Data Integration and Visualisation for Demanding Marine Operations.

Conference Paper · April 2016

DOI: 10.1109/OCEANSAP.2016.7485617

Hao Wang
Norwegian University of Science and Technology, Gjøvik

Girts Strazdins
Norwegian University of Science and Technology

Guoyuan Li
Norwegian University of Science and Technology

Houxiang Zhang
Norwegian University of Science and Technology

Data Integration and Visualisation for Demanding Marine Operations

Hao Wang*, Xu Zhuge*, Girts Strazdins*, Zheng Wei*, Guoyuan Li†, Houxiang Zhang†
* Big Data Lab, Faculty of Engineering and Natural Sciences,
Norwegian Unversity of Science and Technology in Aalesund, Norway.
† Mechatronics Lab, Faculty of Maritime Technology and Operations,
Norwegian Unversity of Science and Technology in Aalesund, Norway.

***Abstract*—Marine operations face the major challenge of increased complexity. Much more data are being collected for diagnostic and monitoring purposes with sensors on board a vessel. However, due to the lack of easy data presentation and interaction, the crew members are overwhelmed by the large volume of information and alert messages, causing serious *information overload problems*. In this paper, we explore data integration and visualisation techniques for maritime operations and present a proof-of-concept prototype using real offline data from our industrial collaborators. This work is an important part of our initial efforts towards a *visual analytics* framework for maritime operations.**

***Index Terms*—Maritime operations, visual analytics, data integration, data visualisation.**

I. Introduction

Norway is ranked as the world's second largest nation in maritime operations. Advanced marine operation is becoming the core activity in the Norwegian maritime industrial cluster, and new areas of marine operations emerge for the so-called "after oil" era. Marine operations face the major challenge of increased complexity in technology and integrated operations involving multiple vessels and autonomous units.

Technologies are being adopted for acquiring monitoring data about how the vehicle and different components are behaving. Recently, with the intention of remote ship monitoring for better services for shipping customers, vessel builders started to adopt new sensor technology by installing different sensors for different components on board a vehicle and transmit data using satellite communications to land-based service centres, e.g., *HEalth MOnitoring System* (HEMOS) by Rolls-Royce Marine AS.

These systems provide more accurate and timely operational data, but they also introduce new danger to the operations: *information overload problem* (IOP) [1] – the crew members receive a large volume of monitoring information and alert messages that s/he can easily overlook important/vital ones. Therefore, it is urgently needed to develop and implement a new framework to integrate and visualise the monitoring data in an informative way to assist the on-board operations and onshore analysis.

The new paradigm *Big Data Analytics* (BDA) is quickly rising. It is highlighted in a statement from the United Nations that "the world is experiencing a data revolution" [2]. As a key step in BDA, data integration combines data from various sources and provides a unified view of these data, including data fusion, data cleansing/purification, and data validation. The monitoring data on a vessel comes from many sources, in different formats and frequencies. i.e., the data are multifaceted [3]. For example, low frequency data include heading, speed, and GPS; and high frequency data include vibration and torque. In addition, a major challenge for vessel data integration is the poor quality of the raw sensor data. Data cleaning aims to improve data quality using for example statistics, integrity constraints. However, it is generally very difficult to guarantee the accuracy of the data cleaning process without verifying it via experts or external sources [4]. Therefore, close collaboration with domain experts is essential for cleansing the monitoring data from vessels.

Another major challenge is the presentation and user-interaction of the monitoring data. Marine operations are complicated and the monitoring data can be used in many different, even unforeseeable, ways. The user-interaction functionality in existing data management and visualising tools are either very limited or involved with complex query commands, which is not applicable for marine operations. Therefore, easy and intuitive data visualisation and interaction for both 1) on-board captains and crew members and 2) onshore online support teams and offline analysts are essential.

Visual Analytics (VA) combines automatic analysis techniques with interactive data visualisations. The seminal papers [1], [5] have shown that visual analytics enables a virtuous cycle of user interaction, parameter refinement for algorithmic analysis methods so as to achieve rapid correction and improvement of human's knowledge and decisions.

