A Visual Analytics Approach for Equipment Condition Monitoring in Smart Factories of Process Industry

Wenchao Wu*
Siemens Ltd. China

Yixian Zheng†
China Telecom Shanghai Ideal Information Industry (Group) Co., Ltd.
Shanghai Engineering Research Center of Internet Big Data

Kaiyuan Chen‡
University of California, Los Angeles

Xiangyu Wang§
University of Southern California

Nan Cao¶
Tongji University

ABSTRACT

Monitoring equipment conditions is of great value in manufacturing, which can not only reduce unplanned downtime by early detecting anomalies of equipment but also avoid unnecessary routine maintenance. With the coming era of Industry 4.0 (or industrial internet), more and more assets and machines in plants are equipped with various sensors and information systems, which brings an unprecedented opportunity to capture large-scale and fine-grained data for effective on-line equipment condition monitoring. However, due to the lack of systematic methods, analysts still find it challenging to carry out efficient analyses and extract valuable information from the mass volume of data collected, especially for process industry (e.g., a petrochemical plant) with complex manufacturing procedures. In this paper, we report the design and implementation of an interactive visual analytics system, which helps managers and operators at manufacturing sites leverage their domain knowledge and apply substantial human judgements to guide the automated analytical approaches, thus generating understandable and trustable results for real-world applications. Our system integrates advanced analytical algorithms (e.g., Gaussian mixture model with a Bayesian framework) and intuitive visualization designs to provide a comprehensive and adaptive semi-supervised solution to equipment condition monitoring. The example use cases based on a real-world manufacturing dataset and interviews with domain experts demonstrate the effectiveness of our system.

1 INTRODUCTION

Equipment maintenance is of vital importance in manufacturing. Inappropriate maintenance and arbitrary failure of equipment will lead to inefficiency and even safety issues, especially in *process industry*[1] which cannot be entertained in this competitive age. Thus, over the past few years, there has been a concerted effort to improve the maintenance strategy. However, the majority of manufacturers are still practising planned schedule maintenance [31], in which the equipment is operated until a predetermined time when maintenance is carried out. This strategy often leads to either over-maintenance as the time tends to be chosen before any potential failure; or lack-of-maintenance since it's almost impossible to cater for all varying failure patterns beforehand. Therefore, an on-line equipment condition monitoring system is in demand which can not only reduce unplanned downtime by early detecting equipment "health" issues but also reduce unnecessary routine maintenance for better availability.

Fortunately, with a worldwide movement of Industrial Internet (or Industry 4.0) [11, 20], the increasing availability of manufacturing data [2] generated within highly digitalized and connected "smart factories" opens up unprecedented opportunities for manufacturers to engage in data-driven science to better understand equipment conditions. However, without an effective method, the data is usually analyzed in aggregated statisics [2]; thus, valuable insights into local details and trends are often missing. Although there have been some attempts in developing algorithms to model and measure equipment conditions [13, 31], it is still difficult for analysts to trust or make use of the results [23] since manufacturing data are highly dynamic and analytical tasks are usually complicated. In another word, a fully automatic equipment condition monitoring is difficult, requiring considerable experience and profound knowledge in various fields. Thus, analysts seek the help of visual analytics to take full advantage of both advanced computational power and human cognitive abilities for a more granular and intelligent monitoring approach.

Despite that the crucial role of visual analytics has been identified [36], to the best of our knowledge, few examples have been reported which apply visual analytics to equipment condition monitoring in manufacturing. In this paper, we work closely with managers and operators in a petrochemical plant, a typical type of factory in process industry, to help them deal with challenges met in equipment maintenance. In particular, we build a visual analytics system with a semi-supervised framework to help managers and operators define health status of online equipment and derive meaningful rules or patterns for effective equipment condition monitoring. Due to the lack of a precise definition of boundary between normal and abnormal equipment conditions, it is hard to be achieved by pure automatic solutions, especially for process industry with complicated and highly-sensitive correlations among a large number of variables. In our work, this is accomplished following a human-in-the-loop approach [8] by which users could interact with context information conveniently and apply their own domain knowledge in the analytical process. The major contributions of this work can be summarized as follows:

- We design and implement a visual analytics system with a semi-supervised framework to address the major challenges of equipment condition monitoring met by real world operators and managers from a factory of process industry (i.e., a petrochemical plant).
- We develop a suite of interactive visualization techniques enhanced with new features to support visual-assisted knowledge discovery and sense making from manufacturing data, thereby helping users define status and train adaptive models for effective equipment life-cycle condition monitoring.
- We showcase an experience of working with target users from manufacturing industry to iteratively design a visual analytics system, deploy on site, and evaluate through case studies and expert interviews.

2 RELATED WORK

2.1 Visualization of Manufacturing Data

In recent years, with the launch of Industrial Internet (or Industry 4.0), more and more people have started to realize the value of data collected through the manufacturing process and tried to analyze them to reveal important insights that can improve manufacturing [2]. With the growing amount and complexity of manufacturing data, it could be anticipated that visual analytics, an effective approach for gaining insight from large and complex data [29], would play a more and more crucial role [36].

However, so far only a few visual analytics solutions target the data analysis tasks in manufacturing scenarios. Jo et al. [18] present

*E-mail: wenchao.wu@siemens.com
†E-mail: arthuryixian@gmail.com
‡E-mail: chenkaiyuan@ucla.edu
§E-mail: wang677@usc.edu
¶E-mail: nan.cao@gmail.com

[1] **Process industry** is the branch of manufacturing industry associated with formulas and manufacturing recipes, which is contrasted with *discrete industry* that is concerned with discrete units, bills of materials and assembly of components. For example, process industries include manufacturing in chemical, petrochemical, pharmaceutical and biotechnology, etc. [21].

2165-8773/18/$31.00 ©2018 IEEE
DOI 10.1109/PacificVis.2018.00026

LiveGantt as an interactive visualization tool for a large manufacturing schedule. Wörner and Ertl [45] introduce a visual analytics system for manufacturing simulation. In this paper, rather than the planning and simulation stage, we focus on the operation stage of manufacturing which has not received much attention from researchers in the field of visualization. Matkovic et al. [32] exploit virtual instruments enhanced with history encoding for process monitoring. Compared with this method, our work integrates in-depth analysis and modeling process with intuitive visualizations into an advanced monitoring solution. In addition, Xu et al. [49] propose ViDX, a visual analytics system, to analyze performance and identify inefficiencies of assembly lines in an automobile factory. This method depends on historical data of processing time of each part on a certain work station, which makes it inapplicable to process industry. To the best of our knowledge, our system is the first to provide an array of visualizations that can be combined for interactive equipment real-time condition monitoring in process industry and facilitating an interpretation to leverage users' domain knowledge and experience.

2.2 Visualization of Time-series Data

Time-series data are ubiquitous in manufacturing industry, which are collected from various sensors, monitoring the status of different parts and aspects of a factory. Thus, understanding and analyzing these time-series data is the basis for condition monitoring of equipment in factories. In the past few years, researchers have spent much effort to develop visualization techniques for time-series data [1]. The most prevalent method is based on the horizon axis [54]. For example, line chart [14] and its variants, like horizon graph [47], show the changes of one attribute over time with a horizon time axis, while stacked graph [48] is used to show the changes of multiple attributes over time simultaneously. Besides, various visualization techniques are proposed to support different analytical tasks on time-series data in real applications. For example, Zhao et al. [53] present the Ringmap that employs subdivided ring segments with different colors to visualize multiple cyclic activities over time. Van et al. [43] present a calendar display to explore data which are aggregated daily, weekly and monthly. Our system is inspired by these techniques and integrates them with several novel designs and interaction techniques. Beyond presentation tools, we propose more comprehensive visual analytics techniques to handle multiple analytical tasks for smart machine condition monitoring.

2.3 Anomaly Detetction

The problem of machine condition monitoring can be approached within the framework of anomaly detection, which identifies abnormal patterns based on a model trained with numerous examples representing "normal" patterns. This is especially suitable for monitoring conditions of on-site machines in factories of process industry, where large amounts of data for "normal" conditions are available, while there are always insufficient data to describe various kinds of "anomalies".

