UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Enhancing industrial process interaction using deep learning, semantic layers, and augmented reality

December 2023

Author: Juan Jesús Izquierdo Doménech

Director: Dr. Jordi Linares Pellicer

To Jordi, my thesis director, for his infinite patience, who gave me support during this process.

To my research colleagues, Isabel and Jorge, for their support, and who also needed an extra "pill" of patience.

To my family and partner, for their unconditional support and, again, their patience.

Abstract

Augmented Reality (AR) and its ability to integrate synthetic content over a real image provides invaluable value in various fields; however, the industry is one of these fields that can benefit most from it. As a key technology in the evolution towards Industry 4.0 and 5.0, AR not only complements but also enhances human interaction with industrial processes. In this context, AR becomes an essential tool that does not replace the human factor but enriches it, expanding its capabilities and facilitating more effective collaboration between humans and technology. This integration of AR in industrial environments not only improves the efficiency and precision of tasks but also opens new possibilities for expanding human potential.

There are numerous ways in which humans interact with technology, with AR being one of the most innovative paradigms in how users access information; however, it is crucial to recognize that AR, by itself, has limitations in terms of interpreting the content it visualizes. Although today we can access different libraries that use algorithms for image, object, or even environment detection, a fundamental question arises: To what extent can AR understand the context of what it sees? This question becomes especially relevant in industrial environments. Can AR discern if a machine functions correctly, or is its role limited to presenting superimposed digital indicators? The answer to these questions underscores both the potential and the limits of AR, driving the search for innovations that allow for greater contextual understanding and adaptability to specific situations within the industry.

At the core of this thesis lies the objective of not only endowing AR with "semantic intelligence" capable of interpreting and adapting to context, but also of expanding and enriching the ways users interact with this technology. This approach mainly aims to improve the accessibility and efficiency of AR applications in industrial environments, which are by nature restricted and complex. The intention is to go beyond the traditional limits of AR, providing more intuitive and adaptive tools for operators in these environments.

The research unfolds through three articles, where a progressive multimodal architecture has been developed and evaluated. This architecture integrates various user-technology interaction modalities, such as voice control, direct manipulation, and visual feedback in AR. In addition, advanced technologies based on Machine Learning (ML) and Deep Learning (DL) models are incorporated to extract and process semantic information from the environment. Each article builds upon the previous one, demonstrating an evolution in AR's ability to interact more intelligently and contextually with its environment, and highlighting the practical application and benefits of these innovations in the industry.

In the first article, an architecture comprising four fundamental layers is presented and evaluated: the interaction layer, the business layer, the physical AR layer, and the semantic layer. This architecture is later expanded in the subsequent articles. The evaluation of this architecture demonstrates its ability to acquire and analyze visual information from the environment, focusing on elements like an on/off button or a pressure valve. To carry out the system evaluation, classic regression and classification models are employed, as well as convolutional neural networks (CNNs). A step-by-step guidance application for plant operators is developed, dividing them into two groups: *AR standard application* and *AR application with semantic layer*. Furthermore, this system benefits from including Transformers, a highly specialized architecture in processing textual information. This enables the user to ask questions in natural language through voice recognition technologies, receiving answers generated based on available documentary information, such as technical documents about a specific machine. An essential feature that differentiates this proposal from other AR use cases in the industry is the ability to verify, through the semantic layer, the actions undertaken by the user (such as activating a button or checking the pressure level of a machine) before proceeding to the next step. Thus, it is possible to reduce operator's cognitive load compared to traditional AR applications.

In the second article, the research advances along the lines of the first, delving deeper into the use of the semantic layer alongside AR, adding the ability to guide the user in a more complex environment using Simultaneous Localization and Mapping (SLAM) techniques. This study is distinguished by its broader evaluation approach, involving three distinct groups: one with AR and semantic layer with natural language interaction, another with a "blind" AR application, and a third without technological assistance, relying solely on technical documentation. This multiple-evaluation structure allows for a comprehensive comparison and clearly reveals the advantages of the proposed architecture. The results underscore a significant increase in comfort and safety, demonstrating how the semantic layer not only improves user-machine interaction but also validates and optimizes task execution in industrial environments.

Finally, the third article focuses is on how the knowledge of the Subject Matter Expert (SME) can be leveraged through the aforementioned technologies and Large Language Models (LLMs). The research stands out for its approach in combining, not only the existing technical documentation on machinery and various processes, but also the knowledge and experience of the SME in the form of "pills" anchored to specific positions in the environment. Thus, the operator can either make natural language questions about any element or consult the SME's annotations at a specific point. At all times, the information from the technical documentation and the expert knowledge are used to provide a response to the operator. The system evaluation was conducted with two groups of users who had to perform a series of tasks. While group A had only access to technical documentation and an SME, group B had access to the developed application. Considering the limitations of having an SME always available on-site, the system evaluation revealed a marked preference for the ability to access expert information anchored, highlighting the advantage of having this immediate expert assistance, as well as a practical and efficient solution to facilitate the transfer of expert knowledge in the industry.

Throughout and after the development of this thesis, the following conclusions have been drawn:

- Semantic layers and LLMs integration with AR greatly improved task efficiency, especially in complex, cognitive tasks, enabling quicker, more accurate outcomes.

- Semantic AR systems enhanced complex tasks and simpler ones through automatic validation and guidance, boosting overall efficiency.

- AI-enhanced AR with Natural Language Processing (NLP) features led to faster information access and more efficient decision-making than traditional methods like consulting manuals or SMEs.

