# An AI-enabled assembly support system for industrial production

Mona Forsman[1][0000−0001−6601−5190], Benjamin Björklund[1], Henrik Björklund[2][0000−0002−4696−9787], and Johanna Björklund[2][0000−0003−0596−627X]⋆

[1] Codemill AB, Umeå, http://www.codemill.se
[2] Umeå University, Umeå, www.cs.umu.se

**Abstract.** In this paper, we present a prototype assembly support system for hydraulic components, tailored to a mechanical industry production plant. The case study for the project is Komatsu Forest, a leading provider of forestry machines. The system uses multimodal data analysis to understand the assembly process and detect errors. In particular, it uses a TensorFlow network to identify hydraulic components, a projective computer vision model to map a CAD drawing against the partial assembly, and natural language processing techniques to recognise patterns in the non-conformance reports.

**Keywords:** Multimodal analysis · Machine learning · Assembly support

## 1 Introduction

We present a prototype AI system for industrial assembly, intended to support the assembly worker and provide automatic quality control, documentation, and data collection. The system takes as its primary input a video stream of an ongoing assembly, and uses computer vision to (i) match the governing blueprint in the form of a CAD drawing against the construction, and (ii) recognise the components used in the assembly. Based on this information, the system can assist the assembly worker with the selection, positioning, and fastening of components. The system communicates with the user by enriching the input video stream with drawings and assembly instructions, using see-through-augmentation [1], in which successfully mounted components are checked off as work progress. This is an extension of previous work, in which projection-based augmentation allows remote experts to guide local users through complex tasks [3].

The case study for the project is a plant for the production of forestry machines, harvesters and forwarders, operated by Komatsu Forest. A typical harvester uses hydraulics for power transmission to steering, crane, and saw aggregate. Each machine is made to the specifications of the buyer. The customisation is motivated by, for example, the regulatory rules of the local market, potential environmental impact, intended usage, and expected working conditions. It is not uncommon that the machine requires more than 10 000 components to

⋆ We thank ProcessIT Innovations for funding that has made this work possible.

build. Given the sheer number of assembly steps needed to fit these components together, the combined risk of a non-conformance is significant, even if the risk associated with each step is minute.

If a non-conformance is detected in the assembly plant before the machine is finished, then the consequence is typically only an offset in the assembly schedule. If, on the other hand, the non-conformance is not detected until the machine has reached its operation site, then the non-conformance could cause a downtime of several days, as the machine may be stationed in a remote forest area where the access to spare parts and professional mechanics is low. Moreover, in some cases the repair work calls for a disassembly of large parts of the machine, and this is difficult to accomplish in the field.

Since the implementation of a full-scale system was not realistic given the limited time frame, we focused our attention on a number of key challenges, and directed our efforts towards these. Moreover, we restricted our attention to the detailed assembly of the service valve shown in Figure 1. We chose the valve because it is of a manageable size, sufficiently generic not to constitute a trade secret, and requires a varied range of operations to build. The valve is also of practical interest: Depending on the configuration of the machine, approximately 20 hydraulic hoses are connected to this valve using different hydraulic couplings. A fault in the assembly of components similar to this service valve can lead to leakage of hydraulic fluid, which can have a severe impact on the local environment. Moreover, the service valve is placed low in the chassis under the engine, and the repair of an assembly non-conformance is likely to cause an extended downtime and a consequent loss of income for the machine's owner. Finally, since similar components are common in advanced machinery, our results regarding the valve are likely to be transferable to other domains.

Recent surveys of support systems for the manufacturing industry have been compiled in [7, 2]. The use of multimodal analysis has recently been investigate in the case of human-guided robot assembly by Wan et al. [6]. The authors consider a two-phase solution. In the first phase, the human operator demonstrates the assembly, and in the second phase, the robot detects geometrical objects and manipulates these to follow the operator's instruction. To realise accurate 3D vision, the system is guided by augmented-reality markers in the training phase, and point clouds in the execution phase.

