

Exploration of Linked Anomalies in Sensor Data for Suspicious Behavior Detection

Ian Turk¹, Matthew Sinda², Xin'an Zhou³, Jun Tao⁴, Chaoli Wang⁴ and Qi Liao²

¹(Department of Computer Science, University of Wisconsin-Eau Claire, Eau Claire, Wisconsin)

²(Department of Computer Science, Central Michigan University, Mount Pleasant, Michigan)

³(School of Computer Science, Fudan University, Shanghai, China)

⁴(Department of Computer, Science and Engineering, University of Notre Dame, Notre Dame, Indiana)

Abstract We present a visual analytics system to understand the operation data of a company, GASTech, from IEEE VAST Challenge 2016. The data include proximity data recording the locations and movements of employees, and heating, ventilation, and air conditioning (HVAC) data recording the environmental conditions in the building. Analyzing the data to detect the suspicious behaviors of some disgruntled employees is of special interest. Our system provides coordinated multiple views to visualize the proximity data and the HVAC data over time. Visual hints and comparisons are designed for users to identify abnormal patterns and compare them. Furthermore, the system automatically detects and correlates the anomalies in the data. We provide use cases to demonstrate the effectiveness of our system.

Key words: visual analytics; sensor data; security

Turk I, Sinda M, Zhou XA, Tao J, Wang CL, Liao Q. Exploration of linked anomalies in sensor data for suspicious behavior detection. *Int J Software Informatics*, Vol.10, No.3 (2016): 000–000. <http://www.ijsi.org/1673-7288/10/228.htm>

1 Introduction

Detecting suspicious behaviors of employees is crucially important for the safety of a company. The state-of-the-art security system in a building may supply various kinds of motion detectors and environmental sensors instrumented in different zones to detect suspicious behaviors. However, these sensors generate large amounts of data, and it is difficult for human users to perceive without effective analysis and visualization. IEEE VAST Challenge 2016^[1] targeted at this problem, and provided a data set consisting of various kinds of sensor data in a building of the GASTech company. The participants were required to design visual interfaces to assist the detection of the typical patterns and suspicious behaviors in this data.

In this paper, we focus on the Mini-Challenge 2 of IEEE VAST Challenge 2016. The Mini-Challenge 2 provides two weeks of operation data from which the typical patterns and suspicious behaviors should be detected. The operation data consist of

Corresponding author: Jun Tao, Email: jtao1@nd.edu

Received 2016-08-15; Revised 2016-10-01; Accepted 2016-10-28.

the proximity data and the heating, ventilation, and air conditioning (HVAC) data. The building has three floors, and each floor is divided into multiple HVAC zones. The HVAC data was collected by sensors in each HVAC zone, reporting building temperatures, heating and cooling system status, and concentration levels of carbon dioxide (CO₂), hazium and other chemicals. Among these chemicals, hazium might deserve special attention, since it is possibly dangerous. The proximity data were gathered from the proximity cards carried by all employees. There are two types of proximity data: the passive proximity data and the mobile proximity data. The passive proximity data was generated by the passive proximity card readers located in each of the proximity zones in the building. Note that the proximity zones are numbered differently from the HVAC zones. When a proximity card moves from one zone to another, a record would be generated with the proximity card ID, time, and the zone being entered. The mobile proximity data was collected by a mail delivery robot. This robot follows a specifically designed route and records the nearby proximity cards.

We present a visual analytics system to explore and understand this data. The interface of our system consists of three views: the proximity view, HVAC view, and event view. The proximity view and HVAC view visualize the proximity data and HVAC data in a user-specified time period. Using the proximity view, users may select one employee to observe his/her movement pattern or select multiple employees for comparison. With the HVAC view, users can observe the values of a selected set of variables for all HVAC zones. For each variable, the difference between the current value and the average value at the same time period across all days is visually encoded, so that users can easily identify the abnormal variables for further investigation. More importantly, our system automatically detects the anomalies and builds connections among them according to their time and location information. The anomalies are displayed in the event view along the timeline together with visual links indicating their connections. Starting from one anomaly, users can easily trace back to the previous anomalies that may possibly cause this anomaly. We present the anomalies detected in the results and provide a reasoning on the root cause of the suspicious behaviors.