Given its wide range of applications, anomaly detection has been extensively studied over the past decades [7, 16]. Pimentel et al. [34] summarized previous work and categorized related techniques into five major types, including a) domain-based, b) information-theoretic, c) distance-based, d) reconstruction-based, and e) probabilistic techniques. In particular, domain-based techniques [15, 30] offer a straightforward approach, aiming to define a boundary around the "normal" data that any point falling outside is considered as abnormal. However, this type of techniques is often influenced by outliers in the training set, and the choice of parameters which control the size of the boundary region is also difficult. Information-theoretic techniques [24, 37] detect anomalies based on information-theoretic measures, such as entropy. This type of techniques could be applied to produce numeric results of anomalies, e.g., the abnormal score, but are limited in offering interpretation and understandability. In addition, distance-based techniques [25, 46] measure similarity between data points and identify "anomalies" occurring far from their nearest neighbours, but it's computationally expensive to calculate distance between all pairs of data points, especially in a high-dimensional space and a large dataset, which makes these techniques hard to be applied for real-time monitoring. Stojanovic et al. [40] try to tackle this challenge by employing clustering method and parallel computing techniques. However, it still suffers poor interpretability which makes it difficult for users, especially those in the traditional process industry, to trust or make use of the results. In comparison, reconstruction-based techniques [17, 27] perform data-mapping based on constructed models and estimate a reference "normal" condition to enable comparisons with the actual observed value in a much easier way, while probabilistic techniques [19, 41] characterize data with more interpretable statistic models, which is essential to support an interactive analysis in real-world traditional industrial applications [42]. Our system integrates these two types of techniques and present a unified visual analytics framework that enable interactive and adaptive anomaly detection situated in the streaming, traditional manufacturing data scenarios.

Moreover, in recent years, a few visual analytics systems are designed for anomaly detection [5, 6, 33, 52], which are also related to our work. For example, Zhao et al. presented FluxFlow [52] to identify and interpret anomalous information spreading patterns. Cao et al. introduced TargetVue [5] employing intuitive glyph design to facilitate detection of users with abnormal behaviors on social media. Different from all these previous work, we introduce a unique, adaptive real-time anomaly investigation method, which incorporates human judgement to guide the detection algorithm to produce interpretable results that can be accepted and made use of by users in traditional manufacturing industry. The proposed visual interactive framework not only respects users' working habits in process industry but also enables monitor and investigate with rich context information in general big manufacturing data scenarios - where real-time analysis and intuitive visualization are desirable.

3 System Design and Overview

3.1 System Design

3.1.1 Data Description and Transformation

Since 1970s, the third industry revolution with the main characters of automation has swept across the world. General-purpose computing devices, such as programmable logic controllers (PLCs), are widely deployed to control and monitor industrial processes automatically, especially in process industry. These devices could be viewed as a series of sensors. The data involved in this research is a set of time-series data collected from streaming data sources enabled by these sensors, which record the status of different parts of a factory from time to time. The data format is $< Timestamp, ID_{sensor}, Value >$.

In our system, we first transform the raw data from these streaming data sources into a vector time series as shown in Fig. 1a to support a near-real-time analysis. We note that our system runs in near-real-time as the data analysis and visualization are lagging slightly behind the data collection process. In particular, due to different sampling rate of sensors, within each time span in the time series, the system collects and calculates the average value of all records for each sensor, thus producing a vector. When all sensors get updated values, the system continues analyzing and visualizing the data collected at time span $(t-1)$, while collects the new data arriving at the current time span t. The granularity of a time span could be chosen depending on the application requirements and computational capacity of the analysis modules (e.g., 30 seconds, 1 minute, or 5 minutes).

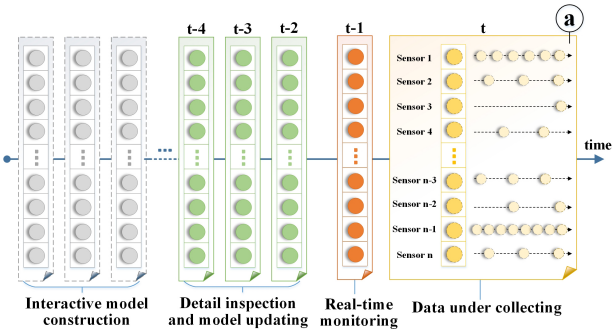


Figure 1: The pipeline for data processing and analysis.

3.1.2 Design Process

We followed a user-centered design process to develop and improve our visual analytics system iteratively. In the past ten months, we worked closely with four domain experts from a petrochemical plant, who were our target end-users as well. Two of them are the manager and operator responsible for equipment operation and

management, and the other two specialize in design and implementation of the "Industrial Big Data Services" program in the plant. Frequent video conferences, email discussions, and regular on-site meetings (monthly) were held at different design and implementation stages, aiming to close the gap between industrial needs and available technologies for a practical and effective solution of data-driven equipment condition monitoring.

3.1.3 Task Analysis

Through frequent exchange of views with the experts, we identified three major challenges faced by our target users during equipment condition monitoring, which helped us better understand the problem domain.

Firstly, how to define trustworthy observation rules for normal or abnormal conditions **(C1)**? The domain experts found that the traditional rules based on thresholds of sensor values were too "rigid" for reflecting the complexity of real manufacturing conditions, while engineers nowadays highly rely on experience on equipment conditions in the past and struggle to make comparative judgements. However, human activity often brings delays and unstableness [40], which in turn degrades the capability of monitoring and its efficiency. To make matters worse, there exist complicated correlations among these sensor values, while understanding such correlations is essential for effective monitoring. Thus, the emerging problem is how to integrate engineer's experience with advanced computing capability for more effective and stable equipment condition monitoring.

Secondly, the domain experts complained that the volume of manufacturing data, that needs to be monitored, are large in scale and even unmanageable, owing to a considerable number of sensors in the plant and high sampling frequency. It leads to poor transparency of equipment conditions and difficulties in making decisions based on real data for time-critical tasks. Therefore, the challenge met by our collaborators is how to facilitate engineers in exploring and exploiting potentials in manufacturing data to improve the capability and agility of sensing and responding to potential risks **(C2)**?

Last but not least, equipment's conditions keep changing during its entire life-cycle. Then another challenge is how to cope with such changes so as to provide an adaptable and flexible condition monitoring approach **(C3)**?

To tackle these challenges, a set of tasks are identified and compiled with the consideration of sense making models for intelligence analysis [22, 35], which should be supported by our system:

T.1. **Interactive Feature Extraction:** It would be extremely beneficial and important to engage the users with domain knowledge and experiences on operating equipment in the model construction process $(C1)$, which is essential to help users understand how our system works, thus making it trustable for users in traditional manufacturing industry. To this end, we should provide interactive feature extraction functionalities which should be easy to use even without understanding of algorithm details. Therefore, as the first step, our system needs to measure and visualize correlations between these sensors in an intuitive way, and enables users to identify important target sensors, extract their correlated ones, and make adjustments based on experience **(T.1.1)**. In this way, plausible monitoring ***feature vectors*** can be generated for users. Meanwhile, as suggested by our collaborators, target sensors need to be further grouped into ***modules*** **(T.1.2)** for an efficient and comprehensive monitoring. This also matches users' working habits.

T.2. **Real-time Monitoring:** Based on the extracted feature vectors, users need to perform a real-time or near-real-time monitoring based on the streaming manufacturing data in a large scale so as to sense and respond to potential risks in a timely manner $(C2)$. Instead of relying on "subjective" expert experience, our system should support monitoring via a condition reference library (hereinafter referred to as ***training set***) constructed based on the "objective" past conditions **(T.2.1)**, which is then used to train the monitoring model. After that, for the real-time monitoring, besides generating alarms, users should be able to grasp a general idea of detected suspicious or abnormal cases so that they could respond in a timely manner **(T.2.2)**.

T.3. **Detail Inspection and Model Updating:** Appropriate exploration of details need to be done for practical applications. In particular, users need to explore detected suspicious or abnormal cases within context so as to formulate their own judgement on potential causal relations **(T.3.1)**. Meanwhile, in order to adapt to various changing conditions during the entire life-cycle of equipment $(C3)$, the system should support users to inspect and update the training set in an intuitive and flexible way **(T.3.2)** so as to guide the system to refine the monitoring model according to the user feedback.

3.1.4 Design Goals

Based on the identified analytical tasks, we further compiled a set of design requirements with our collaborators which guided the subsequent design of our system.

R.1. **Interpretability: Intuitive visual metaphor and narrative structure.** Visualizations with intuitive visual metaphors and an easy-to-understand narrative structure are desired by our collaborators. In this way, the system can support users from traditional process industry, who may not have much background on information technology, to understand analytical process, integrate with their domain knowledge, and discover abnormal patterns with interpretable visual evidence and related details. Therefore, our system conveys information through well-established and insightful visualizations, posting relatively fewer challenges with respect to interpretation. These techniques are further tailored or extended with new features to address the specific needs for our problem. Moreover, multiple well-coordinated views are employed, thus enabling users to perform various analytical tasks step-by-step.

R.2. **Insightfulness: Dual-scale encoding for a full picture of equipment conditions**. For online monitoring (T.2.2) in real-world applications, a full picture of equipment conditions at two different scales is often required. On one hand, users want to examine the detected suspicious or abnormal case in a timely manner so as to understand "when and what happens" and enable them to take prompt actions. On the other hand, it is also crucial for users to grasp an intuitive overview of long-term trends of equipment conditions which can shed more light on several important issues for equipment monitoring. For example, users may wish to know which sensor comes up with more abnormal conditions in the last hour, or whether there're more anomalies detected today than yesterday. In this way, they can get a quick idea of potential risks conveniently.