- Users of AR systems with semantic layers experienced increased satisfaction and ease of use.
- The effective use of AR systems in various challenging settings, like textile labs and shop floors, showed their broad applicability in industry.
- The AR applications were user-friendly, allowing even novices to operate unfamiliar machinery effectively.

Resumen

La Realidad Aumentada (Augmented Reality, AR) y su capacidad para integrar contenido sintético sobre una imagen real proporciona un valor incalculable en diversos campos; no obstante, la industria es uno de estos campos que más se puede aprovechar de ello. Como tecnología clave en la evolución hacia la Industria 4.0 y 5.0, la AR no solo complementa sino que también potencia la interacción humana con los procesos industriales. En este contexto, la AR se convierte en una herramienta esencial que no sustituye al factor humano, sino que lo enriquece, ampliando sus capacidades y facilitando una colaboración más efectiva entre humanos y tecnología. Esta integración de la AR en entornos industriales no solo mejora la eficiencia y precisión de las tareas, sino que también abre nuevas posibilidades para la expansión del potencial humano.

Existen numerosas formas en las que el ser humano interactúa con la tecnología, siendo la AR uno de los paradigmas más innovadores respecto a cómo los usuarios acceden a la información; sin embargo, es crucial reconocer que la AR, por sí misma, tiene limitaciones en cuanto a la interpretación del contenido que visualiza. Aunque en la actualidad podemos acceder a diferentes librerías que utilizan algoritmos para realizar una detección de imágenes, objetos, o incluso entornos, surge una pregunta fundamental: ¿hasta qué punto puede la AR comprender el contexto de lo que ve? Esta cuestión se vuelve especialmente relevante en entornos industriales. ¿Puede la AR discernir si una máquina está funcionando correctamente, o su rol se limita a la presentación de indicadores digitales superpuestos? La respuesta a estas cuestiones subrayan tanto el potencial como los límites de la AR, impulsando la búsqueda de

innovaciones que permitan una mayor comprensión contextual y adaptabilidad a situaciones específicas dentro de la industria.

En el núcleo de esta tesis yace el objetivo de no solo dotar a la AR de una "inteligencia semántica" capaz de interpretar y adaptarse al contexto, sino también de ampliar y enriquecer las formas en que los usuarios interactúan con esta tecnología. Este enfoque se orienta particularmente a mejorar la accesibilidad y la eficiencia de las aplicaciones de AR en entornos industriales, que son por naturaleza restringidos y complejos. La intención es ir un paso más allá de los límites tradicionales de la AR, proporcionando herramientas más intuitivas y adaptativas para los operadores en dichos entornos.

La investigación se despliega a través de tres artículos de investigación, donde se ha desarrollado y evaluado una arquitectura multimodal progresiva. Esta arquitectura integra diversas modalidades de interacción usuario-tecnología, como el control por voz, la manipulación directa y el feedback visual en AR. Además, se incorporan tecnologías avanzadas basadas en modelos de aprendizaje automática (Machine Learning, ML) y aprendizaje profundo (Deep Learning, DL) para extraer y procesar información semántica del entorno. Cada artículo construye sobre el anterior, demostrando una evolución en la capacidad de la AR para interactuar de manera más inteligente y contextual con su entorno, y resaltando la aplicación práctica y los beneficios de estas innovaciones en la industria.

En el primer artículo, se presenta y evalúa una arquitectura compuesta por cuatro capas fundamentales: la capa de interacción, la capa de negocios, la capa física de AR y la capa semántica. Esta arquitectura se ve ampliada en los artículos subsiguientes. La evaluación de esta arquitectura demuestra su capacidad para adquirir y analizar información visual del entorno, centrándose en elementos como un botón de encendido/apagado o una válvula de presión. Para llevar a cabo la evaluación del sistema, se emplean modelos clásicos de regresión y clasificación, así como redes neuronales convolucionales (Convolutional Neural Networks, CNN). Se desarrolla una aplicación de guía paso a paso para operarios de planta, dividiéndolos en dos grupos: *AR standard application* y *AR application with semantic layer*. Además, este sistema se beneficia de la inclusión de Transformers, una arquitectura altamente especializada en el procesamiento de información textual. Esto posibilita que el usuario realice preguntas en lenguaje natural mediante tecnologías de reconocimiento de voz, obteniendo respuestas generadas en función de la información documental disponible, como por ejemplo, documentos técnicos sobre una máquina específica. Una característica esencial que diferencia esta propuesta de otros casos de uso de AR en la industria es la posibilidad de verificar, mediante la capa semántica, las acciones que acomete el usuario (como podrían ser activar un botón o verificar el nivel de presión de una máquina) antes de continuar con el siguiente paso. De este modo, es posible reducir la carga cognitiva del operario respecto a aplicaciones de AR tradicionales.

El segundo artículo avanza en la línea de investigación del primero profundizando en el uso de la capa semántica junto a la AR, añadiendo la posibilidad de realizar una guía al usuario en un entorno más complejo mediante técnicas de localización y mapeo simultáneo (Simultaneous Localization And Mapping, SLAM). Este estudio se distingue por su enfoque de evaluación más amplio, involucrando a tres grupos distintos: uno con AR y capa semántica con interacción en lenguaje natural, otro con una aplicación de AR "ciega", y un tercero sin asistencia tecnológica, dependiendo únicamente de la documentación técnica. Esta estructura de evaluación múltiple permite una comparación exhaustiva y revela claramente las ventajas de la arquitectura propuesta. Los resultados subrayan un incremento notable en la comodidad y seguridad, demostrando cómo la capa semántica no solo mejora la interacción usuario-máquina, sino que también valida y optimiza la ejecución de tareas en entornos industriales.

