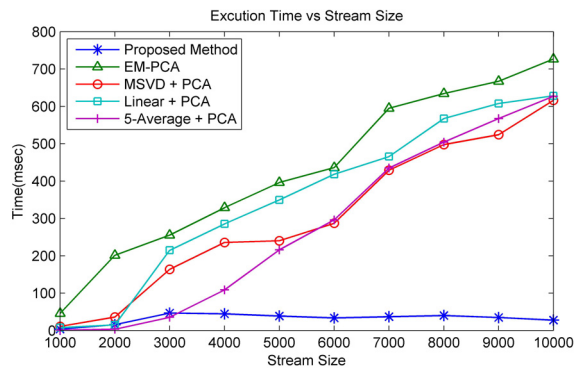


Fig. 5. Comparison of execution time

In our experiment, we apply PCA to extract features from the recovered dataset, while measuring time complexity. Fig. 5 shows the plot of execution time vs. the stream size. As a result, the execution times of the existing methods grow exponentially because PCA processes in batches. Although traditional EM-PCA is not the extra feature selection method, it is performed too in batches. However, the execution time of our method only maintains non-increasing operations since it is based on the gradual method. Our method updates a few variables such as weights and hidden variables, without re-computation of the overall data matrix, when a new signal is entered in each time point. Our method does not increase computation time even though the dataset size is gradually increased.

Table 4. Classification accuracy

Method to recover missing value	Feature Selection	Energy range	Number of Hidden Values (on average)	Classification method (Binary classification)	
				5-NN(%)	SVM(%)
Proposed Method	-	95 - 98%	3	54	63
	PCA	> 95 %	5	56	76
	ICA	> 95 %	5	60	65
EM-PCA	-	> 95 %	4	46	60
	PCA	> 95 %	5	50	65
	ICA	> 95 %	5	61	53
MSVD	PCA	> 95 %	5	50	75
	ICA	> 95 %	4	60	55
Linear Interpolation	PCA	> 95 %	6	60	65
	ICA	> 95 %	6	66	61
5-average	PCA	> 95 %	5	53	66
	ICA	> 95 %	6	63	63

3.5 Classification Accuracy

In this paper, we adjunctively evaluated the classification accuracy of the proposed method in the physiological dataset. To measure the classification accuracy, existing methods need additional feature selections or feature reduction steps to detect hidden variables except the traditional EM-PCA. Therefore, we used PCA and ICA to detect hidden variables for the existing methods. For comparison, traditional EM-PCA, MSVD, Linear Interpolation, and 5-average are used to recover missing values, and hidden variables from the recovered signal are detected using PCA and ICA. Our method and traditional EM-PCA can recover missing values and can detect a few hidden variables at the same time.

We used the recorded real datasets from 923 Patients to measure classification accuracy. For the experiments, we classified two classes: those that included AHE (314 Patients), and those that did not include AHE (609 Patients). Our method is able to offer real time processing by updating the weight vector at every time point. This weight vector is used to derive the hidden variables of a new input. However, the existing methods should be processed in the last time point because these methods are processed in batches [19].

In PCA and ICA, the number of components can be determined by the energy rate. In our experiment, we used 95% energy rates to detect the principal components and independent components for PCA and ICA because our method used energy rates from 0.95 to 0.98. We measured the classification accuracy by 10-fold cross-validation with 5-NN and SVM classifier. Table 4 is the classification results from the detected hidden variables by PCA, ICA and our method. Our method achieves the 76% best result in SVM classifier when the recovered dataset is projected by PCA. In the case of 5-NN classifier, Linear Interpolation shows the highest classification rate when the hidden variables were detected by ICA.

The proposed method itself without a feature selection shows 54% and 63% of classification accuracy on 5-NN and SVM, respectively. Our method can recover missing values and detect hidden values in real time. In addition, *PhysioCover* extracts a smaller number of hidden variables than other approaches. On average, our method extracts 3 hidden variables, traditional EM-PCA has 5 hidden variables, MSVD has 5 and 4 hidden variables, and Linear Interpolation extracts 6 hidden variables for both PCA and ICA. Finally, 5-average detected 5 and 6 hidden variables on PCA and ICA, respectively. A smaller number of hidden variables can reduce memory usage and computation time by the classifiers.

4. DISCUSSION

PhysioCover combines the gradual method and EM-based PCA. It automatically recovers the contained missing values. The dataset we used includes acute hypotensive episodes (AHE). If the blood flow is too low to deliver enough oxygen and nutrients to vital organs, it can cause dangerous situations, such as fainting, visual impairment and multiple organ damage. Thus, if not promptly treated, AHE may result in irreversible vital organ damage and death. For this reason, the prediction or detection of AHE is a significant challenge.

Chen et al. [7] developed a method to predict which patients would experience an AHE prior to the occurrence of the AHE based on the weighted average of ABP. Henriques et al. [5] proposed the application of generalized regression neural network multi-models, which most effectively predicted AHE in intensive care units (ICU). Lehman et al. [20] carried out classification and forecasting tasks, using a similarity-based searching and pattern matching algorithm. More recently, Rocha et al. [3] proposed neural network multi-models to predict adverse AHE occurring in ICU.