R.3. **Interactivity: Interactive pattern unfolding.** Since inspection of detected suspicious or abnormal cases in equipment (T.3.1) requires a trial-and-error process, it is crucial for analysts to interact with data directly. Meanwhile, to carry out an efficient and in-depth analysis, a multi-facet filtering based on different properties should also be enabled in our system.

3.2 System Workflow

Fig. 2 illustrates the interactive analysis workflow of our system.

1) The workflow starts with the data preprocessing module which leverages Apache Hadoop[1] on a cluster with 19 data nodes and 342 cores to support collecting and parallel processing of big manufacturing data based on Map-Reduce. It transforms manufacturing data collected from on-site streaming data sources into a series of vectors (Fig. 1) as described in Section 3.1.1. The streaming pipeline facilitates the online monitoring and analysis by our system, and the processed results are stored in MongoDB[2] database to support online queries.

2) The system then enables users to interactively set target sensors and configure modules for monitoring. It could either be a data-driven decision based on an overview of correlations among sensors in the Correlation View (e.g., users could choose sensors highly correlated with others), or be determined based on users' domain knowledge. For these chosen target sensors, users can further extract correlated sensors to generate feature vectors $(T.1)$.

3) After these configurations, users now can initialize a real-time monitoring by training with a few selected training time periods (i.e., training set) during which the equipment is in good and stable conditions (as reference conditions) based on their experience $(T.2.1)$. During the monitoring process, two visualization

[1] http://hadoop.apache.org/

[2] https://www.mongodb.com/

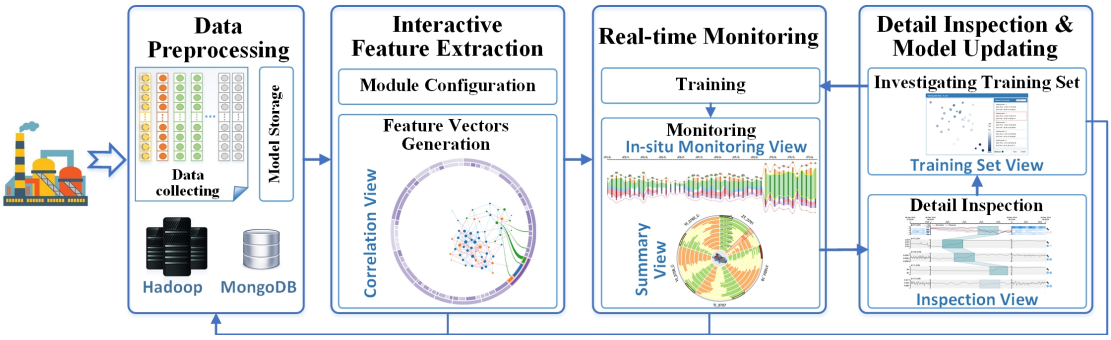


Figure 2: System workflow: After data preprocessing and interactive feature extraction/configuration (Section 4.1), our system can support real-time monitoring (Section 4.2) and detail inspection & model updating (Section 4.3).

views, namely In-situ Monitoring View and Summary View, are provided to help users understand equipment conditions in a timely manner (T.2.2) by revealing detected suspicious and abnormal instances in two different scales intuitively.

4) Users can conduct an in-depth analysis of detected suspicious or abnormal cases (T.3.1) in the Inspection View, based on which users can verify their hypotheses as well as decide whether and how to update the training set. For example, users can simply brush to choose a few time periods in the Inspection View to add into the training set. Then the Training Set View is designed to provide an online user feedback mechanism by incorporating users' judgements and observations on distributions of the training set for an iterative refinement (T.3.2), thus making the monitoring model adaptable in real manufacturing of process industry.

.

4 Interactive Equipment Condition Monitoring via Visual Analytics

Our system consists of three visualization-based modules, which work together to support three major steps of the human-in-the-loop analysis process described in Section 3.1.3. In designing the visualization techniques, we follow the design rationales discussed above to present abstract quantitative analysis results. In this section, we describe the visual analytics techniques designed for these three modules in details respectively.

4.1 Interactive Feature Extraction:

In the first step, we want to allow users to get an overview of various sensors and support them in model construction for monitoring, including identifying target monitoring sensors, extracting feature vectors, and configuring modules (Task T.1). To this end, we need appropriate methods for correlation analysis of all sensors, which is not a trivial task. Thus, in this section, we start with the description on how to calculate the correlations between time-series generated by two different sensors and then present the visualization designs to support an interactive analysis.

4.1.1 Calculating Correlations

In order to show an overview of all sensors, the relationship between each pair of sensors should be measured based on a proper calculation of correlations between time-series generated by these sensors. In our implementation, we adopt Pearson Correlation Coefficient [28] which is widely used in various applications for its simplicity and efficiency. Meanwhile, considering the characteristics of manufacturing data and analytical tasks in our case, we extend the calculation of correlation coefficients with an alignment process. The general procedure is described as follows:

Step 1. Alignment of time-series. When analyzing correlations of time series generated by different sensors on-site, time lags exist ubiquitously, which requires serious considerations to avoid misjudgment. Therefore, two sub-steps are carried out. **1)** As time-series are generated by various sensors with different scales, it is difficult to make a direct comparison for a proper alignment. Thus, we first reconstruct the time series through normalization. Meanwhile, for equipment condition monitoring, as pointed out by our collaborators, we need to focus on the pattern of changes rather than numerical values, thus the first-order difference [3] is further adopted. In particular, a time series $X = \{(x_1,t_1),(x_2,t_2),...,(x_i,t_i),...,(x_n,t_n)\}$ is reconstructed as $X' = \{(0,t_1),(x'_2,t_2),...,(x'_i,t_i),...,(x'_n,t_n)\}$, where

$$x'_i = \frac{x_i - x_{i-1}}{max_i\{\|x_i - \bar{x}\|\}}$$

2) After that, accounting for both positive and negative correlations, we apply dynamic time warping [44] on the absolute value of reconstructed time series. In this way, we get a set of $< i,j >$ pairs which matches each element (x_i,t_i) in the time-series $X = \{(x_1,t_1),(x_2,t_2),...,(x_i,t_i),...,(x_n,t_n)\}$ with each element (y_j,t_j) in the time series $Y = \{(y_1,t_1),(y_2,t_2),...,(y_j,t_j),...,(y_n,t_n)\}$.

Step 2. Calculation of Pearson Correlation Coefficient. After alignment, we now have a set of pairs $< i,j >$, based on which we calculate Pearson Correlation Coefficient as follows:

$$r = \frac{\sum_{<i,j>}(x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{<i,j>}(x_i - \bar{x})^2}\sqrt{\sum_{<i,j>}(y_j - \bar{y})^2}}$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $\bar{y} = \sum_{j=1}^{n} y_j/n$. Furthermore, t-test is employed with Fisher z-transform to calculate a p-value to measure the statistical significance of non-zero correlations (Null hypothesis: correlation coefficient is 0). Thus, the correlation between each pair of time-series can be described by a tuple $\{r,p\}$.

4.1.2 Visual Design and Interactions - Correlation View

Now that we have the correlation value r between any two sensors and also its corresponding statistical significance p, we further define the *degree* d_A of a sensor A as the number of sensors sharing *strong correlations* with A (i.e., $r > 0.4$, $p < 0.05$ as suggested by the experts in our implementation). The threshold could be adjusted for different applications. Then we need a visualization to show the results to users and support an interactive analysis. In our system, we design the Correlation View with two concentric rings (Fig. 3a), with outer layer R_1 encoding all the equipment, and the inner one R_2 representing all the control loops contained by each equipment. Each arc in these two rings encodes a subset of sensors contained by a control loop or equipment. The central angle of an arc encodes the total number of sensors in the corresponding subset, and the opacity conveys the average degree of these sensors. The darker the color, the more sensors share strong correlations with these sensors.