The city Aalesund is located at the heart of the northwestern Møre cluster of maritime industry. The SFI[1] Centre for marine operations has been set up in Norwegian University of Science and Technology in Aalesund in 2015 to support innovation for demanding marine operations. In the framework of our Innovation Norway project "GCE Blue Maritime Big Data" [6], we obtained the monitoring data from HEMOS by Rolls-Royce Marine AS covering three years' operation of one vessel, including high frequency machinery and low frequency

Corresponding addr.: Hao Wang, hawa@hials.no, Norwegian Unversity of Science and Technology, Aalesund, Postboks 1517, 6025 Aalesund, Norway

[1]SFI is a prestigious research program in Norway to build up a Centre for *Research-based Innovation* for a certain application area

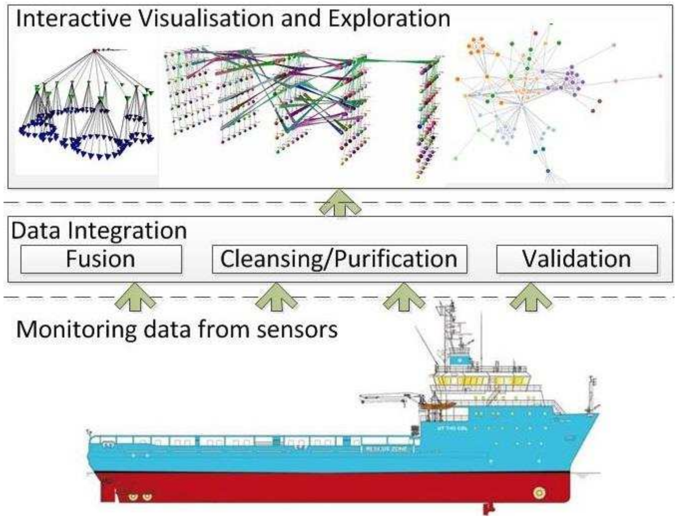

Figure 1. Integration and visualisation of monitoring data for demanding marine operations

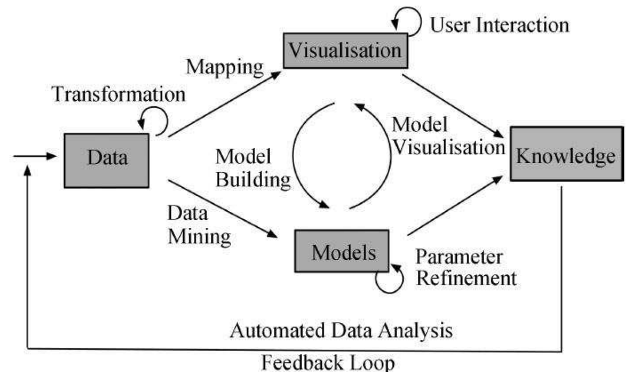

Figure 2. Visual Analytics Framework [11, Fig.1]

vessel behaviour monitoring data.

In the Big Data Lab in NTNU Aalesund (BDL), we aim to develop a visual analytics framework for maritime operations. This paper and papers [7], [8] represent our initial results in data integration and visualisation, efficient pattern identification, and prediction respectively. The overview of data integration and visualisation of this paper is depicted in Figure 1. Monitoring data is collected from various sensors on board a vehicle, the data will go through the integration layer for data fusion, cleansing/purification, and validation. The output from the integration layer is visualised with intuitive user interaction, including data correlations exploration, spatial and temporal zooming, and etc.

We have built a proof-of-concept prototype which integrates and visualises the monitoring data along with a 3D animation of vessel motions. More importantly, the prototype allows easy data exploration w.r.t. spatiotemporal features, data correlations, and etc. The prototype has received positive feedback from our industrial partners and is undergoing further improvements.

The remainder of the paper is structured as follows: Section II reviews existing literature on visual analytics with a focus on data visualisation. In Section III, we discuss in details the challenges we faced in integrating the real monitoring data and potential solutions and strategies. In Section IV, we discuss several aspects of the interactive data visualisation for the monitoring data. Section V presents our prototype and Section VI concludes the paper and presents some future directions.

II. Background

There has always been "too much" data to analyse, why only in very recent years people claim that we are now in the "new era of big data"? Dean [9] gives a good observation on this phenomenon:

> The large data volume does not solely classify this as the big data era… What sets the current time apart as the big data era is that companies, governments, and nonprofit organizations have experienced a shift in behavior. In this era, they want to start using all the data that it is possible for them to collect, for a current or future unknown purpose, to improve their business.

Keim et al., in their seminal paper [1], explored the definition, process, and challenges of visual analytics. They stated that the goal of VA is the creation of tools and techniques to enable people to:

- *synthesise* information and *derive* insight from data;
- *detect* the expected and *discover* the unexpected;
- *provide* timely and understandable assessments;
- *communicate* assessment effectively for action.