These researches carried out pre-processing to recover missing values using linear interpolation, which is one of the traditional methods because their purpose was focused on predicting AHE. The signal at least one hour before occurrence of AHE is regarded as an important element for predicting an AHE event [15]. However, if the signal of this time section is missing, prediction will be less accurate. In the study for future values, prediction such as air pollution, gene expression, and traffic data, missing values are recovered through observed data, interpolation [21], support vector regression (SVR) [22], Bayesian-based PCA [23], neural networks, an autoregressive

integrated moving average (ARIMA), and regression model [24] besides the methods that are used in our experiment.

However, these methods require many input parameters, long execution time by batch processing, or complete data sets. Our method recovers missing values quickly and accurately from real time processing. We compared recovery methods of missing values. As the experimental result, our *PhysioCover* provided more robust recovering results when compared with existing methods. In addition, *PhysioCover* can summarize multi-dimension data by detecting a few hidden variables. Our method can detect hidden variable simultaneously with missing value recovery in each time tick.

Our proposed method has the advantage of real-time processing considering the characteristics of time series data. Although PCA [25], [26] and ICA [2], [27], [28] are robust dimension reduction methods or hidden variable detection methods, they require long execution times by batch processing, as shown in Figure 5. Because our method recovers missing values and detects hidden variables in real time, it can be scaled for various types of time series data, such as econometrics, mathematical finance, weather forecasting, and earthquake prediction data, as well as physiological data.

5. CONCLUSIONS

We propose *PhysioCover*, which automatically recovers missing values, and summarizes physiological time series data consisting of multiple dimensions. It computes the optimal missing values, and identifies a specific pattern using detected hidden variables. The reduction of data based on hidden variables can be used as learning data. Moreover, the proposed method can reduce the processing complexity and memory requirements. This method provides better results than other recovery methods such as interpolation and MSVD. The proposed approach is a valuable method for the accurate analysis of physiological signals. Its effectiveness is demonstrated by the accurate recovery of missing values and the automatic detection of hidden variables from physiological signals.

In this paper, we only evaluated physiological time series data for missing value recovery. For further work, we will assess various types of multiple time series data, and will use a robust method that can treat multi-way physiological data or multi modal physiological data.

ACKNOWLEDGMENTS

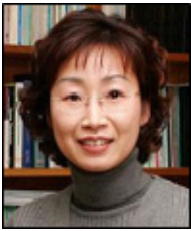
“This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST)(2013-056480)”, “This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1014) supervised by the NIPA(National IT Industry Promotion Agency)”.