Once an arc is selected, the relationships of all sensors in the corresponding subset are displayed with a force-directed layout in the center. Each sensor is presented as a node whose size encodes its degree. The larger the node, the more sensors sharing strong correlations with the corresponding sensor. Moreover, if two sensors share strong correlations, we add an edge between the corresponding nodes, whose length encodes the correlation coefficient $\|r\|$, color (i.e., blue and red) indicates positive or negative correlations, and opacity encodes the statistical significance p. In addition, if a node in the center shares strong correlations with sensors in other subsets, the corresponding arc will exert an attraction force on it. Furthermore, users can interactively paint arcs and their corresponding nodes with colors so as to facilitate comparisons (Fig. 3a-1). In this way, by observing these nodes, users can make data-driven choices of target monitoring sensors besides choices based on their domain knowledge (Task T.1.1). Meanwhile, a module

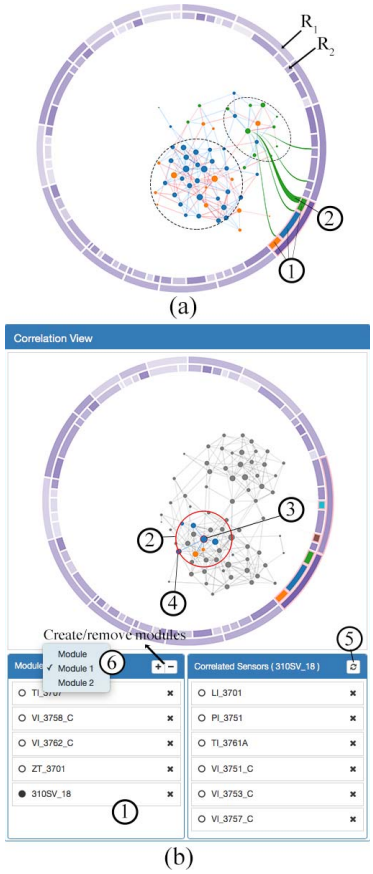


Figure 3: (a) In the Correlation View, two concentrated rings (i.e., R_1 and R_2) and a node-link diagram with force-directed layout are designed to provide an overview of all sensors. Users can interactively choose target monitoring sensors. (b) For a certain target monitoring sensor, interactive analysis is supported to identify correlated sensors and facilitate feature extractions.

configuration panel (Fig. 3b-1) is also provided, enabling users to group target monitoring sensors into ***modules*** interactively (Task $T.1.2$). For example, users can choose an existing module in the pull-down list (Fig. 3b-6) or click on the "+" to create a new module. And then users can choose a target sensor for the selected module by double clicking on a node of large size, indicating a sensor highly correlated with many other sensors, which requires close monitoring.

In addition, the Correlation View can support users to extract correlated sensors for a target monitoring sensor, thus generating corresponding feature vectors conveniently (Task $T.1.1$). Users can select a target monitoring sensor in the module configuration panel, and system will then highlight the corresponding node by red contour in the node-link diagram (Fig. 3b-3), while those nodes without strong correlations to the selected node will be turned into grey. Furthermore, by right clicking on the node, it will emit bands towards arcs, whose widths encode the number of sensors with strong correlations with the target sensor in the corresponding subsets (Fig. 3a-2). Users can further select these arcs with wider bands to add those sensors into the node-link diagram for further explorations. Moreover, as suggested by the domain experts, here we further introduce a *sample-based filtering* technique. Users can click on a node (Fig. 3b-4) to specify a certain correlated sensor based on experience, then the system will filter out those nodes with weaker correlations (Fig. 3b-2). Finally, once all correlated sensors are chosen in the node-link diagram, users can click on the "update" button (Fig. 3b-5) to save the result which is exactly the generated feature vector for the corresponding target monitoring sensor.

4.2 Real-time Monitoring

Based on the extracted feature vectors, the system can support a real-time monitoring of equipment conditions based on past reference conditions (Task $T.2$). Meanwhile, two visualization views are presented to facilitate an intuitive monitoring process.

4.2.1 Monitoring Algorithm

First, we describe the monitoring algorithm [50] employed in our system which is set up in a Bayesian framework. It is composed of a training stage and a monitoring stage, which enables a semi-supervised process to address the major challenges of equipment condition monitoring in a real plant of process industry.

Training Stage In the training stage, we model "normal" equipment conditions based on a training set, which consists of a set of vector time-series describing past reference conditions during certain time periods chosen by users based on experience. Thus, the current condition of equipment can be evaluated through a comparative analysis. In particular, for a certain target monitoring sensor A, based on the extracted feature vectors (i.e., a set of sensors correlated to sensor A) {Sensor A_1, Sensor A_2, ..., Sensor A_k}, its past reference conditions can be characterized by a vector time-series $\mathfrak{V}_A = \{\mathcal{V}_{A,t} | t = 1,2,...,N\}$ where $\mathcal{V}_{A,t} = \{Value_{A,t}, Value_{A_1,t}, Value_{A_2,t}, ..., Value_{A_k,t}\}$; $Value_{A,t}$ and $Value_{A_i,t} (i = 1,2,...,k)$ indicate the value of target Sensor A and correlated Sensor A_i at time span t. Then we choose to fit a regressive Gaussian Mixture Model (GMM) with Dirichlet Process [26, 38] on the chosen training set (i.e., vector time-series $\mathfrak{V}_A$) using Expectation Maximization (EM) algorithm, since GMM is widely used in modeling complex distributions [9]. Different from the traditional GMM, we assign an auxiliary weight β_t to each vector $\mathcal{V}_{A,t}$ of $\mathfrak{V}_A$. The weight follows an exponential decay with respect to time, considering that the vectors of recent equipment conditions should weight more than old ones in our scenario. Thus, in the M-step of the EM algorithm during the training process of GMM, given the probability $w_t(l)$ that vector $\mathcal{V}_{A,t}$ belongs to the l_{th} Gaussian component obtained in the E-step, the mean μ_l can be calculated as

$$\mu_l = \frac{1}{N} \sum_{t=1}^{N} w_t(l) \beta_t \mathcal{V}_{A,t}$$

The covariance σ_l is given by

$$\sigma_l = \frac{1}{N} \sum_{t=1}^{N} w_t(l) \beta_t (\mathcal{V}_{A,t} - \mu_l)(\mathcal{V}_{A,t} - \mu_l)^T$$

The weight of the l_{th} Gaussian component is

$$\pi_l = \frac{\sum_{t=1}^{N} w_t(l) \beta_t}{\sum_{t=1}^{N} \beta_t}$$

In this way, we can get a GMM with L components $\{s_l | s_l \sim N(\mu_l, \sigma_l), l = 1,2,...,L\}$, where L is estimated adaptively via approximate Bayesian criteria [38], thus characterizing past reference conditions of target monitoring sensor A.

Monitoring Stage The task for the monitoring stage is to estimate the normal value that a certain target monitoring sensor A should have at time span t if the equipment operates normally and evaluate risks for current conditions. To this end, in our system, we introduce a Gaussian random vector ε with zero mean and diagonal covariance matrix $\Theta = diag(\theta_0, \theta_1, ..., \theta_k)$ to map the observed vector $\mathcal{V}_{A,t} = \{Value_{A,t}, Value_{A_1,t}, Value_{A_2,t}, ..., Value_{A_k,t}\}$ to the corresponding $\mathcal{V}'_{A,t}$ based on the constructed GMM in the training stage. Formally,

$$\mathcal{V}_{A,t} = \mathcal{V}'_{A,t} + \varepsilon$$

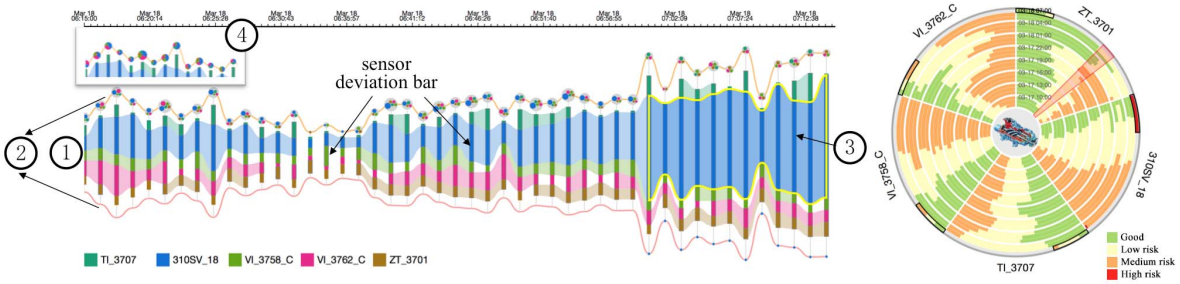
For each component s_l of GMM, we have the joint distribution

$$\begin{bmatrix} \mathcal{V}'_{A,t} \\ \mathcal{V}_{A,t} \end{bmatrix} | s_l = N\left(\begin{bmatrix} \mu_l \\ \mu_l \end{bmatrix}, \begin{bmatrix} \sigma_l & \sigma_l \\ \sigma_l & \sigma_l + \Theta \end{bmatrix} \right)$$

Then, our task can be transformed to estimate

$$\mathcal{V}^*_{A,t} = E(\mathcal{V}'_{A,t} | \mathcal{V}_{A,t}, \Theta) = \sum_{l=1}^{L} P(s_l | \mathcal{V}_{A,t}, \Theta) E(\mathcal{V}'_{A,t} | \mathcal{V}_{A,t}, s_l, \Theta)$$

