Visual Anomaly Detection in Spatio-Temporal Data using Element-Specific References

Daniel Alcaide, Jansi Thiyagarajan, Houda Lamqaddam, Jaume Nualart Vilaplana, and Jan Aerts

Abstract—The analysis and exploration of dynamic spatio-temporal data presents particular challenges. The VAST 2016 contest provided the opportunity to explore solutions in this space, focusing on the identification of patterns and anomalies. In this paper, we present an approach based on element-level references that allows for the exploration of individual movement data as well as sensor readings. This method earned the VAST 2016 Award for Robust Support for Visual Anomaly Detection.

Index Terms—Visual data analysis, anomaly detection, pattern exploration, interactive user interfaces

1 Introduction

Faced with complex datasets, it can be particularly hard to identify anomalies if no prior hypotheses can be defined. The field of Visual Analytics (VA) combines the power of computer-driven data analyses with that of the human for identifying unexpected patterns visually [1]. In this paper, we describe a visual analytics interface for the detection of anomalies in spatio-temporal data using element-specific references. This interface was created within the context of the VAST 2016 (`http://vacommunity.org/VAST+Challenge+2016`) mini-challenge 2. In this challenge, we were asked to identify patterns, anomalies, and relationships in proximity and sensor-data covering two weeks in a given building. Data consisted of a building layout, list of employees, proximity sensor data (i.e. which employee is close to which sensor), proximity sensor data for a roaming robot (i.e. which employee is close to the robot), as well as HVAC and Hazium sensor reading. The interactive version of the visuals presented here is available at `http://vda-lab.be/portfolio.html#vast2016`.

2 Data Preprocessing

To enrich the given dataset, several variables were combined, transformed, and derived. These include mapping the coordinates of the mobile proximity data with the closest room or office, adding complementary information of the employee, and transforming data for detecting when employees enter or exit a particular zone. The detection of anomalies used derived metrics to detect unusual variations.

2.1 Anomaly definition in the proximity dataset

To detect anomalies in the proximity data, we computed two individual-specific scores (Sequence-score and Time-score) representing how unusual the trajectories of that employee in a specific day are.

- *Daniel Alcaide is with Visual Data Analysis Lab, ESAT/STADIUS, KU Leuven, Belgium, and iMinds HI2 Data Science, KU Leuven, Belgium. E-mail: daniel.alcaide@kuleuven.be.*
- *Jansi Thiyagarajan is with Visual Data Analysis Lab, ESAT/STADIUS, KU Leuven, Belgium, and iMinds HI2 Data Science, KU Leuven, Belgium. E-mail: jansi.thiyagarajan@kuleuven.be.*
- *Houda Lamqaddam is with Visual Data Analysis Lab, ESAT/STADIUS, KU Leuven, Belgium, and iMinds HI2 Data Science, KU Leuven, Belgium. E-mail: houda.lamqaddam@kuleuven.be.*
- *Jaume Nualart Vilaplana is with Visual Data Analysis Lab, ESAT/STADIUS, KU Leuven, Belgium, and iMinds HI2 Data Science, KU Leuven, Belgium. E-mail: jaume.nualart@kuleuven.be.*
- *Jan Aerts is with Visual Data Analysis Lab, ESAT/STADIUS, KU Leuven, Belgium, and iMinds HI2 Data Science, KU Leuven, Belgium. E-mail: jan.aerts@kuleuven.be.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

The sequence score evaluates the level of monotony in the employee movements. We generated a reference sequence for each employee based on their daily routine. The number of variations within this data for a day provides a normalized indicator between 0 and 1, where 0 is equal to the reference sequence and 1 completely different.

The time score evaluates whether the time spent by an individual in a specific location is longer or shorter than what is considered "normal" for that individual. Here the reference is computed as the median time spent in all the locations along the days which we have data for. As in the sequence score, the time score is normalized between 0 and 1, 0 being equal to the reference and 1 completely different. Notice that this score only evaluates the time spent in a location independently of the number of times or the order during the day.