Shneiderman [10] proposed a well-known information seeking mantra: "Overview first, zoom/filter, details on demand". However, when the data are too large or complicated, more iterations with the human analyst will be necessary. Therefore, Keim et al. [1] extended the mantra to be: "Analyze first, show the important, zoom/filter, analyze further, details on demands" and introduced the seminal VA framework, depicted in Figure 2.

The use of visualisation can steer the analytical process [12]: the human analyst can interactively choose parts of data or change parameters of automatic analysis methods, the results of which can be further visualised and analysed with other methods.

One key element of VA, compared to the currently dominating approach of automatic (algorithmic) analysis, is the recognition of the importance of *information visualisation* in the human understanding and analysing process. Fekete et al. [5] has rigorously explored the value and benefits of information visualisation. Even as the fully-automated algorithmic methods can quickly identify useful information and provide more accurate prediction, they lack the ability to interact with human and deliver effectively the knowledge. The combination of visualisation and automated algorithmic methods enables a virtuous cycle of user interaction, parameter refinement for algorithmic models so as to achieve rapid correction and improvement of human's knowledge and decisions.

Visual analytics is still relatively new for the maritime community. Maria Riveiro's Ph.D. thesis (2011) [13] on the detection of anomalous vessel behaviour in traffic and some following-up research on maritime traffic are the closest work we can find. The subject area, requirements, and sources of data of their work are different from our project and visual analytics framework, but we did get plenty of information and inspirations from them.

III. Data Integration

Data collected from ship on-board systems presents a set of challenges. In this section we identify the challenges based on our experience with real ship monitoring data sets. We discuss potential solutions and mitigation strategies.

A. Data in raw formats

Data logging systems used on vessels are not necessarily optimised for analytics and visualization. Rather the goal for these tools is to record as much data as possible with minimal loss of accuracy. As a result, the data are usually stored in formats well suited for fast sequential logging, without any indexing or other analytics capabilities. Typical formats include plain-text files in Comma-Separated-Value (CSV) format, Matlab data files, or custom binary files. These formats have their advantages: relatively small needed storage space and fast load time for applications analysing the whole data set. However, these formats require linear algorithms for basically every operation: $O(n)$, where n is the number of data samples. These formats are not optimal for interactive visualizations or analytics. Therefore data analysts should be ready to convert data to other formats more appropriate for integration, analytics and visualization. Choices of appropriate format are specific to the data volume and application requirements.

B. Relational databases versus NoSQL storage

One research direction in the last decade has been the so-called *NoSQL* data storage - data storage models not relying on relational databases and tables. While NoSQL includes a large variety of databases, the term is used to describe massively scalable solutions that can run on clusters of thousands of machines, such as MapReduce [14], BigTable [15] and Dynamo [16]. Another niche for NoSQL data storage is the so-called document stores, such as MongoDB [17], efficient in scenarios with data having dynamic and not-predefined structure. We argue that NoSQL databases are useful in two scenarios: (1) for real-time analysis of live data streams where record pre-processing and indexing are not possible; and (2) for storage of data without a specified schema. Yet for marine operation data, especially for post-analysis and visualization, traditional relational databases are more appropriate due to several reasons:

1) Relational databases are well established and the implementations are well tested. SQL query language provides rich feature set and it is better known among programmers and data analysts. NoSQL requires learning specific query languages and even a different mindset. Tools (e.g. Pig latin [18], HiveQL [19]) are appearing to provide SQL-like language on top of MapReduce NoSQL engines;
2) Vessel monitoring data come from many different sources, including on-board sensors. While the data types and value represent a large range, the structure of data is very well defined, at least at the individual vessel and operation level. Therefore constructing a schema for such data sets is straight forward.
3) Considering limited bandwidth of vessel communication channels during maritime operations, it is not practically feasible to transfer all the data to offshore data centers in real-time. Also, there is no room for huge data centers on board vessels due to strive for energy-efficiency (and hence space-efficiency). Therefore, typically the on-board systems log as much data as possible in raw format for post-processing, while uploading only a limited and pre-processed data set for near-real-time analysis onshore.
4) Parallel processing algorithms, such as MapReduce, divide data in sections which can be processed independently. However, maritime operations are described with spatiotemporal multivariate data. Although each sample can be seen as independent (e.g., calculation of min/max values can be done in parallel), in many cases we are interested to work with derivatives, analyse and visualise the sequence of samples. Therefore, possibilities of parallel processing are reduced.

Based on the above arguments we conclude that relational databases with SQL interface are powerful tools satisfying wide range of data integration and visualization needs for marine operation analysis.

