A Scheme for Sensor Data Reconstruction in Smart Home

Yegang Du(✉)

Dalian Maritime University, 1 Linghai Road, Dalian, China
yg.du@dlmu.edu.cn

Abstract. Today, a large number of sensors are spatially deployed to monitor the environment in smart home. Through sensed data, home automation systems can perceive the surroundings and further provide corresponding services. However, such systems rely on highly available data which cannot be guaranteed by sensors. If some sensors report unavailable data to home automation systems without verification, the systems might malfunction and even affect the safety of residents. In this paper, a generalized highly available data interpolation (HADI) scheme is proposed to serve as a guarantor for sensed data in smart home. HADI takes advantage of the relationship between the faulty sensor and the other heterogeneous sensors to reconstruct the highly available data. Experiments reveal that our proposed scheme can achieve high data availability with less computation cost.

Keywords: Smart home · Data restoration · Data availability · Data interpolation · Home sensor

1 Introduction

Nowadays, numerous advanced Internet of Things (IoT) technologies and devices have been implemented in the smart home environment. Due to the remarkable sensing, communication, processing technologies and devices, the interconnection be-tween physical and virtual things is successfully achieved. In IoT-enabled applications, sensor networks are the most important component. Critical information from both external surroundings and inner systems is sampled by networked sensors [1]. As a typical research field in IoT, the smart home makes full use of sensor networks to sense the ambient physical information and even detect human activities [2]. With the information collected from sensor networks, several home automation 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 [3], experiments revealed that sensors in smart home environment are facing various problems which result in sensor faults or even failures. Fault sensors will generate unavailable data, and these data will be imported into home automation systems. Unavailable data may cause undesired

M. Yang et al. (Eds.): NSS 2021, LNCS 13041, pp. 304–313, 2021.
https://doi.org/10.1007/978-3-030-92708-0_19

control, which not only cost additional energy consumption but also carry risk to normal operation of actuators. For example, in the home energy control system [4], faulty sensors transfer unavailable data to the controller, which may trigger temporary invalidation. Furthermore, actuators that receive continual unavailable data probably result in system failure. Consequently, the efficiency of the control system is challenged.

In this paper, a novel data interpolation scheme HADI is proposed to restore accurate and available data to maintain the regular operation of home automation systems. Temperature, relative humidity, solar irradiance, and wind speed are detected through the data obtained from our experiment platform. This paper mainly focuses on restoring highly available solar irradiance data. 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.

Three main contributions are achieved by HADI. First, this paper summarizes the format and pattern of unavailable data and illustrates the definition of data availability in smart home. Then HADI algorithm is proposed, and analysis in theory is given to explain the effectiveness of our algorithm. Moreover, several experiments have been done to compare the performance of our HADI with the state-of-the-art method. Experiments show that high availability of data is guaranteed, meanwhile, the processing time and training samples are reduced dramatically.

The rest of the paper is structured as follows. Section 2 shows the background and related works on data restoration. Section 3 illustrates the definition and categories of data availability in smart home. Section 4 details the HADI models and mathematical expressions. Section 5 demonstrates solar irradiance data restoration using HADI, and shows the evaluation of the HADI scheme. The paper is concluded in Sect. 6.

2 Related Work

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

In [5], PCA first achieves data recovery for HVAC system, however, this approach merely considers the temporal data of target data, which results in a weak response by data variation. Given by the progress of 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 increases the burden on processing units. Meanwhile, R-PCA costs a 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) [6] integrate Auto Regressive Integrated Moving Average (ARIMA) model with spatiotemporal data, furthermore, realized the dynamic model identification and accurate intermittent data recovery.

But the 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 the processor are doubted as well.

