A Templating Approach to Digitisation of Instrumentation Panel Readouts

A. M. Fagan, G. M. West, S. D. J. McArthur

University of Strathclyde, Glasgow, United Kingdom

andrew.fagan@strath.ac.uk

Abstract

Instrumentation panel readouts are one of many types of paper documents which are still in regular use as part of the operation of nuclear power plants. In order to utilise the information held in these documents, they must be converted into a digital format. This is currently performed manually, but could be improved by automation. For some classes of documents, this problem can be solved easily by the direct application of existing techniques, but in order to begin generalising to many classes of documents without creating a bespoke solution for each, a new approach is required. In this work, we present a case study on digitising instrumentation panel readouts, and show how this template-based method might be generalised to additional classes of document. By taking this approach, a user can easily turn a single well-formed instance of a document into a template which can be matched on to other instances to perform or shortcut the segmentation process. This work shows how this class of document can be digitised from start to finish, and presents a step towards being able to digitise more complex document formats such as engineering drawings, while reducing data entry in the short term.

1. Introduction

As many nuclear power plants have been in continuous operation for several decades, they were originally designed with analogue Instrumentation & Control (I&C) systems. In the course of modernisation, digital systems have been introduced incrementally rather than monolithically, and as such there are many analogue systems still in place, which are limited in the ways they can be interfaced with. As a result, some processes which intuitively seem like they should be simple can often have multiple stages to them. The utilisation of Instrumentation Panel Readouts (IPR) for reactor simulations are a good example of such a process.

Simulations require input from operation plant parameters. In an ideal case of a fully integrated digital I&C system, these parameters might be recorded and transferred to the simulation directly, but instead the best available option is to generate paper printouts of the readout, such as in the example shown in Figure 1. These are generated on site, scanned, collated and then emailed to be have the parameters manually entered into the simulation.

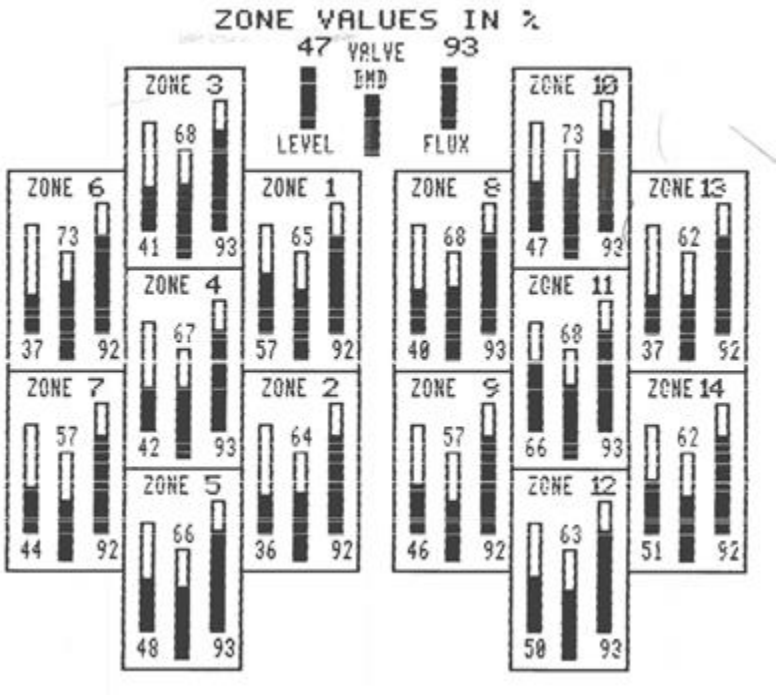

Figure 1 Sample Instrumentation Panel Readout

This process could benefit from refinement, but the early stages of printouts and scanning - which would have the highest impact on the throughput of data - are limited by the analogue systems on which they rely. A refinement then of the manually intensive data entry part of the process would be potentially impactful, but must not compromise the reliability of the data.

Additionally, while this is the output format of one form of IPR, other closely related classes of document, and even other instances of the same document from different sources, have different layouts and combinations of graphics with textual data. It is desirable then to attempt to solve this problem in such a way that the solution devised for this class of document is as transferrable as possible to others that may be of interest, and to maintain a degree of modularity which will allow this solution to be applied elsewhere in the future. By examining the current state of the Document Digitisation landscape, and incorporating a view inspired by Human-in-the-loop Artificial Intelligence systems, a new method for approaching this problem is proposed.