## 2 System description

The assembly support system receives its input video streams from two consumer-grade web cameras. A Racer Kiyo (henceforth; Camera 1) with a built-in lighting ring was used for the identification of components, and a Logitech C925E (Camera 2) was used for the collecting the video stream of the assembly and video analysis. The system was implemented in Python 3.6 with OpenCV for video analysis and run on a laptop with Ubuntu 18.04. We trained the neural networks on a dedicated machine with two ASUS RTX2080ti graphics cards. The

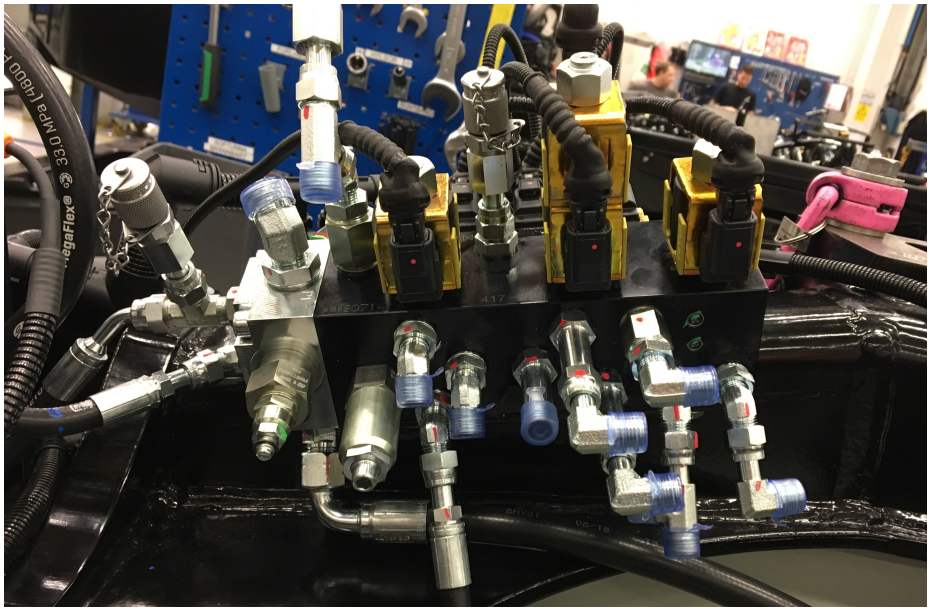

**Fig. 1.** The object in focus for this project, a service valve for hydraulics.

component identification was based on TensorFlow and the pre-trained image recognition network SSD-ResNet50.

The principal functionalities of the system are as follows:

- Identification of hydraulic connectors using a TensorFlow network
- Enumeration of possible mounting points for selected components, indicated with an overlay drawn on top of a video stream.
- Detection of successfully mounted components
- Aggregation of assembly information, primarily image frames, for documentation and future reference

### 2.1 User interface

The support system is designed to guide a human worker, not to control a robot. For this reason, it is important that the worker retains their autonomy and is free to make professional decisions. For example, a fixed workflow planned for a right-handed worker is likely to be inconvenient for a left-handed person. The type of small-series assembly represented by our use case involves a large amount of domain knowledge, and it is important that the workers find their work satisfying and can take pride in their skills. An overly controlling system is likely to cause more irritation than benefit.

The main workflow of the assembly support system is as follows:

1. Initialise the system:

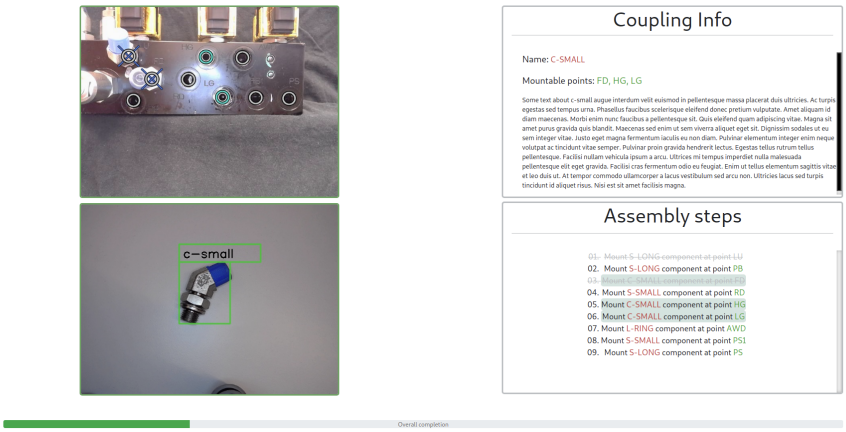


**Fig. 2.** A screenshot of the user interface. Upper left: A video stream of the assembly with suggested positions for the selected component (green circles) and marking of already assembled components (blue crosses). Lower left: Component identification with a separate camera. The component is tagged with the identification made by the neural net. Upper right: Information about the component such as possible assembly positions. Lower right: List view of the assembly stems. The finished ones are stroked out. Possible options for the selected component are marked. At the bottom: A progress bar that shows how much of the assembly is done.

   (a) Check and confirm the orientation of the projected CAD drawing in the video.
2. Main loop:
   (a) Identify component using Camera 1.
   (b) Suggest mount points for the component by augmenting in the video stream from Camera 2.
   (c) Verify the assembly using Camera 2 and give feedback to the user.
   (d) Repeat from 2a until assembly is finished.

### 2.2 Component identification

The prototype system recognises ten different hydraulic couplings; Figure 3 depicts some of these. To identify the components an *object-detecting convolution neural network* was used. During development of the prototype, different models were evaluated based on performance. As previously mentioned this was done with TensorFlow and the different models all used pre-trained weights publicly available online.[3] A data set of annotated images was also created in order to

[3] `https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md`

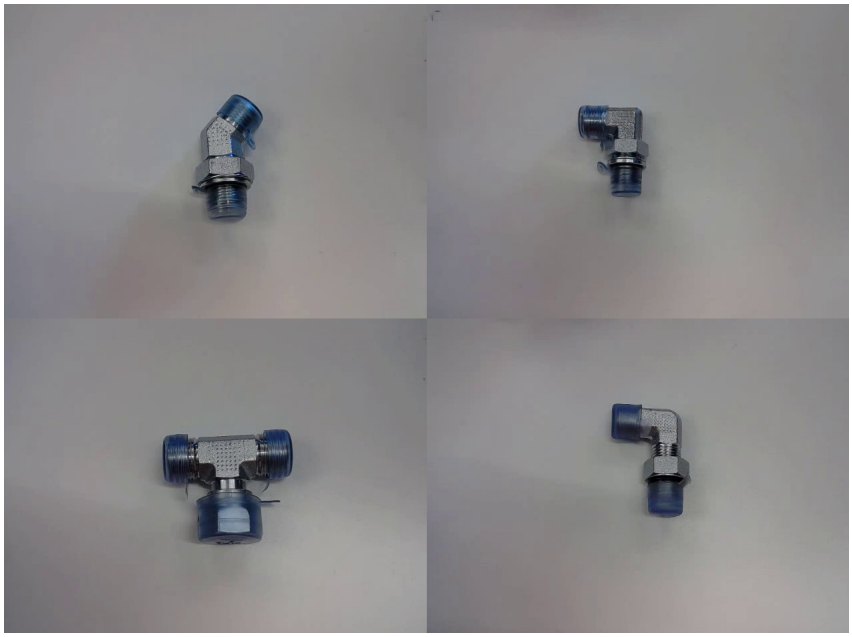

**Fig. 3.** A subset of the hydraulic couplings included in the prototype. Notice how some couplings (the *'L-shaped'* ones in this example) are relatively similar and easily mistaken for each other. Therefore, attention to detail is important and the classifier should be able to detect also small variations in the geometry.

train and test the models. The training examples were created through an automated process, where frames from simple video recordings of the couplings were used to (i) find contours in frame and (ii) create bounding box annotations. The images presented in Figure 3 are examples from the data set without bounding boxes. This automated process gave us a large, but relatively homogeneous, data set available with little effort.