2 Related Work

Visual analytics is essential to handle data sets that are growing quickly in both size and complexity. These techniques are required to gain an understanding of the data, uncover patterns, and further knowledge growth^[5]. As we progress, there is a strong need for visual analytics to combine views from multiple source types to include both static data that could be collected from sensors, and mobile data that could be represented as trajectory data. There are numerous works related to visualizing individual source types, but little has been done to showcase the challenge in visualizing these sources together.

Sensor Data Visualization. Due to the decrease of cost for installing and maintaining sensor networks, the quantities of high dimensional sensor data has increased over the years. However, if the improvements in the ability to visualize this information do not continue at the same pace, one's ability to sense the world around us diminishes. Forlines and Wittenburg^[3] developed a tool named Wakame

for displaying information gathered from a building's environmental sensors. Wakame takes 2D shapes representing multidimensional sensor readings and transforms them into 3D views. These 3D views reflect not only the changes in a sensor's readings, but also the time the reading is taken and the location the sensor exists. While much can be said for the novelty of their approach, work still needs to be completed to generate solid conclusions.

The ability to detect anomalies from sensors is critically important. A simple approach to identifying abnormal data is to separate them from normal data via some classification method. However, this requires that the ground truth of what is considered normal, to be known. Thus, the effectiveness of anomaly detection models is strongly influenced by dynamic changes in the environment they operate. Rassam, *et al.*^[7] developed an approach in which the ground truth is not required in labeling data as either abnormal or normal. Additionally, success of anomaly detection, relies heavily on how the neighborhood is defined. Janeja, *et al.*^[4] worked with movement data to determine anomalies in 2D space for water and highway traffic monitoring. They identified spatial properties in a neighborhood leading to a well refined outlier discovery via clustering.

Trajectory Data Visualization. Trajectory is the most common form of traffic data which have been studied extensively in visualization^[2]. In this context, the trajectory data could come from aircraft, automobile, shipping, train/metro, or pedestrian trajectories. Previous work has focused on the visualization and analysis of trajectory data as well as associated information such as movement directions, change of direction, movement speed, and change of speed. Willems, *et al.*^[8] described various approaches (animation, space-time cube, and density charts) to visualize movement data and show their effectiveness in allowing a user to draw accurate and efficient conclusions. Their study showed that no visualization technique was any better at showing movement data than the others. In fact, their research suggests that depending on the application, more than one visualization method could be better at allowing the user to draw accurate conclusions about applicable features. While their work was specifically tested against the movement of ocean going vessels, it could be easily applied to any situation in which there is a route with different densities and various stopping points. Meghdadi and Irani^[6] presented selective Video Summarization and Interaction Tool (sViSIT) that supports an interactive and exploratory view of surveillance video data. Their system visualizes each object's movement path using a single action shot image, a trajectory in a space-time cube, and an overall timeline view. Using their tool, experts were able to identify items of interest 88% easier than other traditional commercial tools.

3 Our Visual Analytics System

Our system consists of a data analysis component that detects anomalies in both the proximity data and HVAC data, and a visual interface that displays the data together with the detected anomalies for user exploration and reasoning. To detect the suspicious events, the data analysis component focuses on the anomalies in both the proximity data and HVAC data as well as their connections. Users can specify a set of variables to investigate in the interface. Our visual interface shows the anomalies

detected in these variables and link them to the related anomalies in other variables that appear beforehand. This builds the connection between the anomalies and their possible causes, so that users can trace the links backward to infer when, where and how the problem originates.

3.1 Data analysis

The core of our data analysis is anomaly detection. Anomalies in the proximity data and HVAC data are detected using different strategies due to their different natures. We describe these strategies separately as follows.

Anomalies in Proximity Data. An anomaly in the proximity data is a 2-tuple (employee, time interval), indicating that the employee behaves differently from other employees during the same time period or from his/her own movement pattern on the other days. To represent the movement pattern of an employee during a time interval, we accumulate the duration of time this employee spends in each zone and record that in an employee length of stay histogram. Abnormal behaviors are identified through comparing different histograms using the Jensen-Shannon divergence (JSD) and reporting those with large JSD values. Given two distributions P and Q , JSD is defined as

$$\text{JSD}(P||Q) = -\sum M \log M + \frac{1}{2}(\sum P \log P + \sum Q \log Q), \quad (1)$$

where M is the average distribution of P and Q .