C. Different formats and sampling frequencies

Extracted data formats may differ from vessel to vessel, depending on manufacturer of different sensors and software modules. As a result, semantically the same data is encoded in different syntax. Several key points are important here.

First, the visualization should support generic data model. The goal of the integration system is to provide a conversion layer, which takes input data in different raw formats, stores them in a common format for visualisation and analysis. In addition, an automated data assessment module can generate statistics about the data: gaps, number of records, mean values, value distributions, and etc.

Second, the data analysts should be in close cooperation with domain experts to establish common standards and practices for data logging and streaming formats. This aspect is especially challenging considering the degree of conservatism in the industry. Introducing new concepts and standards is challenging. However, more and more maritime experts recognise the importance of data analytics for improvement of ship design and operations.

Data come from different sources. Even inside one system there might be several modules with different sampling frequencies. Therefore they need to be normalized to be comparable [20] to calculate correlations or detect anomalies. Several steps can address the challenge. First, it is important to have an exact time stamp for every data sample. Depending on the data format it may or may not be included in raw data. Second, clock synchronisation of all the modules must be ensured. It is a complex problem in general, simplified solutions may exist in specific cases. Although synchronising clocks before data logging is preferred, this could be outside the competence of data analysts. Therefore post-processing phase requires synchronization of modules with drifted clocks. Third, extrapolation of recorded data is needed to calculate missing records and generate samples with the same frequency for all modules.

D. Black box

When analysts receive the data, to a great extent it is a *black box*. We may have different assumptions about the data. Yet considering the challenging conditions during maritime operations and complexity of the systems, different unpredicted results can be found in recorded data. We have to validate all our assumptions and get statistics to understand the data. Questions include: are there overlaps of data? If file names include date and time, do they represent time of first or last record in the file? Are there gaps in data and how large? Although these questions may sound trivial, one can discover surprising facts while answering them. For example, our experience shows that some data logging systems may have overlaps in data samples, which could be caused by internal buffering and data flushing procedures of the logging systems.

Two aspects are important in this regard. First, domain expert advice is essential. Especially, if these experts are close to the particular vessels and know the data logging systems. Second, the data integration system should include a generic framework for automated testing of data. It is similar to *unit testing* in software development: we define assumptions (or assertions) about the data and the system should be able to automatically detect and report violations. For example, the system can detect data overlaps, wrong timestamps, huge gaps, and etc.

E. Overlaps and gaps in data

As mentioned above, the data can bring different surprises. One typical surprise is gaps in logged data. That can happen due to different reasons: lack of storage space on board[2], system errors or operational policies. Although less likely, data overlaps can also happen, especially when data from different sub-systems are merged and the sub-systems are not aware of the global state of the system.

Four aspects are important in this regard. First and most importantly, the gaps and overlaps should be detected and reported. The data integration system should do that automatically. Second, the statistics and rare incidents should be reported to the analysts. Third, automatic elimination of duplicates should be done. This step is trivial in SQL, showing yet another advantage of relational databases. Forth, extrapolation of data could be useful to fill in short gaps. In this case every record should be marked: either original data, or calculated value. In this way we can filter necessary samples in queries, and exclude extrapolated values if necessary.

[2]In Subsection III-A, we mentioned that the data are typically stored in plain-text formats to reduce size. However, continuous streaming can still generate large volume of data, which can easily exceed the size of on-board storage. Therefore, intermittent sampling is commonly practiced.

IV. Interactive Data Visualisation

From our discussions in SectionIII, we can see that the data we obtained from HEMOS are *multifaceted*, they are *spatiotemporal*, *multivariate*, and *multimodal* data [3]. If we also consider weather and oceans data, then they will become multimodel data, which result from coupled simulation models that represent physically interacting phenomena. In this paper, we only consider the former three facets, leaving the multimodel facet as our next step.

A. Visualisation Techniques

The monitoring data from vessels contain multiple attributes in different spatiotemporal frequencies. The visualisation of multidimensional multivariate data remains a challenge [21], even after several decades of development [22].

For multidimensional data, the technique of *coordinated multiple views* [3] enables that different data variables can be shown, explored, and analysed in multiple linked views placed side by side. The views can be histograms, scatter-plot matrices, parallel coordinates, or function graphs. Data can be selected or *brushed* [23] in a view, the related data items are instantly highlighted in all *linked views*. Views can be enabled or disabled and different parts of data can be filtered, *reduced*, and etc.