In the end the SSD model with the ResNet50 feature extractor was chosen to be used in this project. Other models tested was Faster-RCNN with varying feature extractors and Yolo v3. Model selection was done by a grid search and cross validation. Specifically 5x5 cross validation was used to measure the *mean average precision* across the following hyperparameters: (i) Batch Size, (ii) Regularization term, (iii) Optimizer function, (iv) Learning rate, and (v) Anchor box aspect ratios. To further increase the stability of trained models, *data augmentations* was also used during training.

### 2.3 Video analysis and augmented video

The video analysis serves to project a CAD drawing in the video stream generated by Camera 2, in order to guid and instruct to the assembly worker. The

same drawing is used to find regions of interest in the video, to detect progress through the planned the assembly steps. A transformation matrix between the drawing coordinate system and the video coordinate system is needed to make the projection possible.

When the system is initialised, the position and the orientation of service valve is detected in the video by searching for circles, which are the holes for the connectors. The Hough circles function in OpenCV is used for this detection. When the correct number of circles has been found in a number of subsequent frames, a transformation matrix from the positions of the connector holes in the drawing to the detected set of circles is calculated. The transformation is evaluated by visual inspection from the user ("Is the model drawing correctly displayed in the video?"). New transformations are tested until the user accepts the transformation. When the transformation is accepted, the assembly work can begin. That step of finding the holes is very light sensitive, and adversely affected by, e.g., reflections – hence the need for the user to verify the transformation.

The system saves reference frame of the service valve as it looked before the assembly. During the assembly process, the CAD drawing is projected onto the video. For the most recently identified component, the possible positions are indicated in the video with green circles. The video frames are analysed in areas of interest around the positions for the couplings. When the subsequent video frames are sufficiently similar (which most probably indicates that there are no hands working in the image), the mean intensity values in the areas of interest are compared to the same areas in the reference frame. If the intensity differs more than a set value, it is assumed that something is mounted in that hole and the hole is crossed out.

## 3 Analysis of non-conformance reports

Within the scope of the project, we have analysed data from Komatsu Forest's system for non-conformance reporting. The aim was to find patterns that can help production management to prioritise among quality assurance measures.

### 3.1 Data

The data set consists of a spreadsheet matrix where each row represents one non-conformance report (ncr). It has 21 577 rows. The 60 columns represent different aspects of the non-conformance in question. The data types for the columns vary, including text strings, category labels, numerical values, and values representing choices from drop-down menus in the graphical user interface.

### 3.2 Delay times

A column of particular interest is 'Störningstider' ('Delay times'), that represent how much delay, measured in minutes, the non-conformance issue has caused. Of the 21 577 ncr:s, 13 878 has a reported delay time. Finding factors that can

**Table 1.** The number of available characters for each column and the average number per ncr.

| | **Characters** | **Average** |
|---|---|---|
| **Describe the nc** | 274436 | 13 |
| **How shall we...** | 389123 | 18 |
| **Detailed descr.** | 1922466 | 89 |
| **Detailed descr. 2** | 1823324 | 85 |

predict long delays would be of great benefit. A problem in the analysis was that some reported values do not seem to conform to the reality of the production process. For example, the value 1 000 000 appears a number of time, corresponding to almost two years. We here choose to disregard all reported values larger than 10 000, leaving 13 770 ncr:s of interest.

We investigated the correlation between a number of other columns and the delay times, but without finding statistically significant correlations. The two exceptions were, unsurprisingly, the two columns 'AvvikelseorsakId' ('Cause for non-conformance') and 'AvvikelsekodId' ('Non-conformance code').

