一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
ASN2017-89 (2018-01)

Highly Available Data Interpolation (HADI) Scheme for Automated System in Smart Home Environment

Cheng LI† Yuan FANG†† Yuto LIM† Yasuo TAN†

†Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi City, Ishikawa Prefecture, 923-1292 Japan

††Dalian Polytechnic University, Qinggongyuan 1#, Dalian, Liaoning, China

E-mail: †{s1510216, ylim,ytan}@jaist.ac.jp, ††fangy@dlpu.edu.cn

Abstract Today, hundreds or even thousands of sensors are spatially deployed in the smart home environment. Through these sensed data, an automated system, e.g., home energy management system (HEMS) requires high availability data to be processed and controlled. If some sensors result fault data, this can lead to a low availability data to the automated system. As a result, the automated system can produce an undesired output parameters and led to the entire system malfunction. In this paper, a generalized highly available data interpolation (HADI) scheme is proposed to take advantages of other heterogeneous data. Numerical simulation reveals that our proposed scheme can achieve high data availability for the automated system with minimum cost computation.

Keywords Smart Homes, Sensor, Data Restoration, Data Availability, Data Interpolation

1. INTRODUCTION

Today, numerous advanced Internet of Things (IoT) technologies and devices have been implemented in smart home environment. Due to the remarkable sensing, communication and processing technologies and devices, the interconnection between physical and virtual things is successfully achieved. In IoT- enabled applications, sensor networks are the most important component, since critical information from both external surroundings and inner systems is sampled by networked sensors [1]. As a typical research field in IoT, smart home researches make full use of sensor networks, which are sensing the ambient physical information or even detecting the human activities. With the information collected from sensor networks, several automated systems such as home energy management systems (HEMS) or heating, ventilation and air conditioning (HVAC) systems have been achieved. Therefore, sensors are playing a significant role in smart home.

However, according to [2], experiments revealed that sensors in smart home environment are facing various problems which result in the sensor faults or even failures. Fault sensors will generate unavailable data, and these data will be imported into automated systems. Unavailable data can cause undesired control, which not only cost additional energy consumption, but also carry risk to normal operation of actuators. For EETCC system, faulty sensors transfer unavailable data to EETCC controller, which may trigger temporary invalidation. Furthermore, actuators which receive continual unavailable data probably result in system failure. Consequently, efficiency of EETCC is challenged.

In this paper, a novel data restoration scheme HADI is proposed, and HADI will focus on restoring relative accurate and absolute available data to maintain regular operation of automated system. Two main contributions are achieved through HADI. First, this paper illustrates the definition of availability in smart home environment. Through the real experiment data which is obtained from iHouse, the availability of temperature, relative humidity, solar irradiance and wind speed are detected. This paper mainly focuses on restoring low available solar irradiance and simulation results show that high availability is guaranteed. Second, the novelty is mainly reflected in taking advantage of spatiotemporal heterogeneous data, hence, processing time and training samples are reduced dramatically which differs from the general approaches.

The rest of the paper is structured as follows. Section 2 shows the background on definition of availability and related works on data restoration. Section 3 details the HADI models and mathematical expressions. Section 4 shows the evaluation of the HADI scheme, including the verification of equations, numerical analysis and representative comparisons which is followed by conclusions in Section 5.

2. BACKGROUND

2.1. iHouse and EETCC System

iHouse is an advanced facility of smart home environment which is located at Nomi city, Ishikawa

prefecture. In iHouse, over 300 objects including various sensors, appliances and electronic devices are connected with a specialized communication protocol ECHONET for smart home issues. In this paper, the raw data is almost obtained from iHouse.