2. Background

2.1 Document Digitisation

Document digitisation is a very mature topic in academic literature; for as long as computer systems have seen industrial use there has been interest in capturing knowledge that is held in paper documents and being able to process it [1]. This refers to the problem of turning an image of a paper document, in part or whole, into useful data which can be processed or traversed by a computer [2]. The advantages of this include the ability to directly pipe data into subsequent data processing applications, such as the simulations mentioned previously, or to allow users to access this data more quickly by indexing or formatting information for ease of traversal. In the

nuclear power industry, some work has already been done on digitisation of specific classes of document, such as civil engineering structural documents [3].

While the terminology has varied, in all applications there are two major steps to the digitisation process [4]. The first stage is a layout analysis or segmentation stage, wherein the document is divided into regions and these regions are classified as being of a specific format, such as text, images, diagrams or other document specific classes.

This is followed by a translation stage, which turns these regions into useful information by the application of more specific subsystems tailored to digitise different data types, with the most obvious example being Optical Character Recognition (OCR) for text. This step of the process is now very advanced, and there is a high level of crossover with other fields of research such as computer vision, as well as a great deal of research being done to build increasingly refined classifiers.

While this step is extremely important, it will not be the focus of this work, which will instead be on making the process of applying these mature techniques to new classes of document quicker and easier by allowing the user to more easily interface with the layout and segmentation of a document.

2.2 Layout Analysis

In some primarily textual documents, layout analysis can be largely ignored. Modern OCR tools such as Tesseract [5] are sufficiently advanced that, as long as the textual and non-textual parts of a document are clearly separated and the text is well formed and oriented correctly, an off-the-shelf solution can be applied directly to them and perform well, due in part to the added benefit of performing textual analysis on the completed corpus and being able to identify words which might have been misread. In many cases however, tables and images can look enough like text to interfere with the output of simple OCR systems. Aside from this, in many cases of digitisation, the layout within the document is directly relevant to the information being extracted, and the removal of such context makes the output less useful than the original document. In cases such as these, the layout analysis step is just as important as the actual translation of information, if not more so.

Layout analysis and segmentation can be approached in two broad ways. The first of these being the careful digitisation of highly variable documents. Examples of these include documents of historical significance, or from fields where mistakes are extremely costly [6], [7]. This is generally a more bespoke process, where many tools are used to analyse a document and create a custom layout for that class of document which can only be applied to very similar instances of the same class.

This family of approaches is slow and inefficient, requiring a user with knowledge of the digitisation process, as well as a working understanding of the document being digitised, to spend time analysing and segmenting each document. It results in a carefully segmented document with few, if any, mistakes being passed to the translation step, but also has very limited capability for transferring work done to similar classes of document, beyond the expertise gained by the user. In addition, the manually intensive nature of this process means that large

amounts of data are not well suited to digitisation, even when documents are extremely similar to one another.

The second approach is to apply intelligent classification tools at the layout analysis. By creating classifiers that detect features characteristic of a region, it is possible to automatically identify them to analyse the layout of a document. This works especially well on extremely generic formats, such as reports and books, where the majority of the document is textual [8]. In these, there is a clear separation of blocks allocated to non-textual data, such as figures or images which are unlikely to need further processing but are instead simply identified and saved. The majority of such documents are in the form of prose text, which can be processed easily. Since such documents are ubiquitous across industries, these tools are extremely transferrable, and can be applied to new, similar classes of documents with relative ease. While the accuracy will not be as high utilising this process, the speed at which documents can be digitised can outweigh this in many circumstances.

This approach is less effective on document classes which are less generic, or which have more of a mixture of textual and graphical data, such as IPRs or Engineering Drawings. Even for these types of document, it is possible to utilise this automatic segmentation approach, but it requires additional steps which increase the complexity of the operation.

Many instances must be analysed to create a generic view of the class of document, specifying what types of region might appear and where those regions might appear in the layout. Automatic identifiers must be devised for each of these types of regions, and while some may be transferrable from other classes of document, some will have to be designed for the new application. This means that for a given class of document, a great deal of initial design work is required, and to transfer this work to a new class means beginning the process again.

To improve this workflow and allow mixed graphical and textual documents to be digitised, we propose a hybrid dynamic templating approach. Using such an approach, a user without detailed knowledge of the digitisation process can make meaningful strides towards creating a template for a new class of document, and get useful results straight away with as little investment of time or design work as is feasible.

