IEICE TRANS. INF. & SYST., VOL.E104–D, NO.2 FEBRUARY 2021

LETTER

Effective Anomaly Detection in Smart Home by Analyzing Sensor Correlations

Giang-Truong NGUYEN[†a)], **Van-Quyet NGUYEN**[††b)], **Van-Hau NGUYEN**[††c)], ***Nonmembers***, ***and*** **Kyungbaek KIM**[†d)], ***Member***

SUMMARY In a smart home environment, sensors generate events whenever activities of residents are captured. However, due to some factors, abnormal events could be generated, which are technically reasonable but contradict to real-world activities. To detect abnormal events, a number of methods has been introduced, e.g., clustering-based or snapshot-based approaches. However, they have limitations to deal with complicated anomalies which occur with large number of events and blended within normal sensor readings. In this paper, we propose a novel method of detecting sensor anomalies under smart home environment by considering spatial correlation and dependable correlation between sensors. Initially, we pre-calculate these correlations of every pair of two sensors to discover their relations. Then, from periodic sensor readings, if it has any unmatched relations to the pre-computed ones, an anomaly is detected on the correlated sensor. Through extensive evaluations with real datasets, we show that the proposed method outperforms previous approaches with 20% improvement on detection rate and reasonably low false positive rate.

key words: *smart home, sensors, anomaly, spatial correlation, dependable correlation*

1. Introduction

In a Smart Home environment, pervasive sensors placed in many areas and objects (e.g., on the couch) play an important role to capture any activities of residents and environmental entities and provide benefits for residents based on the captured sensor events such as turning on the light automatically or taking a health report. Unfortunately, these sensors could capture activities inaccurately, which generates abnormal events, due to some factors (e.g., malware compromising carelessly configured sensors). These anomalies could cause problems such that the electrical devices can be controlled in an unexpected way, or the health status of residents could be incorrectly reported. Hence, detecting sensor anomalies has attracted much attention [2], [3], [6].

Detecting anomalies has attracted much attention in a traditional wireless sensor network [4], [5]. Specifically, multiple homogeneous sensors are deployed spatially close to each other [4], [5]. In this homogeneous approach, anomalies are detected based on considering data from some specific sensors which are physically close to each other. However, in a smart home environment, heterogeneous approach is usually employed [2], [3], and different types of sensors are installed in different places and objects to capture almost all the activities of residents. The homogeneous approach could be applied in a smart home, and multiple sensors are deployed on the same objects or places. However, it may cause high cost of deployment and maintenance [3]. Also, in a smart home environment, pervasive binary sensors, whose output is 0 or 1, are employed more popularly [2], [3] than the sensors generating time series analog data [4]. Since this binary output is simple and does not carry much information itself [2], it requires different approaches to handle the anomaly detection.

Ye et al. [2] have proposed a cluster-based local outlier factor technique to detect anomalies. Specifically, DBSCAN clustering technique is applied on generated sensor events and the distance metric between two events is calculated based on locations of sensors and attached objects types. After clustering, outliers or very small sized clusters are considered as anomalies. This method is effective when anomalies are minority of generated events and they happen in long distance from normal events. However, if anomalies are blended into normal events or they forms a similar sized cluster to other normal clusters, this method is difficult to recognize them.

Choi et al. [3] have introduced a snapshot-based detection method, which employs event-based correlations between sensors. Specifically, a snapshot is a set of binary states of all sensors in a given duration, and it could indicate the correlation between sensors if they have generated events together in any duration. In the pre-computation phase, a list of all possible snapshots and the transition probability between any two possible snapshots are prepared. Then, in the detection phase, anomalies are detected if any snapshots are not found from the pre-computed list, or the transition probability of any two continuous snapshots is zero. However, the snapshot-based method does not consider the physical proximity between sensors, and it may list some possible snapshots between sensors which should not be correlated (e.g., they are too far from each other, or they belong to different activities). For example, if there is only one person in a house, sensor events in a living room should not be correlated to sensor events in a kitchen.