### 3.3 Linguistic data

A number of columns contain text entered by the person who created the ncr. These are the columns 'BeskrivAvvikelsen' ('Describe the non-conformance'), 'HurSkaViLösaAvvikelsenSåAttDenInteÅterkommer' ('How shall we handle the non-conformance so that it is not repeated'), 'DetBeskrivning' ('Detailed description'), and 'DetBeskrivning2' ('Detailed description 2'). The free-text fields are, relatively often left empty or filled in with short comments not suitable for further analysis, e.g., 'ok'. Table 1 shows how many characters in total were available for the four columns and the average number of characters per ncr.

The amount of available data turned out to be a problem. We tried to train models to predict delay times using the textual data. The results, however, were not encouraging. It is hard to determine the exact cause of this, but we conjecture that the amount of data is a major factor, as well as the poor quality of the data.

We also used Latent Dirichlet Allocation topic modelling to try to find patterns in the textes. This method takes the text, or rather the word frequencies in them, as well as a number $k$, and computes $k$ "topics", where each topic is a probability distribution over the vocabulary. The results can only be qualitatively evaluated, but the method showed some results of interest for production management. Table 2 shows the 10 most heavily weighted words for the topics we got for the column 'Describe the non-conformance', when setting $k = 5$. The first topic is clearly concerned with assembly errors. The second one has more emphasis on problems with external deliveries and material. The third topic has almost all the probability mass assigned to the tree words 'inga' ('no'),

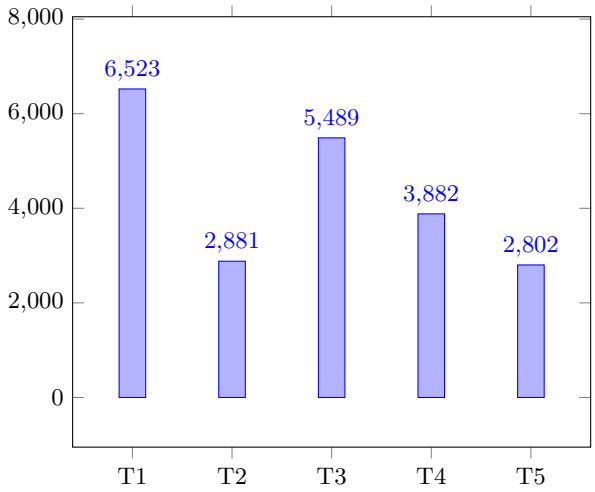


**Fig. 4.** Distribution of ncr:s over the five "topics" mined from the texts from the column 'Describe the non-conformance'.

'avvikelser' ('non-conformancies'), and 'funna' ('found'). The reason for this is that the phrase "no non-conformancies found" is a very common entry. The fourth topic is less unambigouos, while the fifth clearly focuses on damaged parts. Figure 4 shows how the model distributes the ncr:s across the five topics.

Our judgement is that topic modelling has a potential to be useful for finding new patterns in the ncr data. The amount of data available, as well as its quality, is, however crucial. Also, further studies of the usability of topic modelling in this setting would require close cooperation with domain specialists who can assess how valuable the extracted information is to the target organisation.

## 4 Discussion

During the construction of the system prototype, we have earned insights about what is technologically feasible, and a deeper understanding of the requirements related to user experience and system design. It is clear that to make the system successful, it is important to take a user-centric approach. The user should find the system supportive and helpful – if it is perceived as too restrictive, it will become a work environment problem.

In the long term, we think the system would be most easily accessed if integrated as part of a protective head visor, complete with headphones and microphone for natural-language interaction. Preliminary experiments with the Microsoft HoloLens are promising, but the technology is still to immature to be a viable alternative. Especially the weight of the headset and the narrow field of view poses problems [5]. In the long term, retinal displays (which draw a raster display directly onto the retina of the eye) are an interesting alternative [4].

**Table 2.** The ten most heavily weighted words for each of the five topics mined from the column "Describe the non-conformance".

