MORENO-GARCIA, C.F. and ELYAN, E. 2019. Digitisation of assets from the oil and gas industry: challenges and opportunities. In Proceedings of 2019 International conference on document analysis and recognition workshops (ICDARW), 22-25 September 2019, Sydney, Australia. Piscataway: IEEE [online], 7, pages 2-5. Available from: https://doi.org/10.1109/ICDARW.2019.60122

Digitisation of assets from the oil and gas industry: challenges and opportunities.

MORENO-GARCIA, C.F., ELYAN, E.

2019

Digitisation of Assets from the Oil & Gas Industry: Challenges and Opportunities

Carlos Francisco Moreno-Garcia
School of Computing Science
and Digital Media
Robert Gordon University
Aberdeen, AB10 7GJ, UK
Email: c.moreno-garcia@rgu.ac.uk

Eyad Elyan
School of Computing Science
and Digital Media
Robert Gordon University
Aberdeen, AB10 7GJ, UK
Email: e.elyan@rgu.ac.uk

***Abstract*—Automated processing and analysis of legacies of printed documents across the Oil & Gas industry provide a unique opportunity and at the same time pose a significant challenge. One particular example is the case of Piping and Instrumentation Diagrams (P&IDs). These are complex engineering drawings that are extensively used in the Oil & Gas industry, which contain critical information for risk assessment, and require highly skilled people to provide an accurate interpretation and analysis of their contents. This paper provides an overview of the P&IDs digitisation problem. We outline the opportunities and key challenges, discuss recent relevant work and state of the art and outline possible future direction to solve the problem. During a two-years collaborative project with an industrial partner from the Oil & Gas sector, we have encountered three main challenges other than traditional inherent image and document related challenges. These are, documents quality, skewed distribution of data and topology. In this paper, we discuss these challenges in depth and survey the main state-of-the art methodologies that may solve them.**

I. Introduction

The conversion of printed information into digital assets has become a priority for the Oil & Gas industry [1]. This problem has been identified in multiple scenarios, such as reading log records and photography analysis, amongst others. In particular, through a two-year collaborative project with Det Norske Veritas Germanischer Lloyd (DNV GL), we have identified a particular problem which is challenging yet interesting; the digitalisation and contextualisation of a class of engineering drawings known as Piping and Instrumentation Diagram (P&ID).

In 2018, we published a comprehensive review on the new trends for the conversion of complex engineering drawings into digital format [1], putting particular emphasis on the case of P&IDs. In this work, we first presented the elemental workflow for P&IDs, which is composed of two steps: digitalisation and contextualisation. By digitalisation we refer to all the functions and methods used to detect, extract the features and classify the three elemental shapes contained in almost every engineering drawing: symbols, text and connectors. In counterpart, by contextualisation we refer to the process of formatting the digitalised data in a way that makes sense to a field expert, such as a petroleum engineer or a risk analyst. We noticed that whilst much literature has been published for digitalisation (the reader is refered to [1]), it is rare that the issue of contextualisation is fully solved. Moreover, we found out that, although there has been a huge advance in deep learning (DL) for image recognition, engineering drawing image analysis has not been really benefited by such advancements.

The aim of this paper is to describe and offer insights into the three main challenges that we have faced during our collaboration with the industrial partner; two in digitalisation (i.e. quality and skewed data distribution challenge) and one in contextualisation (the topology challenge). Although these problems may have been encountered in the past in previous work and for different domains, to our knowledge this is the first time that these problems have been clearly identified. Moreover, we have reviewed literature from different domains to have an insight of a solution to such challenges.

The rest of this paper consists on the following sections. In Section II we discuss the most common challenges found during our collaboration with an industrial partner on the digitalisation and contextualisation of P&IDs, which are quality, skewed data distribution and topology challenges, along with the potential solutions to each of these problems. In Section III, we discuss the conclusions and our insights for future work and collaborations with the scientific community.

