Multimedia Tools and Applications (2023) 82:15875–15901
https://doi.org/10.1007/s11042-022-13803-1

TRACK 4: DIGITAL GAMES, VIRTUAL REALITY, AND AUGMENTED REALITY

Towards achieving a high degree of situational awareness and multimodal interaction with AR and semantic AI in industrial applications

Juan Izquierdo-Domenech[1] · Jordi Linares-Pellicer[1] · Jorge Orta-Lopez[1]

Received: 12 May 2022 / Revised: 29 August 2022 / Accepted: 5 September 2022 /
Published online: 29 September 2022

Abstract

With its various available frameworks and possible devices, augmented reality is a proven useful tool in various industrial processes such as maintenance, repairing, training, reconfiguration, and even monitoring tasks of production lines in large factories. Despite its advantages, augmented reality still does not usually give meaning to the elements it complements, staying in a physical or geometric layer of its environment and without providing information that may be of great interest to industrial operators in carrying out their work. An expert's remote human assistance is becoming an exciting complement in these environments, but this is expensive or even impossible in many cases. This paper shows how a machine learning semantic layer can complement augmented reality solutions in the industry by providing an intelligent layer, sometimes even beyond some expert's skills. This layer, using state-of-the-art models, can provide visual validation and new inputs, natural language interaction, and automatic anomaly detection. All this new level of semantic context can be integrated into almost any current augmented reality system, improving the operator's job with additional contextual information, new multimodal interaction and validation, increasing their work comfort, operational times, and security.

Keywords Augmented reality · Semantics · Deep learning · Industry · CNN · Transformers · Multimodal interaction

Jordi Linares-Pellicer and Jorge Orta-Lopez contributed equally to this work.

✉ Juan Izquierdo-Domenech
juaizdom@upv.es

Jordi Linares-Pellicer
jlinares@dsic.upv.es

Jorge Orta-Lopez
jororlo2@upv.es

[1] Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Camí de Vera, s/n, València, 46022, València, Spain

1 Introduction and related work

The use of augmented reality (AR) and its advantages in industrial settings has been nothing new since the introduction of its possibilities in the field [8]. Different AR solutions are currently successfully applied in nearly any industrial sector in production lines, operation, and work in various industrial environments, maintenance tasks, reconfiguration, and others.

Several authors have already applied it for assembly tasks [25, 29], as a step-by-step guide [33] or maintenance tasks [5, 16]. The main advantage AR provides in these environments is safety and comfort to the operator. Using different AR solutions, industrial operators can be assisted in the diverse maintenance, repair, and control processes through additional synthetic elements anchored on the physical elements.

Nowadays, there are solutions that, in addition to the automatic assistance of traditional AR systems, allow the participation of a real expert to aid the operator in specific tasks remotely. It is especially interesting when the nature of the actions cannot be carried out with current AR solutions alone due to their difficulty, risk, or other issues. In these conditions, the expert can maintain bidirectional oral communication with the operator and create indications about the elements or areas of interest. These indications or synthetic elements are perfectly anchored in the physical environment using an AR device manipulated by the operator. For example, Mourtzis, Siatras, and Angelopoulos use the approach of a remote expert and uses the Microsoft Hololens as the AR device [26]. However, the need for an expert and depending on their availability and cost limits the general use of this type of solution.

The evolution of systems based on Deep Learning (DL) in areas such as vision, image interpretation, and natural language processing (NLP) permits the development of solutions to the necessity for expert assistance in AR environments. DL's new possibilities allow new situational awareness possibilities for the operator. Situational awareness in this context refers to the perception of the elements, their meaning, and the projection of their status in the near future [14]. DL also provides potential users with new possibilities, such as mechanisms for detecting anomalous patterns, a task sometimes beyond the reach of an expert through visual inspection and in real-time. Some examples of the use of Machine Learning (ML) and DL techniques applied to the detection of anomalies in the industrial field can be found in [22] and [42].

The use of architectures such as Convolutional Neural Networks (CNN) can assist the operator in visual validation tasks with capabilities comparable to an expert providing remote assistance. For instance, Lai, Tao, Leu, and Yin use an R-CNN, a network specialized in detecting regions and classifying objects inside these regions, for the detection of tools in developing a multimodal AR system for intelligently aiding in assembly tasks [23]. For this work, the main focus is on different controls distributed over several machines the operator interacts with.

The new opportunities, thanks to the evolution in NLP, by architectures based on transformers such as BERT (Bidirectional Encoder Representations from Transformers), ([37] and [11]), can provide the operator with answers to their questions in a natural language format. These questions can be asked not only to Supervisory Control And Data Acquisition (SCADA) systems, Enterprise Resource Planning (ERP), or Human-machine interface (HMI) but also to extensive technical manuals via Question Answering (QA). For example, Coli, Melluso, Fantoni, and Mazzei use natural language to retrieve information from technical documents through a conversational agent (also known as a Chatbot [10]) based

on MultiWordNet [28], and Yu et al. uses natural language to retrieve answers based on previous questions to the system [39].

The possible detection of anomalies or unusual patterns by integrating multimodal information makes using techniques based on ML and DL conceivable candidates to overcome the limitations of expert assistants when facing significantly complex patterns, where the response speed is essential.

The hybridization of AR with the possibilities offered by image understanding through neural networks, NLP systems, and models for anomaly detection and predictive maintenance allows a semantic AI extension (semantic layer) by providing meaning and identity to the elements of the 3D geometry of the environment (physical layer). Providing meaning and identity to the different elements will allow operators a higher cognitive level of interaction with them. The present work proposes an architecture based on multimodal interaction. Combining DL techniques for image interpretation, NLP, and anomaly detection and using AR as the central axis for integrating these new possibilities makes it feasible to offer great comfort and assistance to operators in industrial environments. A general architecture is presented, and particular solutions are tested in a real production chain.

This article is structured as follows: Section 2 gives a detailed explanation of the proposed architecture, Section 3 explains the followed approach for validating the operators' actions, Section 4 describes how a chatbot can help the operator in retrieving industrial data, Section 5 focuses on the problem of asking questions on technical documentation, Section 6 proposes the usage of AR for indicating the operator the position of an anomaly. Finally, Section 7 explains the application developed to test the proposed architecture, Section 8 evaluates the system in an industrial environment, and Section 9 presents the conclusions.

2 Architecture overview

Since the concept of Industry 4.0 appeared in 2013 [21], the operator's role has been questioned. Process automation and the communication between the different industrial elements represent a radical change and a challenge for those companies that do not have the most modern machines [19]. However, thanks to technologies such as AR, the operator gains protagonism; this happens after going through a process of adaptation and learning, hence being able to give solutions to more complex problems and providing a more decisive role to the company's value chain [18]. Therefore, based on the three key elements that make up a Cyber-physical system (CPS) [12]:

- A physical object, such as a machine or a production line.
- A data model, accessible through the network, for accessing information from that machine.
- A service to allow accessing the data.

This work proposes an architecture that integrates the operator in an automated process through AR and combines technologies of different nature, all ML or DL based, such as NLP to promote a more natural interaction, CNNs to help the operator understand the environment, and ML techniques for anomaly detection.

In Fig. 1, a general overview of the architecture is shown, where it is possible to distinguish four layers that improve the integration and the work of an operator in an industrial plant.