We consider two kinds of comparison between histograms. First, for each employee, we compare the histograms aggregated during the same time interval on different days (e.g., everyday from 8:00am to 12:00pm). The histograms with large average JSD are considered to be anomalous. This kind of anomalies indicates that the employees behave differently from their normal movement patterns. Second, for each department, we compare the histograms aggregated for different employees on the same day, and also report the ones with large JSD. This kind of anomalies captures those employees that behave differently from their colleagues.

Note that different employees may visit different zones for the same purpose. For example, the office of one employee may be located in Floor 2 Zone 1, but the office of another employee may be in Floor 2 Zone 3. In addition, they may visit different pantry rooms and restrooms due to the spatial closeness to their offices. In this case, different trajectories may indicate the same behavior (e.g., moving from their offices to the closest restrooms). Since employees in the same department may share similar movement patterns, we assume that the correspondence of zones for different employees can be inferred from the amount of time they spend in the zones. For example, employees in the administration department may spend most of the time in their offices. Under this assumption, we sort the histograms in the decreasing order of the length spent in each zone before computing the JSD, so that the two zones in which two employees spending the most amount of time are paired in the computation, and so forth. We do not compare employees from different departments, since they have different roles which probably leads to different movement patterns. Finally, for each kind of comparison, we report k employees and time intervals whose corresponding histograms have the largest average JSD values to others.

Anomalies in HVAC Data. An anomaly in the HVAC data is a 3-tuple

(variable, zone, time interval), indicating that the variable behaves differently in that zone during that time interval. Anomalies in the HVAC data are automatically detected and displayed as bars in the event view. We detect the anomalies using a combination of correlations and integrals of differences. After reading all the data from the files, the program goes through them and calculates the min, max, standard deviation, and the average values of weekday and weekend for each variable in each HVAC zone. Weekdays and weekends are separated because most variables have significantly different values between Monday-Friday and Saturday-Sunday. Min and max values for each variable are stored and used later to create the scales in the dot-matrix plot and line graph.

The program then searches through the HVAC data and performs a series of tests on small sections (e.g., hourly) of the data. A cross-correlation with a small time delay and the integral of the difference are calculated with respect to the average values of weekday and weekend that were calculated earlier. If the correlation value is below a certain threshold and the integral of the difference is above a certain threshold (which is scaled with the duration of the section and the max value of the variable), that section is considered anomalous, and an anomaly is created and associated with the HVAC zone where it occurred. This anomaly is displayed as a bar in the event view.

3.2 Visual interface

Our visual interface consists of three views: the proximity view, HVAC view, and event view, as shown in Figure 1. The proximity view displays several selected histograms of the proximity data, and the HVAC view displays the values of HVAC variables in a user-specified time interval T . This interval is specified by a time instance t and a time window size w using two sliders (i.e., $T = [t - w/2, t + w/2]$). The event view displays the anomalies and their connections, and it is linked to the proximity view and HVAC view. For instance, users can select anomalies in the event view and observe the data in the other two views. They can also select variables of interest in the HVAC view to filter the anomalies displayed in the event view. We refer the readers to the video at <https://vimeo.com/176912286> for a more intuitive understanding of the interface and interaction.

Proximity View. The proximity view visualizes, for one or more employees, the histograms of the duration spent in each zone during the user-specified time interval T . When a single employee is investigated, we display a circle in each zone. The size of the circle indicates the duration this employee stays in the corresponding zone. For a zone not visited by this employee, the circle will be transparent with only its border visible. When multiple employees are selected for comparison, the size of a circle in a zone indicates the total duration these employee stay in the zone. The circle will be further divided into multiple fans and visualized as a pie chart, where the size of each fan is proportional to the duration one employee stays in the zone. The fans are color coded, so that fans sharing the same color in different zones correspond to the same employee. Users can add or remove employees to be investigated in the proximity view to observe the duration they spend in each zone, or they can select the employees exhibiting abnormal behaviors in the event view. By selecting an anomaly in the proximity data in the event view, the corresponding histogram will be automatically displayed in the proximity view, together with a histogram representing

the normal behavior for comparison.