Manuscript received April 15, 2020.
Manuscript revised August 27, 2020.
Manuscript publicized November 9, 2020.

[†]The authors are with the Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea.
[††]The authors are with the Faculty of Information Technology, Hung Yen University of Technology and Education, Hung Yen, Vietnam.
a) E-mail: truongnguyengiang.bk@gmail.com
b) E-mail: quyetict@utehy.edu.vn
c) E-mail: haunv@utehy.edu.vn
d) E-mail: kyungbaekkim@jnu.ac.kr (Corresponding author)

DOI: 10.1587/transinf.2020EDL8056

In order to mitigate the limitations, we propose a novel method of detecting sensor anomalies under smart home environment by considering spatial correlation and dependable correlation between sensors. The spatial correlation considers the physical proximity between sensors and it is used to sort out possible pairs of sensors which have a dependable correlation. The dependable correlation indicates how a pair of sensors correlated, and it is calculated based on the likelihood of coexistence of sensor events. The proposed method contains a pre-computation phase and a detection phase. In the pre-computation phase, the list of possible pairs of sensors with tags of suitable dependable correlation is prepared. During a detection phase, possible sensor activities are measured based on spatial and dependable correlation, and if an activity is not matched into the prepared list, it is considered as an anomaly. The extensive results show that the proposed method is able to successfully detect sensor anomaly and outperforms previous approaches with 20% improvement on detection rate and reasonably low false positive rate.

2. Detecting Anomalies by Analyzing Sensor Correlation

Our proposed method is carried out in two phases: (1) Pre-computation phase: pre-computing correlations between sensors by measuring spatial correlation and dependable correlation; (2) Detection phase: gathering possible correlations between any sensors generating events at every time window, and detecting anomalies on a sensor if having any unmatched correlations to the pre-computed ones related to the sensor. These two phases are described in details as follows.

2.1 Pre-Computation Phase

The main purpose of this phase is to prepare correlations between any pair of sensors based on previous history of sensor events. To do this, we first calculate the spatial correlation based on the physical proximity of sensors. We then measure the dependable correlation by considering likelihood that two sensors generate events together in a duration (i.e., a time window). Finally, we categorize the sensor correlations with the combination of the spatial correlation and the dependable correlation.

2.1.1 Calculating Spatial Correlation

In a smart home environment, it is assumed that the positions of all the sensors are fixed, which means that their positions do not change most likely during both the pre-computation and detection phases. In this case, if two sensors are physically near each other, they may generate events simultaneously and have a correlation. According to this, the spatial correlation between sensor i and j (P_{ij}) is depended on their placed positions, which can be organized by a hierarchical structure of the house. For example, a sensor in a bed room is presented by $bedroom \sqsubseteq sleepingArea \sqsubseteq leftArea \sqsubseteq house$ and a sensor in a kitchen is presented by $kitchen \sqsubseteq cookingArea \sqsubseteq leftArea \sqsubseteq house$. Here, *house* is the root node of this hierarchy and others are location nodes. Note that P_{ij} can be calculated based on their hierarchical similarity as Eq. (1):

$$P_{ij} = P_{ji} = \frac{2 * depth(LCS)}{depth(i) + depth(j)} \tag{1}$$

where $depth(i)$ is the depth from the root to the location node of sensor i. LCS (least common sub-summer) node is the nearest common place concept of both sensors [7]. In the example, the hierarchical similarity between bedroom and kitchen becomes $2 * 2/(4 + 4) = 0.5$.

The spatial correlation P_{ij} has a value from 0 to 1, and higher value means that two sensors are closer and have higher probability of generating event simultaneously.

2.1.2 Calculating Dependable Correlation

The dependable correlation considers dependency of a sensor j to i, that is, the likelihood that a sensor j generates an event while sensor i generates an event in a time window.