2.2 Anomaly definition in the building dataset

The building dataset contains 419 temporal variables (including Hazium sensors) along the different zones and floors of the building. When these variables are measured by different metric units, it is difficult to detect when a variable or a set of them are out of the normal range. The approach presented in this report is based on computing the number of standard deviations from a reference value for each variable in the system.

The reference value was defined as the usual behavior of each variable. This value takes into account all variables per zone and per hour along the 14 days of data. Due to the general absence of employees during the weekend, we distinguished two kinds of references values: weekdays and weekends. The computation of this reference value is described as follows per zone and hour: [1] The initial 5-minute intervals of data were aggregated into hours to increase the robustness of the value; [2] The statistical median of each variable was computed; [3] The standard deviation (SD) for every variable used the units of the original variable. As it is not possible to compare variables that use different unit metrics, the resulting SD was divided by the reference value; [4] The final unit-less result was grouped by zone.

3 Data Exploration Interfaces

In this section, we introduce four interfaces created to discover patterns and abnormalities in both the proximity and the building datasets. These interfaces have been designed following the Shneiderman Overview first basic principle for visual design [3]. The graphical language used is shared along the presented views.

3.1 Interfaces for Proximity dataset

To visualize the patterns in the proximity dataset, we developed the Proximity Pattern Explorer Interface (PPEI) for showing the occupancy of each zone in each floor throughout the 14 days. It allows to zoom in on a particular day to get more detail on daily patterns or visualize line-charts for the different departments. The Proximity Anomalies Detections Interface (PADI; Fig. 1) presents an interactive scatterplot matrix for all the days available (Fig. 1 A). Each circle represents an

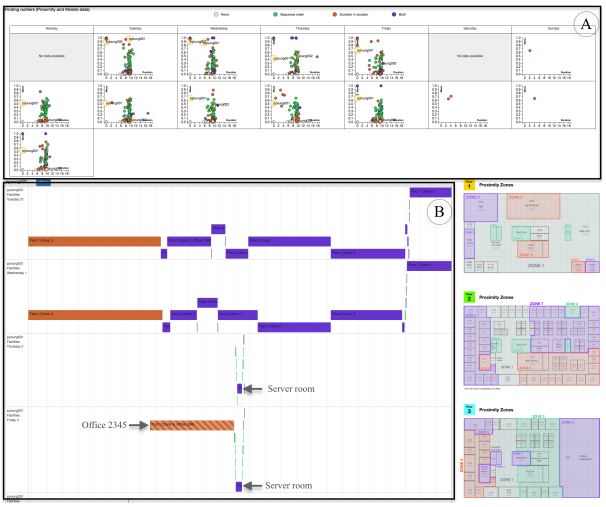

Fig. 1. Detail of Proximity Anomalies Detections Interface (PADI). A) Scatterplot matrix for all the days available. Each circle represents an employee in a specific day. B) List of timelines of the movements of a selected employee.