2.3 Human-in-the-loop Systems

In applications areas such as the nuclear power industry, the reliability of a system is of paramount concern. Systems whose output cannot be guaranteed to be effective at all times are risky to deploy in fields where the cost of failure is high, and so Artificial Intelligence systems are often supervised by a human, negating much of the benefit in speed and employee-time cost of deploying such a system. This restriction takes on a different light however by viewing the presence of the human supervisor not as an auditor, but as a part of the system. By leveraging the human expertise as part of a feedback loop to continuously improve the AI system. Human-in-the-loop (HITL) systems are those where there is continuous interaction between an AI system and a human operator in order to accomplish a task more quickly or effectively than either could alone[9].

HITL has already been applied to the process of augmenting the design process of complex systems in HELIX[10], a human AI interaction tool which allows for the rapid design and

iteration of AI system design. This process is very similar at this level to the problem of document digitisation, with a user skilled in an application area but not in the technical discipline at hand able to design complex tools using their domain knowledge is an extremely useful idea to this problem. HELIX approaches this by providing the user with graphical tools to assist with visualisation and interaction, and allows all HELIX systems to work to a similar interface for rapid iteration across many domains. These ideas will all be transferrable when considering how the user should be positioned within the proposed templating approach.

3. Dynamic Templating Approach

To digitise IPR and other documents, the high-level structure shown in Figure 2 is proposed. The user generates a simple template, as will be elaborated on in Section 3.1, by using a graphical toolset on a single instance of the class of document they are attempting to digitise. The template amounts to a segmentation of the document into labelled blocks with specific types.

These blocks are used to match the template onto other candidate members of the class using the method explained in Section 3.2, before passing the identified features to type specific translation modules to be converted into useful data while keeping it in the context it originally arose from, allowing it to be utilised for all manner of subsequent applications.

To transfer these ideas to another class of document, the user needs only create a template for the new type of document. Some information, templates, and other developed modules might be transferrable onto the new class, but a new bespoke system design is not required.

3.1 Generation of Templates

To allow a user to generate a template quickly and easily, we present them a view on the document which allows them to interact with it using intuitive graphical tools. The primary method of interaction is to draw a box around a region, then flag the type of that region and give it a name. The coordinates and sizes of regions will be retained, which allows processing to be performed based on a region's relative location to another. Figure 3 shows the same example from Figure 1, marked up as a template in the way described here.

The primary types are textual, which by default feeds into OCR for translation, and graphical, which will be saved as is. Regions can also be denoted as being "anchors", meaning that they are expected to be the same across all instances of the class, and can therefore be used to attempt to map the template to more difficult instances, such as those which have been scanned badly and ended up askew, or are otherwise distorted. In the example of the IPRs, the lines surrounding each zone readout would be useful anchors, as would the title text "ZONE VALUES IN %", as they are presented in the same way on all examples of the class. Any document which does not have features which match with a reasonable degree of certainty onto these anchors is likely to be either severely distorted, or of a different class.