Finalmente, en el tercer artículo se analiza cómo el conocimiento del experto en la materia (Subject Matter Expert, SME) puede ser aprovechado gracias a las tecnologías anteriormente mencionadas y los modelos de lenguajes masivos (Large Language Models, LLMs). La investigación destaca por su enfoque en combinar, no solo la documentación técnica existente sobre la maquinaria y los diferentes procesos, si no junto a esto, el conocimiento y experiencia del SME en forma de "píldoras" ancladas a posiciones concretas en el entorno. De este modo, el operario bien puede realizar consultas en lenguaje natural de cualquier elemento, bien puede consultar las anotaciones del SME en un punto en concreto. En todo momento, la información de la documentación técnica y el conocimiento experto son usados para devolver una respuesta al operario. La evaluación del sistema se llevó a cabo con dos grupos de usuarios que deben realizar una serie de tareas. Mientras que el grupo A tenía único acceso a documentación técnica y a un SME, el grupo B disponía de acceso a la aplicación desarrollada. Teniendo en cuenta las limitaciones que supone el disponer de un SME en todo momento en planta, la evaluación del sistema reveló una preferencia marcada por la capacidad de acceder a información experta anclada, resaltando la ventaja de disponer de esta asistencia experta inmediata, así como una solución práctica y eficiente para favorecer la transferencia de conocimiento experto en la industria.

A lo largo y tras el desarrollo de esta tesis, se han extraído las siguientes conclusiones:

- La integración de capas semánticas y LLMs con la AR mejoró significativamente la eficiencia de las tareas, especialmente en tareas complejas y cognitivas, permitiendo resultados más rápidos y precisos.

- Los sistemas de AR con capa semántica no solo mejoraron tareas complejas sino también tareas más sencillas a través de validación automática y guiado, aumentando la eficiencia general.

- La AR mejorada con IA y características de procesamiento de lenguaje natural (Natural Language Processing, NLP) condujo a un acceso más rápido a la información y a una toma de decisiones más eficiente que los métodos tradicionales como la consulta de manuales o expertos.

- Los usuarios de sistemas de AR con capas semánticas experimentaron un aumento en la satisfacción y facilidad de uso.

- El uso efectivo de sistemas de AR en entornos complejos variados, como laboratorios textiles y plantas industriales, demostró su amplia aplicabilidad en la industria.

- Las aplicaciones de AR eran fáciles de usar, permitiendo incluso a inexpertos operar maquinaria desconocida de manera efectiva.

Resum

La Realitat Augmentada (Augmented Reality, AR) i la seua capacitat per integrar contingut sintètic sobre una imatge real ofereix un valor incalculable en diversos camps; no obstant això, la indústria és un d'aquests camps que més pot aprofitar-se'n. Com a tecnologia clau en l'evolució cap a la Indústria 4.0 i 5.0, l'AR no només complementa sinó que també potencia la interacció humana amb els processos industrials. En aquest context, l'AR es converteix en una eina essencial que no substitueix al factor humà, sinó que l'enriqueix, ampliant les seues capacitats i facilitant una col·laboració més efectiva entre humans i tecnologia. Esta integració de l'AR en entorns industrials no solament millora l'eficiència i precisió de les tasques, sinó que també obri noves possibilitats per a l'expansió del potencial humà.

Existeixen nombroses formes en què l'ésser humà interactua amb la tecnologia, sent l'AR un dels paradigmes més innovadors respecte a com els usuaris accedeixen a la informació; no obstant això, és crucial reconéixer que l'AR, per si mateixa, té limitacions quant a la interpretació del contingut que visualitza. Encara que en l'actualitat podem accedir a diferents llibreries que utilitzen algoritmes per a realitzar una detecció d'imatges, objectes, o fins i tot entorns, sorgeix una pregunta fonamental: fins a quin punt pot l'AR comprendre el context d'allò veu? Esta qüestió esdevé especialment rellevant en entorns industrials. Pot l'AR discernir si una màquina està funcionant correctament, o el seu rol es limita a la presentació d'indicadors digitals superposats? La resposta a estes qüestions subratllen tant el potencial com els límits de l'AR, impulsant la recerca d'innovacions que permeten una major comprensió contextual i adaptabilitat a situacions específiques dins de la indústria.

En el nucli d'esta tesi jau l'objectiu de no solament dotar a l'AR d'una "intel·ligència semàntica" capaç d'interpretar i adaptar-se al context, sinó també d'ampliar i enriquir les formes en què els usuaris interactuen amb esta tecnologia. Aquest enfocament s'orienta particularment a millorar l'accessibilitat i l'eficiència de les aplicacions d'AR en entorns industrials, que són de naturalesa restringida i complexos. La intenció és anar un pas més enllà dels límits tradicionals de l'AR, proporcionant eines més intuïtives i adaptatives per als operaris en els entorns esmentats.

La recerca es desplega a través de tres articles d'investigació, on s'ha desenvolupat i avaluat una arquitectura multimodal progressiva. Esta arquitectura integra diverses modalitats d'interacció usuari-tecnologia, com el control per veu, la manipulació directa i el feedback visual en AR. A més, s'incorporen tecnologies avançades basades en models d'aprenentatge automàtic (ML) i aprenentatge profund (DL) per a extreure i processar informació semàntica de l'entorn. Cada article construeix sobre l'anterior, demostrant una evolució en la capacitat de l'AR per a interactuar de manera més intel·ligent i contextual amb el seu entorn, i ressaltant l'aplicació pràctica i els beneficis d'estes innovacions en la indústria.