In addition, Artificial Neural Network (ANN) [7] has been applied on temperature recovery for HVAC systems in 1996. And neural network-based model is optimized by Z. Liu et al. [8] by deep multimodal encoder (DME) framework, which has excel-lent performance on high unavailability. However, either ANN or DME requires the reliable data as training sample s, besides, the iterative process of the 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 Data Availability in Smart Home

3.1 Definition of Data Availability

In [9], availability is general purposed as the following equation:

$$\lim_{t\to\infty} = A = \frac{MTTF}{MTTF + MTTR} \tag{1}$$

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:

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

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

3.2 Unavailable Data Description

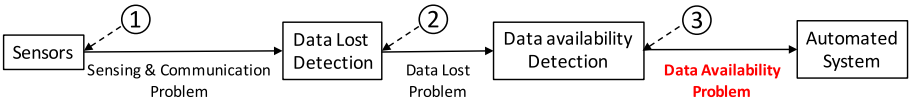

Fig. 1. Process of unavailable data investigation.

In this paper, we investigated availability of the temperature, relative humidity, solar irradiance and wind speed data. Note that, we don't consider the data loss and data delay in our model as shown in Fig. 1. In addition, we classify the unavailable for-mat for single data as shown in Table 1 according to [10].

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
Calibration	Sensor reports values that are offset from the ground truth

Table 2. Pattern of unavailable data and description.

Duration	Description
Intermittent	Data act as unavailable in one or several seconds. Most intermittent unavailable data are mainly caused by outlier or spike
Continual	Unavailable data last for a long period, a few minutes, even hours. Most continual unavailable data are related with "Stuck-at" and calibration

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

Through the investigation, we define threshold [0,1360] (W/m^2) as the range of available solar irradiance values. Through availability investigation in the year 2016, we find unavailable data in the daytime last 574.2 h. It means that the pyranometer is unavailable for nearly 1.6 h every single day. Meanwhile, most of the unavailable data reveals a continual pattern. Hence, this paper will focus on solar irradiance interpolation.

4 Models of HADI Scheme

4.1 HADI Structure

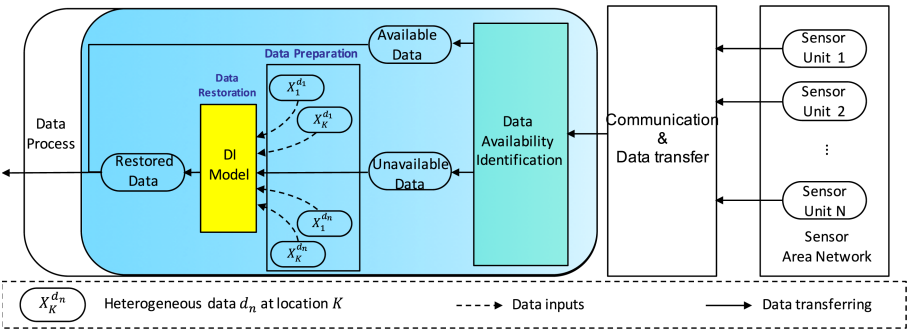

Fig. 2. Structure of HADI scheme.

Figure 2 shows the structure of the HADI scheme, we suppose HADI is located in the data process module. Raw data from each sensor unit is identified as available or unavailable data in the 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 to the HADI scheme, high availability is guaranteed.

4.2 HADI Algorithm

Algorithm 1. Highly available data interpolation (HADI) scheme.

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
  // Aggregate 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 $u = 1 : K$ **do**
    $X_u(C(1:m)) = [X_u(C(1)), \cdots, X_u(C(m))]$
    // Substitute $X_u(C(1:m))$ for Data Interpolation(DI) model
    $\widetilde{Y}_u(C(1:m)) = \left[\widetilde{Y}_u(C(1)), \cdots, \widetilde{Y}_u(C(m))\right]$
    // Calculate the Root Mean Square Error
    $\varphi(u) = \sqrt{\frac{\left(\tilde{Y}_u(C(1)) - \hat{Y}_u(C(1))\right)^2 + \cdots + \left(\tilde{Y}_u(C(m)) - \hat{Y}_u(C(m))\right)}{m}}$
  end for
  $\varphi(p) = min(\varphi(1:K))$
  // Calculate current time $\tilde{Y}_i(n,t)$
  $\tilde{Y}_i(n,t) = f(X_p(n,t))$
else
  // $Y_t(n,t)$ is available.
  $\check{Y}_t(n,t) \leftarrow Y_t(n,t)$
end if

As shown the Algorithm 1, 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 the horizon from suffering too much unavailable data in raw data, the horizon of target data is reconsidered with a set of available and restored data when unavailable data occupy more than 50%.