Exploring distributions or correlations among different data dimensions is the key to visualise the multivariate data. A lot of methods have been proved to effectively display and understand these correlations, e.g. projection-based methods [20] and factor generation methods [24].

Many application areas can treat the temporal and spatial data just like any other data attributes. However, the monitoring data for vessels are mainly collected from different temporal data sources along with geospatial data sources like GPS coordinates, which play a central role in the analytical tasks. In addition, an important observation is that in the context of maritime operations, visualising the movement of the vessel w.r.t. the time and location axes is often useful for analysts to understand the behaviour and identify patterns and problems, e.g., propeller ventilation [7].

B. Interaction Techniques

Kerren and Schreiber [12] presented a taxonomy of interaction techniques, based on which we present as follows a customised taxonomy according to the needs of our framework.

(A). Data and View Specification

1. Encode/Visualise: Users can choose the visual representation of the data records.
2. Reconfigure: Some interaction techniques allow the user to map specific attributes to graphical entities.
3. Filter: This technique is of great importance for visual analysis as it allows the user to interactively reduce the data shown in a view. For temporal monitoring

variables, dynamic queries by using time-range sliders are commonly used.
4. Derive: The completion of additional computations based on the primary input data.
5. Adjust: Related to the previous interaction type is the modification of parameters for automatic analyses (incl. simulations).

(B). View Manipulation
1. Select: The aim is to select an individual object or a set of them in order to highlight, manipulate, or filter out them.
2. Navigate/Explore: This important class of interaction techniques typically modify the level-of-detail.
3. Coordinate/Connect: Linking a set of views or windows together to enable the user to discover related items.
4. Organise: Large systems often consist of several windows and workspaces that have to be organised on the screen.

(C). Process and Provenance
1. Record: Methods that store and visualise the interaction history performed by the user help to facilitate the iterative analysis process.
2. Annotate: Graphical or textual annotations help the analyst to point to elements or regions within the visual representation.
3. Share: Collaboration in VA often occur in practice, but it is still not very well researched.
4. Guide: It would be beneficial if a VA system would support "guided analytics to lead analysts through workflows for common tasks".

V. Prototypes and Discussion

A. A Data Integration Prototype

We have implemented a data integration prototype in Python and tested it using part of the HEMOS data.

The prototype includes several useful functions:

- Converting data samples to a generic format and importing them to a MySQL database.
- Detecting and removing duplicate data samples. The source data were distributed among many small files. The integration script collects them all together and eliminates duplicates.
- Generating statistics of data records: answering questions like how long the periods of uninterrupted (consecutive) data are and how large the gaps where we do not have any data are, as well as statistics on overlaps in data. In addition to calculating the numbers, the system can generate histogram plots.
- A script that can select a projection of the data samples in a specified time window, with several resolution options (one second, one minute, ten-minutes). The data is exported in JSON format, as a generic Web API, which can be used by a wide range of applications. Currently, the visualisation with D3.js is using this as a data source.

Histogram plots of the *number of occurrences* of three different types of interesting time intervals in our test-data set are depicted in: the consecutive data periods (Figure 3), gaps (Figure 4) and overlaps (Figure 5). Most of the consecutive periods with data are shorter than 90 minutes, while majority of gaps are less than 30 hours long. The length of data intervals is more spread while the gaps are very strictly limited: only three outliers are longer than 30 hours.

We can also see interesting facts about overlaps. It is very important to see that there are overlaps. When our research group initially received the test-data set, it was not expected to have any overlaps in the data. The performed analysis suggests that there is a large number of overlaps in the range up to 1 minute (with several outliers 3 minutes long). Interestingly, the number of overlaps increases if we focus closer to one minute range: there are clearly much more overlaps with length from 30 to 60 seconds, compared to overlaps of 30 seconds or less. The 60-second boundary makes the analysts concerned. This topic will clearly be discussed with domain experts to find out potential reasons for the overlaps.