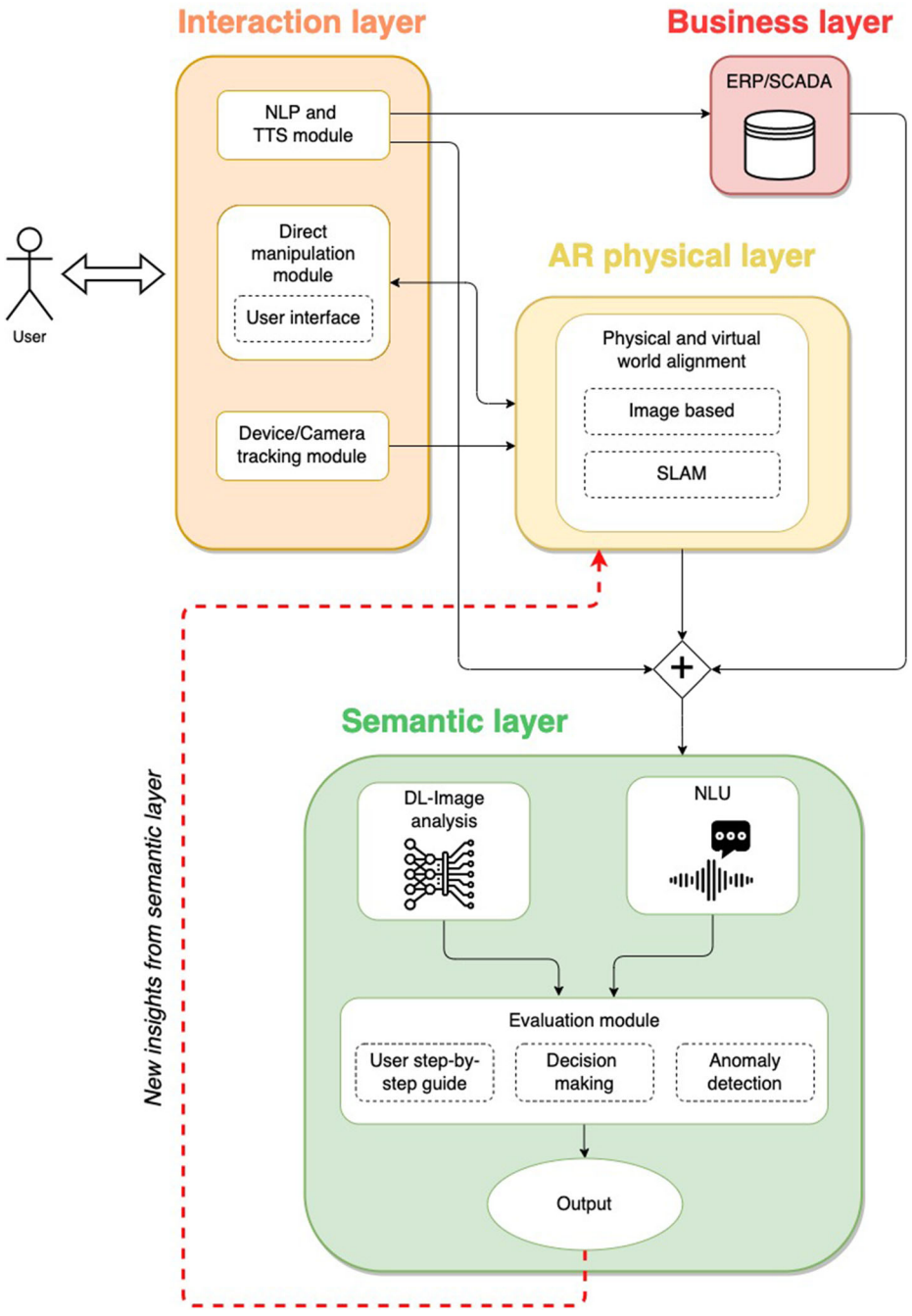


Fig. 1 General architecture overview

The main characteristic of the proposed architecture is to achieve a synergy of the different elements that allow going beyond an isolated use of an AR system, reading and interpretation of values of industrial components through CNNs models, interaction through natural language, and anomaly detection.

The AR system acts as a central hub:

- The AR system shows in context, and anchored to the elements in question, the information obtained by CNNs (i.e., values and states). In turn, the AR system provides context and layout of the controls that simplify the work of the CNNs in providing the necessary elements for a geometric correction (inverse perspective), greatly simplifying their training.
- By obtaining values in a vector of features (from the CNNs and the ERP/SCADA), the results of the anomaly detection model used can also be displayed as visual guides in the AR environment for a better interpretation of the problem by the operator.
- The NLP system also benefits from the feedback provided by the AR system, which allows knowing the operator's location and narrowing down the context of the possible questions asked.

The architecture's different components, characteristics, and interactions will be detailed in the following sections. Although some details will be provided about the implementation used in evaluating this approach, it is worth noting that the system allows the use of different types of components in their different layers and solutions, always maintaining the advantages of their interaction and synergy.

2.1 Interaction layer

In this layer, all the interaction methods and communication possibilities of the operator with the machine are centralized, either through natural language, direct manipulation (e.g., touches on the screen, gesture recognition, and others), or through the camera and other sensors (e.g., LIDAR, RGBD cameras and others) on a mobile device or specific devices such as AR glasses. The camera and other specific sensors will allow the AR system to analyze the environment to understand its location and spatial mapping. The AR physical layer later explored will take care of this detection.

The interaction between the operator and the machine via the AR system is intended to take place in situ because, in this way, the understanding of the context, especially in scenarios in which the main objective is learning the system, is enhanced [17].

Additionally, the so far common user interaction styles in AR-based applications are extended, with three additional elements that allow multimodality:

1. The interaction in natural language
2. The automatic validation actions that arise from obtaining the spatial mapping of the environment, typically found in AR systems
3. The ability to give meaning to the captured elements (i.e., what they are and what their status is) through DL techniques

2.2 AR physical layer

One of the essential layers of the proposed system's architecture is the AR physical layer. This layer ensures that the user's device can understand its environment and superimpose

synthetic information over the real environment. This layer is defined as the standard mechanism in most of the current AR systems and that, in one way or another, allows a spatial mapping of the environment and the augmentation of reality with new synthetic elements to help the operators in their work.

Mobile devices and smart glasses are the most used in the industrial field; and although the focus of this work is on mobile devices, given their cheap availability to most companies, they are not the only devices, and it would be convenient to carry out an evaluation of which device is more suitable according to the context of use [13].

In this layer, it is possible to use any solution based on image tracking [36], surface tracking [34], or even Simultaneous Localization and Mapping (SLAM) techniques, which allow the device to discover its position in an unknown environment, and in real-time [20]. Any of these approximations are valid; even a mixed implementation would be feasible if it allows for improving the positioning of the device in space and the geometrical understanding of the environment.

The implemented solution uses the two AR techniques that provide the necessary elements for the semantic layer: image-based tracking and SLAM. The SLAM possibility is convenient in cases when the industrial panel or machine is not unique or not easily distinguishable based on its image. In both cases, these two techniques provide the necessary elements to facilitate the development of the semantic layer:

- To determine the operator's position, allowing the generation of helpful context information in the semantic layer.
- To provide the necessary parameters to apply a geometric correction to the captured images that simplify the training of the CNNs and maximize their accuracy.

The AR physical layer is the main input element of the semantic level, which, as it will be discussed, will give additional meaning to the elements of the environment in order to improve the performance of current systems. The images captured by the AR system need a perspective correction before going to the semantic level to facilitate their subsequent analysis by a neural network (e.g., operators are not necessarily facing an orthogonal position in front of the machine due to some obstacles). This problem is solved by applying the inverse of the geometric perspective transformation, which is feasible from the information provided by standard AR systems. This stage is described in Fig. 2 as the last image adaptation before the semantic layer.

Also, Fig. 2 shows that the Optical value extraction module corresponds to the sequence of steps necessary for the correct training of a neural network, either for the classification or regression of possible values from an image; in this case, the different controls of interest. The AR system can detect which machine or element the operator is facing, which allows a preliminary knowledge and location of which controls may be interesting to analyze using a neural network to obtain their possible status and other values.

As can be seen in Fig. 2, to read the images captured by the device, the system relies on two elements:

- Data augmentation
- Geometric transformation

Our solution for understanding images is based on using CNN architectures. These neural networks are widely used and allow image classification (e.g., if a switch is on/off)

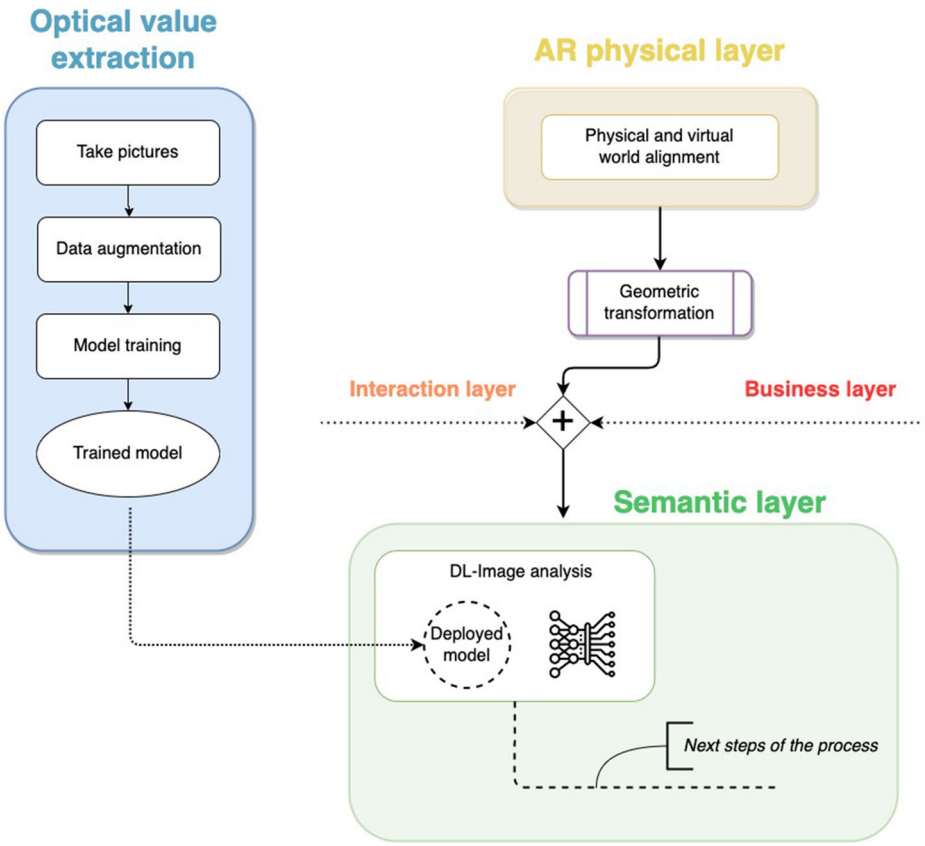


Fig. 2 Detailed diagram of the physical and semantic layers

and regression (e.g., obtaining a specific value from the image of an analog control with continuous values).