Figure 1. The overview of our visual analytics system. The HVAC view (including dot-matrix plot and line graph), proximity view, and event view are shown at the top, middle, and bottom, respectively. Anomalies related to the hadium concentration spikes at F3Z1-HVAC are linked in the event view.

The circles are displayed on top of the floor map of the proximity zones or the HVAC zones specified by users. When the map of the HVAC zones is used, we further visualize the values of the first selected variable at the time instance t by adjusting the color of each zone. A zone in gray indicates that its value of the first selected variable is zero, and its color gradually transits from gray to blue, then to red, when the value of variable increases. The location of an employee is indicated by a fan with larger radius (as highlighted in the dashed circles in Fig. 1), so that users can spot the suspicious employee(s) causing the change of hadium concentration in a zone.

HVAC View. The HVAC view consists of a dot-matrix plot and a line graph. The dot-matrix plot visualizes each variable at the time instant t in each zone as a dot. Each column represents a zone and each row represents a variable, as shown the top left region in Fig. 1. The color of a dot scales from blue to red as the value of the variable increases. The transition is scaled based on the overall min and max values of the variable in every zone. The size of a dot indicates the difference between the current value and the average value computed at the same time on different days. The difference is normalized using the standard deviation, so that the smallest dot indicates that the current value is exactly the same as the average value, and the largest dot indicates that the current value is at least one standard deviation away from the average. This allows users to easily notice the variables whose values are

different from their averages.

The line graph shows the values of a few selected variables in the user-specified time interval T . Each variable can be selected by clicking on the associated dot in the dot-matrix plot, and the corresponding line graph will be assigned a unique color. The values of each variable over time are displayed as a solid line, and the corresponding average values are optionally displayed as a dashed line with the same color. The average values of weekday and weekend are calculated beforehand, which are useful for analyzing anomalies noticed by users in the dot-matrix plot. The time instance t corresponding to the dot-matrix plot is displayed in the line graph as a black vertical dashed line.

Event View. We show the anomalies of proximity and HVAC data in the same time range T at the bottom and top of the event view, respectively. To reduce the number of displayed HVAC anomalies for clear observation, we only display the anomalies corresponding to the variables selected in the HVAC view and the anomalies that can be traced back from these anomalies within two links. Each anomaly is shown as a bar which is linked to the related anomalies by a smooth curvy edge. An anomaly of a selected variable is displayed in the same color as the corresponding line in the line graph. The other HVAC anomalies are colored in orange and the proximity anomalies are colored in gray.

4 Results

We implement our system using D3.js and JavaScript. We investigate the two-week operation data (from May 31 to June 13) of GAStech using our system and identify multiple anomalies in both the proximity and HVAC data. In this section, we denote a HVAC or proximity zone on a floor as F_iZ_j -HVAC or F_iZ_j -PROX, where i and j denote the floor ID and zone ID, respectively. We analyze the relationships between the concentration level of the dangerous chemical hazium and other sensor data. We explain our findings using the data in F3Z1-HVAC and on Floor 2 as examples.

Anomalies in F3Z1-HVAC. Figure 1 shows our visualization with the hazium concentration in F3Z1-HVAC. We specify the time interval to the entire two weeks in order to obtain an overview of anomaly connections that may be related to the hazium concentration in F3Z1-HVAC. The time instance to investigate is set to 9:05am on June 1, several hours before the hazium was detected for the first time when multiple events in other variables were active. In the event view, we find that most anomalies are identified in the variables related to the temperature, such as thermostat temperature, heating setpoint, cooling setpoint, and reheat damper position. The reheat damper position is especially suspicious, since anomalies are detected in all the neighboring HVAC zones of F3Z1-HVAC over the entire two weeks, as highlighted in the blue rectangles in Fig. 1. This can also be perceived in the second row of the dot-matrix plot (highlighted in the red rectangle), which corresponds to the reheat damper positions in all HVAC zones. Most of the dots are relatively large at 9:05am on June 1, indicating that the reheat damper positions in those zones are different from their average values. In the column corresponding to F3Z1-HVAC, we also find that the dots associated with the reheat damper position, thermostat temperature, heating setpoint, and cooling setpoint are large. We select the thermostat temperature and