II. Challenges

A. The Quality Challenge

Most of the literature related to engineering drawing digitisation has been focused on systems for a particular type of printed representation, such as [2] for telephone manholes from a certain company or [3] for mechanical drawings designed with a certain standard. In the case of P&IDs, the majority of the literature is focused on high quality drawings, where the image is clear in general, text is fully legible by a human expert, symbols are well-defined and connectivity between components is understandable, with all lines orthogonal and perpendicular to the x- and y-axis. This is true for all screend P&ID systems, both in scientific literature [4]–[8] and

[1] https://www.dnvgl.com/oilgas/download/digitalization-in-oil-and-gas-sector.html

in the ones proposed by the industry [2,3,4,5]. Although theses systems are good for a given purpose, they fail to acknowledge that one of the biggest concerns of the industry is to digitalise drawings of poor quality. Figure 1 shows an example provided by our industrial partner of a P&ID with a very poor quality.

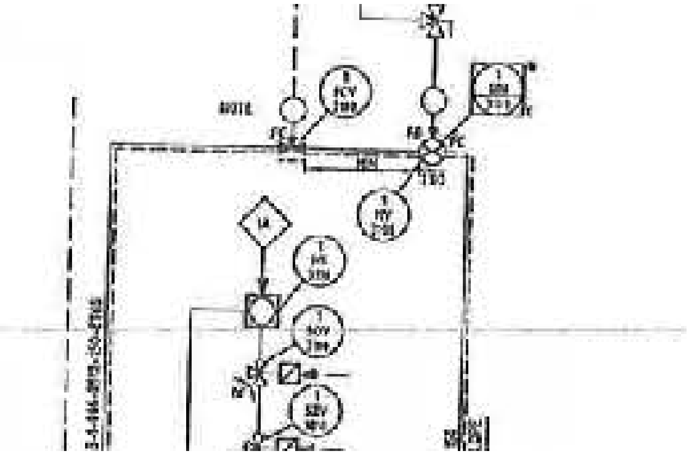

Fig. 1. Sample of a P&ID with poor quality.

This drawing has been printed and scanned multiple times, and thus we can appreciate the following challenges for the digitalisation:

1) The CAD file used to create this drawing has been lost and thus, creating a new copy is not possible.
2) The drawing has been printed and scanned, and thus the lines and symbols are neither orthogonal nor perpendicular to the x- or y-axis.
3) Given that the sheet of paper was folded, we can appreciate a semi-horizontal line of pixels which is aggregated noise.
4) The test is very hard to read. Some of the strings are not legible even for the human eye.
5) Symbols are also very degraded, and there is no key drawing which can aid to determine which symbols are being depicted.

To locate the symbols in this drawing, we consider that the most plausible solution is to train a DL algorithm capable of finding the key shapes with training samples of reduced quality. For this purpose, a viable alternatice to represent the symbols could be the one proposed by Guyomard et al. [9], where authors were capable of detecting drawn symbols in old maps by using low level descriptors on a small-neighbourhood local context. Afterwards, the connectivity can be deducted by analysing the running pixels between each symbol detected. Finally, the text detection is the greatest challenge in this image, as we have noticed that not even a human expert is able to recognise the text. In terms of applicability in the industry, we could rely on experts which manually label symbols in a subset of engineering drawings using any conventional tool for such purpose such as Sloth [6], and then using these symbols to train the neural network. Furthermore, this approach could be replicates to find the text shapes.

B. The Skewed Data Distribution Challenge

Even for drawings with a better quality, we still find issues to classify the shapes have been detected. In particular, symmbols found on P&IDs tend to present the notable characteristic of class imbalance, regardless of the standard. Figure 2 shows the class distribution of the symbols which were detected and manually labelled in a series of P&IDs from a particular standard, which was presented in the work by Elyan et al. [8]. Notice that there is an evident imbalance in the distribution of the symbols obtained.

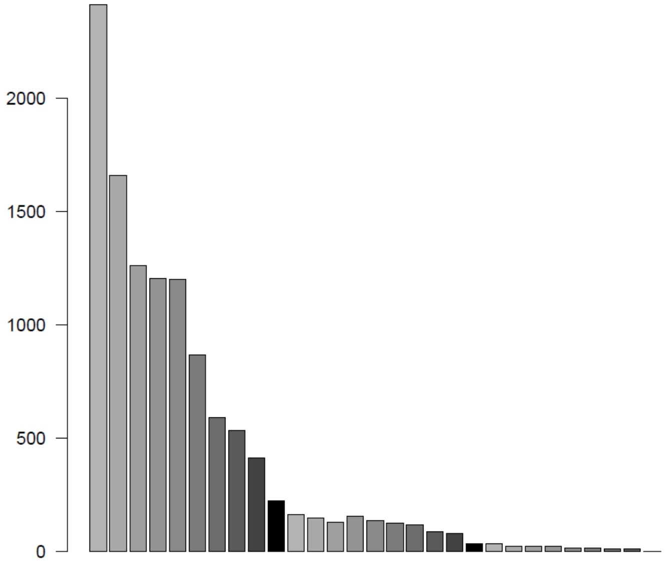

Fig. 2. Example of a typical class distribution of P&ID symbols.