Unfortunately, both $\mathcal{V}'_{A,t}$ and Θ are unknown and should be estimated based on $\mathcal{V}_{A,t}$, thus EM algorithm [51] is applied where $\mathcal{V}'_{A,t}$ is regarded as the hidden variable and Θ is the parameter



(a) (b)

Figure 4: (a) In-situ Monitoring View shows the risks detected by the monitoring algorithm for all target sensors of a module in real-time; (b) Summary View provides an overview of long-term trends of equipment conditions, which also keeps updating in real-time.

to be estimated. After iterations of E-step and M-step, we can get the corresponding "normal" $Value'_{A,t}$ of the target monitoring sensor A and the estimated diagonal covariance matrix $\Theta = diag(\theta_0, \theta_1, ..., \theta_k)$, with which we provide an adaptive approach to measure the different deviation levels of observed value $Value_{A,t}$ with respect to each observed input vector $\mathcal{V}_{A,t}$. In particular, we define four deviation levels to quantify corresponding risks (i.e., **Good**: $|Value_{A,t} - Value'_{A,t}| < \theta_0$; **Low risk**: $\theta_0 \leq |Value_{A,t} - Value'_{A,t}| < 2\theta_0$; **Medium risk**: $2\theta_0 \leq |Value_{A,t} - Value'_{A,t}| < 3\theta_0$; **High risk**: $|Value_{A,t} - Value'_{A,t}| \geq 3\theta_0$). Furthermore, we define a ***high-risk case*** as a set of continuous time spans (more than two time spans) with high risk.

Scalability of Algorithm In order to support a real time monitoring, especially with a large number of correlated sensors, we try to alleviate the scalability issue in two ways, including the use of isotropic Gaussian model [9] and selection of a limited number of Gaussian components during EM iterations without much loss of precision ($> 95\%$) [50]. In this way, the proposed algorithm is capable of monitoring real time data with waits of seconds.

4.2.2 Visual Design and Interactions

For visual design, the design goal is to allow users to efficiently capture a general picture of equipment conditions (Task *T*.2.2). Thus, two designs, namely In-situ Monitoring View and Summary View, are adopted to visualize related information in two different scales (*R*.2).

In-situ Monitoring View In this view, we want to show the risks detected by the monitoring algorithm (i.e., suspicious or abnormal conditions), which can be represented by *deviations* of real sensor values from estimated "normal" values over time. Based on iterative discussions with our domain experts, we found that what users from manufacturing sites want to understand is how the deviations of chosen target sensors evolve over time and which sensor needs further attentions. Thus, a composite design is adopted to visualize related information in a compact way.

1) Module Flow: First, a *module flow* is generated for each module, consisting of a group of target monitoring sensors. For aesthetics and legibility, we employ a stacked graph, a well-established and intuitive visualization method for time-varying data, in which layers represent evolving deviations of those target sensors and are stacked in a symmetrical shape to facilitate comparisons. Fig. 4a-1 shows an example of the module flow where the x-axis represents time. The deviations of target monitoring sensors for each time span are presented by vertical bars in different colors, namely *sensor deviation bar*, which are aligned vertically at the corresponding time point. The height of a bar indicates the deviation amount (i.e., $|Value_{A,t} - Value'_{A,t}|/\theta_0$). The longer the bar, the higher the risk with the sensor. For visual clearness, instead of connecting all bars over the whole time period, we only connect those bars of sensors with medium or high risks detected by the monitoring algorithm. Thus, users can observe the temporal patterns of detected risks intuitively. In addition, as suggested by our collaborators, the time periods with high risks are further highlighted (Fig. 4a-3) for more attentions. As time goes by, the flow keeps generating for new time spans and shifts left grid by grid for an efficient in-situ monitoring.

2) Alert Sensor Strand: To enhance the understanding of the overall condition of a monitoring module, we overlay two *alert sensor strands* along the contour of the module flow (Fig. 4a-2). In this way, by observing the shape of the strand curves, users can easily grasp a general idea of overall risks of a module. Furthermore, to help users easily track sensors with risks, we visualize those sensors with medium and high risk as dots in a circle embedded within two strands respectively (red lower strand for high risk; orange upper strand for medium risk). The sizes of these dots encode deviations (i.e., risks), and the colors are consistent with layers in the flow. However, this design might lead to visual clutter when there are too many sensors with medium or high risk. Then users can switch dots to pie-charts (Fig. 4a-4) or track via interactive highlighting.

Summary View In-situ Monitoring View shows the live status of target monitoring sensors for a limited time period (e.g., a few minutes), while investigation on the long-term trends of equipment conditions is also desired for a comprehensive online monitoring (R.2). To this end, our system further integrates a Summary View to provide users with an overview of a module's conditions during a certain past time period. Radial visualizations, such as ring maps, are commonly used for the analysis of temporal patterns. Inspired by these designs, we also adopt an intuitive radar metaphor with radial layout to present a series of monitoring sensor's status simultaneously. As shown in Fig. 4b, we create multiple concentric rings. Each of these rings corresponds to a certain time period in the past (e.g., one hour in our implementation), and the time is assigned from the inner most ring to the outer most ring so as to provide more space to reserve more details for recent time periods. Each ring is divided into a number of sectors, each representing a target monitoring sensor in the module. The sensors are linearly arranged around the ring based on their order within the manufacturing process. In each sector, the lengths of arcs in different colors (i.e., green, yellow, orange, and red) indicate the durations of the corresponding sensor with different deviation levels (i.e., good, low risk, medium risk, and high risk; refer to Section 4.2.1).

As time goes by, we simulate a radar scanning display to automatically update the arcs in the outer most ring in real-time, where the red cursor is rotating with a frequency decided by the granularity of a time span during data transformation (Section 3.1.1). Thus, by observing the lengths of arcs of different colors within different rings or sectors, users can quickly see the trends and understand which sensor is in a riskier condition; and whether the condition of corresponding module is becoming worse and needs more attentions. Furthermore, to facilitate comparisons, users can rotate the view and reorder the sectors interactively.

Discussion on Scalability Due to the limited screen space and capability of users for comparison tasks, these designs will be faced with scalability issues when there are a large number of sensors. In such cases, we consider that the goal of our system lies in detecting risks of several key sensors of a certain equipment in an early stage; therefore, we address scalability issues in two ways. On one hand, the module of too many target monitoring sensors is not recommended. On the other hand, when such cases come up, we would follow the mantra "overview first and detail on demiand". We can first group sensors into modules, then modules into super modules, and encode the statistical information of these super modules. When users click on a super module of interest, the contained modules will be expanded and visualized for further analysis.

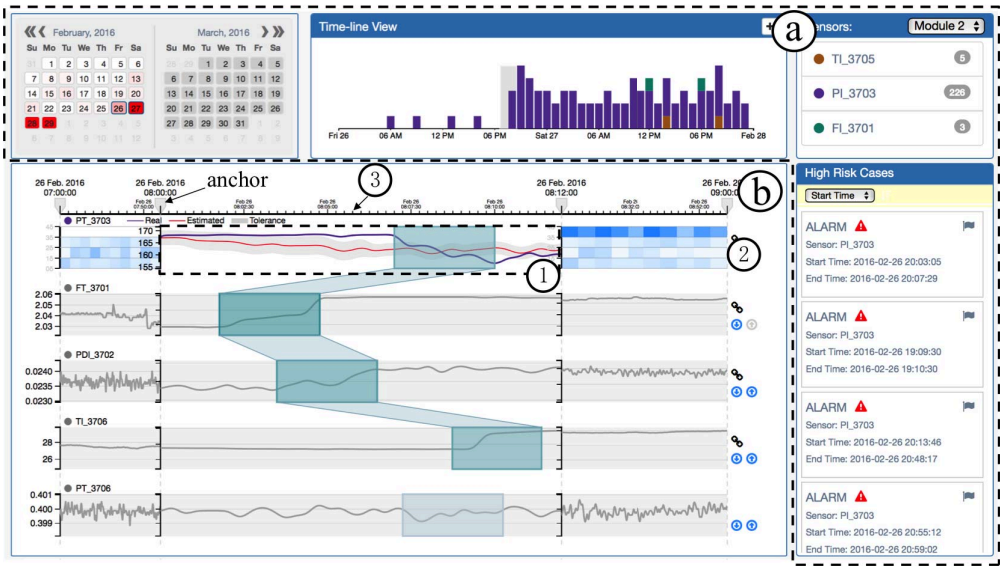


Figure 5: The system interface for detail inspection: (a) Multi-facet Filter Panel enables efficient exploration of detected high-risk cases; (b) Inspection View supports a closer inspection of vector time-series and analyzes their correlations.