En el primer article, es presenta i avalua una arquitectura composta per quatre capes fonamentals: la capa d'interacció, la capa de negocis, la capa física d'AR i la capa semàntica. Esta arquitectura es veu ampliada en els articles subsegüents. L'avaluació d'aquesta arquitectura demostra la seua capacitat per a adquirir i analitzar informació visual de l'entorn, centrant-se en elements com un botó d'encesa/apagada o una vàlvula de pressió. Per a dur a terme l'avaluació del sistema, s'emprenen models clàssics de regressió i classificació, així com xarxes neuronals convolucionals (Convolutional Neural Networks, CNN). Es desenvolupa una aplicació de guia pas a pas per a operaris de planta, dividint-los en dos grups: *AR standard application* i *AR application with semantic layer.* A més, este sistema es beneficia de la inclusió de Transformers, una arquitectura altament especialitzada en el processament d'informació textual. Això possibilita que l'usuari realitze preguntes en llenguatge natural mitjançant tecnologies de reconeixement de veu, obtenint respostes generades en funció de la informació documental disponible, com per exemple, documents tècnics sobre una màquina específica. Una característica essencial que diferencia esta proposta d'altres casos d'ús d'AR en la indústria és la possibilitat de verificar, mitjançant la capa semàntica, les accions que emprén l'usuari (com podrien ser activar un botó o verificar el nivell de pressió d'una màquina) abans de continuar amb el següent pas. D'aquesta manera, és possible reduir la càrrega cognitiva de l'operari respecte a aplicacions d'AR tradicionals.

El segon article avança en la línia d'investigació del primer, profunditzant en l'ús de la capa semàntica junt amb l'AR, afegint la possibilitat de realitzar una guia a l'usuari en un entorn més complex mitjançant tècniques de localització i mapatge simultani (Simultaneous Localization And Mapping, SLAM). Este estudi es distingeix pel seu enfocament d'avaluació més ampli, involucrant a tres grups distints: un amb AR i capa semàntica amb interacció en llenguatge natural, un altre amb una aplicació d'AR "cega", i un tercer sense assistència tecnològica, depenent únicament de la documentació tècnica. Esta estructura d'avaluació múltiple permet una comparació exhaustiva i revela clarament els avantatges de l'arquitectura proposada. Els resultats subratllen un increment notable en la comoditat i seguretat, demostrant com la capa semàntica no sols millora la interacció usuari-màquina, sinó que també valida i optimitza l'execució de tasques en entorns industrials.

Finalment, en el tercer article s'analitza com el coneixement de l'expert en la matèria (Subject Matter Expert, SME) pot ser aprofitat gràcies a les tecnologies anteriorment mencionades i els models de llenguatge massius (Large Language Models, LLMs). La recerca destaca pel seu enfocament en combinar, no sols la documentació tècnica existent sobre la maquinària i els diferents processos, sinó junt amb això, el coneixement i experiència de l'SME en forma de "píndoles" ancorades a posicions concretes en l'entorn. D'aquesta manera, l'operari bé pot realitzar consultes en llenguatge natural de qualsevol element, bé pot consultar les anotacions de l'SME en un punt en concret. En tot moment, la informació de la documentació tècnica i el coneixement expert són usats per a tornar una resposta a l'operari. L'avaluació del sistema es va dur a terme amb dos grups d'usuaris que han de dur a terme una sèrie de tasques. Mentre que el grup A tenia únic accés a documentació tècnica i a un SME, el grup B disposava d'accés a l'aplicació desenvolupada. Tenint en compte les limitacions que suposa el fet de disposar d'un SME en tot moment en planta, l'avaluació del sistema va revelar una preferència marcada per la capacitat d'accedir a informació experta ancorada, ressaltant l'avantatge de disposar d'aquesta assistència experta immediata, així com una solució pràctica i eficient per a afavorir la transferència de coneixement expert en la indústria.

Al llarg i després del desenvolupament d'esta tesi, s'han extret les següents conclusions:

- La integració de capes semàntiques i LLMs amb l'AR va millorar significativament l'eficiència de les tasques, especialment en tasques complexes i cognitives, permetent resultats més ràpids i precisos.

- Els sistemes d'AR amb capa semàntica no solament van millorar tasques complexes sinó també tasques més senzilles a través de validació automàtica i guiatge, augmentant l'eficiència general.

- L'AR millorada amb IA i característiques de processament de llenguatge natural (Natural Language Processing, NLP) va conduir a un accés més ràpid a la informació i a una presa de decisions més eficient que els mètodes tradicionals com la consulta de manuals o experts.

- Els usuaris de sistemes d'AR amb capes semàntiques van experimentar un augment en la satisfacció i facilitat d'ús.

- L'ús efectiu de sistemes d'AR en entorns complexos variats, com laboratoris tèxtils i plantes industrials, va demostrar la seua àmplia aplicabilitat en la indústria.

- Les aplicacions d'AR eren fàcils d'usar, permetent fins i tot a inexperts operar maquinària desconeguda de manera efectiva.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Augmented Reality (AR) stands at the forefront of a technological renaissance, blending the digital and physical worlds in a way that enhances real-world experiences with virtual overlays. Pioneered by researchers such as Caudell (Caudell and Mizell 1992) and Azuma (R. T. Azuma 1997 and R. Azuma et al. 2001), AR has evolved from a novel concept into a robust tool with wide-ranging applications across industries. While Caudell initially introduced AR in the context of aiding manufacturing processes (see Figure 1.1 for a prototype diagram), Azuma's works provide a comprehensive definition of AR as a system that combines real and virtual environments, is interactive in real-time, and aligns virtual with real objects. Milgram et al. have established a comprehensive taxonomy that classifies applications across a spectrum extending from actual reality to Virtual Reality (VR). This continuum also encompasses AR and Augmented Virtuality (AV), as delineated in their work (Milgram et al. 1995). The corresponding graphical representation of this continuum is illustrated in Figure 1.2.