As all horizons are completed, the positions of available data in the 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 a 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, interpolation of target data is accomplished by inputting correlated heterogeneous data at location p into the DI model.

5 Di Model for Hourly Solar Irradiance

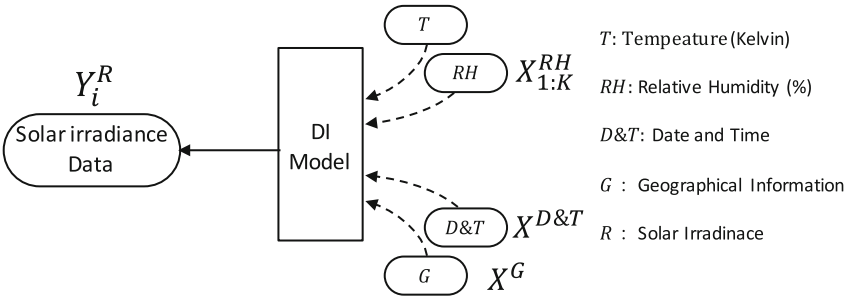

Fig. 3. DI model for solar irradiance interpolation.

In this section, we will apply our HADI scheme to solve the low availability of solar irradiance. As shown in Fig. 3, heterogeneous data inputs include the time information, geographical information, temperature, and relative humidity data from sensors distributed in 11 spots of the experiment platform.

5.1 Hourly Solar Irradiance

In this paper, we apply an improved hourly solar irradiance method for conversion. According to [11], 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 parameters. n/N represents relative sunshine duration. Moreover, R_{clear} (W/m^2) consists of surface beam irradiance $R_{b,clear}$ (W/m^2) and solar diffuse irradiance $R_{d,clear}$ (W/m^2):

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

$$R_{b,clear} = R_0 \overline{\tau}_{b,clear} \tag{6}$$

$$R_{d,clear} = R_0 \overline{\tau}_{d,clear} \tag{7}$$

where R_0 (W/m^2) is the solar irradiance on a horizontal surface at the extraterrestrial level [11]. The broadband solar beam radiative transmittance $\overline{\tau}_{(b,clear)}$ and radiative transmittance $\overline{\tau}_{(d,clear)}$) are able to be described as:

$$\overline{\tau}_{b,clear} \approx \max(0, \overline{\tau}_{oz}\overline{\tau}_w\overline{\tau}_g\overline{\tau}_r\overline{\tau}_a - 0.013) \tag{8}$$

$$\overline{\tau}_{d,clear} \approx 0.5[\overline{\tau}_{oz}\overline{\tau}_g\overline{\tau}_w(1 - \overline{\tau}_a\overline{\tau}_r) + 0.013] \tag{9}$$

$$\overline{\tau}_g = \exp(-0.0117(m')^{0.3139}) \tag{10}$$

$$\begin{aligned}\overline{\tau}_r = \exp[-0.008735(m')(0.547 + 0.014(m') - 0.00038(m')^2 \\ +4.6 \times 10^{-6}(m')^3)^{-4.08}]\end{aligned} \tag{11}$$

$$\overline{\tau}_w = \min[1.0, 0.909 - 0.036ln(mv)] \tag{12}$$

$$\overline{\tau}_{oz} = \exp[-0.0365(ml)^{0.7136}] \tag{13}$$

$$\overline{\tau}_a = \exp\{-m\beta[0.6777 + 0.1464(m\beta)^2]^{-1.3}\} \tag{14}$$