4.3 Detail Inspection and Model Updating

4.3.1 Investigating Anomalies

Besides real-time monitoring, our system also enables an interactive investigation of identified suspicious or abnormal cases to help users explore potential causes and form their own judgement (e.g., whether to update the training set) (Task *T*.3.1). In this section, we describe the interactive visual interface in our system to support this exploration process.

Multi-facet Filter Panel Following the information seeking mantra "Overview first, zoom and filter, then details on demand" [39], our system integrates a Multi-facet Filter Panel (Fig. 5a) to support an efficient exploration. The panel consists of a calendar view, a timeline, a sensor list and a high-risk case list. First, the sensor list on the top right presents the number of high-risk cases detected for each target monitoring sensor, based on which users can choose a few sensors for inspection. Then the calendar panel on the top left shows the intensity of risks with the chosen sensors on a daily basis. The color opacity is determined by the number of detected high-risk cases. The darker the color, the more high-risk cases our system detected on that day. Users can select a set of continuous days on the calendar, and then the timeline on the right will be updated to show the number of high-risk cases for each hour during the selected days. By brushing on the timeline, users can choose a specific time range for further exploration. Meanwhile, the detailed information of those high-risk cases detected during the chosen time range will also be shown in the high-risk case list on the bottom right. Users can also select to focus on a particular case for further investigation.

Inspection View After the time range being chosen, a closer inspection of the vector time-series is required to offer an intuitive understanding of certain high-risk cases. In particular, our system needs to support an in-depth exploration of multi-variant time-series (i.e., time-series for a certain target monitoring sensor and its correlated sensors) and analyze the correlations between them. Well established visualization techniques are preferred by our collaborators for better interpretability and scalability. Thus, we employ a multi-line chart design as shown in Fig. 5b, which is a typical juxtaposition visual comparison technique [12] by presenting items side-by-side for a straight-forward analysis. In this multi-line chart, we show the time-series of the target monitoring sensor on the top and the correlated sensors below. For the target monitoring sensor, we first come up with a river plot [4] design (Fig. 5b-1), where the purple curve indicates the real observed value; the red curve indicates the estimated normal value by the monitoring algorithm; and the light grey field indicates the range within high-risk threshold (i.e., deviations $< 3\theta_0$). However, when a long time range is chosen, a direct application of river plot may result in visual clutter and poor legibility due to high-frequency and large-amplitude value changes of some sensors (e.g., vibration sensors). To tackle this problem, we extend it with a *pixel-based visualization* and an *interactive time axis*. The pixel-based visualization (Fig. 5b-2) provides an overview of estimated risks. Each cell represents a certain level of risks (on vertical axis) during a certain time period showing on the horizontal axis. The color opacity of each cell encodes the number of time spans with the corresponding risk level during that time period. Then users can use the interactive time axis (Fig. 5b-3) to set anchors and stretch to get more space for certain time intervals, where the corresponding river plot will be embedded. And the remaining parts will be dynamically compressed by further aggregation to adapt to the available space.

Furthermore, we include computational correlation analysis methods in the Inspection View and design an interactive slider for users to steer the algorithm. Users can brush on the river plot to choose a certain time interval, then the system will identify and highlight most correlated part in each time-series of correlated sensors, following a similar procedure in Section 4.1.1. The opacity of the highlighting part indicates how the correlation is. After that, the correlated sensors will be reordered, and sensors with stronger correlations will be ranked closer to the target monitoring sensor at the top. Moreover, users can choose a few sensors with strong correlations and sort them based on the temporal sequence of identified correlated parts in order to explore potential causal relations. Meanwhile, manually adjusting the order of these sensors by the "floating" and "sinking" buttons on the right is also supported for convenient comparisons.

4.3.2 Updating Training Set

Based on the investigation of monitoring results in the Inspection View, users can decide whether the training set needs to be refined for model updating. Then, an on-line user feedback mechanism for an iterative updating process by incorporating users observations and experience is desired (Task *T*.3.2). To this end, a Training Set View is designed to allow users observe the distribution of the chosen reference conditions in the training set and make adjustments interactively. The detailed design is described as follows:

Training Set View First of all, the training set can be viewed as a set of vector time-series indicating past reference conditions during certain time periods. To reveal the distribution, we first need to measure their similarity. Our system employs the *sum of minimum distances* [10] whose basic idea is to find a mapping that minimizes the sum of the distances between each vector of two time-series. To preserve the temporal information, the order of mapping should be monotonic.

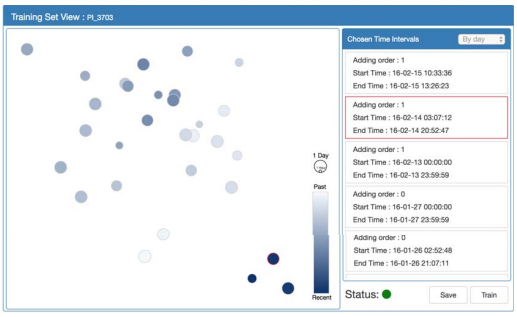


Figure 6: Training Set View shows the distribution of chosen time-series in the training set and supports an interactive updating.

Now that we have measured the similarity between any two vector time-series in the training set, then we need a visualization to show the results to users. In our system, we employ Multidimensional Scaling (MDS) to generate a scatter-plot display (Fig. 6) with each point representing a vector time series during a certain time period and the relative distance between each pair of points encoding the similarity of the corresponding vector time-series. Meanwhile, we extend the point-based display with an attribute glyph, where the size of each point encodes the length of the time period and the color encodes the temporal order. Particularly, a darker point of a larger size represents a vector time series of a longer time period in more recent times. In addition to the scatter plot, chosen time intervals are listed on the right. To support a more convenient exploration of distribution with different levels of detail, our system supports automatic splitting of chosen time-series by days or hours, and the points in the scatter plot can then be updated accordingly. In this way, users can easily get a general idea of a newly added time-series and also remove unnecessary time-series interactively. Finally, users can click "save" and "train" buttons at the bottom to save the updated training set and start training/retraining process.

5 Evaluation

To evaluate the usability and effectiveness of our system, we conducted in-depth interviews with four domain experts invited by our collaborators, including two equipment operators (Expert A and B) and one program manager (Expert C) of the "Industrial Big Data Services" from their petrochemical plant, and a professor (Expert D) from a research consortium focusing on smart manufacturing. In this section, we first describe a series of case studies based on real-world manufacturing data, which demonstrates the usage of our system. Then we report experts' feedback on our system.

5.1 Case Studies

The dataset we used contains over 0.42 billion records of 791 sensors in a petrochemical plant collected from January to March of 2016. As suggested by our collaborators, we made full use of this three-month dataset to simulate the complete using process of our system. In particular, 1) we used the data of January for feature extraction and model construction. After that, the monitoring algorithm was applied on the data of February and March respectively. 2) For February, we inspected the monitoring results for model validation and updating; 3) Then for March, we simulated the real-time monitoring process.

First of all, we started the case study with a tutorial explaining the major features of our system and demonstrating the system's functionalities. To make it easier to understand, we illustrated encoding schemes with example visualizations prepared earlier. The tutorial lasted for about 1 hour, and the participants could ask any questions during the tutorial to ensure fully understanding on how to use the system and avoid misunderstanding. After that, our experts were invited to use the system to perform three major steps described above with the help of a moderator who was available to answer any question on system usage to avoid confusions. During the study, Expert A and B, who are our target users, operated the system on a large screen, and sat together with Expert C and D for frequent group discussions. In the following part, we report the process in detail.

5.1.1 Correlation Identification for Monitoring Configuration

After loading the data of January into the system, the correlations of all sensors were calculated and an overview was shown in the Correlation View (Fig. 3a). By observing the rings, experts identified one arc of the outer layer R_1 with relatively darker purple, which indicated potential importance of the corresponding equipment (in this case, it is a compressor) since more sensors sharing strong correlations with the sensors on this equipment. By clicking on this arc, the corresponding three control loops (i.e., arcs of the inner layer R_2) were highlighted and their containing sensors were shown with a node-link diagram in the middle. Our experts then interactively assigned different colors to different control loops for better differentiation. In the node-link diagram, we could see that there were two major clusters (highlighted with dotted circles in Fig. 3a), one with more blue and orange nodes and the other with more green nodes. Expert A commented that orange and blue nodes correspond to sensors in two highly correlated control loops for rotating axes (*mechanical loop* L_1) and lube (*lube loop* L_2), while green nodes correspond to sensors in a relatively independent control loop for compression process (*compress loop* L_3), which verified our observations in the node-link diagram. Based on this observation, experts picked a few nodes of large size, indicating more correlations with other sensors, as target monitoring sensors and interactively grouped them into two modules (Fig. 3c), where Module 1 contained 5 sensors in loop L_1 and L_2; and Mondule 2 contained 3 sensors in loop L_3.