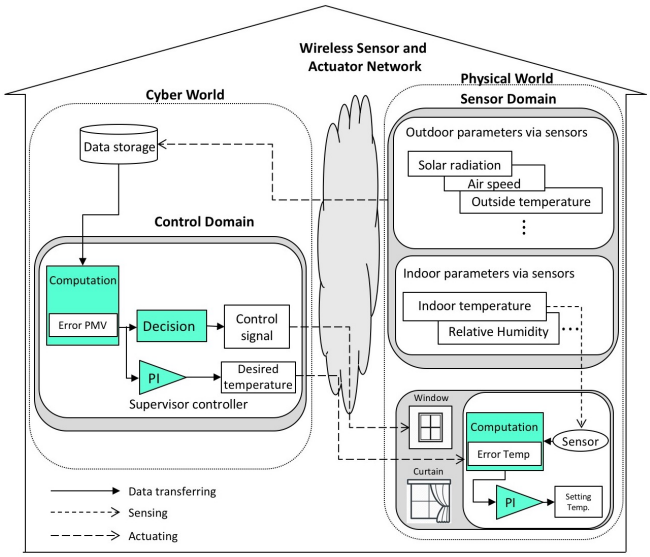

Fig. 1: EETCC system architecture

Energy efficient thermal comfort control (EETCC) system [3] is an advanced automated system that achieved demands on occupants' thermal comfort, furthermore, EETCC also succeed in relative high energy efficiency. As the principal inputs data to perform EETCC algorithm, temperature, relative humidity, solar radiation and air speed are sensed by sensors distributed in smart home environment. Thus, available data play a vital role in performance of EETCC system. Thus, this paper proposes a scheme that concentrates on maintaining high data availability for EETCC system. **Fig .1** shows architecture of EETCC system.

2.2. Data Availability

In [4], availability is general purposed as following equation :

$$\lim_{t\to\infty} A(t) = A = \frac{MTTF}{MTTF + MTTR} \qquad (1)$$

where t denotes the time of item, moreover, $MTTF$, $MTTR$ is the mean time to failure and to repair, respectively. Therefore, for sensor x, we have the sensor availability A_x defined by faults in this scheme:

$$\lim_{t\to T} A_x(t) = A_x = \frac{IAD}{(IUD + IAD)} \qquad (2)$$

where T is the operation time. Similarly, IAD, IUD is the interval of available and unavailable data, respectively.

2.3. Unavailable Data Description

In this paper, we investigated availability of the temperature, relative humidity, solar irradiance and wind speed data. Besides, we don't consider the data lost and data delay as unavailable, and the process of unavailable data investigation is shown in **Fig .2**. In addition, we classify the unavailable format for single data according to [5], and as shown in **Table.1**

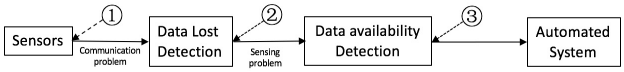

Fig. 2: Process of unavailable data investigation

Table 1: Unavailable format and description

Format	Description
Outlier	Isolated data point or sensor unexpectedly distant from models
"Stuck-at"	Multiple data points with a much greater than expected rate of change
Cailbration	Sensor reports values that are offset from the ground truth

Table 2: Pattern of unavailable data and description

Pattern	Description
Intermittent	Data shows unavailabe in one or several seconds. Most intermittent unavialable data are caused by outlier.
Continual	Unavailabe data last for a long period, a few minutes, even hours. Most continual unaviable data are related with "Stuck-at" and cailbration.

In **Table.2**, we define the pattern of unenviable data as intermittent and continual. It will help us to recognize the unavailable data and figure out the restoration method.

Through the investigation, we define threshold $\theta = [0,1360\ W/m^2]$ as range of available solar irradiance value. Through availability investigation in year 2016, we find unavailable data in daytime last 574.2 hours. It means that pyranometer is unavailable nearly 1.6 hours every single day. Meanwhile, most of unavailable data reveal the continual pattern. Hence, this paper will focus on solar irradiance restoration.

2.4. Related work

Researches related with data restoration have been carried out in the last 20 years. Related works are mainly divided in three kinds of mechanism: Principle component analysis, Linear regression, Artificial Neural Network.

In [6], PCA first achieve data recovery for HVAC system, however, this approach merely considers the temporal data

of target data, which results in weak response by data variation. Given by the progress of T. Yu et al. [1], a recursive principal component analysis (R-PCA) is proposed. R-PCA represents a remarkable efficiency on data fault detection, data aggregation and recovery accuracy, whereas recursion increase burden on processing units. Meanwhile, R-PCA costs longer processing time due to the high complexity.