$$m = 1/[\sin h + 0.15(57.296h + 3.885)^{-1.253}] \tag{15}$$

$$m' = mp/p_0 \tag{16}$$

$$p = p_0 \exp(-z/H_T) \tag{17}$$

where $\overline{\tau}_{oz}, \overline{\tau}_w, \overline{\tau}_g, \overline{\tau}_r, \overline{\tau}_a$ are the radiative transmittance due to ozone absorption, water vapor absorption, permanent gas absorption, Rayleigh scattering, and aerosol extinction, respectively. In addition, h (rad) denotes the solar elevation, m refers to relative air mass, m' is the pressure-corrected air mass, p_0 (Pa) is the standard atmospheric pressure, p (Pa) is the surface pressure. l is the thickness of ozone and β in Eq. 14 is the Ångström turbidity coefficient. z is surface elevation from the mean sea level, H_T is the scale height of an isothermal atmosphere, and H_T is 8430.

5.2 Evaluation of HADI Scheme

In this section, we evaluate the efficiency of HADI. 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 a length of N, RMSE can be written as:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N} \left(d(n) - \widetilde{d}(n)\right)^2}{N}} \tag{18}$$

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

$$MAE = \frac{1}{N} \sum_{n=1}^{N} \left| d(n) - \widetilde{d}(n) \right| \tag{19}$$

In Eq. 18 and Eq. 19, $d(n)$ denotes the original measured data, and $\widetilde{d}(n)$ 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.

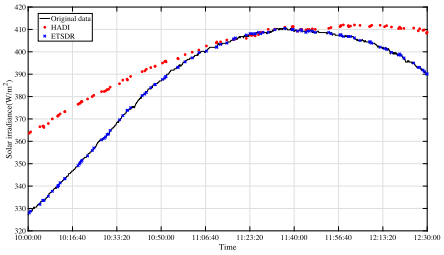

Fig. 4. Intermittent unavailable data interpolation.

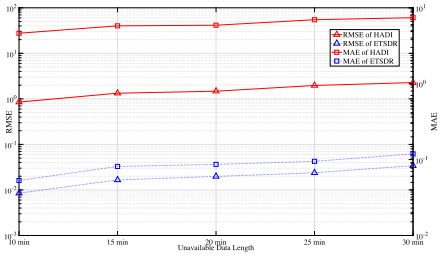

Fig. 5. Performance of intermittent unavailable data interpolation.

Intermittent Unavailable Data Interpolation. Figure 4 shows an example of intermittent unavailable data interpolation. In 100 min' dataset, we interpolate 10 minutes' unavailable data at a random time. Although ETSDR performs a higher accuracy, however, high availability is guaranteed with HADI as well as ETSDR.

In addition, we increase the percentage of unavailable data from 10% to 30% as shown in Fig. 5, and results reveal that ETSDR shows extremely high accuracy with-out the influence of unavailable data increase.

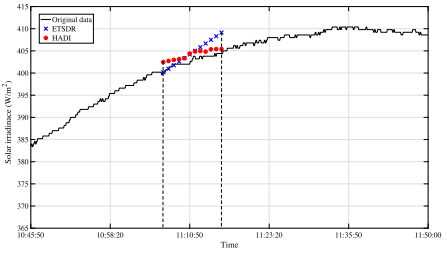

Fig. 6. Continual unavailable data interpolation.

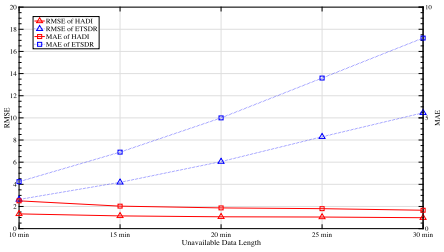

Fig. 7. Performance of continual unavailable data interpolation.

Continual Unavailable Data Interpolation. On the other hand, we consider the circumstance of continual unavailable. Figure 6 shows continual unavailable interpolation of same length unavailable data in Fig. 4. Results reveal that HADI performs better than ETSDR on dealing with continual unavailable data. Without real-time data to update the ARIMA model in ETSDR, the interpolation data regress to straight line rapidly, and consequent enormous error gradually.