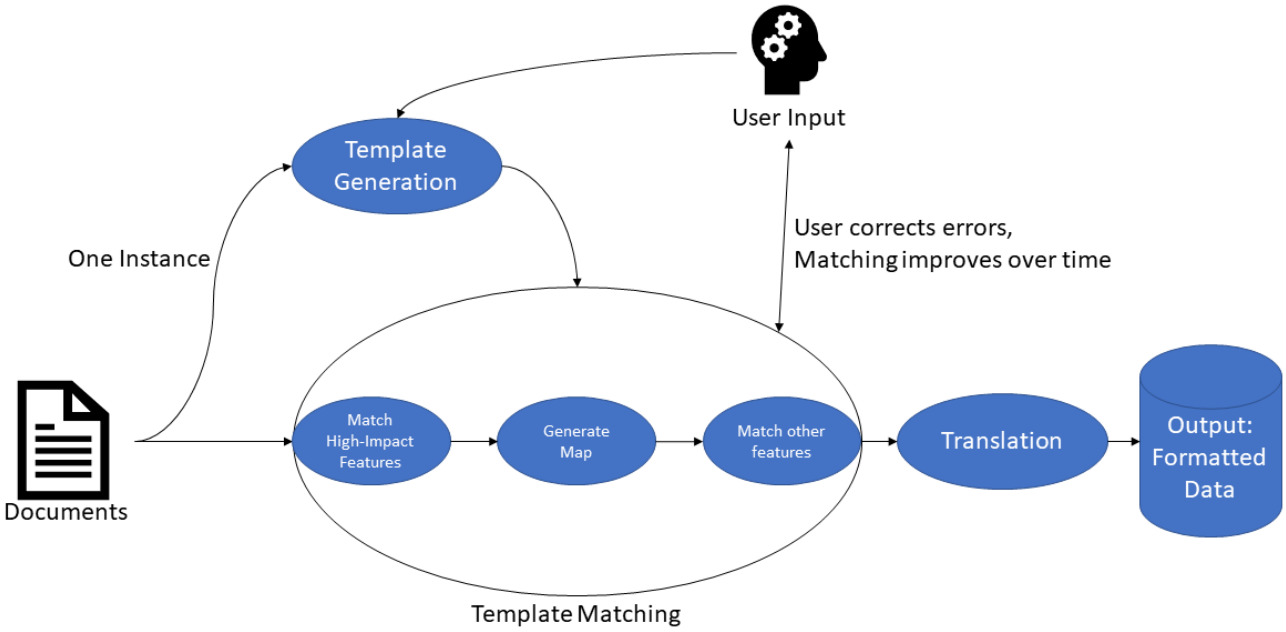

Figure 2 Overview of Templating Approach

The user can denote a region with a custom type name, allowing them to flag a more complex region which may need a bespoke classifier to process. A simple example might be the three vertical bars in each of the zones of the IPR, for which any of a number of simple image processing steps might be taken to return a percentage filled. In cases where a custom region contains critical information, there will be no easy way to circumvent the development of a bespoke matcher, but by allowing it to be flagged for later consideration, and capturing the coordinates within the template, this can be delayed while still partially digitising the document immediately.

It is also possible to denote a region as being a self-contained reproducible unit within the document. In the continuing IPR example, each zone block can be processed using a single template, and the IPR template can consist of 14 of these. This means that in addition to their position within the document as a whole, it is also possible to examine features within a sub-region and compare them against each other. It also makes it easier to create bespoke classifiers for these regions if required, since there will be many more labelled examples of them.