heating setpoint to observe their patterns in the line graph, and we find that the heating setpoint was set to a lower value than its average at that time, leading to a lower value of thermostat temperature. Since the hazium events often coincide with the anomalies in the variables that are used to control the temperature in the building, we suspect that the generation, release and propagation of hazium may depend on the temperature condition, and the malicious employee might use the HVAC system to adjust the temperature.

We further investigate the proximity anomalies related to these HVAC anomalies. The proximity anomalies are detected at the daily level (i.e., the histograms are aggregated for each day). Three employees (Padraic Ratigan, Patrick Young, and Pamala Beebe) exhibited abnormal behaviors at 9:05am on June 1, as indicated by the bars with green, blue and purple borders in the event view shown in Fig. 1. Their corresponding proximity data histograms over the two weeks and locations at 9:05am on June 1 are visualized in the proximity view using the same color correspondence, as highlighted in the dashed circles in Fig. 1. We find that all these three employees appeared on Floor 3 at that time and they all entered the server room, which contains the HVAC control system, in these two weeks. We narrow the time window size to focus on the few hours before and after 9:05am on June 1, and find that these three employees actually entered the server room as well during that short time period. The above clues indicate that all these three employees may be related to the abnormal pattern of the HVAC variables.

Among these three suspicious employees, we find that Patrick Young's behavior is of particular interest. Six anomalous events are found for Young. The first two events each covers working hours of an entire day, but the other four events are short in time, covering only around ten minutes. Since the histograms are generated for each day, this indicates that his proximity data could only be found for a very short period of time on each day. Investigating the histograms of his movement data, we find that Young exhibited two different movement patterns in these six events. His movement pattern of the first two events can be observed in Fig. 2 (a). He spent most of the time in F3Z2-PROX, where his office is located, and spent short periods of time in the zones containing the hallways and the server room. Since Young belongs to the facilities department, visiting the server room seems to be reasonable. But in the other four short events, he visited several zones in a few minutes, including the server room. He even only appeared on Floors 2 and 3 without any record of entering the building, as shown in Fig. 2 (b). Strangely, he stayed in F2Z2-PROX for the longest time on June 2 and 3. In addition, the proximity view shows no proximity record of him on the other days. By examining the original data, we find that Young had reported the loss of his proximity card and was assigned a new card on June 1. The proximity data with strange pattern were generated from his lost proximity card. Since this proximity card generated no data on most of these days, his normal behavior in the first two days is considered to be different from the other days and detected as an anomaly. We suspect that someone might have stolen his card for malicious behaviors, since proximity information was recorded for his new card. In this case, the employees in F2Z2-PROX are suspicious, since the old proximity card appeared in that zone for quite a while on June 2 and 3, but Young did not usually visit that zone.



Figure 2. Different movement patterns on (a) June 1 and (b) June 2 of a suspicious employee Patric Young.

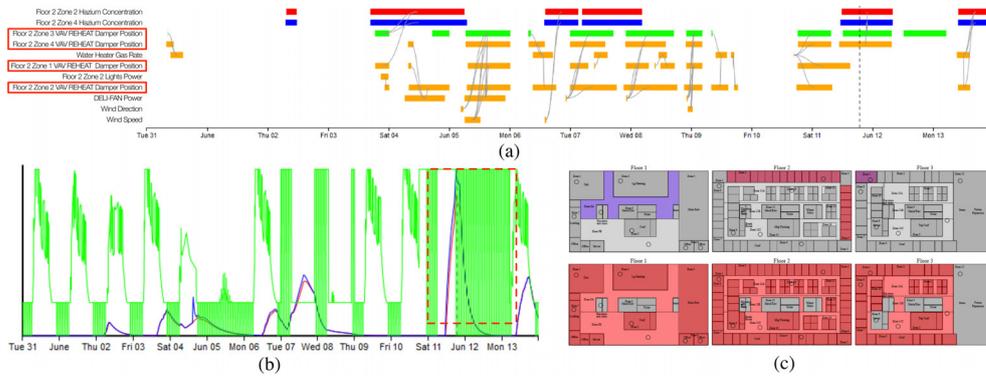


Figure 3. Haziium concentration and reheat damper positions on Floor 2. (a) shows the event view. (b) shows the line graph. (c) shows the proximity view with the HVAC zones.