In the case at hand, and after the perspective correction of the original image, CNNs with a straightforward architecture to obtain good results and metrics are the only requirement, without needing more complex CNNs or transfer learning. It is essential to use image augmentation techniques to generate a set of variations that allow the CNN a correct generalization and subsequent good prediction metrics for each control. In particular, the synthetic generation of variations is based on rotations, zooms, noise, contrast, and lighting changes. These alterations are essential for the correct detection of the element to be interpreted.

The perspective correction and image augmentation process greatly simplify the necessary preliminary work in preparing the required images of the different controls in the training of the CNNs. In the tests, it has simply been necessary to capture a single image per control and state, and in the case of analog controls, several pictures with the range of possible values between the two extremes. An image augmentation process (e.g., rotations, zooms, noise, contrast, and lighting) generates the required datasets to give accurate final results.

2.3 Semantic layer

The semantic layer of the proposed architecture couples the information received by the previous layers to extract relevant information to transfer to the operator. This interaction will be given using the AR system and its inherent benefits.

For this, the information from the AR physical layer and the already trained CNNs are used for the analysis and extraction of meaning from the visual information, being able to read the value of one or several analog or discrete controls, as can be seen in Fig. 8.

The operator also benefits from this semantic layer, given the possibility of interacting in natural language. Chatbots and NLP advanced QA systems can work together with any visual information captured by the CNNs or other real sensors. The visual identification of an element can provide valuable context for possible queries the operator can send to an ERP system, as observed from Listing 1. Furthermore, it is also possible to ask specific questions about technical documentation, as seen in Table 1. This synergy with visuals, sensory, and natural language interaction will be described in further detail later.

2.4 Business layer

Today, most industrial plants are partially or fully sensorized and adapted through ERP, SCADA, or HMI control systems; however, access to this information usually requires the operator to move to a computer or an HMI system, which might be inconvenient when accessing the information is periodic or urgent. For this reason, and based on the three key points listed above for a CPS, this data access service can be derived so that the user can make requests in natural language to any device used in the AR solution, such as a mobile device or some smart glasses.

One of the most significant benefits of this approach is the relief of the operator from having to learn specific commands or actions in complex menus. This common way of interacting with SCADA or ERP systems requires essential training time; otherwise, they are only within reach of experts. In the proposed approach, the queries the operator wants to make are given in natural language, a very intuitive way of interacting that reduces the learning time compared to more traditional approaches. It should be noted that this approach requires the post-processing of the information to translate the requests into the language or query expected by the system as it will be described.

3 Visual interpretation and validation

The definition of a *generic model* for reading, interpreting, and extracting values or states from images of industrial controls is still a challenge to be solved due to the great variety of elements used in industrial environments, their different features, models, ranges, scales, and manufacturers; however, the use of the information from the AR system regarding the location and spatial layout of the control to be interpreted significantly facilitates the necessary training in the most advanced techniques based on CNNs (i.e., geometric correction using the inverse of the perspective transformation).

Focusing on Fig. 2, in this work, the use of simple CNN architectures for the interpretation of values based on images captured by the device is proposed, as can be seen in Fig. 3, where the architecture is capable of interpreting the value of analog controls. The potential of this approach lies in its combination with the AR physical layer.

```
{
  "text": "Rollers in stock?",
  "intents": [
    {
      "id": "1606940483084759",
      "name": "get_stock",
      "confidence": 0.9984
    }
  ],
  "entities": {
    "element:element": [
      {
        "id": "1087430018514134",
        "name": "element",
        "role": "element",
        "start": 0,
        "end": 7,
        "body": "Rollers",
        "confidence": 0.9995,
        "entities": [],
        "value": "rollers",
        "type": "value"
      }
    ]
  }
}
```


Listing 1 The question is "Rollers in stock?" with an intent of getting the stock of a specific item, identified by the entity "rollers". These elements can be easily translated to a formal query to an ERP

As has already been mentioned, AR can be used for many tasks such as product design [27], process control [40], and maintenance and training tasks [16]. Suppose the opportunities offered by understanding these images are added on top of these functionalities. In that case, it is possible to obtain systems that conduct the operator in a much more intelligent and safe way through the tasks that make up a process, reduce errors, and even increase security and comfort in tasks with a high-risk component [6]. In this way, it is possible to develop a virtual expert able to help the operators.

In the experiments carried out, different CNN architectures for classification and regression tasks based on the images captured by the device are used. In Fig. 4, it is possible to detect the state of a button (i.e., on/off) and ensure that the button is in the correct state before continuing with any operator's task; and in Fig. 5, the system can interpret the value through an analog control that uses a pointer to indicate the current pressure value. In the case of Fig. 5, the instrument is a pressure gauge that allows measuring the pressure of fluids contained in a closed container. Regardless of whether the operator knows if a pressure value is appropriate or not, the semantic layer can interpret and communicate that information to the operator through elements in AR.

The plainness of the architecture used for this regression problem can be analyzed in Fig. 12 in Appendix, a simplified CNN based on [1] that gives great precision in estimating

Table 1 Using a transformer to ask questions in NL in technical documentation

Question	Score	Predictions
What type of screw should I use to set the limit indicator?	0.624583	flat head screwdriver
What is the next step after setting the limit indicator?	0.255615	replace the cover
How do I replace the cover?	0.015868	aligning the cutout in the cover to the groove
In which direction do I have to turn the cover?	0.953776	clockwise
How many millimeters do I turn the cover?	0.627026	6 to 7mm
What screwdriver width do I need?	0.698608	2.9mm
How do I decrease the press?	0.979576	counterclockwise
What color is the case cover?	0.952443	black
What is the first step for assembling the cover ring?	0.681499	remove the small screw (1 position) from the pressure gauge
What is the second step for assembling the cover ring?	0.431261	place the cover ring on the pressure gauge
What is the last step for assembling the cover ring?	0.219696	remove the small screw (1 position) from the pressure gauge

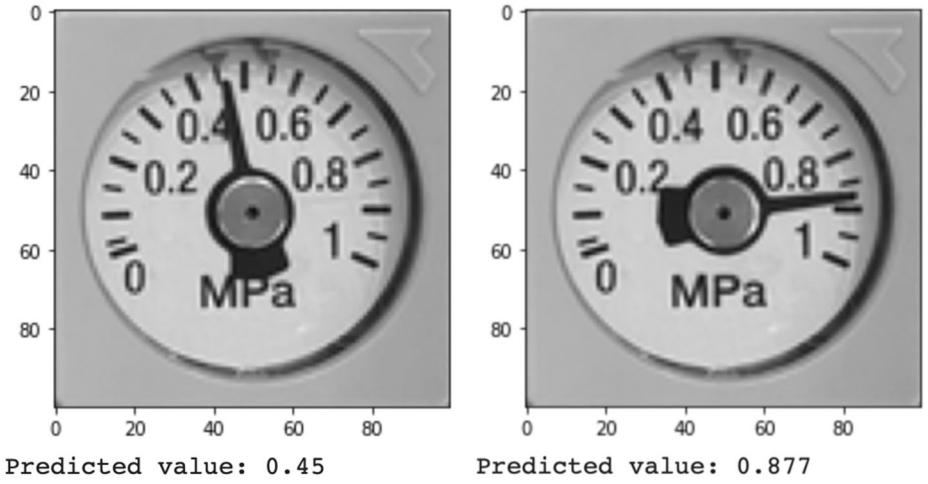


Fig. 3 Using a CNN with regression to interpret the values of a pressure gauge

the value from the control image, with a regression coefficient close to 0.95, with Nadam optimizer and around 100 epochs with mean squared error loss function.