Technologically, AR operates on a spectrum of platforms, leveraging various libraries and frameworks. Key among these is the Unity 3D engine, renowned for its versatility in creating immersive AR experiences. Additionally, libraries

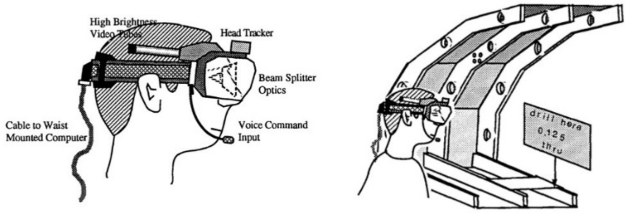


Figure 1.1: Early AR head-mounted display system, showcasing components for visual display, head tracking, and voice command input. (Caudell and Mizell 1992)

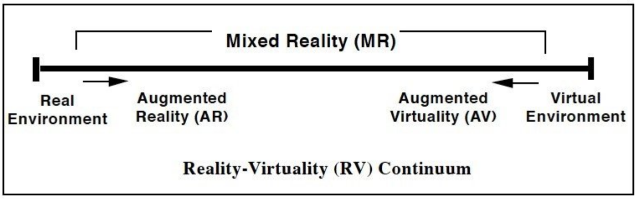


Figure 1.2: Virtual Reality Continuum. (Milgram et al. 1995)

such as ARCore by Google (Google 2018) and ARKit by Apple (Apple 2017) have democratized AR development, enabling creators to build applications that are accessible to a broader audience through smartphones and tablets. These technologies facilitate a more interactive user experience and contribute significantly to the ease of implementing AR solutions in existing industrial systems. In the domain of AR content display, Billinghurst et al. identify several innovative methods (Mark Billinghurst, Clark, and G. Lee 2015). Video-see-through technology involves capturing the real world through cameras and overlaying digital content onto this feed, displayed on a screen, which is beneficial for precise content control but may encounter latency issues. Optical see-through, another method, employs transparent displays on glasses or head-mounted displays to overlay digital information directly onto the user's view of the real world, offering a more immediate AR experience, though aligning virtual and real objects can be challenging. Projection-based AR, on the other hand, projects digital images onto physical surfaces, making any surface an interactive display suitable for settings like advertising and education, though its effectiveness varies with lighting and surface properties. Pejsa et al. (Pejsa et al. 2016) proposed a telepresence system using this technology, as seen in Figure 1.3. Lastly, eye multiplexed AR introduces a unique layer of augmented content that is deliberately not aligned with the user's direct line of sight to reality, ensuring no interference with their natural vision. This approach requires users to adjust their gaze away from their immediate view to access and interact with the augmented content, effectively segregating the AR experience from the real-world view and enabling a distinct, yet non-intrusive, augmentation of their surroundings. An example of an application can be seen in Figure 1.4, where the user must look to a side of the screen to access the synthetic, non-aligned information.

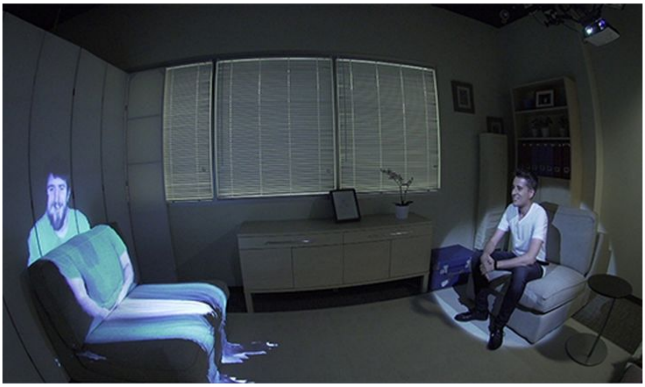


Figure 1.3: Projection-based AR. (Pejsa et al. 2016)

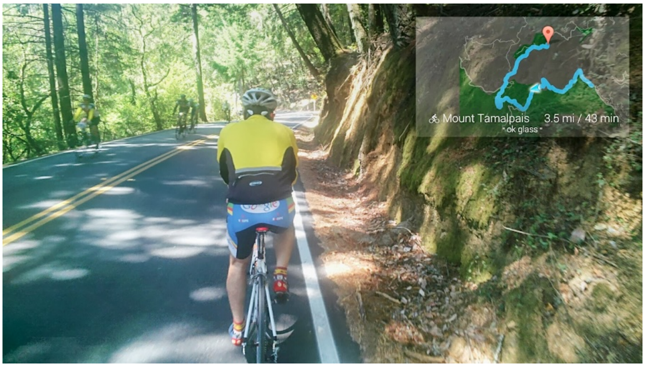


Figure 1.4: Example of and Eye multiplexed AR application. (Google)

The advent of Industry 4.0 marks a transformative era in the realm of manufacturing and industrial processes, characterized by an unprecedented integration of advanced technologies. As Kagermann et al. elucidate (Kagermann et al. 2013), the incorporation of Big Data, the evolution of Digital Twins, and the advancements in Additive Manufacturing, collectively signify a monumental shift in operational paradigms. Figure 1.5 illustrates the nine pillars of Industry 4.0. Notably, the integration of AR as a paradigm in Human-Computer