To address this issue, Elyan et al. [8] proposed the use of class decomposition [10] to increase the classification accuracy and to improve sample distribution. The idea is to apply a clustering algorithm (such as k-means, FC-means or DBSCAN) on each class. This is done in an attempt to balance the number of samples per class by properly configuring the number of centroids used on each class. In addition, it is also likely that sub-classes with elements that have more interrelation are generated, and thus a better class distribution is achieved. For instance. It may be possible that when labelling symbols manually, a human expert may consider that an open valve (Figure 3 left) and a closed valve (Figure 3 right) are the same symbol, and therefore numerous valve samples are collected for the class *valve*. By applying class decomposition using any clustering algorithm with a value of $k = 2$, it may be possible

[2] http://www.pidpartscount.com

[3] http://www.radialsg.com/viewport

[4] https://www.rolloos.com/en/solutions/analytics-documents/viewport

[5] https://www.bilfinger.com/en/media/news/bilfinger-introduces-new-technology-for-the-digitalization-of-plant-documents/

[6] https://sloth.readthedocs.io/en/latest/

to automatically generate two sub-classes: $valve_a$ with the open valves and $valve_b$ with the closed valves.

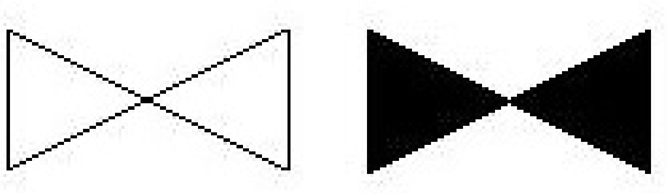

Fig. 3. An example of an open valve (left) and a closed valve (right).

Class decomposition has been proven useful for assessing the majority classes (i.e. the classes where a large majority of samples has been obtained). Nonetheless, the problem of classes with a low number of samples still remains. We have observed that it is common in practice that for a given symbol there is only one sample present in a collection of drawings, and thus the classification is harder to achieve. To increase the number of samples for these minority classes, a generative adversarial network (GAN) specialised on few-shot classification may come as a viable solution [11]. In principle, a GAN is defined as a model where two networks are trained in a way that one acts as the generator and the second one as the discriminator. The first one attempts to generate new samples for a specific class, while the second one intends to distinguish between real and fake samples. By learning this distinction, better fake samples can be generated. A few-shot GAN is capable of performing labelling conditioning even for unlabelled samples. This means that, in spite of a symbol found for the first time, it is still possible to augment the number of samples.

Even when considering all of these factors, there is still a possibility that as the stream of P&IDs to digitise increases in volume, new symbols are contained in these drawings and thus, the likelihood of miss-detection increases. To that aim, Faria et al. [12] proposed MINAS, which is a multi-class learning algorithm which specialises in novelty detection. based on the labelled data set, a decision model is built to try to classify new examples marked as *unknown*. Such unknown examples can be latter used to create new valid patterns which are added to the current model, thus increasing the chance of detecting previously unseen symbols.

Note that the applicability of methods of this nature in the current industrial scenario have been proven in our experience with the industrial partner, as our systems currently make use of these techniques to enhance classification rates.

C. The Topology Challenge

Once all main shapes of the drawing have been digitised, it is important to understand that the conversion of the P&ID into a fully -functional digital format is not finished. As mentioned in Section I, contextualisation is the part of the task that deals with the meaning of the digitised data. The most important requirement of the system that our industrial partner has requested is the understanding of the connectivity of the symbols found on the main pipeline and the identification of each symbol as an *event*. This means that the output of the system should not only be a parts count, but also a structure depicting how all of these symbols are interconnected. This task is already hard by itself given the large amount of intertwined lines that exist in the drawing. Moreover, a parts count would need to comply with a certain format where each symbol corresponds to an area, section and composition within the drawing [1]. In order for a system to properly deduct the event properties, the connectivity of the drawing needs to be fully understood.