Let us assume that S_i and S_j are the states of the events generated by sensor i and sensor j at a specific time unit (e.g. a second) in a time window, respectively. If sensor i generates an event at that time unit, then $S_i = 1$; otherwise, $S_i = 0$. Then, we consider c_{11} as the number of time units on a time window where $(S_i = 1, S_j = 1)$. Similarly, c_{10} and c_{01} are considered for the cases of $(S_i = 1, S_j = 0)$ and $(S_i = 0, S_j = 1)$, respectively. With this concept, dependable correlation of sensor i and sensor j is calculated by using Jaccard correlation [8] under the view of each sensor:

$$e_{ij} = \frac{c_{11}}{c_{11} + c_{10}} \text{ and } e_{ji} = \frac{c_{11}}{c_{11} + c_{01}} \tag{2}$$

where e_{ij} and e_{ji} are the dependable correlation under the view of sensor i and j, respectively. After getting all e_{ij} from all the time windows that they generate events together, the overall dependable correlation, E_{ij}, is obtained by:

$$E_{ij} = \frac{\sum(e_{ij})}{N_{ij}} \text{ and } E_{ji} = \frac{\sum(e_{ji})}{N_{ij}} \tag{3}$$

where $\sum(e_{ij})$ is the sum of e_{ij} for every time windows where both of sensor i and j generate events, and N_{ij} is the number of time windows where both of sensor i and j generate events.

The dependable correlation E_{ij} has a value from 0 to 1, and higher value means that the sensor j generates an event with higher probability if the sensor i generates an event.

2.1.3 Finding Correlation between Sensors

In a smart home environment, it is often observed that a sensor generates events if and only if another sensor is generating events (e.g., TV sensor with couch sensor). Also,

sometimes a sensor generates events alone (e.g., just couch sensor). Based on this observation, we define three types of correlation between sensors:

- *sole*: Sensor i generates events without any constraints from any other sensors.
- *main-main*: Two sensor i and j generate events simultaneously with similar magnitude during a time window.
- *main-sub*: During a time window, sensor j generates only a few events while sensor i generates events dominantly. In which, sensor j is called as a sub sensor and sensor i is a main sensor.

A correlation between sensor i and j is represented with 3-tuple as $(i, j, type)$, where $type$ can be s,mm and ms for *sole*, *main-main*, and *main-sub*, respectively (e.g. $(1, -, s)$,$(1, 2, mm)$,$(1, 3, ms)$). For *sole* correlation, the second element of 3-tuple is null and presented with dash.

During the pre-computation phase, the history of sensor events is analyzed and a set of correlations, S_C, is prepared. Basically, every sensor can has a sole correlation and a sole correlation $(i, -, s)$ for every sensor is inserted into S_C.

First of all, it is conducted to find all of the sensor pairs which have strong spatial correlation. For each sensor i, we obtain pairs of sensor, $(i, j, -)$, whose $P_{ij} \geq \Delta$; where Δ is a spatial correlation threshold (e.g., $\Delta = 0.8$) and add the pairs of sensors into S_C.

Then, we identify a type of correlation for every 3-tuple in S_C with a dependable correlation threshold, α, as followings:

- for each 3-tuples $(i, j, -)$ in S_C, we check:
  - if $E_{ij} \geq \alpha$ and $E_{ji} \geq \alpha$ (e.g., $\alpha = 0.8$), then set (i, j, mm), that is, it has *main-main* correlation;
  - if $E_{ij} < \alpha$ and $E_{ji} \geq \alpha$, then then set (i, j, ms), that is, it has *main-sub* correlation;
  - otherwise, remove $(i, j, -)$ from S_C;

After conducting these procedures above, we obtain a set of correlations, S_C, for a given event history under a given smart environment.