REFERENCES

- [1] F. Canento, A. Fred, H. Silva, H. Gamboa, and A. Lourenço, "Multimodal biosignal sensor data handling for emotion recognition," *Proc. IEEE Sensors*, 2011, pp. 647-650.
- [2] S. Chiappa and D. Barber, "EEG classification using generative independent component analysis," *Neurocomputing*, vol. 69, 2006, pp. 769-777.
- [3] T. Rocha, S. Paredes, P.D. Carvalho, and J. Henriques, "Prediction of acute hypotensive episodes by means of neural network multi-models," *Computers in Biology and Medicine*, vol. 41, 2011, pp. 881-890.
- [4] I. Stanimirova, M. Daszykowski, and B. Walczak, "Dealing with missing values and outliers in principal component analysis," *Talanta*, vol. 72, 2007, pp. 172-178.
- [5] J. Henriques and T. Rocha, "Prediction of acute hypotensive episodes using neural network multi-models," *Computers in Cardiology*, vol. 36, 2009, pp. 549-552.
- [6] J. Paalasmaa, D. J. Murphy, and O. Holmqvist, "Analysis of Noisy Biosignals for musical performance," *Proc. IDA'12*, 2012, pp. 241-252.
- [7] X. Chen, D. Xu, G. Zhang, and R. Mukkamala, "Forecasting acute hypotensive episodes in intensive care patients based on a peripheral arterial blood pressure waveform," *Computers in Cardiology*, vol. 36, 2009, pp. 545-548.
- [8] S. Papadimitriou, J. Sun, and C. Faloutsos, "Streaming Pattern Discovery in Multiple Time-Series," *Proc. VLDB'05*, 2005, pp. 697-708.
- [9] E. Adams, B. Walczak, C. Vervaet, P. G. Risha, and D. L. Massart, "Principal component analysis of dissolution data with missing elements," *International Journal of Pharmaceutics*, vol. 234, 2002, pp. 169-178.
- [10] S. Roweis, "EM algorithms for PCA and SPCA," *Proc. NIPS'97*, 1997, pp. 626-632.
- [11] L. Smith, *A Tutorial on Principal Components Analysis*, Cornell University, USA, 2002.
- [12] J. Sun, S. Papadimitriou, and C. Faloutsos, "Online Latent Variable Detection in Sensor Networks," *Proc. ICDE'05*, 2005, pp. 1126-1127.
- [13] J. Y. Pan, H. Kitagawa, M. Hamamoto, and C. Faloutsos, "AutoSplit: Fast and Scalable Discovery of Hidden Variables in Stream and Multimedia Databases," *Proc. PAKDD'05*, 2005, pp. 519-528.
- [14] <http://physionet.org/challenge/2009/>
- [15] F. Chiarugi, I. Karatzanis, V. Sakkalis, I. Tsamardinos, Th. Dermitzaki, M. Foukarakis, and G. Vrouchos, "Predicting the Occurrence of Acute Hypotensive Episodes: The PhysioNet Challenge," *Computers in Cardiology*, vol. 36, 2009, pp. 621-624.
- [16] L. Zhao, T. Chai, and Q. Cong, "Operating Condition Recognition of Predenitrification Bioprocess Using Robust EMPCA and FCM," *Proc. WCICA'06*, 2006, pp. 9386-9390.
- [17] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, B. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, 2001, pp. 520-525.
- [18] K. S. Ng, H. J. Yang, and S. H. Kim, "Hidden pattern discovery on event related potential EEG signals," *Biosystems*, vol. 97, 2009, pp. 15-27.
- [19] X. T. Doan, R. Srinivasan, P. M. Bapat, and P. P. Wangikar, "Detection of phase shifts in batch fermentation via statistical analysis of the online measurements: A case study with rifamycin B fermentation," *Journal of Biotechnology*, vol. 132, 2007, pp. 156-166.
- [20] L. Lehman, M. Saeed, G. Moody, and R. Mark, "Similarity-based searching in multi-parameter time series databases," *Computers in Cardiology*, vol. 35, 2008, pp. 653-656.
- [21] M. N. Norazian, Y. A. Shukri, R. N. Azam, and A. M. M. Al Bakri, "Estimation of missing values in air pollution data using single imputation techniques," *ScienceAsia*, vol. 34, 2008, pp. 341-345.
- [22] X. Wang, A. Li, Z. Jiang, and H. Feng, "Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme," *BMC Bioinformatics*, vol. 7:32, 2006, pp. 1-10.
- [23] G. N. Brock, J. R. Shaffer, R. E. Blakesley, M. J. Lotz, and G. C. Tseng, "Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes," *BMC Bioinformatics*, vol. 9, 2008, pp. 1-12.
- [24] S. Sharma, P. Lingras, and M. Zhong, "Effect of missing values estimations of traffic parameters," *Transportation Planning and Technology*, vol. 27, 2004, pp. 119-144.
- [25] I. Milovanovic and D. B. Popovic, "Principal Component Analysis of Gait Kinematics Data in Acute and Chronic Stroke Patients," *Computational and Mathematical Methods in Medicine*, vol. 2012:649743, 2012, pp. 1-8.
- [26] J. Lee, and R. G. Mark, "An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care," *BioMedical Engineering OnLine*, vol. 9:62, 2010, pp. 1-17.
- [27] M. P. S. Chawla, "Detection of Indeterminacies in Corrected ECG Signals Using Parameterized Multidimensional Independent Component Analysis," *Computational and Mathematical Methods in Medicine*, vol. 10, 2009, pp. 85-115.
- [28] X. Jiang, L. Zhang, Q. Zhao, and S. Albayrak, "ECG Arrhythmias Recognition System Based on Independent Component Analysis Feature Extraction," *Proc. TENCON'06*, 2006, pp. 1-4.

**Sun-Hee Kim**

She received the B.S in Multimedia from Korean Educational Development Institute in 2004 and the M.S. degree in Computer Science from Dongguk University, Korea in 2006. She received the Ph. D. degrees in Computer Science from Chonnam National in 2011. She recently works in Chonnam National University as a researcher. Her research interests include Data Mining, Machine Learning and Bioinformatic.

**Hyung-Jeong Yang**

She received her B.S., M.S. and Ph. D from Chonbuk National University, Korea. She is currently an associate professor at Dept. of Electronics and Computer Engineering, Chonnam National University, Gwangju, Korea. Her main research interests include

multimedia data mining, pattern recognition, artificial intelligence, e-Learning, and e-Design.

**Soo-Hyung Kim**

He received his B.S. at Dept. of Computer Engineering, Seoul National University, and M.S. and Ph.D. at Dept. of Computer Science, Korea Advanced Institute of Science and Technology, Korea. He is currently a professor at Dept. of Electronics and Computer Engineering

and a vice-Dean of the Engineering College, Chonnam National University, Gwangju, Korea.

**Guee-Sang Lee**

He received his BS in electrical engineering and his MS in computer engineering from Seoul National University, Seoul, Rep. of Korea, in 1980 and 1982, respectively. He received his PhD in computer science from Pennsylvania State University,

University Park, PA, USA, in 1991. He is currently a professor of the Department of Electronics and Computer Engineering at Chonnam National University, Gwangju, Rep. of Korea. His main research interests are image processing, computer vision, and video technology.