Interaction (HCI) emerges in the shape of a pivotal aspect, enhancing rather than replacing the human element in industrial processes and trying to make the compute "invisible" (Mark Billinghurst, Clark, and G. Lee 2015). Figure 1.6 illustrates the distinct characteristics of the AR paradigm in comparison to the widely recognized "Desktop metaphor" and VR paradigms, as delineated by Rekimoto et al. (Rekimoto 1995). This symbiosis of human expertise and technological advancement heralds a promising future for manufacturing workflows. Research on the application of AR in industrial contexts is extensive. Notable contributions include Shen et al., who developed a collaborative product design approach utilizing AR (Shen, Ong, and Nee 2010), and Ng et al., who introduced the GARDE project, an innovative AR system that uses gesture-based interaction (Ng et al. 2011). In the realm of process design, Yuan et al. have contributed by developing an AR-based assembly guidance tool (Yuan, Ong, and Nee 2008), and Ong et al.'s work on bare-hand assisted assembly processes facilitated through AR (Ong and Z. B. Wang 2011). Furthermore, the advancements in AR for maintenance processes have been significantly shaped by authors such as Mourtzis et al. (Mourtzis, Siatras, and Angelopoulos 2020) and Palmarini et al. (Palmarini, Fernández, et al. 2022).

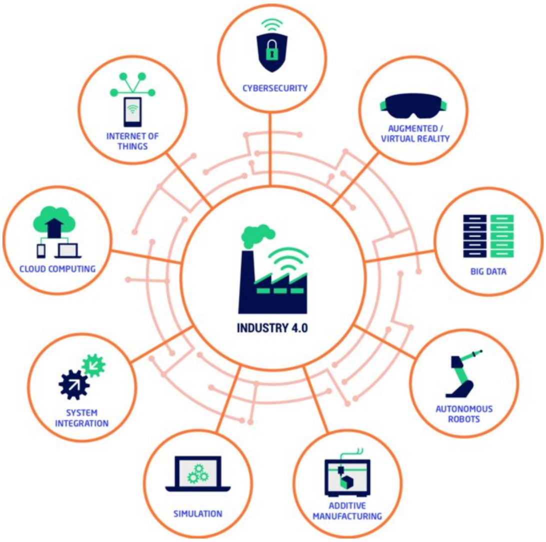


Figure 1.5: The nine pillars of Industry 4.0. (Kadir 2020)

Nevertheless, the rapid automation associated with Industry 4.0 and the prospective Industry 5.0 raises concerns regarding the displacement of human roles. In this context, AR plays a crucial role in augmenting human capabilities, offering

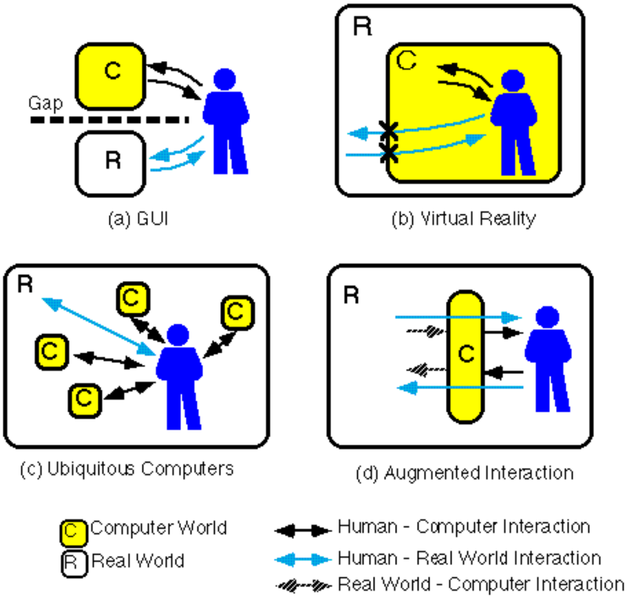

Gap
C
R
(a) GUI
R
C
(b) Virtual Reality
R
C
C
C
C
(c) Ubiquitous Computers
R
C
(d) Augmented Interaction
C Computer World
R Real World
Human - Computer Interaction
Human - Real World Interaction
Real World - Computer Interaction

Figure 1.6: HCI paradigms. (Rekimoto 1995)

enhanced operational possibilities and capacities. Despite these technological strides, the transition to Industry 4.0 is not without challenges, particularly for existing enterprises. Guerreiro et al. introduces the concept of 'Smart Retrofitting', advocating for the adaptation of current machinery and processes to Industry 4.0 standards, with minimal time and cost implications (Guerreiro et al. 2018). The challenges of adopting Industry 4.0 are multifaceted. Ing et al. highlight obstacles in data management and integration, knowledge-driven processes, security concerns, capital investment, workforce dynamics, and educational needs (Ing Tay et al. 2019). Prause identifies market uncertainty, relative competitive advantage, and top management support as key determinants in the adoption process, especially for small and medium enterprises, underlining that external factors have a higher impact than internal ones when adopting the new industry standards (Prause 2019). Furthermore, Sevinç et al. underscore the unique difficulties small and medium-sized enterprises encountered during this transition, suggesting employing multi-criteria decision-making methods to streamline the process (Sevinç, Gür, and Eren 2018). One potential strategy for navigating these challenges is the application of Lean Philosophy, emphasizing incremental changes and continuous improvement, as articulated by Womack et al. (Womack, Jones, and Roos 1992). Several studies suggest that AR plays a significant role in helping small and medium companies adopt Industry 4.0 (Pierdicca et al. 2017 and Martin, Bohuslava, and Igor 2018). In line with this, Garza et al. emphasize the value of AR in circumventing the impracticality of having a Subject Matter Expert (SME) available on the shop floor at all times (Garza et al. 2013). This perspective underscores the importance of AR in providing critical expertise and guidance, essential for the effective implementation of Industry 4.0 technologies.