Linear regression is a widely-used approach in data analysis. Efficient temporal and spatial data recovery (ETSDR) [7] integrate Auto Regressive Integrated Moving Average (ARIMA) model with spatiotemporal data, furthermore, realized the dynamic model identification and accurate intermittent data recovery. But performance of dealing with continual unavailable data by using ETSDR is a great challenge. Since the ETSDR update the linear model for every single data for each sensor, the processing time and burden on processor are doubted as well.

In addition, Artificial Neural Network (ANN) [8] has been applied on temperature recovery for HVAC system in 1996. And neural network based model is optimized by Z. Liu et al. [9] by deep multimodal encoder (DME) framework, which has excellent performance on high unavailability. However, either ANN or DME request the training sample as reliable data, besides, the iterative process of neural network is time-consuming for dynamic systems. Therefore, an approach is expected to reduce processing time and achieve accurate data restoration will be presented in this paper.

3. MODELS OF HADI SCHEME

3.1. HADI Structure

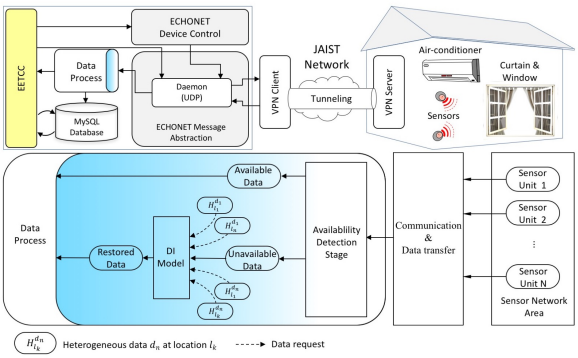

Fig. 3: Structure of HADI scheme

Fig. 3 shows the structure of HADI scheme, we suppose HADI is located at data process module. Raw data from each sensor unit is identified as available or unavailable data in availability detection stage. Available data will continue the data process. However, Unavailable data are supposed to be restored by Data Interpolation (DI) model with spatiotemporal heterogeneous data. Therefore, due the HADI scheme, high availability is guaranteed.

3.2. Symbols and Description

Table 3: Symbols and description

Symbol	Description
n	Number of data
t	A certain time instance
C	Set of sequence number of all available data in the horizon
m	Number of available data in horizon
K	Number of snesor's locations
i	A set of location of target data, $i \in [1, K]$
$Y_i(n,t)$	Target data at time t
$\check{Y}_i(n,t)$	Available target data at time t
$\hat{Y}_i\ (n,t)$	Unavailable target data at time t
$\check{Y}_i(C(1:m)$	Available target data in horizon
$\tilde{Y}_i$	Restored data by hetergenous data
$X_{1:K}(n,t)$	Set of other heterogeneous data at time t
$X_{1:K}(C(1:m))$	Heterogeneous data in horizon at the same moment with target data
A_{H_Y}	Percentage of available data in horizon
$H_{X(1:K)}$	Horizon of other heterogeneous data
$H_{Y(i)}$	Horizon of target data
θ_y	Threshold of data
h	Determined length of horizon (10min)
p	Set of locations determind the minimum root mean square error (RMSE)

3.3. HADI Algorithm

Fig. 4 shows the HADI algorithm. We classify the raw data as available and unavailable by threshold. Then, an appropriate horizon with temporal raw data is determined for training. Meanwhile, we determine the same length horizon of correlated spatial heterogeneous data. For preventing horizon suffering from too much unavailable data in raw data, horizon of target data is reconsidered with set of available and restored data when unavailable data occupy more than 50 percent. As all horizons are completed, the positions of available data in target data horizon compose the set $C(1:m)$. Besides, to keep time synchronization, we need to find the corresponding data $X_{1:K}(C\ (1:m))$ in heterogeneous horizon. DI model contains the correlations between X and Y, these correlations can be described as $Y = f(X)$, hypothetically.

With spatial heterogeneous data as inputs, K groups of simulated data are generated. Through calculating RMSE between K groups simulated data and available data in target data horizon, we can find sensors at locations p determine the minimum RMSE. Finally, with current heterogeneous data input, restoration of target data is accomplished by inputting correlated heterogeneous data at location p into DI model.