When testing discrete elements, the architecture shown in Fig. 12 in Appendix gives accuracy, precision, and recall values close to 1 in the tests. Again, the Nadam optimizer was used with less than 100 epochs.

The previous knowledge of the position of each control, thanks to the AR system that allows knowing with certainty the machine the operator is working with and the perspective correction, are fundamental elements in the great precision obtained by the CNNs and an important simplification of the training process.

This simplification is achieved thanks to the perspective correction that can be calculated from the internal parameters of the location of the elements in the real world and their relative position with respect to the operator. This allows, starting from only one image per

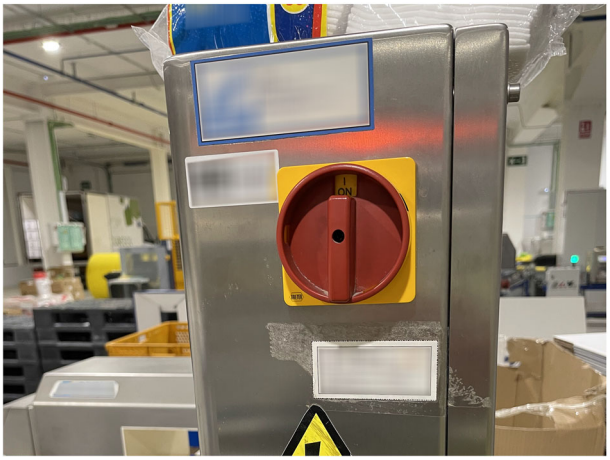


Fig. 4 Classification example

Fig. 5 Regression example

state, to apply image augmentation techniques that only consider lighting variations, small rotations, and zooms. In the case of the discrete control of two states, on/off, two images have been used, of which 1000 variations have been generated with image augmentation of each one, using 1500 as training and 500 as test. In the case of analog control, 25 images of intermediate positions of the analog gauge have been used, which have generated 1000 images each with image augmentation, with again 75% for training and 25% for testing.

4 NLP using chatbots

Chatbots are Natural Language Understanding (NLU) platforms that make designing and integrating a conversational user interface easy and help aid the operator's daily tasks [10]. With rule-based grammar and ML matching, chatbots detect the intents and entities from the input utterances. It is convenient to use rule-based grammar with few examples and ML matching when many examples are available for better accuracy. The chatbot must be trained using a collection of examples or utterances, where the user manually labels a collection of intents and entities. After some examples, the chatbot can accomplish the intent and entity recognition with high accuracy and be further trained with real questions after deployment. Intents and values are generally returned in a JSON format that can be easily converted into a formal query to a database, ERP, or SCADA system. An example of how to get the remaining stock about a specific item in the facility is shown in Listing 1 with the AR solution facing the example in Fig. 6. The flexibility of this approach enables the possibility of making the same query/intent for different elements/entities.

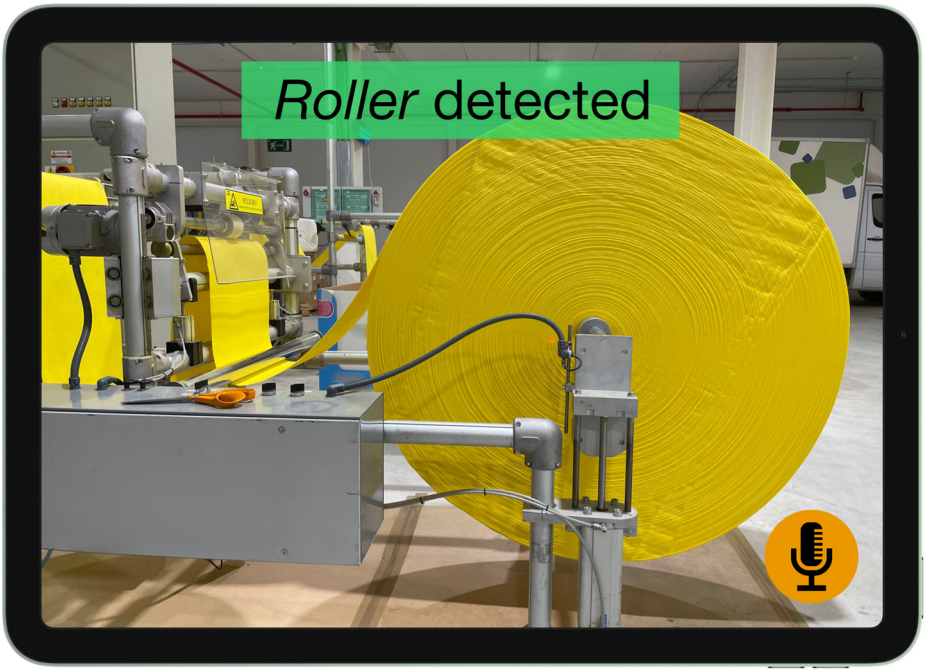


Fig. 6 The operator can ask questions in natural language about this machine. The AR system gives information regarding what element is the operator in front of, so the question is complemented with the required context

One of the additional benefits of using chatbots is that the use of natural language not only favors interaction more intuitively and naturally with the interface but also helps the integration of staff with functional diversity [2]. In general, the semantic elements that assist the operator described in this proposal can facilitate the inclusion of operators with functional diversity in new tasks that were previously out of their possibilities.

Many tools permit the implementation of chatbots easily. It is possible to use cloud services such as Dialogflow or Wit.ai, although using tools like Rasa is also possible if an independent local server-based system is planned.

5 NLP using transformers with questions and answers

The substitution of an expert in all their functions implies the assistance through the perception of the environment for interpreting visual controls, the validation of the operator's actions, and the possibility of answering possible questions of technical nature.

Traditionally, obtaining additional information relevant to an operator's work is either through an HMI or queries to SCADA or ERP systems. The operator can interact and obtain relevant information by interacting with menus and screens that, perhaps, are far away from the element's position to be consulted. Direct interaction with an expert can significantly facilitate this task, but it does not eliminate the eventual translation of the operator to other areas where the elements they can use to retrieve the information are located. Experiments have been conducted to evaluate the possibility of generating a virtual expert, as seen in [4].

Recent NLP technologies involve a new step in the capability to receive questions in natural language that can be converted into queries to databases or SCADA/ERP systems, as has already been mentioned in Sections 1 and 4.

Apart from providing this possibility, the new capabilities derived from transformers are explored in the present work. After an unsupervised training process with large corpora, these recent neural network architectures are capable of various high-level NLP functionalities, such as text classification, chatbot generation, or text summarization. Some of the most widely used architectures today are RoBERTa [24], DistilBERT [32], and Google's T5 [30]. Specifically, the current work has explored transformers' use in resolving QA tasks on technical manuals.

The lack of need for the operator to consult paper technical manuals during their activities saves them valuable time. Not having to carry this information with him or move to another part of the facility to consult is a new step to provide a high degree of assistance on traditional AR systems.

Although training transformers from scratch using a corpus of specific technical documents is a possibility, it is typical to use pre-trained transformers. Pre-training is the first step of transfer learning in which a model is trained on a self-supervised task on vast amounts of unlabeled data. The model is then fine-tuned on smaller labeled datasets specialized on specific tasks, resulting in a more significant performance than simply training on the small, labeled datasets without pre-training. In this case, the tests were done with pre-trained transformers with a final fine-tuning process to improve the results in QA, and their metrics were finally evaluated with SQuAD (Stanford Question Answering Dataset) [31].

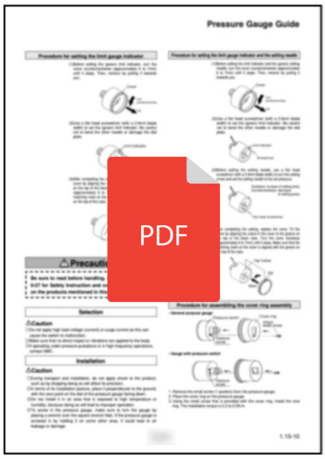
Different architectures have been explored in this respect, choosing to use an extractive open QA (the answers come strictly from the context) Intel/bert-large-uncased-squadv1.1-sparse-80-1x4-block-pruneofa [41] for the experiments (with an f1-score of 91.174 on SQuADv1.1). Some significant tests have been carried out on this model to validate the possibilities of this new interaction. Examples of these tests can be seen in Table 1.

Suppose technical manuals are available in natural language and with due length and depth in their explanations. In that case, current transformers can respond in natural language to many problems that, even if they need to be solved in natural language, can compete not only in speed of response but also in precision with the operator or expert using technical documentation. Figure 7 shows a brief view of some of the answers/predictions provided by the transformer, whose context of the search for answers is the technical documentation for the assembly and adjustment of a pressure gauge.