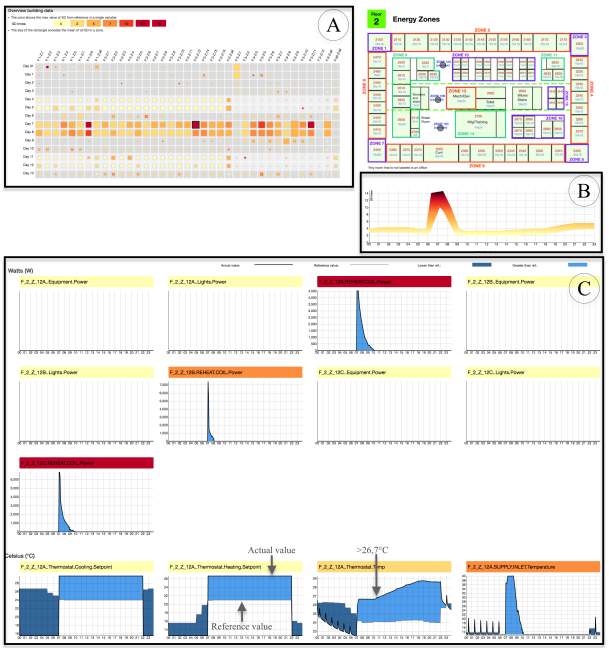

Fig. 2. Detail of Building Anomalies Detections Interface (BADI). A): Heatmap for building dataset represented per day and zone. Each square is encoded by size and color. Size: mean of the number of SDs of all variables; Color: highest SD of the variable in a specific zone. B) Line-chart showing the time in the X-axis, and the mean value of the number of SDs in the Y-axis. C) Comparative line-charts for individual variables. Each one represents the mean of the variables and the actual values. Dark-blue is for values lower than the reference value; light-blue for values greater than the reference value.

employee in a specific day. A circle will be green if the deviation from the reference is caused by the sequence of movements, orange if it is caused by the time spent in the locations, and purple if it is caused by both of the above reasons. In each scatterplot, the X-axis represents the time spent in the building by the employee, and the Y-axis represents the mean between sequence-score and time-score. If a circle is selected, we can see the other days for the same employee highlighted, and a timeline of the movements of the selected employee (Fig. 1 B). The color of the boxes uses the same color scheme as described above. When using the robot proximity data, the number of the offices appear in the timeline. A hashed box shows that an employee is not in their assigned office.

3.2 Interfaces for Building dataset

The Building Pattern Explorer Interface (BPEI) interface help the users to identify patterns by categories such as HVAC system, water heating, power consumption, control system and Hazium concentration. This interactive visualization provides an overview of the data, and allows detailed evolution of a single variable of the system enabling filter by floor, by zone, by day and by hour. The Building Anomalies Detections Interface (BADI; Fig. 2) presents an initial interactive matrix that displays data per day and per zone (Fig. 2 A). Each square in this matrix is encoded by size and color. Size represents the mean of the number of SDs of all variables; color represents the value of the highest SD of the variables. When a zone is selected, the floorplan and the complementary plots (Fig. 2 B and C) are displayed. The line-chart on Fig. 2 B represents time in the X-axis, and the mean value of the number of SDs in the Y-axis. The remaining area gives a list of the diagrams for each variable as comparative line charts. These show the mean of the variables and their actual values. Dark-blue is for values lower than the reference value; light-blue for values greater than the reference value.

4 Anomalies Detected

Different anomalies were detected in this dataset, detailed explanations of which will be available in the VAST Visual Analytics Benchmark Repository (`http://hcil2.cs.umd.edu/newvarepository/benchmarks.php/`). These included security risks and possibly faulty and/or tampered-with sensors. In addition, we identified a progressive increase in concentration of the (fictitious) toxic Hazium, which may be linked to the presence of one particular individual.

5 Conclusion

The presented suite of visual analysis interfaces provides interactive visualizations specifically designed to identify patterns and anomalies for the given GAStech data. Moreover, these visualizations allow us to focus on a variety of tasks, as described by Munzner [2]. The proposed anomaly detection approach shows how data aggregation can help to enrich data, and eventually to navigate through a high dimensional system guiding the user to the most relevant indicator and subjects. Although the approach presented was exclusively designed for this contest (i.e. using individual/building specific references), we strongly believe that it could be applied to other scenarios with similar datasets.

Acknowledgments

The work presented here is supported by H2020 Virogenesis Grant nr 634650, IWT SBO ACCUMULATE Grant nr 150056 and iMinds ICON MECOVI.

References

[1] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Visual analytics: Definition, process, and challenges. In *Information visualization*, pp. 154–175. Springer, 2008.

[2] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

[3] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pp. 336–343. IEEE, 1996.