Algorithm 1 Highly available data interpolation (HADI)

if $Y_i(n,t) > \theta_y$ **then** // $Y_i(n,t)$ is unavailable.
  $\hat{Y}_i(n,t) \leftarrow Y_i(n,t)$
  // Determine spatial and temporal horizon as:
  $H_{Y_i} = Y_i(n-h) : Y_i(n-1)$
  $H_{X_{1:K}} = X_{1:K}(n-h) : X_{1:K}(n-1)$
  if $A_{H_{Y_i}} < 50\%$ **then**
    // Available data are rare, reconsider horizon with restored data
    $H_{Y_i} = \left\{\check{Y}_i \bigcup \tilde{Y}_i\right\}(n-h) : \left\{\check{Y}_i \bigcup \tilde{Y}_i\right\}(n-1)$
  end if
  // Acquire the available data set in horizon $H_{Y_{(i)}}$
  $\hat{Y}_i(C(1:m)) = \left[\hat{Y}_i(C(1)), \cdots, \hat{Y}_i(C(m))\right]$
  for each $j \in [1, K]$ **do**
    $X_j(C(1:m)) = [X_j(C(1)), \cdots, X_j(C(m))]$
    // Substitute $X_j(C(1:m))$ for Data Interpolation(DI) model
    $\tilde{Y}_j(C(1:m)) = \left[\left(\tilde{Y}_j(C(1))\right), \cdots, \left(\tilde{Y}_j(C(m))\right)\right]$
    // Calculate the Root Mean Square Error
    $RMSE(j) = \sqrt{\frac{(\tilde{Y}_j(C(1)) - \hat{Y}_i(C(1)))^2 + \cdots + (\tilde{Y}_j(C(m)) - \hat{Y}_i(C(m)))}{m}}$
  end for
  $RMSE(p) = min(RMSE(1:K))$
  // Calculate current time $\tilde{Y}_i(n,t)$
  $\tilde{Y}_i(n,t) = f(X_p(n,t))$
else
  // $Y_t(n,l_i)$ is available.
  $\check{Y}_t(n,l_i) \leftarrow Y_t(n,l_i)$
end if

Fig. 4: HADI algorithm

4. NUMERICAL ANALYSIS

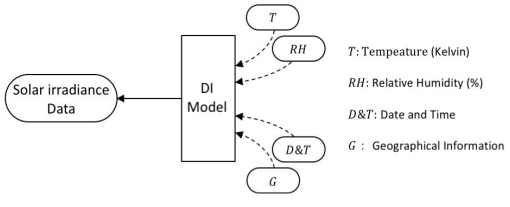

Fig. 5: DI model for solar irrdiance restoration

In this section, we will evaluate the efficiency of HADI scheme and solve the low availability of solar irradiance. As shown in **Fig. 5**, heterogeneous data inputs include the time information, geographical information, temperature and relative humidity data from sensors distributed in 11 spots of iHouse.

4.1. DI Model for Hourly Solar Irradiance

In this paper, we apply an improved hourly solar irradiance method for conversion. According to [10], the equations are described as follows:

Solar irradiance R (W/m^2) and global solar irradiance R_{clear} (W/m^2) can be expressed as:

$$R = \tau_c R_{clear} \tag{3}$$

where τ_c denotes a radiative transmittance coefficient, which is supposed to be an empirical function of relative sunshine duration:

$$\tau_c = a + bn/N + c(n/N)^2 \tag{4}$$

where a, b, c are the model parameter. n/N represents relative sunshine duration. Moreover, R_{clear} is consist of surface beam irradiance $R_{b,clear}$(W/m^2) and solar diffuse irradiance $R_{d,clear}$ (W/m^2):

$$R_{clear} = R_{b,clear} + R_{d,clear} \tag{5a}$$

$$R_{b,clear} = R_0 \bar{\tau}_{b,clear} \tag{5b}$$

$$R_{d,clear} = R_0 \bar{\tau}_{d,clear} \tag{5c}$$

where R_0(W/m^2) is the solar irradiance on a horizontal surface at the extraterrestrial level [10]. The broadband solar beam radiative transmittance $\bar{\tau}_{b,clear}$ and radiative transmittance $\bar{\tau}_{d,clear}$ are able to described as:

$$\bar{\tau}_{b,clear} \approx \max(0, \bar{\tau}_{oz}\bar{\tau}_w\bar{\tau}_g\bar{\tau}_r\bar{\tau}_a - 0.013) \tag{6a}$$

$$\bar{\tau}_{d,clear} \approx 0.5[\bar{\tau}_{oz}\bar{\tau}_g\bar{\tau}_w(1 - \bar{\tau}_a\bar{\tau}_r) + 0.013] \tag{6b}$$

$$\bar{\tau}_g = \exp(-0.0117(m')^{0.3139}) \tag{6c}$$

$$\bar{\tau}_r = \exp[-0.008735(m')(0.547 + 0.014(m') - 0.00038(m')^2 + 4.6\times10^{-6}(m')^3)^{-4.08}] \tag{6d}$$

$$\bar{\tau}_w = \min[1.0, 0.909 - 0.036ln(mw)] \tag{6e}$$

$$\bar{\tau}_{oz} = \exp[-0.0365(ml)^{0.7136}] \tag{6f}$$

$$\bar{\tau}_a = \exp\{-m\beta[0.6777 + 0.1464(m\beta) - 0.00626(m\beta)^2]^{-1.3}\} \tag{6g}$$

$$m = 1/[\sin h + 0.15(57.296h + 3.885)^{-1.253}] \tag{6h}$$

$$m' = m\,p/p_0 \tag{6i}$$

where $\bar{\tau}_{oz}, \bar{\tau}_w, \bar{\tau}_g, \bar{\tau}_r, \bar{\tau}_a$ are the radiative transmittance due to ozone absorption, water vapour absorption, permanent gas absorption, Rayleigh scattering and aerosol extinction, respectively. In addition, h (rad) denotes the solar elevation, m refers to relative air mas, m' is the pressure-corrected air mass, p_0 (Pa) is the standard atmospheric pressure, p (Pa) is the surface pressure, then:

$$p = p_0 \exp(-z/H_T) \tag{7}$$

where z(m) is surface elevation from the mean sea level, $H_T =$ is the scale height of an isothermal atmosphere and $H_T = 8430$ (m). To get thickness of ozone layer l (cm), an empirical formula is used:

$$l = 0.44 - 0.16\times(\sqrt{[(|\emptyset| - 80)/60]^2 + [(d - 120)/(263 - |\emptyset|)]^2}) \tag{8}$$

where $\emptyset$(degree) denotes the latitude, d is determined by Julian day J_d as follows:

$$d = \begin{cases} J_d & if\ J_d < 300 \\ J_d - 366 & if\ J_d > 300 \end{cases} \quad (9)$$

In(6e), w(cm) is the precipitable water, which use the relative humidity rh (%) and air temperature T (Kelvin). Ångström turbidity coefficient β in(6g) is calculated as follows:

$$\beta = (0.025 + 0.1\cos^2\emptyset)\exp(-0.7z/1000) \quad (10)$$

4.2. Verification of Solar Irradiance Equations

Simulation results are different due to meteorological data and geographical parameters. Therefore, before we implement equations in DI model, we verify the equations with observed data in [10]. Coefficients and parameters are listed in **Table 4**. And in **Fig. 6**, result of verification proves that equations fit observed solar irradiance data dramatically.

Table 4: Coefficient and Parameters for verification.

Coefficient	Value	Parameter	Value (unit)
a	0.456	z	1219 (m)
b	0.3566	$\emptyset$	31.8 (°)
c	0.1874		

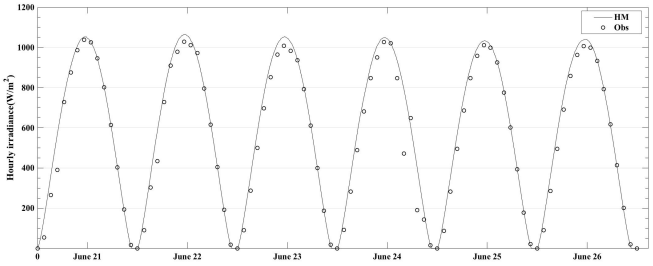

Fig. 6 Verification of Solar Irradiance Equations

4.3. Evaluation of HADI Scheme

Table 5: Coefficient and Parameters for DI Model

Coefficient	Value	Parameter	Value (unit)
a	0.2976	z	132 (m)
b	0.4119	$\emptyset$	36.44 (°)
c	-0.0254		