In the transition towards Industry 4.0, a significant challenge persists in how technical documentation is provided. Despite technological advancements, most machinery documentation in industrial settings still relies on traditional formats such as printed paper, as authors such as Ventura (Ventura 2000) and Abramovici et al. (Abramovici, Krebs, and Schindler 2013) observe. These formats are hardly updatable, accessible, or translatable, and they lack portability, thus leading to inefficiencies in their practical use. The rise of digital platforms like WikiHow or YouTube illustrates a shift towards more dynamic and user-friendly documentation methods, addressing the mentioned limitations. However, a more tailored solution for industrial applications is emerging through the development of AR technologies. As detailed by Gattullo et al. (Gattullo et al. 2019) and further supported by the research of Quint et al. (Quint and Loch 2015), and Kollatsch (Kollatsch and Klimant 2021), AR offers a revolutionary approach to accessing documentation. By overlaying digital information directly onto the physical machinery, AR reduces the cognitive load associated with interpreting technical jargon and complex step-by-step tutorials. This in-place provision of guidance enhances the understanding of complex machinery and streamlines maintenance and operational processes, marking a significant leap forward in the documentation practices for Industry 4.0. Further augmenting this technological evolution, the integration of advanced natural language processing (NLP) tools like Transformers (Vaswani et al. 2017), exemplified by BERT (Devlin et al. 2018) or Google's T5 (Raffel et al. 2020), could revolutionize information retrieval in these settings. These models have shown exceptional capabilities in tasks such as Question Answering (QA) and summarization, potentially enabling more efficient and context-aware access to technical information.

Artificial Intelligence (AI) is rapidly evolving and gaining maturity as it integrates into various industrial domains. On one spectrum, Machine Learning (ML) demonstrates exceptional proficiency in both classification and regression tasks across a wide array of applications. These include predictive maintenance, quality control, supply chain optimization, manufacturing process optimization, fraud detection, customer service, and workplace security, to name a few. Rai et al. emphasize the impact of ML in the manufacturing industry, enabling smart factories and offering benefits such as predictive maintenance, process optimization, and supply chain management (Rai et al. 2021). Paolanti et al. focus on using ML, precisely the Random Forest approach, for predictive maintenance in Industry 4.0. The study demonstrates high accuracy in predicting different machine states, leading to improved system reliability (Paolanti et al. 2018). Karrupusamy discusses the importance of predictive maintenance in the manufacturing industry and highlights the extensive application of ML approaches. The paper presents a comparative study and highlights the superiority of the Random Forest model in accuracy and precision for predicting machine failures (P 2021). At the same time, Deep Learning (DL), adept in handling unstructured data such as images, videos, and audio, is making significant inroads into industrial applications. Its capabilities are particularly evident in areas such as image analysis, NLP, speech recognition and generation, facial recognition, and anomaly detection. Malaiya et al. evaluate DL models for network anomaly detection and find that models based on Seq2Seq with LSTM structures show promising performance, achieving high accuracy in identifying network anomalies (Malaiya et al. 2019). Rushe et al. explore anomaly detection in raw audio and demonstrate that autoregressive DL architectures, such as WaveNet, outperform baseline models in detecting anomalies (Rushe and Mac Namee 2019). Munyua et al. provide a survey of DL solutions for anomaly detection in surveillance videos, highlighting the superiority of DL over traditional ML methods in this domain (Gatara Munyua, Wambugu, and Njenga 2021). Zamora et al. use the YOLO architecture (You Only Look Once, Redmon and Farhadi 2018) for object detection and classification (Zamora-Hernández et al. 2021). While not exhaustive, this list underscores the expansive and diverse range of applications where DL is making a transformative impact. Despite its potential, challenges persist in integrating AI into industry. Peres emphasizes that although AI shows promise in aiding manufacturers with the digital transformation of Cyber-Physical Systems (CPS) (Griffor et al. 2017), its widespread adoption beyond initial pilot experiments remains limited (Peres et al. 2020). Briefly, a CPS can be conceptualized as an integrated framework consisting of a physical entity, such as a machine; a corresponding data model, which is network-accessible; and a dedicated service for data retrieval and management (Drath and Horch 2014).

Integrating AR with ML models presents a transformative approach to perceiving and interacting with our environment. While AR is incapable of understanding their surroundings, by merging it with ML, shop floor operators will have the opportunity to access environments where information is not just overlaid but is contextually embedded, understanding and responding to the subtleties of their surroundings. This innovation paves the way for applications that will reshape industries, enhancing learning experiences and aiding operators in their work routines. At the heart of this advancement lies the potential to turn physical spaces into interactive ones. The implications are vast, offering new paradigms for accessing, processing, learning, and utilizing information. Our investigation into this novel integration of AR and environmental semantics lays the groundwork for a new way of processing and interacting with information, promising a more integrated and intuitive user experience in the industry.