Furthermore, another challenge of P&ID contextualisation is related to linking different pages to conform a single structure. It is important to understand that a P&ID page is not a standalone representation, but part of a collection of interconnected pages. A P&ID usually has arrow-like symbols called *continuity labels* which specify to which page is the drawing linked to. Figure 4 shows an example of the problem at hand for a set of eight drawings. Notice that each page aligns in an irregular way with the rest of pages. Moreover, it may be the case that multiple pages link between themselves simultaneously. Considering that a collection of P&IDs describing a single process, rig or plant may require between 100 and 1000 pages, this problem becomes computationally demanding.

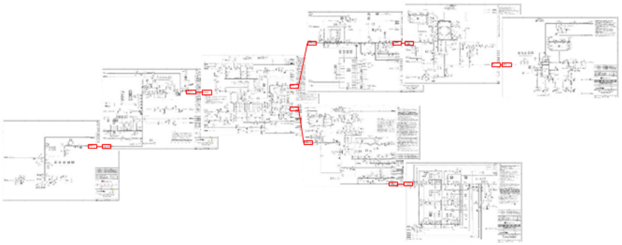

Fig. 4. Linking P&IDs.

The solution for the topology problem may reside in graph representations. Graphs are data structures which consist in a set of attributed nodes and edges. These have been widely used in literature to represent information in diverse domains, such as images, biometrics, chemical compounds and social media interactions, amongst others [13]. In fact, Howie et al. [4] proposed the use of graph representations to store each symbol as a structural graph. In the case of the P&ID as a whole, each page can be represented as a graph, with symbols as nodes and connectors as edges. Moreover, graphs can be directed, meaning that the flow of the pipeline can be inserted as an attribute of an edge. This way, the properties of a certain node can be deducted by considering a *seed* node and inheriting the properties throughout the graph. Graph representations have also proven to be effective to match different type of P&ID representations. Such is the case of Wen et al. [14], who presented a graph matching based system to find the relation between 2D and 3D process plant drawings.

It is interesting to note that in the current industrial context, the representation of assets using graph-based topologies is in use and shows a promising direction. Such is the case of the work presented by Rantala et al [15], who recently presented a system (in collaboration with an industrial partner) to represent process plants and refinery systems from 3D models or spreadsheets in order to apply graph matching between different plants, thus being able to support the reuse of previous designs.

III. Conclusion

In this paper, we have presented the three main challenges that surged from a two-year collaborative project with an industrial partner, in which the main goal is to digitise and contextualise a class of engineering drawings known as P&IDs. We have identified three major challenges: quality, imbalance and topology, and we have discussed some insights on the solutions and areas of work for each of these challenges, while discussing how these techniques can be successfully implemented in the actual industrial context.

The quality challenge is the backbone of the problem, given that a poor quality drawing (e.g. a paper that has been scanned numerous times) is virtually impossible to digitise unless a robust set of preprocessing methods are applied. To that aim, we have proposed the application of low level features to train a robust DL model capable of finding the shapes of interest. Afterwards, some running-pixel algorithms can be applied to deduce the connectivity, while the text identification would pose the greatest challenge, according to the image quality at hand.

In the case of the skewed data distribution problem, it is key to identify in advance the symbols that may appear in the drawing to increase the success of classification. This can be achieved by obtaining the *key* diagram (i.e. the one containing all symbols and names) or by adequately defining the standard of P&ID to digitise. Moreover, it is desirable to handle the imbalance of the samples obtained, such as re-sampling or class decomposition, before training any classifier. Still, it may be the case that new symbols are found as a higher stream of drawings is processed. For these cases, we propose the use of multi-class learning algorithms specialised in novelty detection.

The topology challenge refers to the need to not only digitise the information depicted on the drawing, but also to put this information into context. As most P&IDs are composed of symbols (i.e. nodes), connectors (i.e. edges) and text (i.e. description of the attributes of both), we believe that graph representations are the best suitable candidate to represent and store the drawings to know the components' connectivity, properties and relation within other pages in the same collection. The conversion of the acquired information into a graph is still done manually, and therefore we consider that the priority of this challenge is to design automated systems to create such graphs.