The results are promising, but the accuracy of the responses is highly variable. This possible variability depends not only on the architecture of the chosen transformer but also on its pre-training process (i.e., main corpus) and fine-tuning (i.e., adjustments for QA). Considering these aspects, it is also essential that the technical manuals themselves, their length and clarity in the explanations, and the characteristics of the questions asked, have greater weight in the accuracy of the possible answers.

The final model's accuracy metrics, capable of answering questions from the operator in front of a technical document describing different processes related to a device or machine, can only be evaluated in a specific context. If there are some manuals, a set of questions, and the answers obtained by the model, the only way to evaluate the model's adequateness is by comparing its responses to the ones given by humans [31].

Again, highlight that, even with the limited experimentation, the results and advantages of using these transformers architectures in front of challenges such as QA of manuals are



- Question: *How many millimeters do I turn the cover?*
- Answer: 6 to 7mm

- Question: *What is the next step after setting the limit indicator?*
- Answer: replace the cover

- Question: *What is the first step for assembling the cover ring?*
- Answer: remove the small screw (1 position) from the pressure gauge

...

Fig. 7 Some examples of real questions using a manual of a pressure gauge

inarguable, particularly when facing decisions that require a quick response and taking into account the extra benefits of integrating this technology into an AR solution.

For questions about the information contained in SCADA systems, ERP, and others, the implementation is even easier to achieve since the only need is to perform preliminary training on a chatbot architecture, as discussed in Section 4.

6 ML for anomaly detection

In the architecture exposed in this paper, the utmost effort is to complement the operator's knowledge, assist their work, and even replace the need for a remote expert.

It is evident that having the assistance of a remote expert integrated into an AR solution is an element of great value, hardly replaceable in its entirety, but it is the purpose of the present work to make use of human assistance only in very justified cases.

There are scenarios where some problems may arise that neither an operator nor a remote expert can solve within a limited time. It is the case of having to detect some complex anomalies that are challenging to see (i.e., when they result from a combination of different values from different sources).

In the scheme proposed in the current work, information from the sensors and other information available in real-time is combined, plus a set of values that can be obtained through CNNs from the image coming from the AR system. Many values may need to be summarized into a feature vector required to train an anomaly identification ML system. There are many and very diverse possibilities depending on the anomalies' characteristics (e.g., point, contextual, or collective) [9]. Not in a few cases, the complexity of these anomalous patterns can escape the most experienced operator or expert and allow, for example, for efficient predictive maintenance (e.g., stopping the production process when an imminent problem is suspected), risk reduction, and operators' integrity, production outside of standardized values and possible defective products, among others.

The synergy of the proposed solution is based on the combination of sensorized information captured from neural networks, its union with ML techniques for detecting anomalies, detecting possible problems, locating these problems spatially, and giving convenient indications in AR to the operators. Therefore, it is not only about identifying possible anomalies but benefiting from the AR by pointing them in the physical environment.

In the presented example in Fig. 8, different unsupervised classification algorithms have been tested for anomaly detection. Some examples have been Isolation Forest [35] or K-Means [3], with very positive results; however, what is beneficial about the architecture is not only the speed of detecting the problem in a potentially complex situation, even for an expert but also the AR-based feedback, which would allow operators to focus their attention right on the spot where the problem lies.

7 General multimodal AR approach

As a consequence of the elements proposed in the suggested architecture from Fig. 1, the final solution achieves a multimodal interaction with AR as the articulating axis, managing to go beyond the traditional possibilities of interaction in AR. In this way, the operator obtains a set of possibilities in maintenance, repair, or reconfiguration tasks, similar to those with the assistance of an expert.

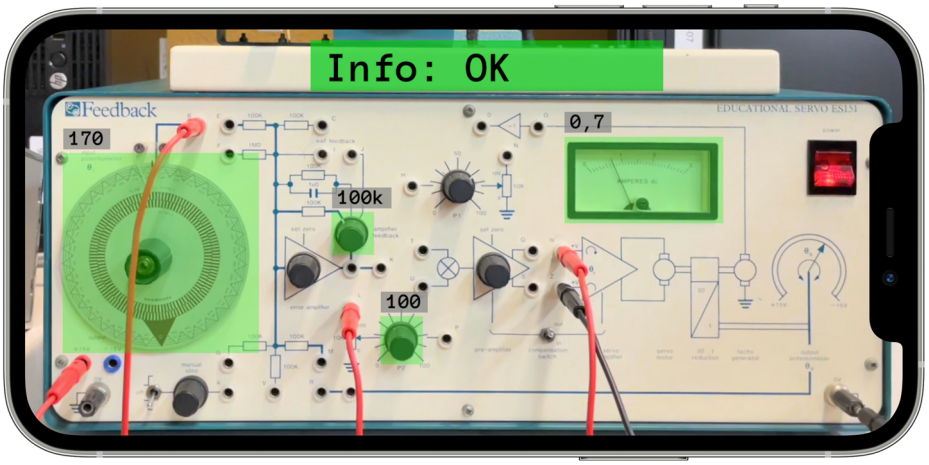


Fig. 8 The combination of several values can be seen as standard or as an anomaly, and visual cues are possible in AR

The operators' workflow is enhanced, not only by the usual interactions AR systems are capable of but also with two new possibilities:

- CNN-based visual validation is carried out reactively by the operator. Suppose that in a specific action, the application receives the positive validation of the CNN (e.g., a specific value in an analog control by regression or the specific position or state of a switch in classification). The application can automatically move on and invite the operator to perform another action from a list of maintenance or reconfiguration tasks.
- Translation of natural language sentences into specific queries to ERP, SCADA systems, and questions to technical documentation and operations manuals with transformer architectures.

All proactive or reactive interactions and their responses are duly transformed into synthetic information of interest to the operator and anchored through the physical layer of the AR on the elements involved. Figure 8 is a real example of the testing process where information of interest to the operator about the factors involved is signaled at all times.

All the tests have been carried out successfully on a real production line and using, in this case, a tablet mobile device; however, as mentioned before, the use of specific AR devices such as smart glasses is also possible.

Figures 9 and 10 show two of these tests in which it has been possible to evaluate the multimodal nature of the solution and the ability to provide solutions and obtain answers in real-time without assistance from a remote expert. Specifically, the steps followed in the sample application are:

1. The operator launches the application, and the AR physical layer determines its position in front of the device or machine.
2. The AR solution invites the operator to perform a specific operation, for example, activating a device such as the switch from Fig. 9. After the operator's action and a perspective correction, the control image is sent to a CNN to classify if its state is on or off, and the result authorizes or not the operator to continue with the next step.
3. In some processes, a specific value may be required in some non-sensorized control, as in the case of the pressure gauge in Fig. 10. If a particular value needs to be reached to continue the task, the AR physical layer is used to lead the operator's focus. In this case, the regression CNN reads the values of the analog control in real-time and permits appropriate decisions to be made.
4. All the values of interest coming either from sensors or visually captured by the different CNNs are sent to anomaly detection ML systems, in this case, using K-Means or Isolation Forest clustering. Again, any anomaly is displayed to the operator in its physical context using the AR layer.

It is necessary to emphasize that the operator can ask questions in natural language to the system during any of the steps mentioned above.

8 Experimental setup and evaluation

The evaluation of the proposed method has been carried out in a company's facilities. The company has a large factory with numerous production lines covering a broad and diverse set of final products. This fact has facilitated the selection of a group of operators with these two characteristics:

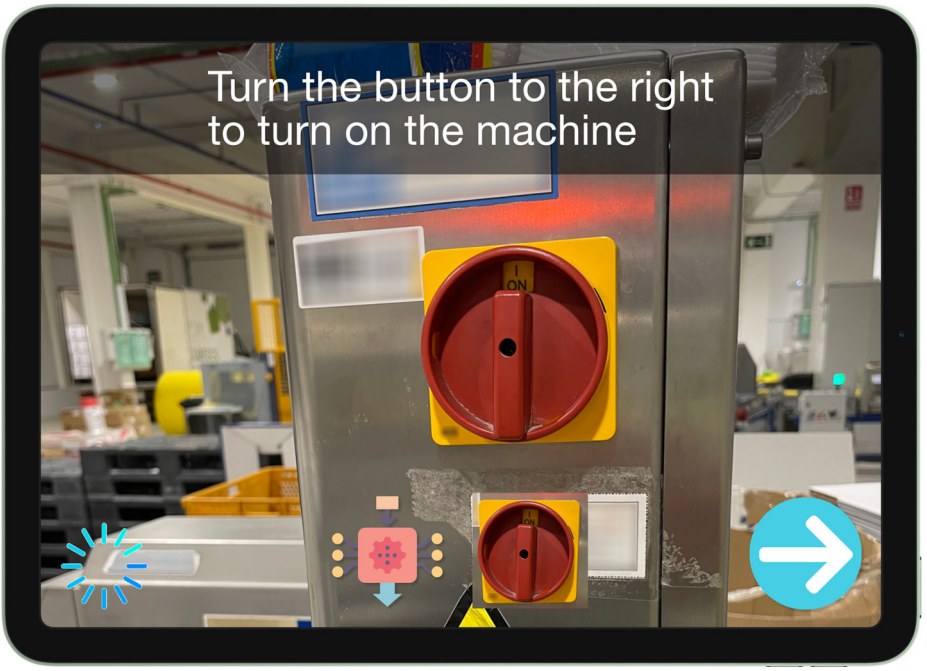


Fig. 9 When the machine is switched on, the app lets the operator move to the next step