Anomalies on Floor 2. Figure 3 shows anomalies related to the haziium concentration events in F2Z2-HVAC and F2Z4-HVAC for the entire two weeks. The haziium events in F2Z2-HVAC are highlighted in red and those in F2Z4-HVAC are highlighted in blue. In Fig. 3 (a), the event view shows that most anomalies connected to the haziium events are related to temperature. We can see that the reheat damper positions in four HVAC zones on Floor 2 are identified and linked to the haziium anomalies, as highlighted in the red rectangles. This reveals similar relationships among HVAC anomalies in our first example. In Fig. 3 (b), the line graph shows that the two haziium sensors detected similar patterns of haziium concentration in F2Z2-HVAC and F2Z4-HVAC, since the red and blue curves mostly coincide with each other. During the highest spikes in these two zones, we find that the reheat damper positions maintained at an abnormally high level, as indicated by the green curves in the dashed red rectangle. The relationship between the haziium concentration and damper positions can be revealed more clearly in the proximity view, as shown in Fig. 3 (c). In the top row, where the haziium concentration is mapped to the zones, we can see that haziium was detected by all the four sensors in

the building. In the bottom row, we map the damper positions to the zones and find that the damper positions in all zones were set to high values. However, no anomaly is found in the heating setpoint or cooling setpoint, which may explain why the thermostat temperature appeared to be normal. It is likely that the suspicious employee only opened the damper to allow hazium to propagate through the air conditioning system.

5 Conclusions

We have presented a visual analytics system that automatically detects and links anomalies for suspicious behavior detection. This is achieved by first analyzing the distribution of data points over time for variables or employees to detect individual anomalies and then correlating multiple anomalies from different variables or employees that occur close to each other in both space and time. The results demonstrate the effectiveness of our approach. In the future, we plan to improve our solution by automatically composing a series of anomalies (could be hierarchically organized) to represent a coherent causal event so that users could spend more of their effort on visual reasoning than user interaction. This would lead to more knowledge gained in a shorter amount of time, which would be very useful for analyzing a catastrophic event with cascade effect.

References

- [1] IEEE VAST challenge 2016. <http://vacommunity.org/VAST+Challenge+2016>.
- [2] Chen W, Guo , Wang FY. A survey of traffic data visualization. *IEEE Trans. on Intelligent Transportation Systems*, 2015, 16(6): 2970–2984.
- [3] Forlines C, Wittenburg K. Wakame: Sense making of multi-dimensional spatial-temporal data. *Proc. of the International Conference on Advanced Visual Interfaces*. 2010. 33–40.
- [4] Janeja VP, Adam NR, Atluri V, Vaidya J. Spatial neighborhood based anomaly detection in sensor datasets. *Data Mining and Knowledge Discovery*, 2010, 20(2): 221–258.
- [5] May R, Hanrahan P, Keim DA, Shneiderman B, Card S. The state of visual analytics: Views on what visual analytics is and where it is going. *Proc. of IEEE Symposium on Visual Analytics Science and Technology*. 2010. 257–259.
- [6] Meghdadi AH, Irani P. Interactive exploration of surveillance video through action shot summarization and trajectory visualization. *IEEE Trans. on Visualization and Computer Graphics*, 2013, 19(12): 2119–2128.
- [7] Rassam MA, Maarof MA, Zainal A. Adaptive and online data anomaly detection for wireless sensor systems. *Knowledge-Based Systems*, 2014, 60: 44–57.
- [8] Willems N, van de Wetering H, van Wijk JJ. Evaluation of the visibility of vessel movement features in trajectory visualizations. *Computer Graphics Forum*, 2011, 30(3): 801–810.