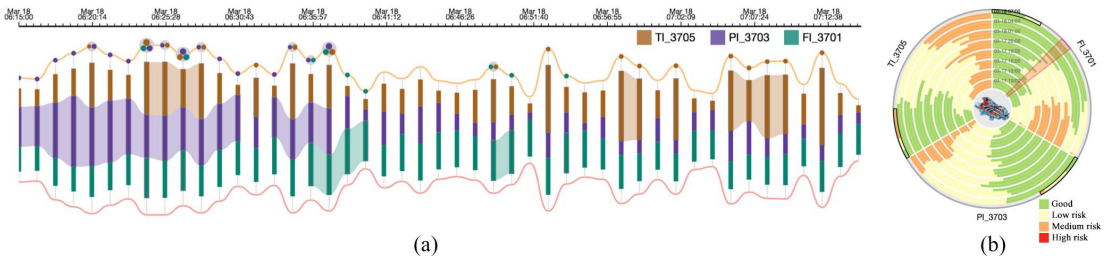
After that, experts had to identify correlated sensors for each target monitoring sensors to extract feature vectors for monitoring. Expert B suggested that, for an effective and comprehensive monitoring of the loop L3, except for those sensors on the equipment, other sensors in the plant may also be considered. Thus he right clicked on a big green node (a chosen target monitoring sensor), and the generated green bands (Fig. 3f) indicated that this sensor shared strong correlations with sensors of its upstream equipment. Thus, sensors of those two correlated control loops of the upstream equipment were added into the node-link diagram and assigned different colors for further exploration (Fig. 3b).

With the interactions supported by our system, experts could combine their domain knowledge with the recommendations by the calculation of our system to define correlated sensors for each target monitoring sensor conveniently. For example, as shown in Fig. 3b, after sample-based filtering, six nodes were highlighted in the node-link diagram as candidates of correlated sensors for the chosen target sensor $310SV\_18$. However, experts found that the edge between one of these nodes (highlighted by an arrow in Fig. 3b) and the target sensor was in very light blue, implying poor statistical significance. Based on experience, Expert A told that the stableness of this sensor was poor which might explain this observation. Thus, this sensor is removed from the list of correlated sensors. In this way, 2 modules of 8 target monitoring sensors were configured with extracted feature vectors for condition monitoring of the compressor. Initial models were constructed based on several time periods in January chosen by experts when the compressor was in good conditions according to their experience and records.

5.1.2 Inspection and Model Validation/Updating

With the constructed models, the monitoring algorithm was applied on the data of February. Then the experts tried to use our system to explore the monitoring results, so as to validate the initial models and update the training set. As shown in Fig. 5, the sensor list on the top right shows that a large number of high-risk cases were detected for Sensor *PI*_3703 in Module 2. According to the calendar panel, there were obviously more high-risk cases after Feb. 26. Therefore, the experts clicked to choose two days, Feb. 26-27, then the timeline on the right was updated. We can see that the high-risk cases were detected continuously after 8:00 p.m. of Feb. 26, which made our experts confusing since there were no accidents or anomalies during that time period in their memory. Thus, they brushed on the timeline to choose a time period of 7:00-9:00 p.m. to activate the Inspection View for further exploration.

An overview of estimated risks of the target monitoring sensor (i.e., *PI*_3703) was shown in the pixel-based visualization (Fig. 5b-2) at the top of the Inspection View. By observing the cells at the top representing high-risk level, the experts interactively added two anchor points at 8:00 p.m. and 8:12 p.m. on the time axis, and stretched to activate the river plot (Fig. 5b-1) for more details of the first detected high-risk case. By inspecting the curves in the river plot, we can clearly see that the estimated normal value of Sensor *PI*_3703 (red curve) calculated by the monitoring algorithm changes first, followed by a more significant change of the real observed value (purple curve), while the grey field grows much wider during that time period. Expert C inferred that the growth



(a) (b)

Figure 7: (a) In-situ Monitoring View and (b) Summary View for real-time monitoring of Module 2 based on the data of March.

of the grey field might imply an unstable manufacturing condition with value changes of other correlated sensors. Thus, he suggested to brush to choose a time period of 8:07-8:10 p.m. on the river plot, covering the entire value changing process of Sensor *PI*_3703. Then the system automatically highlighted the most correlated parts of the time series below for other correlated sensors. The experts chose three sensors with highlighted blocks in darker green, indicating stronger correlations, and sorted them based on the temporal sequence. Obviously, the changes on Sensor *TI*_3706 and *PDI*_3702 happened much earlier which may be the cause for the subsequent high-risk case detected on Sensor *PI*_3703. Based on this observation, Expert A recalled that there was an adjustment of manufacturing parameters due to the switch of raw materials on Feb. 26, which may explain so many high-risk cases detected by our system. Therefore, the experts chose three time periods covering Feb. 27-29, and added them into the training set for model updating. In the Training Set View (Fig. 6), we can easily identify newly added time-series which are represented by three darker points in the scatter plot, lying a bit away from other points of January. This implies the difference between the newly added time-series and the original ones, These could be nomal operating conditions after the manufacturing parameter adjustment which were not captured in the previous training process. Hence, the experts believed that updating models with these three new time-series can indeed help our system more completely cover possible normal operating conditions of the target sensor for better monitoring.

5.1.3 Real-time Monitoring

Based on the updated models, we simulated real-time monitoring using the data of March. Based on two In-situ Monitoring Views generated in our system, the experts could intuitively monitor the current status of all sensors in Module 1 and Module 2 respectively. Compared with Module 2 (Fig. 7a), there are more connected bands on the module flow and more dots on the alert sensor strand of Module 1 (Fig. 4a), indicating a poorer current condition of Module 1. In addition, the growing height of the module flow of Module 1, especially for the blue layer, caught our experts' attention and reminded us to prepare for potential risks on the corresponding sensor. Meanwhile, by observing two generated Summary Views (Fig. 4b and Fig. 7b), a larger number of orange arcs told that Module 1 had been in a poorer condition during the past few hours. After these observations, a high-risk case was detected around 7:00 a.m. of Mar. 18, just as expected (highlighted with yellow contour in Fig. 4a). Expert C commented that, with the real-time information shown in the In-situ Monitoring View and Summary View, we could observe potential risks in a much earlier stage and take necessary measures accordingly.

5.2 Expert Feedback

Interactive Visual Design All experts were impressed by the design of our system. Expert C commented that "*The interface is user friendly, and the visualizations are aesthetically pleasing*". In particular, Expert A and B were fond of the design of the In-situ Monitoring View and Summary View, and Expert B said that "*The flow and radar metaphors provide vivid and compact presentations of real-time equipment conditions from different scales, which is helpful to identify overall trends and make comparisons*". Expert A added "*Compared to the traditional monitoring dashboard in a tabular or line-chart form, I can easily grasp a general idea of equipment conditions by quickly glancing at these two views. It significantly reduces my work burden*". In addition, Expert D considered detail inspection valuable for users to deeply inspect time-series data. He also appreciated the flexible exploration supported by the Multi-facet Filter Panel and Inspection View, and highlighted that "*Flexible navigation is essential to address complex analytical tasks, and side-by-side comparisons provide a straightforward approach for pattern identification*". Last but not least, the experts also acknowledged the usefulness of the Correlation View and Training Set View for an interactive modeling process.

Usability and Improvements Expert C appreciated our system as a pioneering work for exploring the potential of applying visual analytics in smart factories of process industry. Expert A believed that a major advantage of our system is "*It combines intuitive visualization with advanced analytical methods to provide a powerful tool for exploring data and steering the analysis process, which makes users feel more comfortable and confident about the results obtained from the system*". Expert B added "*At first, I was attracted by such a novel and modern dashboard, but it did require a bit of learning curve at the beginning to get familiar with all the views and operations of the system. Once I get used to it, I find these visualizations adopted are quite informative and the nalytical pipeline is easy to follow*". Expert D was interested in the correlation analysis supported by our system and said "*This method should be very useful to extend to other applications in manufacturing, such as fault analysis, because it could help avoid missing potential factors by fully combining advantages of human and computers*". Besides that, valuable suggestions are also raised to improve the system. Expert C suggested "*We could take other manufacturing information, such as maintenance records, into consideration, which could help to refine the training set for more accurate model construction*". In addition, Expert D commented that, except for Pearson Correlation Coefficient, our system could integrate more advanced correlation measurements to characterize relationships among sensors in a more comprehensive way.