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|
| **maladjusted** / 0.25 | **external** / 0.29 | **none** / 0.26 | **restat** / 0.18 | **damaged** / 0.26 |
| **unmarked** / 0.23 | **supplier** / 0.29 | **deviation** / 0.26 | **assembly** / 0.18 | **fits** / 0.18 |
| **assembled** / 0.19 | **unassembl.** / 0.21 | **found** / 0.26 | **marked** / 0.15 | **skratches** / 0.17 |
| **badly** / 0.06 | **leakage** / 0.13 | **additional** / 0.06 | **filled in** 0.09 | **broken** / 0.14 |
| **tightend** / 0.06 | **missing** / 0.04 | **untightend** / 0.04 | **error** / 0.09 | **version** /0.06 |
| **marked** 0.06 | **layout** / 0.03 | **wrongly sel.** / 0.03 | **o-ring** / 0.05 | **level** / 0.06 |
| **wrongly att.** / 0.05 | **material** / 0.01 | **cleanliness** / 0.03 | **damaged** / 0.04 | **transport dmg** / 0.04 |
| **late** / 0.04 | **return** / 0.01 | **unmarked** / 0.03 | **attached** / 0.04 | **unselected** / 0.03 |
| **according to** / 0.02 | **article** / 0.01 | **own** / 0.03 | **badly** / 0.04 | **wrongly** / 0.02 |
| **specific.** / 0.02 | **waste** / 0.02 | **marked** / 0.02 | **unmarked** / 0.04 | **improved** / 0.02 |

The lighting conditions was a recurring problem throughout the project. We experimented with different light sources to facilitate the video analysis. A diffuse, stable light without shadows in the work area is preferable. A future work station should probably have white screens around with indirect lighting. As remarked by the manufacturing management, this is in line with the general requirements on a well-lit workstation. The white balance also poses a challenge, since we are looking for holes in a black object. A dark background gives a better range for the light values in the image and simplifies the task.

There are many other candidate features that could be included in the assembly support system. For practical use, a feedback system should be implemented: At present, the system assumes that the worker mounts the component most recently shown to Camera 1, and not some other component. A next step could be to support the mounting of components that must be placed at a given angle, by drawing a line with the correct angle on top of the video stream and then verifying the the mounted components direction.

A question raised by manufacturing management is "Who should write the assembly instructions?". It is a recurring problem in specialised assembly that the assembly instructions are of low quality. Sometimes they are given is an exploded view of the object without any directions, and the knowledge of how to assemble certain parts must be memorised by the assembly workers. Our vision is that the system should be integrated to the CAD system, with connection to a component database and a standard procedure database. The construction engineer could then, when designing the CAD model, import component and standard-procedure data. The goal should be that the assembly support system can import a simple wire-frame of the object, together with component information to assemble and standard procedures about how the components should be assembled.

## References

1. Ronald Azuma, Yohan Baillot, Reinhold Behringer, Steven Feiner, Simon Julier, and Blair Macintyre. Recent advances in augmented reality. IEEE comput graphics appl. *Computer Graphics and Applications, IEEE*, 21:34 – 47, 12 2001.
2. E. Bottani and G. Vignali. Augmented reality technology in the manufacturing industry: A review of the last decade. *IISE Transactions*, 51(3):284–310, 2019.
3. Pavel Gurevich, Joel Lanir, and Benjamin Cohen. Design and implementation of teleadvisor: A projection-based augmented reality system for remote collaboration. *Comput. Supported Coop. Work*, 24(6):527–562, December 2015.
4. Miwa Nakanishi and Tomohiro Sato. Application of digital manuals with a retinal imaging display in manufacturing: Behavioral, physiological, and psychological effects on workers. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 25(2):228–238, 2015.
5. Rafael Radkowski and Jarid Ingebrand. Hololens for assembly assistance – a focus group report. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 274–282, 05 2017.
6. W. Wan, F. Lu, Z. Wu, and K. Harada. Teaching robots to do object assembly using multi-modal 3D vision. *Neurocomputing*, 259:85 – 93, 2017.
7. X. Wang, S. K. Ong, and A. Y. C. Nee. A comprehensive survey of augmented reality assembly research. *Advances in Manufacturing*, 4(1):1–22, Mar 2016.