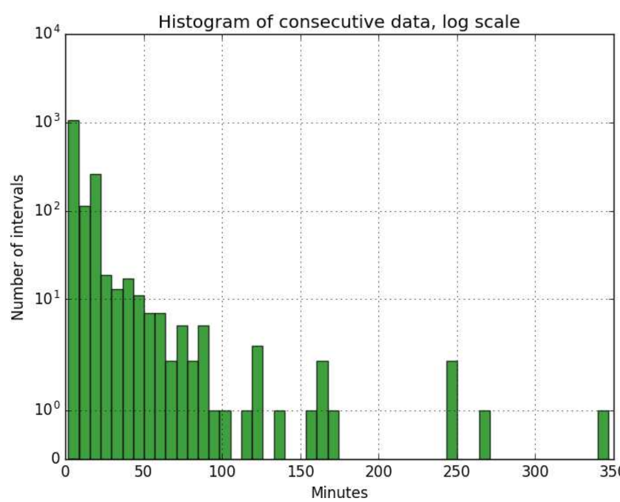

Figure 3. Consecutive periods of uninterrupted data in our test-data set

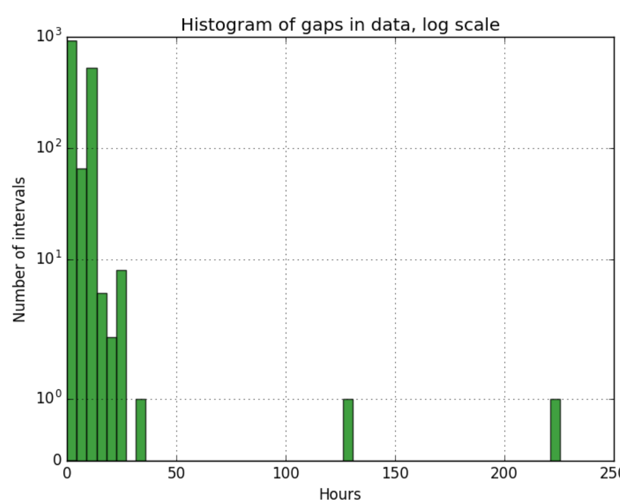

Figure 4. Gaps in our test-data set

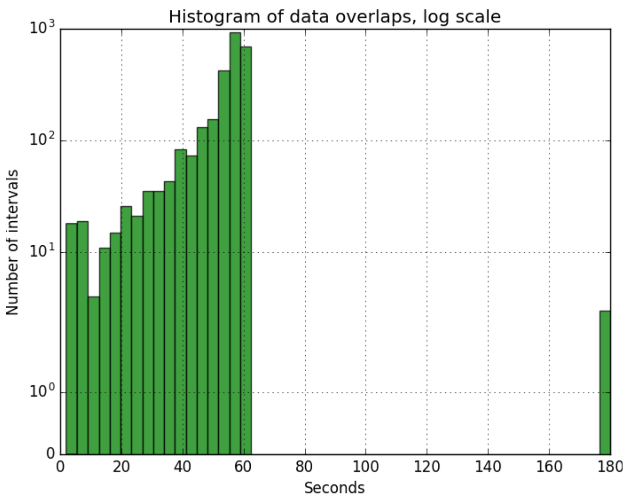

Figure 5. Overlaps in our test-data set

B. An Interactive Visualisation Prototype

The interactive visualisation prototype developed with Javascript and D3.js is depicted in Figure 6. D3.js[3] is an open-sourced JavaScript library. The library provides a large collection of data manipulation and visualisation components, allowing developers to bind arbitrary data to a *Document Object Model* (DOM), without being tied to a proprietary framework.

For the multivariate facet of the data, we used the force-directed graph (depicted in Figure 7) and the chord diagram (depicted in Figure 8) to visualise the correlations between data variables. In the force-directed graph, a physics engine is embedded to enable dynamic interaction. In the graph, 1) each node represents a different variable; 2) the thickness of the link between two nodes represents the correlation coefficient between the two variables; (3) the user can select and drag any node, if a node is more correlated with more nodes, then this variable has a stronger effect or "force" on all other variables; 4) when one node is selected, the links to all variables that are correlated to it are highlighted. The chord diagram is another graph visualising the quantified correlations. When the user select one variable, all variables that are correlated to it are highlighted.

We implemented an interactive line chart, depicted in Figure 9. With WebGL and Three.js[4], we implemented the 3D vessel motion animation (depicted in Fig. 10), in which we embedded an interactive 2D trajectory visualisation using the Google Maps® JavaScript API.

It is important to note that these visualisation components are coordinated multiple views of the same data source. the user can use the force-directed graph or the chord diagram to explore the variables and their correlations, the user can select one or several variables to be shown in the line chart. Once the user selects a time period in the line chart and starts the animation, the changes of variables and the motion of the vessel will be animated synchronously.