6 Discussion and Future Work

In this work, we present an interactive visual analytics system to support an effective and efficient equipment condition monitoring. A semi-supervised framework and a suite of interactive visualizations are proposed to address the major challenges and enable visual-assisted knowledge discovery. Well-established visualization techniques are employed to lower the learning curve for engineers in a plant of process industry without related background. We demonstrate the effectiveness and usefulness of our system through case studies and expert interviews. Our work is still in progress. There exist some limitations of the current prototype that we would like to address in the future. First, it is our plan to further investigate more advanced anomaly detection models for equipment conditions and improve current monitoring algorithm to enable a more effective analysis under complex situations of manufacturing. In addition, we also intend to conduct well-organized and long-term studies with quantitative measurements and collect feedbacks from end users to further improve our system.

Acknowledgments

The authors thank the anonymous reviewers for their valuable comments. This research was supported in part by NSFC Grant 61602306, and a grant from Siemens Ltd. Yixian Zheng is the corresponding author.

References

[1] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.
[2] E. Auschitzky, M. Hammer, and A. Rajagopaul. How big data can improve manufacturing. *McKinsey & Company*, 2014.
[3] L. Brooks and D. Buckmaster. First difference signals and accounting income time series properties. *Journal of Business Finance & Accounting*, 7(3):437–454, 1980.

[4] P. Buono, C. Plaisant, A. Simeone, A. Aris, G. Shmueli, and W. Jank. Similarity-based forecasting with simultaneous previews: A river plot interface for time series forecasting. In *Information Visualization, 2007. 11th International Conference*, pages 191–196. IEEE, 2007.

[5] N. Cao, C. Shi, S. Lin, J. Lu, Y.-R. Lin, and C.-Y. Lin. Targetvue: Visual analysis of anomalous user behaviors in online communication systems. *IEEE transactions on visualization and computer graphics*, 22(1):280–289, 2016.

[6] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 143–152. IEEE, 2012.

[7] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[8] L. F. Cranor. A framework for reasoning about the human in the loop. *UPSEC*, 8(2008):1–15, 2008.

[9] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[10] T. Eiter and H. Mannila. Distance measures for point sets and their computation. *Acta Informatica*, 34(2):109–133, 1997.

[11] P. C. Evans and M. Annunziata. Industrial internet: Pushing the boundaries. *General Electric Reports*, 2012.

[12] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[13] Y. Han and Y. Song. Condition monitoring techniques for electrical equipment-a literature survey. *IEEE Transactions on Power delivery*, 18(1):4–13, 2003.

[14] R. L. Harris. *Information graphics: A comprehensive illustrated reference*. Oxford University Press, 2000.

[15] P. Hayton, S. Utete, D. King, S. King, P. Anuzis, and L. Tarassenko. Static and dynamic novelty detection methods for jet engine health monitoring. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 365(1851):493–514, 2007.

[16] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

[17] S. Jakubek and T. Strasser. Fault-diagnosis using neural networks with ellipsoidal basis functions. In *American Control Conference, 2002. Proceedings of the 2002*, volume 5, pages 3846–3851. IEEE, 2002.

[18] J. Jo, J. Huh, J. Park, B. Kim, and J. Seo. Livegantt: Interactively visualizing a large manufacturing schedule. *IEEE transactions on visualization and computer graphics*, 20(12):2329–2338, 2014.

[19] Z. Ju and H. Liu. Fuzzy gaussian mixture models. *Pattern Recognition*, 45(3):1146–1158, 2012.

[20] H. Kagermann, J. Helbig, A. Hellinger, and W. Wahlster. *Recommendations for Implementing the strategic initiative INDUSTRIE 4.0: securing the future of German manufacturing industry; final report of the Industrie 4.0 working group*. Forschungsunion, 2013.

[21] S. Kalpakjian and S. R. Schmid. *Manufacturing engineering and technology*. Pearson Upper Saddle River, NJ, USA, 2014.

[22] Y.-a. Kang and J. Stasko. Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 21–30. IEEE, 2011.

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

[24] E. Keogh, S. Lonardi, and C. A. Ratanamahatana. Towards parameter-free data mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 206–215. ACM, 2004.

[25] D. Kim, P. Kang, S. Cho, H.-j. Lee, and S. Doh. Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing. *Expert Systems with Applications*, 39(4):4075–4083, 2012.

[26] T. Kimura, T. Tokuda, Y. Nakada, T. Nokajima, T. Matsumoto, and A. Doucet. Expectation-maximization algorithms for inference in dirichlet processes mixture. *Pattern Analysis and Applications*, 16(1):55–67, 2013.

[27] D. Kit, B. Sullivan, and D. Ballard. Novelty detection using growing neural gas for visuo-spatial memory. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 1194–1200. IEEE, 2011.

[28] J. Lee Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[29] S. Liu, W. Cui, Y. Wu, and M. Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, 2014.

[30] J. Ma and S. Perkins. Online novelty detection on temporal sequences. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–618. ACM, 2003.

[31] K. Martin. A review by discussion of condition monitoring and fault diagnosis in machine tools. *International Journal of Machine Tools and Manufacture*, 34(4):527–551, 1994.

[32] K. Matkovic, H. Hauser, R. Sainitzer, and M. E. Groller. Process visualization with levels of detail. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, pages 67–70. IEEE, 2002.

[33] S. McKenna, D. Staheli, C. Fulcher, and M. Meyer. Bubblenet: A cyber security dashboard for visualizing patterns. In *Computer Graphics Forum*, volume 35, pages 281–290. Wiley Online Library, 2016.

[34] M. A. F. Pimentel, D. A. Clifton, C. Lei, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99(6):215249, 2014.

[35] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4, 2005.

[36] J. Posada, C. Toro, I. Barandiaran, D. Oyarzun, D. Stricker, R. de Amicis, E. B. Pinto, P. Eisert, J. Döllner, and I. Vallarino. Visual computing as a key enabling technology for industrie 4.0 and industrial internet. *IEEE computer graphics and applications*, 35(2):26–40, 2015.

[37] S. Radhakrishnan and S. Kamarthi. Complexity-entropy feature plane for gear fault detection. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 2057–2061. IEEE, 2016.

[38] S. J. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1133–1142, 1998.

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

[40] L. Stojanovic, M. Dinic, N. Stojanovic, and A. Stojadinovic. Big-data-driven anomaly detection in industry (4.0): An approach and a case study. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 1647–1652. IEEE, 2016.

[41] S. Sundaram, I. G. Strachan, D. A. Clifton, L. Tarassenko, and S. King. Aircraft engine health monitoring using density modelling and extreme value statistics. In *Proceedings of the 6th International Conference on Condition Monitoring and Machine Failure Prevention Technologies*, 2009.

[42] G. K. Tam, V. Kothari, and M. Chen. An analysis of machine-and human-analytics in classification. *IEEE transactions on visualization and computer graphics*, 23(1):71–80, 2017.

[43] J. J. Van Wijk and E. R. Van Selow. Cluster and calendar based visualization of time series data. In *Information Visualization, 1999.(Info Vis' 99) Proceedings. IEEE Symposium on*, pages 4–9. IEEE, 1999.

[44] K. Wang, T. Gasser, et al. Alignment of curves by dynamic time warping. *The annals of Statistics*, 25(3):1251–1276, 1997.

[45] M. Wörner, T. Ertl, S. Miksch, and G. Santucci. Visual analysis of advanced manufacturing simulations. In *EuroVA 2011: International Workshop on Visual Analytics*, pages 29–32. The Eurographics Association, 2011.

[46] M. Wu and C. Jermaine. Outlier detection by sampling with accuracy guarantees. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 767–772. ACM, 2006.

[47] W. Wu, J. Xu, H. Zeng, Y. Zheng, H. Qu, B. Ni, M. Yuan, and L. M. Ni. Telcovis: Visual exploration of co-occurrence in urban human mobility based on telco data. *IEEE transactions on visualization and computer graphics*, 22(1):935–944, 2016.

[48] W. Wu, Y. Zheng, H. Qu, W. Chen, E. Gröller, and L. M. Ni. Boundaryseer: Visual analysis of 2d boundary changes. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 143–152. IEEE, 2014.

[49] P. Xu, H. Mei, L. Ren, and W. Chen. Vidx: visual diagnostics of assembly line performance in smart factories. *IEEE transactions on visualization and computer graphics*, 23(1):291–300, 2017.

[50] C. Yuan and C. Neubauer. Bayesian sensor estimation for machine condition monitoring. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 2, pages II–517. IEEE, 2007.

[51] C. Yuan and C. Neubauer. Bayesian sensor estimation for machine condition monitoring, July 21 2009. US Patent 7,565,262.

[52] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins. # fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1773–1782, 2014.

[53] J. Zhao, P. Forer, and A. S. Harvey. Activities, ringmaps and geovisualization of large human movement fields. *Information visualization*, 7(3-4):198–209, 2008.

[54] Y. Zheng, W. Wu, Y. Chen, H. Qu, and L. M. Ni. Visual analytics in urban computing: an overview. *IEEE Transactions on Big Data*, 2(3):276–296, 2016.