However, as shown in Fig. 7, HADI shows a steady performance on accuracy whose RMSE and MAE vary at a low value despite the percentage of unavailable data growth.

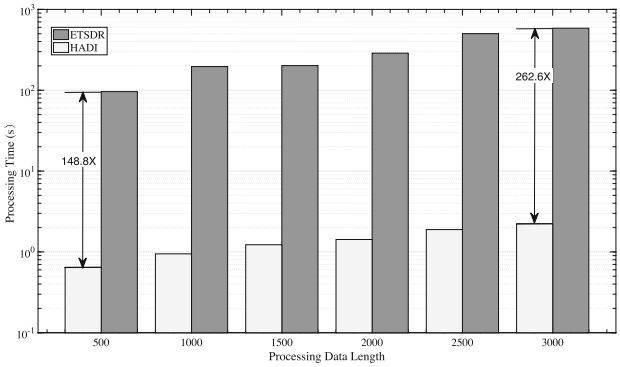

Fig. 8. Performance of processing time.

Processing Time Comparison. Processing time is a significant feature for real-time home automation systems as well, therefore we compare the processing time between two schemes. In each processing data length, we keep the percentage of unavailable data as 10%, and Fig. 8 shows that on a 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.

6 Conclusion and Future Work

In this paper, we purposed a new data interpolation scheme based on spatiotemporal heterogeneous data to solve continual unavailable issues. HADI is able to maintain relatively high accuracy and absolutely available. Furthermore, HADI shows amazing efficiency in processing time. HADI is a generalized scheme that can be widely used not only in smart home but also in other crowded sensor networks. Our future work will focus on attempting to introduce more models and correlations into HADI so that home automation systems will be isolated from unavailable data.

References

1. Yu, T., Wang, X., Shami, A.: Recursive principal component analysis-based data outlier detection and sensor data aggregation in IoT systems. IEEE Internet Things J. **4**(6), 2207–2216 (2017)
2. Du, Y., Lim, Y., Tan, Y.: Rf-arp: Rfid-based activity recognition and prediction in smart home. In: 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), pp. 618–624. IEEE (2019)
3. Choi, J., Jeoung, H., Kim, J., Ko, Y., Jung, W., Kim, H., Kim, J.: Detecting and identifying faulty IoT devices in smart home with context extraction. In: 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), pp. 610–621. IEEE (2018)
4. Radja, M., Hungilo, G., Emmanuel, G., Suyoto, S.: IoT: Improved home energy control system based on consumer. In: IOP Conference Series: Materials Science and Engineering, vol. 1098, p. 042028. IOP Publishing (2021)
5. Hao, X., Zhang, G., Chen, Y.: Fault-tolerant control and data recovery in HVAC monitoring system. Energy Build. **37**(2), 175–180 (2005)
6. Nower, N., Tan, Y., Lim, A.O.: Efficient temporal and spatial data recovery scheme for stochastic and incomplete feedback data of cyber-physical systems. In: 2014 IEEE 8th International Symposium on Service Oriented System Engineering, pp. 192–197. IEEE (2014)
7. Lee, W.Y., Shin, D., House, J.: Fault diagnosis and temperature sensor recovery for an air-handling unit. Technical report, American Society of Heating, Refrigerating and Air-Conditioning Engineers (1997)
8. Liu, Z., Zhang, W., Quek, T.Q., Lin, S.: Deep fusion of heterogeneous sensor data. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5965–5969. IEEE (2017)
9. Rausand, M., Hoyland, A.: System reliability theory: models, statistical methods, and applications, vol. 396. John Wiley & Sons (2003)
10. Ni, K., et al.: Sensor network data fault types. ACM Trans. Sensor Networks (TOSN) **5**(3), 1–29 (2009)
11. Yang, K., Koike, T.: A general model to estimate hourly and daily solar radiation for hydrological studies. Water Resour. Res. **41**(10) (2005)