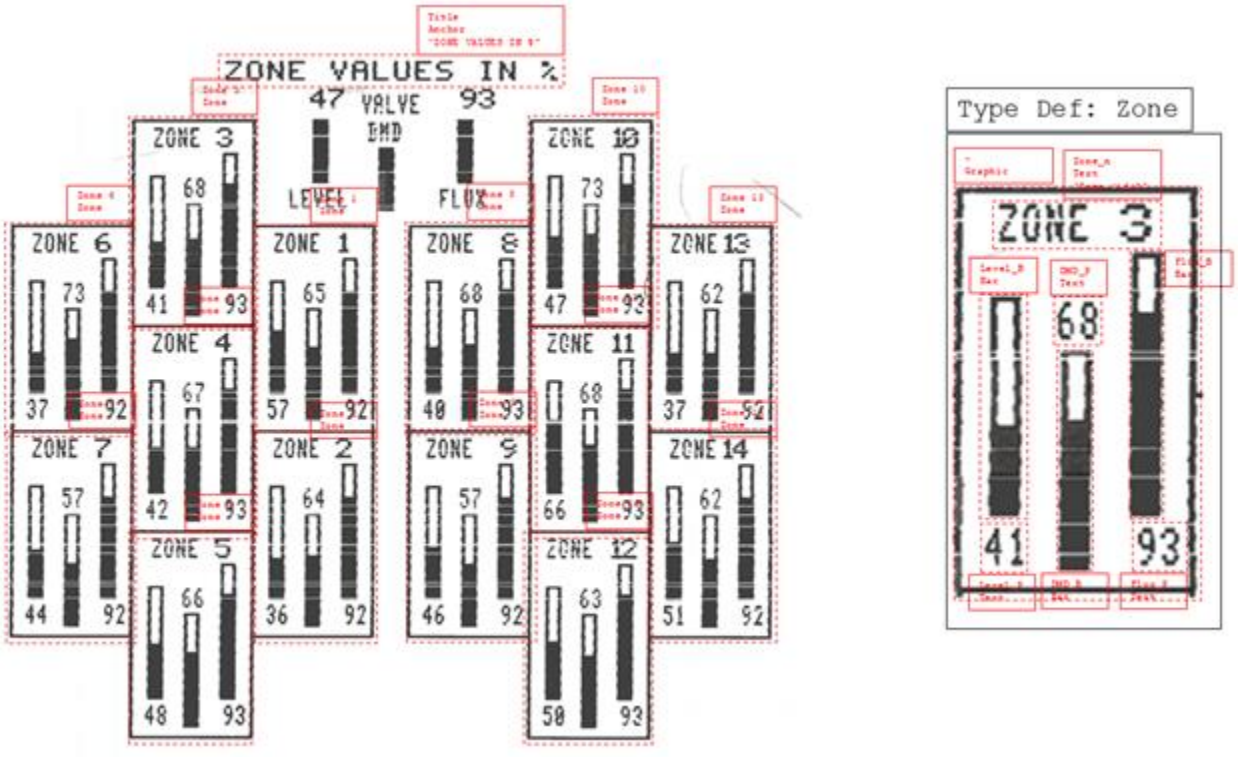

Figure 3 Marked up template

The user should also be able to denote data ranges and relationships between multiple data points within a document in order to allow for cross validation and confidence metrics to be reported on the final digitised output. For example, in the IPR, each number in a zone is a percentage, so if the OCR module outputs something other than a number between 0 and 100, this can be flagged as a likely fault. In addition, each percentage is next to a bar which serves as an approximate graphical representation of it, meaning that if these two figures are not within a small range of each other, it likely means one was misread, lowering confidence in the translation.

3.2 Matching Templates

Once the user has marked up a complete instance of a document, an attempt can be made to match that template onto another instance. Each region flagged during the template generation process will have some method for identifying candidate matches to that region. In cases such as text fields and commonly reoccurring images, these are often the same or similar to their translation step. Many common image processing methods are also excellent at matching instances of the same graphic. These identifiers also have associated statistics on how sensitive they are (how often they activate, regardless of the quality of output), and how accurate they are, recorded from previous attempts to identify the feature. Beginning with the highest accuracy detectable features, which are usually but not always anchors, an attempt is made to match the document onto its template, by searching the approximate region for a feature match, moving gradually out if nothing is found.

In the event that none of the high impact features are identified, they may also be attempted with several skewing, dilation and rotation filters applied to the image. Anchors, being constant between documents, are especially useful at this stage; a distortion which causes an anchor to be identified is likely to be close to the reverse of the distortion on the actual image.

Once several features are identified, their relative positioning is used to create a coordinate map of the matched document relative to the template, which allows features which are less consistent to be identified. A feature, for example with only a 30% accuracy metric but which has been flagged as a match in the correct location is much more likely to be correct than one flagged elsewhere.

In the event that some features still fail to be matched, the document is flagged as requiring user input. By flagging new instances of the unmatched features, and taking the opportunity to approve the matches which were done successfully and flag failures, the feature detectors can consistently improve over time.

Once the document, either by successful matching or by user intervention, is fully marked up, each of the features are passed to the appropriate translation module. Any additional cross validation steps can be performed, asking for user validation in cases where confidence is still low. With this completed, the document has been fully digitised.

This work is still in an early stage of development, and while the template generation and simple matching techniques have performed well as a proof of concept on three document classes with manual intervention, testing of many documents, across more than three classes will require a more fully featured implementation. In particular, testing at scale would require that the system can digitise documents start to finish, with the modules directly passing information between themselves rather than being separate functional units as they currently are.

Since the focus of this work is on segmentation, treating the translational modules as self contained, the actual output of these translation units is not particularly helpful in evaluating the performance of the method. Instead, the most relevant way to verify the performance will be to start by labelling a collection of documents with the class they belong to, and then having the system attempt to match them on to their template, and reporting its success rate. With a large collection of documents across several classes, it should be possible to demonstrate statistics on the overall performance of the system.

4. Conclusions

This work presented a view on a technique which allows a user to begin to digitise new documents, starting from no analysis having been performed, with a particular emphasis on Instrumentation Panel Readouts, of which many are read and manually copied each day. This technique incorporates a hybridisation of manual and intelligent segmentation, together with inspiration from human-in-the-loop AI systems to balance performance, ease of use and scalability. In the short term, this work allows the user to minimise a manually intensive data entry task by allowing them to take steps towards the development of an automatic digitisation solution, without needing to study digitisation at length and take significant development time.

By allowing the user to create a general template, we allow them to quickly and easily undertake what might otherwise be the initial design and analysis step, leveraging their domain knowledge

in a way that might previously have required detailed knowledge capture discussions between user and developer. By keeping the templating system intuitive and graphical, we minimise the effort required for a new user to get started with digitisation, and allow them to get useful results straight away. In many cases, this may immediately result in the desired information being extracted, but even when it does not, the templating structure allows the user to implicitly flag those parts of the process which might require additional development.