[3]https://d3js.org/

[4]http://threejs.org/

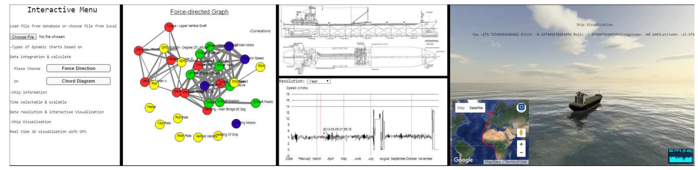

Figure 6. Visualisation prototype with Javascript and D3.js

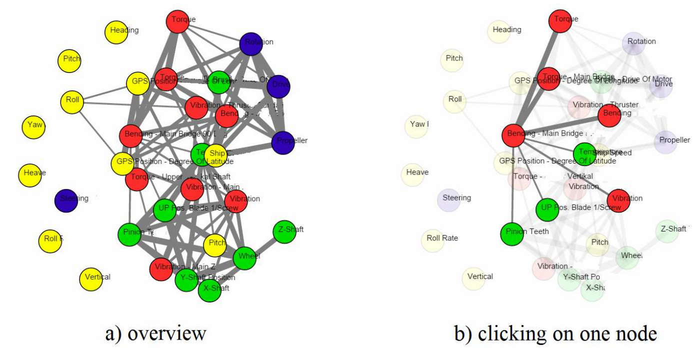

Figure 7. Interactive force-directed graph

VI. Conclusion and Future Work

In this paper, we present a detailed review on data integration and visualisation techniques in the context of monitoring sensor data from maritime vessel operations. Following the principles in our discussions, we present a proof-of-concept prototype for offline data integration and visualisation.

There are still several important aspects that we have to explore before we can build up a solid visual analytics framework. For data integration, although we argue that relational databases have advantages for our current stage, we envision that transition to NoSQL could be necessary in the future. First, cost and size of computing power and data storage are decreasing. It is likely that in the future low-cost, high-performance computing will be feasible on board a vessel. Second, onshore operational centers manage a fleet of vessels in real time. Large amounts of data are fused here, and we can see each vessel as an independent data source. Communication channel bandwidth is also subject to increase. We therefore expect in the future that vessels will send high-frequency multivariate data to onshore centers. It can be in the form of real-time streams, or previously buffered data downloaded on request. Some of the usage scenarios include emergency situation detection where alarms should be raised in near-real-time, therefore high-performance scalable analytics becomes important.

For data visualisation, many visualisation and interaction techniques could be included in our prototype, esp. the ones in categories "(B) View Manipulation" and "(C) Process and Provenance".

Acknowledgements

The authors would like to thank for the valuable inputs from Ibrahim A. Hameed, Bikram Kawan, Rune Valle and other members from BDL; Rune Garen, Leif R. Solaas, Thomas Oksavik, Krzysztof Swider, and Are Folkestad from Rolls-Royce Marine AS; Per Erik Dalen from Aalesund Knowledge