2.2 Detection Phase

During the detection phase, we detect the abnormal dependable correlations, which may not satisfy the physical proximity and may not expected according to the number of residence in a smart home environment. In order to detect anomalies, in every time window, we perform the following steps:
Step 1: We find out the set of *Possible Relational Activities* (called *PRA*) between sensors. Firstly, we add every pair of sensors such as $(i, j, -)$ into PRA and add $(i, -, -)$ for every sensor i into PRA. Then, we identify the type of each correlation and remove unidentified correlations from PRA. To do this, for every sensor i we need to conduct followings:

- if there is no event from sensor i, then remove $(i, *, *)$ from *PRA*;
- otherwise, then set set $(i, -, s)$ and for each other sensor $j \neq i$ do:
  - if $e_{ij} \geq \alpha$ and $e_{ji} \geq \alpha$, then set (i, j, mm);
  - if $e_{ij} < \alpha$ and $e_{ji} \geq \alpha$, then set (i, j, ms);
  - otherwise, then remove $(i, j, -)$ from *PRA*;

Step 2: For each 3-tuple t in *PRA*, we check $t \notin S_C$ (set of correlation obtained from pre-computation phase). If t does not belong to S_C, we notify that an abnormal activity happens at one of sensors in t.
Step 3: We obtain the number of current residents in a house, denoted as N_r. To achieve this, we can use a *people counting sensor* at the entrance-exit door of the house to track the number of residents, but how to count number of residence is out of scope of this paper.
Step 4: We subsume the correlations in *PRA*, and obtain the number of groups of correlations as N_G. The rule of subsuming correlations is following:

- (i, j, mm) and (j, i, mm) can be merged.
- (i, j, mm), $(i, *, ms)$, $(j, *, ms)$ can be merged.

Step 5: We compare the number of group N_G with the number of residents N_r. If N_G ¿, N_r, we notify an abnormal activity happens. To identify which sensor may involve this abnormal activity, we may compare the difference of correlation groups between the current time window and the previous time window.

3. Evaluation

To evaluate our proposed method, we conducted experiments with two main scenarios: 1) separated abnormal events and 2) mixed abnormal events. For the first scenario, it is assumed that abnormal event occurs in the sensors which are physically far from the normal sensors. For the second scenario, it is supposed that abnormal events are mixed in normal events without any constraint. We employ ARAS dataset [6] which is a 25 days log of sensor events generated by 20 sensors with 2 residents. It is assumed that this sensor log has only normal events, and the first 24 days of sensor events are used for pre-computation of our method and snapshot-based method [3]. On the 25th day, 50 abnormal activities are injected on different time with different duration. Abnormal duration varies from 10 to 50 seconds. Also, in this dataset, the activities of the residents are known such as going out and going in, and it is assumed that the number of residents can be correctly measured in any time window under this smart environment.

For comparing the performance of detecting anomalies, our proposed method is compared with clustering-based method [2] and snapshot-based method [3] in the aspect of detection rate and false positive rate. Detection rate is the portion of the time windows that anomalies are detected correctly over the total number of time windows on which anomalies are injected. False positive rate is the portion of time windows that anomalies are reported over the total