In addition, by digitising many document classes over time with the same approach, we maximise the chance that development work done for one class of document might have significant crossover with another, thereby decreasing potential work in the future without additional development time on the current project.

4.1 Future Work and Additional Applications

This work presents a foundation in the template-based digitisation approach that will eventually be applied as a piece of a larger framework for human in the loop document digitisation [11]. Of particular interest going forward will be classes of documents which present significant additional design considerations, such as Engineering Drawings. These complex documents have their own rich libraries of symbols, connections and embedded information which make many document digitisation approaches difficult.

While a great deal of work is done on documents like these already, much of the focus is on the translation step, building better classifiers for individual low-level components of the document [12], while this work focuses more on high level structure and segmentation. Additionally, the existing body of work is fairly specific to individual types of drawing which have rich datasets available, and transferring this work to other classes of drawings is still extremely difficult. This method will present a significant step in deploying digitisation systems which are able to adapt to new classes of document, while being able to leverage existing research for the significant performance advances in translation that have been made already.

5. References

[1] Y. Y. Tang, S. W. Lee, and C. Y. Suen, “Automatic document processing: A survey,” *Pattern Recognit.*, vol. 29, no. 12, pp. 1931–1952, Dec. 1996, doi: 10.1016/S0031-3203(96)00044-1.

[2] X. Lin, “Quality assurance in high volume document Digitization: A survey,” *Proc. - Second Int. Conf. Doc. Image Anal. Libr. DIAL 2006*, vol. 2006, pp. 312–319, 2006, doi: 10.1109/DIAL.2006.33.

[3] N. H. Bui, P. Charles, and H. Blieck, “Nuclear civil engineering towards the simplification and digitalisation,” in *Lecture Notes in Civil Engineering*, vol. 8, Springer, Singapore, 2018, pp. 1134–1141.

[4] G. M. Binmakhashen and S. A. Mahmoud, “Document layout analysis: A comprehensive survey,” *ACM Computing Surveys*, vol. 52, no. 6. ACM PUB27 New York, NY, USA, 16-Oct-2019, doi: 10.1145/3355610.

[5] R. Smith, “An overview of the tesseract OCR engine,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 2, pp. 629–633, 2007, doi: 10.1109/ICDAR.2007.4376991.

[6] W. Ma, H. Zhang, L. Jin, S. Wu, J. Wang, and Y. Wang, “Joint Layout Analysis,

Character Detection and Recognition for Historical Document Digitization," in *Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR*, 2020, vol. 2020-Septe, pp. 31–36, doi: 10.1109/ICFHR2020.2020.00017.

[7] X. Ding, D. Wen, L. Peng, and C. Liu, "Document Digitization Technology and Its Application for Digital Library in China," in *Proceedings - First International Workshop on Document Image Analysis for Libraries - DIAL 2004*, 2004, pp. 46–53, doi: 10.1109/dial.2004.1263236.

[8] X. Zhong, J. Tang, and A. J. Yepes, "PubLayNet: Largest dataset ever for document layout analysis," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 1015–1022, Sep. 2019, doi: 10.1109/ICDAR.2019.00166.

[9] D. Sousa Nunes, P. Zhang, and J. Sa Silva, "A Survey on human-in-The-loop applications towards an internet of all," *IEEE Commun. Surv. Tutorials*, vol. 17, no. 2, pp. 944–965, Apr. 2015, doi: 10.1109/COMST.2015.2398816.

[10] D. Xin, L. Ma, J. Liu, S. Macke, S. Song, and A. Parameswaran, "Accelerating Human-in-the-loop Machine Learning: Challenges and Opportunities," in *Proceedings of the 2nd Workshop on Data Management for End-To-End Machine Learning, DEEM 2018 - In conjunction with the 2018 ACM SIGMOD/PODS Conference*, 2018, doi: 10.1145/3209889.3209897.

[11] A. M. Fagan, G. M. West, and S. D. J. McArthur, "Human-in-the-loop approach to digitisation of engineering drawings," in *Proceedings of the SICSA eXplainable Artifical Intelligence Workshop 2021*, 2021, pp. 48–55.

[12] C. F. Moreno-García, E. Elyan, and C. Jayne, "New trends on digitisation of complex engineering drawings," *Neural Comput. Appl.*, vol. 31, no. 6, pp. 1695–1712, Jun. 2019, doi: 10.1007/s00521-018-3583-1.