The coefficients and parameters of local geographical information for DI model are shown in **Table. 5**. To evaluate the performance, we use the RMSE and mean absolute error(MAE). RMSE reveals the accuracy of simulation results, then, for data series with length of N, RMSE can be written as:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(d(n) - \tilde{d}(n))^2}{N}} \quad (11)$$

Moreover, MAE is used to measure how close the simulated values are to original measured value:

$$MAE = \frac{1}{N}\sum_{n=1}^{N}\left|d(n) - \tilde{d}(n)\right| \quad (12)$$

In (11) and (12), $d(n)$ denotes the original measured data, and $\tilde{d}(n)$ is the simulated data.

To show the performance of HADI better, we compare HADI with the ETSDR scheme. Simulation results consist of intermittent and continual unavailable data restoration by HADI and ETSDR, respectively.

4.3.1. Intermittent Unavailable Data Restoration

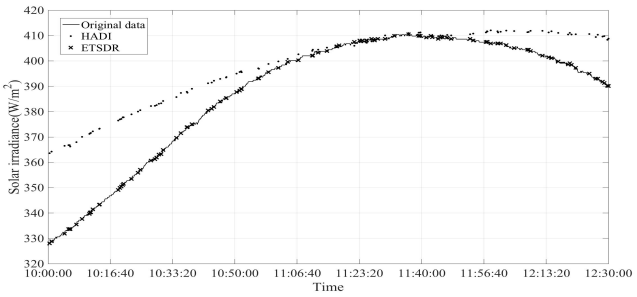

Fig. 7 Intermittent unavailable data restoration

Fig. 7 shows an example of intermittent unavailable data restoration. In 100 minutes' dataset, we interpolate 10 minutes' unavailable data at random time. Although ETSDR performs a higher accuracy, however, high availability is guaranteed with HADI as well as ETSDR.

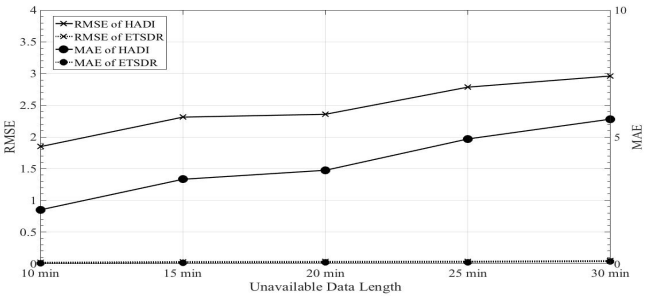

Fig. 8 Performance of intermittent unavailable data restoration

In addition, we increase the percentage of unavailable data from 10% to 30% as shown in **Fig. 8**, and results reveal that ETSDR shows extremely high accuracy without influence of unavailable data increase.

4.3.2. Continual Unavailable Data Restoration

On the other hand, we consider the circumstance of

continual unavailable. **Fig. 9** shows continual unavailable restoration of same length unavailable data with **Fig. 7**. Results reveal that HADI performs better than ETSDR on dealing with continual unavailable data. Without real-time data to update ARIMA model in ETSDR, the restoration data regress to straight line rapidly, and consequent enormous error gradually.

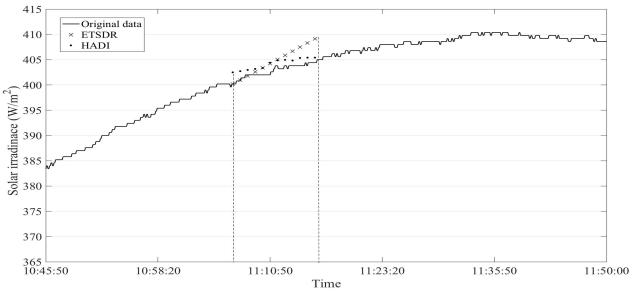

Fig. 9 Continual unavailable data restoration

However, as shown in **Fig. 10**, HADI shows a steady performance on accuracy whose RMSE and MAE vary at a low value despite percentage of unavailable data grow.