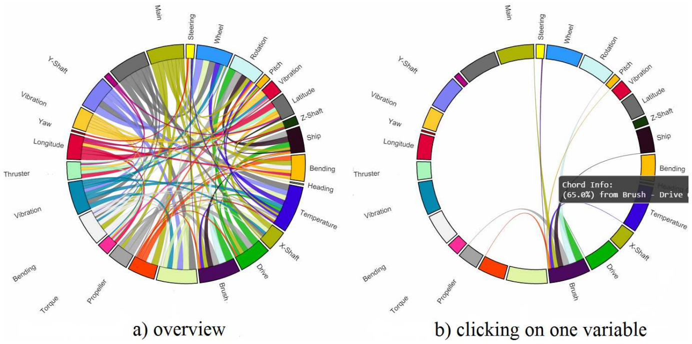

Figure 8. Interactive chord diagram

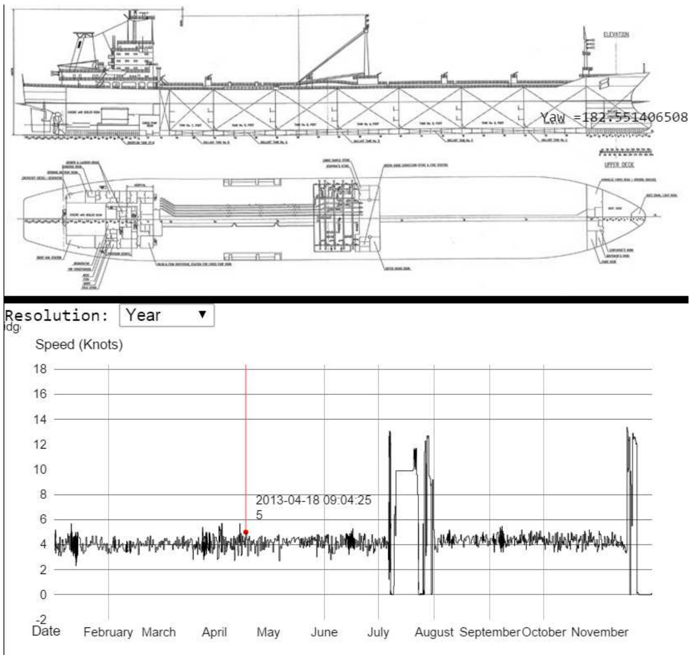

Figure 9. Interactive line charts

Park AS and GCE BLUE Maritime. This research is partially supported by the project "GCE BLUE Maritime Big Data" funded by Innovation Norway and the project "AIS and Maritime Transport" (No. 4506002) funded by MarKom 2020.

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*, ser. LNCS, vol. 4950. Springer, 2008, pp. 154–175.

[2] UN Global Pulse, "Big Data for Development : Challenges & Opportunities," *United Nations*, 2012.

[3] J. Kehrer and H. Hauser, "Visualization and visual analysis of multifaceted scientific data: A survey," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 3, pp. 495–513, 2013.

[4] X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye, "KATARA: Reliable data cleaning with knowledge bases and crowdsourcing," *Proc. VLDB Endow.*, vol. 8, no. 12, pp. 1952–1955, Aug. 2015.

[5] J.-D. Fekete, J. J. Van Wijk, J. T. Stasko, and C. North, "The value of information visualization," in *Information Visualization*, ser. LNCS, vol. 4950. Springer, 2008, pp. 154–175.

[6] H. Wang, O. Osen, G. Li, W. Li, H.-N. Dai, and W. Zeng, "Big data and industrial internet of things for the maritime industry in northwestern norway," in *TENCON 2015: IEEE Region 10 Conference*, 2015.

Figure 10. 3D Animation of vessel motion

[7] H. Wang, S. Fossen, F. Han, and I. A. Hameed, "Data-driven Identification and Analysis of Propeller Ventilation," in *Oceans 2016: MTS/IEEE Oceans Conference*, 2016.

[8] G. Li, B. Kawan, H. Wang, A. Styve, O. L. Osen, and H. Zhang, "Analysis and Modelling of Sensor Data for Ship Motion Prediction," in *Oceans 2016: MTS/IEEE Oceans Conference*, 2016.

[9] J. Dean, *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. John Wiley & Sons, Inc, 2014.

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

[11] G.-D. Sun, Y.-C. Wu, R.-H. Liang, and S.-X. Liu, "A survey of visual analytics techniques and applications: State-of-the-art research and future challenges," *Journal of Computer Science and Technology*, vol. 28, no. 5, pp. 852–867, 2013.

[12] A. Kerren and F. Schreiber, "Toward the role of interaction in visual analytics," in *Proceedings of the Winter Simulation Conference*. Winter Simulation Conference, 2012, p. 420.

[13] M. J. Riveiro, "Visual analytics for maritime anomaly detection," Ph.D. dissertation, Örebro university, 2011.

[14] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.

[15] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," *ACM Trans. Comput. Syst.*, vol. 26, no. 2, pp. 4:1–4:26, Jun. 2008.

[16] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: amazon's highly available key-value store," in *ACM SIGOPS Operating Systems Review*, vol. 41, no. 6. ACM, 2007, pp. 205–220.

[17] K. Banker, *MongoDB in Action*. Greenwich, CT, USA: Manning Publications Co., 2011.

[18] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, "Pig latin: a not-so-foreign language for data processing," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 1099–1110.

[19] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy, "Hive: a warehousing solution over a map-reduce framework," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1626–1629, 2009.

[20] P. Joia, F. V. Paulovich, D. Coimbra, J. A. Cuminato, and L. G. Nonato, "Local affine multidimensional projection," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2563–2571, 2011.

[21] C. Johnson, "Top scientific visualization research problems," *Computer graphics and applications, IEEE*, vol. 24, no. 4, pp. 13–17, 2004.

[22] P. C. Wong and R. D. Bergeron, "30 years of multidimensional multivariate visualization." in *Scientific Visualization*, 1994, pp. 3–33.

[23] R. A. Becker and W. S. Cleveland, "Brushing scatterplots," *Technometrics*, vol. 29, no. 2, pp. 127–142, 1987.

[24] C. Turkay, A. Lundervold, A. J. Lundervold, and H. Hauser, "Representative factor generation for the interactive visual analysis of high-dimensional data," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2621–2630, 2012.