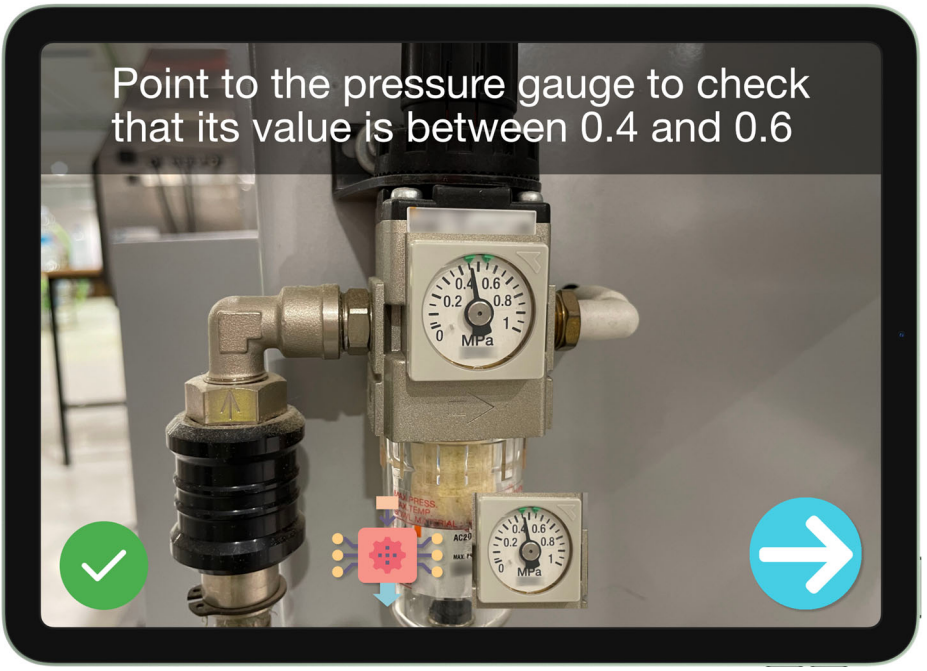


Fig. 10 Automatic value extraction from a pressure gauge

1. The operators already have experience in the work and management of production lines.
2. None of the operators have worked directly on the production line or machines used in this evaluation.

In this way, it has been possible to have eight highly skilled operators who are not directly acquainted with the specific processes to evaluate. This aspect has allowed the division of the operators into two groups to evaluate the advantages of the presented elements.

The additional elements of the proposed enhanced AR system with a semantic layer are evaluated, not the inherent advantages of current AR systems. The workers who operated the system had no previous experience using AR technology. Given the inexperience of the operators with this type of technology, the evaluation of the usability of the system through the System Usability Scale (SUS) [7] has been discarded due to the possible influence that AR could induce on usability in the first use of the technology. AR is used in many fields of industry, and its opportunities and benefits have already been extensively evaluated and demonstrated ([6, 15], and [38]). Therefore, evaluating the times in achieving the proposed tasks is sufficient to determine the benefits of the presented approach.

The evaluated applications, used by each group independently, belong to these two types:

- **Group 1** *(AR standard application)*: An AR application with a series of steps indicates the elements the operator has to interact with. The operator also has technical manuals and access to a terminal to consult an ERP.
- **Group 2** *(AR application with semantic layer)*: An AR application with the elements presented in this work, specifically:
  - visual validation of user actions
  - obtaining values automatically from visual elements
  - voice interaction in natural language to manuals or questions to the ERP
  - a layer of additional anomaly detection

  The system integrates the visual location of a possible incident through the AR physical layer.

The evaluated process has focused on the elements of Figs. 6, 8, 9, and 10. This process consists of the following set of tasks:

Task 1 *Machine activation* (Fig. 9). In group 1, only the step to be carried out is indicated, and the control in the AR environment is highlighted. In group 2, additional validation is performed to check that the machine has been activated effectively, and the new task is automatically triggered when 'on' is visually detected.
Task 2 *Reaching a certain pressure value* (Fig. 10). In group 1, only the control to be monitored is indicated, with the operator checking that the indicator reaches the expected value by direct visual inspection. In group 2, the semantic layer (i.e., a regression CNN) automatically checks that the level has been exceeded, and the AR application automatically notifies the operator.
Task 3 *Tolerable pressure value margin query* (Fig. 10). Group 1 must conduct this consultation on the technical manuals (i.e., on a mobile device). Group 2 can launch this query through a question in natural language by voice.

Task 4 *Stock query* (Fig. 6). Group 1 must make the query in a terminal. Group 2 uses voice interaction to make the query in natural language (i.e., the command is converted from the chatbot response data into an SQL statement).

Task 5 *Anomaly detection* (Fig. 8). For simplicity, a device not directly related to the production line has been used, but it is suitable for evaluating the operators' skills when faced with this type of problem. Group 1 is informed about two combinations considered anomalous by four controls, two analog and two with discrete values. An anomaly occurs when the analog needle exceeds a threshold but only with a particular combination of the other three controls. Group 2 does not know when the anomaly occurs and must only operate with the device and wait for possible automatic detection of the anomaly. Both groups are invited to manipulate the only three possible controls, and changes are artificially induced on the analog control so that the two groups can face high control values with combinations considered either anomalous or permissible.

In task 1, as expected, all the operators of both groups operate correctly, but the shift to the next task is carried out automatically in group 2, which implies a shorter final time in the task since, in group 1, the operators must press the 'next' button after completing their action. The times can be seen in Table 2.

In task 2, the operator's reaction time is assessed when a certain threshold is exceeded in the analog control. Reaching a specific pressure value may depend on other factors unrelated to the experiment. As expected, the reaction times are similar, given that the operators in group 1 were aware of the expected value. However, the greater security provided by having a semantic layer that automatically validates and warns of this situation is evident. In addition, when one of these situations occurs, the AR system can indicate to the operator the control or element that requires their attention to detect a specific circumstance. After the experiment, the operators of group 1 agreed on the clear advantages of having the automatic validation of group 2.

In task 3, the time differences are very notable. Consulting technical documentation takes much longer than formulating a question in natural language and receiving the answer in voice and natural language. In this case, it is essential to note the possible inconveniences when faced with a question erroneously interpreted or answered by the transformer. It was

Table 2 Task completion times without and with semantic layer

	Group 1 (No semantics)				Group 2 (With semantics)			
	Worker 1	Worker 2	Worker 3	Worker 4	Worker 5	Worker 6	Worker 7	Worker 8
Task 1	13s	10s	14s	20s	11s	12s	16s	14s
Task 2	3s	1s	1s	3s	1s	2s	4s	1s
Task 3	46s	123s	32s	43s	5s	8s	* 18s	5s
Task 4	23s	16s	19s	21s	8s	5s	6s	5s
Task 5	98s	110s	76s	134s	2s	3s	1s	1s

*The operator re-phrased its question to be correctly understood by the semantic layer

Workers [1-4] belong to group 1, and workers [5-8] to group 2

necessary to repeat the question on only one occasion when obtaining an incoherent answer in the tests carried out. Even in this case, the final time was less than the average time spent in direct consultations on the technical documentation, accessible through an external terminal near the operated machine.

In task 4, group 1 has a nearby terminal to perform the query. In this way, the time of interacting with the ERP to check the existence of stock of a particular production line component is evaluated. Group 1 times correspond to those of operators familiar with the query tools and the necessary navigation in the corresponding menus; yet, their times in obtaining the answer are much higher than simply asking a question and getting the response through the chatbot used, as performed by group 2.

Finally, in task 5, group 1 took much longer to consider the anomaly as having occurred than group 2, whose interaction is reactive in front of automatic detection by the system. After detecting the anomaly, group 2 times are the minimum associated with a visual and audible signal reaction. The calculated time is the difference between the time the anomaly occurs and how long it takes for the operator to realize it.

Table 2 shows that the improvements obtained by complementing the AR system with the semantic layer and the new NLP possibilities are more than significant.