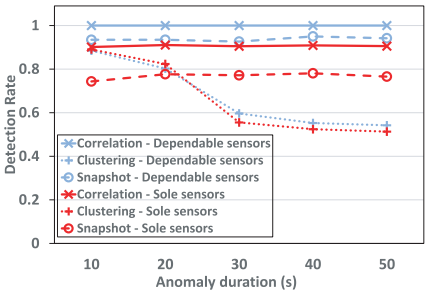

Fig. 1 Detection rate with separated anomalies

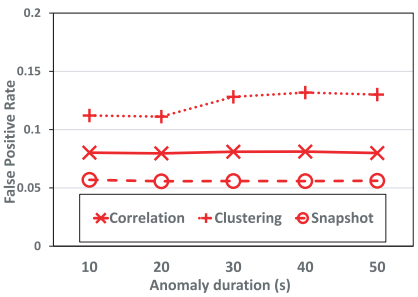

Fig. 2 False positive rate with separated anomalies

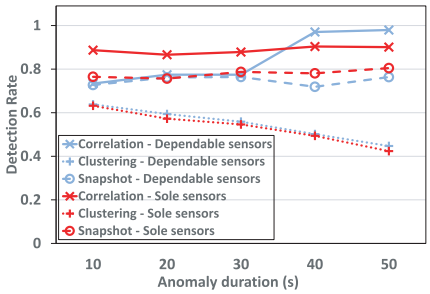

Fig. 3 Detection rate with mixed anomalies

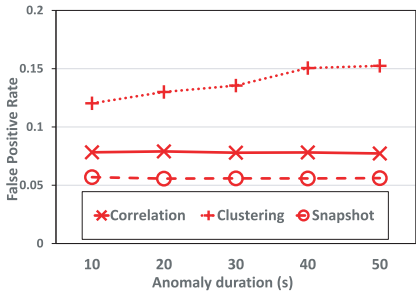

Fig. 4 False positive rate with mixed anomalies

number of time windows on which no anomaly is injected. For every experiment, we keep the size of a time window for all of the methods to 120 seconds, and set the value of spatial and dependable correlation thresholds to 0.8. In the evaluation, our proposed method is denoted as "Correlation", and clustering-based method and snapshot-based method are denoted as "Clustering" and "Snapshot" respectively.

Figure 1 and Fig. 2 shows the performance of different methods under the scenario with separated abnormal events. Overall, our proposed algorithm achieves the best detection rate with a reasonable false positive rate. Especially, our proposed method achieves absolute detection rate for detecting dependable correlation. The main reason is that the generated events occurs in distance and our proposed method can recognize this implicitly by examining both of spatial correlation and dependable correlation. Regarding the clustering-based method, when the anomaly durations increases the detection rate decreases, while the false positive rate increases slightly. That is, the clustering-base method is limited to massive abnormal events. Regarding the snapshot based method, its false positive rate is lowest, but the detection rate is not high especially for the events with sole correlation.

Figure 3 and Fig. 4 shows the performance of different methods under the scenario with mixed abnormal events. This scenario is more difficult case for detecting anomalies occurred in sensors with dependable correlation, because some abnormal events are considered as correlated events. Even though this tough setup, our proposed method still achieves the best detection rate with a reasonable false positive rate. Especially, when the duration of anomaly increases, our proposed method can separate the abnormal events on sensors with dependable correlation.

4. Conclusion

In this paper, a novel method of detecting sensor anomalies under a smart home environment is proposed by considering spatial correlation (i.e. degree of physical closeness) and dependable correlation (i.e. degree of simultaneous event generation). Through the extensive evaluation with real-world dataset, it is shown that the proposed method noticeably improves the performance of detecting complicated anomalies. The natural extension of this paper can be applying the proposed method into other domains such as smart factories and smart buildings.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1A2B4012559). This research is funded by Hung Yen University of Technology and Education under the grant number UTEHY.L.2020.07.

References

[1] F. Viani, F. Robol, A. Polo, P. Rocca, G. Oliveri, and A. Massa, "Wireless architectures for heterogeneous sensing in smart home applications: Concepts and real implementation," Proceedings of the IEEE, vol.101, no.11, pp.2381–2396, 2013.

[2] J. Ye, G. Stevenson, and S. Dobson, "Detecting abnormal events on binary sensors in smart home environments," Pervasive and Mobile Computing, vol.33, pp.32–49, 2016.

[3] J. Choi, H. Jeoung, J. Kim, Y. Ko, W. Jung, H. Kim, and J. Kim, "Detecting and Identifying Faulty IoT Devices in Smart Home with

Context Extraction," 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), IEEE, 2018.

[4] M.C. Jun, H. Jeong, and C.-C.J. Kuo, "Distributed spatio-temporal outlier detection in sensor networks," Digital Wireless Communications VII and Space Communication Technologies, vol.5819, International Society for Optics and Photonics, 2005.

[5] X. Luo, M. Dong, and Y. Huang, "On distributed fault-tolerant detection in wireless sensor networks," IEEE Transactions on computers, vol.55, no.1, pp.58–70, 2005.

[6] H. Alemdar, O.D. Incel, H. Ertan, and C. Ersoy, "ARAS human activity datasets in multiple homes with multiple residents," Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2013.

[7] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp.133–138, 1994.

[8] B. Zhang and S.N. Srihari, "Properties of binary vector dissimilarity measures," Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing, vol.1, 2003.