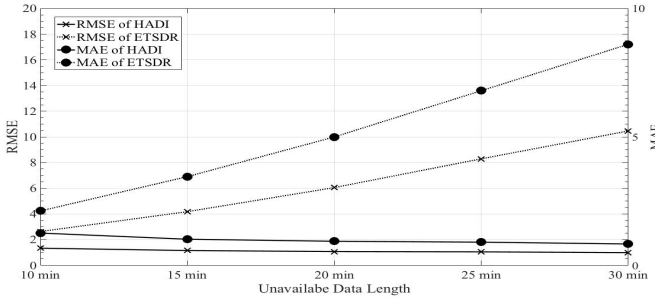

Fig. 10 Performance of continual unavailable data restoration

4.4. Processing Time Comparison

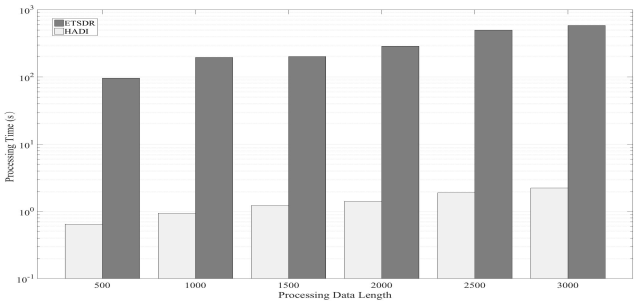

Fig. 11 Performance of processing time

Processing time is a significant feature for real-time automated system as well, therefore we compare the processing time between two schemes. In each processing data length, we keep percentage of unavailable data as 10%, and **Fig. 11** shows that in logarithmic scale, it is obvious that ETSDR costs much more processing time than HADI. Hence, we can conclude that HADI achieves a dramatic efficiency on continual unavailable data problems.

5. CONCLUSION

In this paper, we purposed a new data restoration scheme based on spatiotemporal heterogeneous data to solve continual unavailable issues. HADI is able to maintain relative high accuracy and absolute available. Furthermore, HADI shows amazing efficiency on processing time. Our future work will focus on attempting to introduce more models and correlations into HADI so that automated system will be isolated from unavailable data.

References

[1] T. Yu, X. Wang and A. Shami, "Recursive Principal Component Analysis based Data Outlier Detection and Sensor Data Aggregation in IoT Systems," Internet of Things Journal. IEEE, vol.4, no.3, pp.2207-2216, December, 2017.

[2] H. Timothy W, S. Vijay and J. Lu, "The hitchhiker's guide to successful residential sensing deployments, " Proc. The 9th ACM Conf. on ENSS., pp. 232-245, 2011.

[3] Z. Cheng, S. Wai Wai, Y. Tan and A.O. Lim, "Energy efficient thermal comfort control for cyber-physical home system," In *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on*, pp.797-802, 2013.

[4] M. Rausand and A. Hoyland, "Failure Models," in *Models*, L. Leemis, 2nd Ed. New York: Wiley, pp.15-75, 2004.

[5] K. Ni, et al. "Sensor network data fault types," *ACM Transactions on Sensor Networks (TOSN)*, vol.5, no.3, pp.25, May.2009.

[6] X. Hao, G. Zhang, Y. Chen, "Fault-tolerant control and data recovery in HVAC monitoring system," *Energy and buildings*, vol.37, no.3, pp.175-180, Feb. 2005.

[7] N. Naushin, Y. Tan and A.O. Lim, "Efficient Temporal and Spatial Data Recovery Scheme for Stochastic and Incomplete Feedback Data of Cyber-Physical Systems," *Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on*, pp. 192-197,2014

[8] W.Y. Lee, J.M. House and D.R. Shin, "Fault diagnosis and temperature sensor recovery for an air-handing unit," *Transactions-American Society Of Heating Refrigerating And Air Conditioning Engineers*, no.103, pp.621-633, December, 1997.

[9] Z. Liu, et al. "Deep fusion of heterogeneous sensor data," In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp.59-65, 2017.

[10] K. Yang, and T. Koike, "A general model to estimated hourly and daily solar radiation for hydrological studies," *Water Resources Research*, vol.41, no.10, October, 2005.