Table 3 and Fig. 11 show the result of the ANOVA test of two factors with repeated measures in one of them to determine if the effect of the group influences the execution time of the tasks. The result shows a statistically significant difference between the groups, regardless of the task. However, the interaction between the group and the task was substantial, so its execution time depends on the group that performs them. Thus, in tasks 1 and 2, no differences were observed between the operators of groups 1 and 2, while in tasks 3, 4, and 5, the execution times of the operators of group 2 were significantly lower than those of group 1 ($p = 0.039$, $p < 0.001$ and $p < 0.001$, respectively).

We can conclude that adding the semantic layer proposed in this work reduces the completion time of specific tasks. Even though time reduction is not significant in tasks 1 and 2, where the cognitive load given by the nature of the task is low, the semantic layer can be a helpful assistant when the operator needs more guidance. In tasks 3, 4, and 5, as the complexity of the task grows, we can observe that the distance between groups 1 and 2 increases significantly.

9 Conclusions

AR is becoming a central axis in many processes requiring interaction with the physical environment, which can benefit from various assistance processes. Even though the evolution of associated AR device technologies does not yet reach all the demanding requirements for their use in any domain, it is evident that it is already possible to improve many industrial processes in maintenance, repair, and others.

In a preliminary stage, AR focused on solving problems such as spatial mapping, 3D registration, and the anchoring and alignment of synthetic elements with real elements. This technology provides precise instructions on the elements on which to act, minimizing errors and risks.

However, this AR physical layer can be complemented to solve a new range of problems. Presently, some AR systems complement their features with the possibility of incorporating

Table 3 Descriptive and statistical contrasts

	Task					Tests within-subjects effects	
	1	2	3	4	5	Group	Group*Task
	Mean (Sd)	*Mean (Sd)*	*Mean (Sd)*	*Mean (Sd)*	*Mean (Sd)*	F(4;24); p-value (η^2)	F(4;24); p-value (η^2)
Time						14.32; p< 0.001(0.705)	16.42; $p < 0.001$(0.732)
Group 1	14.25 (4.19)	2.00 (1.15)	61.00 (41.77)	19.75 (2.99)	104.50 (24.19)		
Group 2	13.25 (2.22)	2.00 (1.41)	9.00 (6.16)	6.00 (1.41)	1.75 (0.96)		

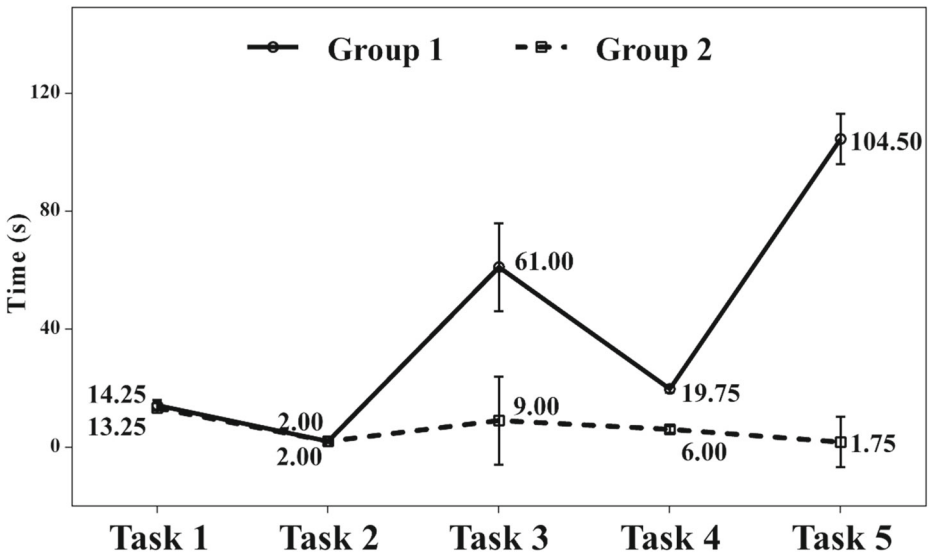


Fig. 11 Tasks comparison box plot

a remote expert capable of visualizing the remote work environment, making annotations and anchoring synthetic elements on the operators' display, and communicating with the operator in the event of unexpected, complex problems, with risk or a high degree of uncertainty.

Many of these possibilities provided by a remote expert can be solved by adding a semantic layer. The evolution of neural networks and their different architectures and opportunities allow that, in an AR environment, the device itself can retrieve 'meaning' from the environment, such as reading states or values from non-sensorized controls. It is also possible to validate the operators' actions (e.g., checking that the operator has activated or not a specific switch before moving on to the next step). On the other hand, the evolution of NLP techniques, Chatbots, and new architectures based on transformers allow the operator to access valuable context information in natural language. The responses can also be returned in natural form to comprehend better the actions carried out.

ML anomaly detection techniques can go beyond the problems or situations that can be solved using a real expert. ML-based anomaly detection techniques can accelerate and determine errors or risk situations, problems, or irregularities in scenarios with a large amount of information from sensors and images retrieved by AR devices.

This paper presents a general scheme of how this new semantic layer, based on visual interpretation and NLP techniques that complement the AR physical layer, gives many responses to changing situations, risk, high uncertainty, and challenging answers in real-time.

Finally, an example has been presented and evaluated, with promising results yielded from adding these layers to current AR systems in industrial environments.

Appendix : CNN architectures

The following diagram represents the architecture for both deep neural networks.:

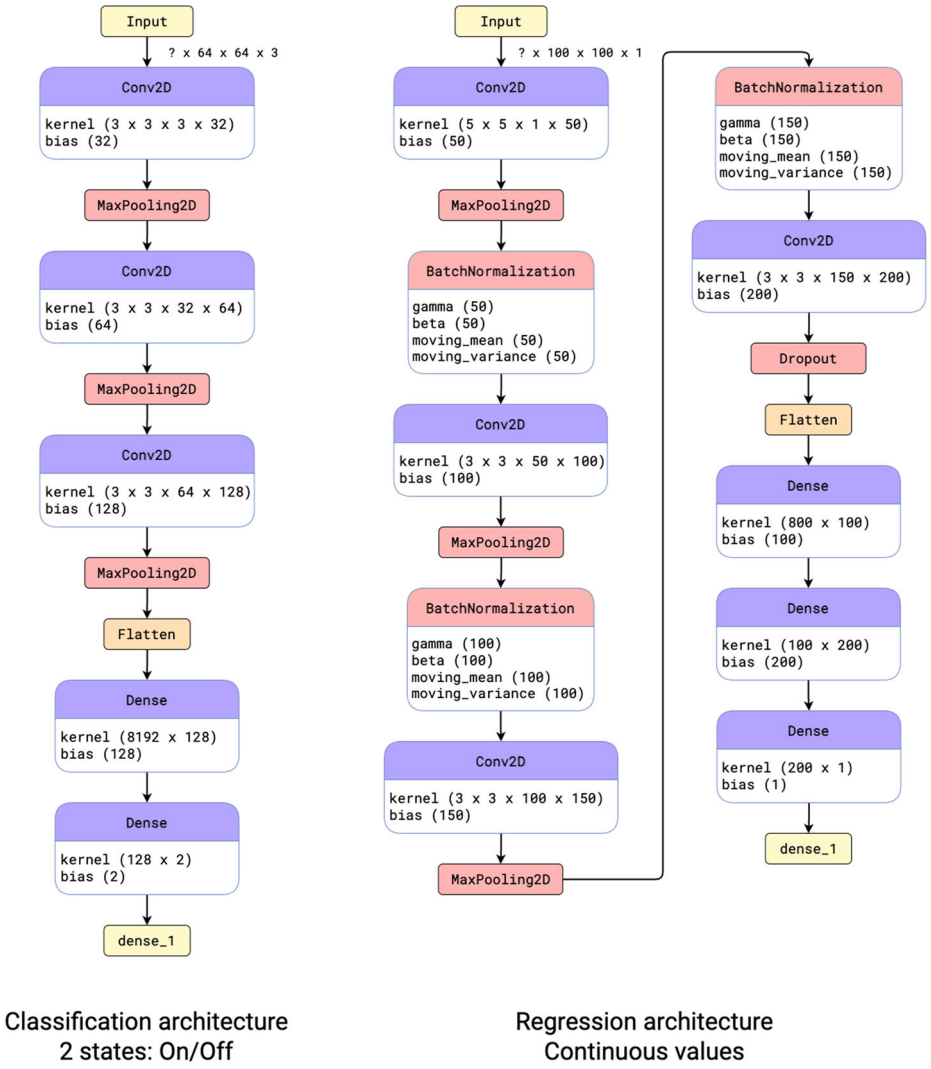


Fig. 12 Classification and regression CNN architectures

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of Interests The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Alexeev A, Kukharev G, Matveev Y, Matveev A (2020) A highly efficient neural network solution for automated detection of pointer meters with different analog scales operating in different conditions. Mathematics 8(7):1–12. https://doi.org/10.3390/math8071104
2. Baldauf M, Bösch R, Frei C, Hautle F, Jenny M (2018) Exploring requirements and opportunities of conversational user interfaces for the cognitively impaired. In: MobileHCI 2018 - beyond mobile: the next 20 years - 20th international conference on human-computer interaction with mobile devices and services, conference proceedings adjunct. Assoc Comput Mach, Inc, pp 119–126
3. Ball GH, Hall DJ (1965) ISODATA, a novel method of data analysis and pattern classification. Technical report NTIS AD 699616. Stanford Research Institute, Stanford, CA
4. Barakonyi I, Psik T, Schmalstieg D (2004) Agents that talk and hit back: animated agents in augmented reality. In: ISMAR 2004: proceedings of the third IEEE and ACM international symposium on mixed and augmented reality, pp 141–150
5. Benbelkacem S, Belhocine M, Bellarbi A, Zenati-Henda N, Tadjine M (2013) Augmented reality for photovoltaic pumping systems maintenance tasks. Renew Energy 55:428–437. https://doi.org/10.1016/j.renene.2012.12.043
6. Bottani E, Vignali G (2019) Augmented reality technology in the manufacturing industry: a review of the last decade. IISE Trans 51(3):284–310. https://doi.org/10.1080/24725854.2018.1493244
7. Brooke J (1996) SUS - a quick and dirty usability scale. Usability Eval Ind 3:3
8. Caudell TP, Mizell DW (1992) Augmented reality: an application of heads-up display technology to manual manufacturing processes. In: Proceedings of the Twenty-Fifth Hawaii international conference on system sciences, pp 659–669. http://ieeexplore.ieee.org/document/183317/
9. Chalapathy R, Chawla S (2019) Deep learning for anomaly detection: a survey, 1–50. https://doi.org/10.48550/arXiv.1901.03407
10. Coli E, Melluso N, Fantoni G, Mazzei D (2020) Towards automatic building of human-machine conversational system to support maintenance processes. https://doi.org/10.48550/arXiv.2005.06517
11. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805
12. Drath R, Horch A (2014) Industrie 4.0: hit or hype? IEEE Ind Electron Mag 8(2):56–58. https://doi.org/10.1109/MIE.2014.2312079
13. Elia V, Gnoni MG, Lanzilotto A (2016) Evaluating the application of augmented reality devices in manufacturing from a process point of view: an AHP based model. Expert Syst Appl 63:187–197. https://doi.org/10.1016/j.eswa.2016.07.006
14. Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. Hum Factors 37(1):32–64. https://doi.org/10.1518/001872095779049543

15. Fraga-Lamas P, Fernández-Caramés TM, Blanco-Novoa O, Vilar-Montesinos MA (2018) A review on industrial augmented reality systems for the industry 4.0 shipyard. IEEE Access 6:13358–13375. https://doi.org/10.1109/ACCESS.2018.2808326
16. Garza LE, Pantoja G, Ramírez P, Ramírez H, Rodríguez N, González E, Quintal R, Pérez JA (2013) Augmented reality application for the maintenance of a flapper valve of a fuller-kynion type m pump. Procedia Comput Sci 25:154–160. https://doi.org/10.1016/j.procs.2013.11.019
17. Gonzalez AV, Koh S, Kapalo K, Sottilare R, Garrity P, Billinghurst M, Laviola J (2019) A comparison of desktop and augmented reality scenario based training authoring tools. In: Proceedings - 2019 IEEE international symposium on mixed and augmented reality, ISMAR 2019, pp 339–350
18. Gorecky D, Schmitt M, Loskyll M, Zühlke D (2014) Human-machine-interaction in the industry 4.0 era. In: 2014 12th IEEE international conference on industrial informatics (INDIN), pp 289–294
19. Guerreiro BV, Lins RG, Sun J, Schmitt R (2018) Definition of smart retrofitting: first steps for a company to deploy aspects of industry 4.0. In: Advances in manufacturing. Springer, pp 161–170
20. Jinyu L, Bangbang Y, Danpeng C, Nan W, Guofeng Z, Hujun B (2019) Survey and evaluation of monocular visual-inertial SLAM algorithms for augmented reality. Virt Real Intell Hardware 1(4):386–410. https://doi.org/10.1016/j.vrih.2019.07.002
21. Kagermann H, Helbig J, Hellinger A, Wahlster W (2013) Recommendations for implementing the strategic initiative INDUSTRIE 4.0: securing the future of German manufacturing industry; final report of the Industrie 4.0 Working Group. Forschungsunion
22. Kamat P, Sugandhi R (2020) Anomaly detection for predictive maintenance in industry 4.0-A survey. In: E3S Web of Conferences, vol 170, pp 1–8
23. Lai ZH, Tao W, Leu MC, Yin Z (2020) Smart augmented reality instructional system for mechanical assembly towards worker-centered intelligent manufacturing. J Manuf Syst 55:69–81. https://doi.org/10.1016/j.jmsy.2020.02.010
24. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: a robustly optimized BERT pretraining approach. https://doi.org/10.48550/arXiv.1907.11692
25. Makris S, Karagiannis P, Koukas S, Matthaiakis AS (2016) Augmented reality system for operator support in human-robot collaborative assembly. CIRP Ann Manuf Technol 65(1):61–64. https://doi.org/10.1016/j.cirp.2016.04.038
26. Mourtzis D, Siatras V, Angelopoulos J (2020) Real-time remote maintenance support based on augmented reality (AR). Appl Sci (Switzerland) 10:5. https://doi.org/10.3390/app10051855
27. Ong SK, Shen Y (2009) A mixed reality environment for collaborative product design and development. CIRP Ann Manuf Technol 58(1):139–142. https://doi.org/10.1016/j.cirp.2009.03.020
28. Pianta E, Bentivogli L, Girardi C (2002) MultiWordNet: developing an aligned multilingual database. In: First international conference on global WordNet, pp 293–302
29. Radkowski R, Herrema J, Oliver J (2015) Augmented reality-based manual assembly support with visual features for different degrees of difficulty. Int J Human-Comput Interact 31(5):337–349. https://doi.org/10.1080/10447318.2014.994194
30. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2019) Exploring the limits of transfer learning with a unified text-to-text transformer. https://doi.org/10.48550/arXiv.1910.10683
31. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ questions for machine comprehension of text. https://doi.org/10.48550/arXiv.1606.05250
32. Sanh V, Debut L, Chaumond J, Wolf T (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://doi.org/10.48550/arXiv.1910.01108
33. Scurati GW, Gattullo M, Fiorentino M, Ferrise F, Bordegoni M, Uva AE (2018) Converting maintenance actions into standard symbols for augmented reality applications in industry 4.0. Comput Ind 98:68–79. https://doi.org/10.1016/j.compind.2018.02.001
34. Simon G, Fitzgibbon AW, Zisserman A (2000) Markerless tracking using planar structures in the scene. In: Proceedings - IEEE and ACM international symposium on augmented reality, ISAR 2000. Institute of Electrical and Electronics Engineers Inc, pp 120–128
35. Tony Liu F, Ming Ting K, Zhou Z-H (2008) Isolation forest. In: Eighth IEEE international conference on data mining. IEEE, pp 413–422
36. Tsai RY (1987) A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf TV cameras and lenses. IEEE J Robot Autom 3(4):323–344. https://doi.org/10.1109/JRA.1987.1087109
37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, vol 30, pp 6000–6010
38. Wang X, Ong SK, Nee AYC (2016) A comprehensive survey of augmented reality assembly research. Adv Manuf 4(1):1–22. https://doi.org/10.1007/s40436-015-0131-4

39. Yu W, Wu L, Deng Y, Mahindru R, Zeng Q, Guven S, Jiang M (2020) A technical question answering system with transfer learning. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp 92–99. https://github.com/wyu97/TTQA
40. Yuan ML, Ong SK, Nee AYC (2008) Augmented reality for assembly guidance using a virtual interactive tool. Int J Prod Res 46(7):1745–1767. https://doi.org/10.1080/00207540600972935
41. Zafrir O, Larey A, Boudoukh G, Shen H, Wasserblat M (2021) Prune once for all: sparse pre-trained language models. https://doi.org/10.48550/arXiv.2111.05754
42. Zonta T, da Costa CA, da Rosa Righi R, de Lima MJ, da Trindade ES, Li GP (2020) Predictive maintenance in the Industry 4.0: a systematic literature review. Comput Ind Eng 12:150. https://doi.org/10.1016/j.cie.2020.106889