For all of the aforementioned challenges, we conclude that there are insufficient efforts of applying DL-based methodologies, and therefore the huge leap that has been noticed in image recognition area has not been translated to the context of engineering drawing image analysis. We encourage the scientific community to collaborate and provide their insights, in order to bridge this gap in the near future.

Acknowledgement

The authors would like to thank the Oil and Gas Innovation Centre (OGIC), The Data Lab Scotland and Det Norske Veritas Germanischer Lloyd (DNV GL) for supporting this work. In addition, The authors would like to Dr. Brian Bain for his contributions to this work.

References

[1] C. F. Moreno-García, E. Elyan, and C. Jayne, "New trends on digitisation of complex engineering drawings," *Neural Computing and Applications*, pp. 1–18, 2018.

[2] J. F. Arias, R. Kasturi, and A. K. Chhabra, "Efficilent Techniques for Telephone Company Line Drawing Interpretatio," in *Proceedings of the Third IAPR Conference on Document Analysis and Recognition - ICDAR'95*, 1995, pp. 795–798.

[3] P. Vaxiviere and K. Tombre, "Celesstin: CAD Conversion of Mechanical Drawings," *IEEE Computer Magazine*, vol. 25, no. 7, pp. 46–54, 1992.

[4] C. Howie, J. Kunz, T. Binford, T. Chen, and K. H. Law, "Computer Interpretation of Process and Instrumentation Drawings," *Advances in Engineering Software*, vol. 29, no. 7-9, pp. 563–570, 1998.

[5] M. K. Gellaboina and V. G. Venkoparao, "Graphic symbol recognition using auto associative neural network model," in *Proceedings of the 7th International Conference on Advances in Pattern Recognition, ICAPR 2009*, 2009, pp. 297–301.

[6] W. C. Tan, I. M. Chen, and H. K. Tan, "Automated identification of components in raster piping and instrumentation diagram with minimal pre-processing," *IEEE International Conference on Automation Science and Engineering*, vol. November, pp. 1301–1306, 2016.

[7] C. F. Moreno-García, E. Elyan, and C. Jayne, "Heuristics-Based Detection to Improve Text / Graphics Segmentation in Complex Engineering Drawings," in *Engineering Applications of Neural Networks*, vol. CCIS 744, 2017, pp. 87–98.

[8] E. Elyan, C. F. Moreno-García, and C. Jayne, "Symbols classification in engineering drawings," in *International Joint Conference in Neural Networks (IJCNN)*, 2018.

[9] J. Guyomard, N. Thome, M. Cord, and T. Artieres, "Contextual detection of drawn symbols in old maps," *Proceedings - International Conference on Image Processing, ICIP*, pp. 837–840, 2012.

[10] R. Vilalta, M.-K. Achari, and C. Eick, "Class decomposition via clustering: a new framework for low-variance classifiers," *Third IEEE International Conference on Data Mining*, pp. 673–676, 2003.

[11] A. Ali-Gombe, E. Elyan, Y. Savoye, and C. Jayne, "Few-shot classifier GAN," in *International Joint Conference on Neural Networks (IJCNN)*, 2018.

[12] E. Faria, A. C. Ponce de Leon Ferreira Carvalho, and J. Gama, "MINAS: multiclass learning algorithm for novelty detection in data streams," *Data Mining and Knowledge Discovery*, vol. 30, no. 3, pp. 640–680, 2016.

[13] C. F. Moreno-García, X. Cortés, and F. Serratosa, "A Graph Repository for Learning Error-Tolerant Graph Matching," in *Structural, Syntactic, and Statistical Pattern Recognition*, vol. 10029, 2016, pp. 519–529.

[14] R. Wen, W. Tang, and Z. Su, "Topology based 2D engineering drawing and 3D model matching algorithm for process plant," *Graphical Models*, vol. 92, no. C, pp. 1–15, 2017.

[15] M. Rantala, H. Niemistö, T. Karhela, S. Sierla, and V. Vyatkin, "Applying graph matching techniques to enhance reuse of plant design information," *Computers in Industry*, vol. 107, pp. 81–98, 2019.