1.2 Scientific goals and research hypotheses

This thesis is centered on the innovative concept of enhancing industrial process interactions by integrating AR, ML, and semantic layers. The main objective is to explore and develop methodologies that elevate the utility and efficiency of AR in industrial settings by embedding intelligent, semantic content. This involves creating a relationship between the physical aspects of industrial environments and a semantic layer facilitated by DL techniques. By doing so, the thesis aims to outdo traditional AR applications, offering a more intuitive and context-aware interaction within industrial processes. The approach is not just to overlay digital information onto a physical environment but to ensure that this information is linked with the environment's characteristics. This research posits that by enriching AR with a layer of semantic understanding, we can significantly improve the accuracy, relevance, and usability of information presented to users, thereby transforming how industrial processes are monitored, controlled, and optimized.

This thesis specific objectives are:

- To design and implement an architecture comprising AR and ML to enhance situational awareness and interaction in industrial environments.
- To integrate NLP capabilities into the architecture, enabling more intuitive and flexible interaction mechanisms, including chatbots and transformers, to facilitate a more natural human-machine dialogue.
- To explore and refine multimodal AR methods in industrial scenarios. This entails the development of different approaches for integrating data in physical scenarios and retrieval, such as direct manipulation, Text-to-speech (TTS), and NLP.

- To facilitate the contribution of SMEs in the environment by allowing them to embed "pills" of knowledge within the system. This seeks to bridge the gap in knowledge transfer, enabling efficient dissemination and retrieval of expert knowledge in industrial settings.
- To ensure the consistency of the retrieved information, focusing on resolving issues such as redundant or contradictory information.
- To evaluate the proposed systems in real-world scenarios and measure task efficiency, information accessibility, and overall operator performance.

The following three studies were conducted in order to attain the mentioned objectives:

Study 1 - *Towards achieving a high degree of situational awareness and multimodal interaction with AR and semantic AI in industrial applications*

This study involved 8 participants, with no distinction on age or gender. The participants belonged to the company where the evaluations were carried out; all had prior experience with production line management but needed to gain experience working on the machines used in the system evaluation and using AR applications.

Participants were divided into two groups. On the one hand, Group 1 (AR standard application) had access to a standard AR application for highlighting the elements to interact with. On the other hand, Group 2 (AR application with semantic layer) had access to the same application, with additional features driven by AI models, such as visual automatic action validation (e.g., machine activation), visual reading of metrics (e.g., pressure levels), voice-based information retrieval (e.g., from documents or ERP systems) and anomaly detection (e.g., identifying potential malfunctions based on machine value anomalies).

The primary objective of this study was to assess the effectiveness of integrating semantic capabilities into an AR application. This assessment measured the time each group took to complete five distinct tasks. The key finding of this study was that while both groups performed similarly in straightforward tasks, notable differences emerged in more complex tasks, highlighting the enhanced efficiency offered by the semantic-enhanced AR application.

Study 2 - *Environment awareness, multimodal interaction, and intelligent assistance in industrial augmented reality solutions with deep learning*

As an evolution from the first article, this second article focuses on providing a guide for implementing the proposed system regardless of the underlying technology. Additionally, Simultaneous Localization and Mapping (SLAM) technology allows the application to guide the operator step-by-step through the shop floor with dispersed machines in the space.

This study included 18 participants without any specific selection based on age or gender. All participants were employees of the company where the evaluation took place. While they all had experience managing production lines, they had yet to gain experience with the specific machines used in this system evaluation.

The participants were segregated into three distinct groups to ensure a thorough evaluation. The primary difference between Groups A and B lay in integrating a semantic layer within Group A's AR application, a feature absent in Group B's app. Conversely, Group C did not utilize any application; instead, they were provided access to electronic documentation, specifically in PDF format, and a list of tasks to perform.

In assessing the system's effectiveness, participants performed eight specific tasks involving an extruder and an injector machine. The evaluation measured the time taken to complete these tasks across different groups and examined the efficiency of QA using transformers. Operators were required to respond to three distinct questions, with their access to information varying according to their group assignment. Moreover, operators in Groups A and B were provided with a Likert-scale questionnaire to measure the system's user-friendliness. The findings revealed that Group A, which utilized the semantic layer, experienced a significant reduction in the time taken to perform tasks that demanded higher cognitive effort. Furthermore, Group A demonstrated superior results and higher satisfaction levels than the other groups. Notably, the resolution of queries was markedly quicker in Group A, underscoring the added efficiency of using the semantic layer for information access.

Study 3 - *Large Language Models for in situ knowledge documentation and access with Augmented Reality*

The third article employs a novel approach to effectively leverage expert knowledge dissemination. This method distinctly delineates the roles of SMEs and operators. It introduces an innovative system where SMEs can embed "knowledge pills" into physical elements within a work environment. These "knowledge pills" are strategically designed for subsequent retrieval and utilization by shop floor operators, thereby facilitating a more streamlined and efficient transfer of expertise. The developed app has a dual-purpose design. It allows SMEs to embed "knowledge pills" into the work environment. Simultaneously, it enables operators to easily retrieve this information by either marking out specific areas of interest directly within their workspace or by asking questions in natural language, enhancing the interaction with their surroundings.

In this study, thirty participants aged between 22 and 28 cooperated, ensuring a balanced representation of both men and women. Similar to the methodologies in the other articles, these participants were familiar with the general environment but lacked specific knowledge about the machines used during the evaluation phase.

Participants were categorized into two groups for the study. Group A was provided access to technical documentation and direct support from the SME. In contrast, Group B relied on an AR application that allowed access to expert knowledge by asking questions in natural language through speech recognition. If a Group B member encountered task completion challenges, they were permitted to consult the SME for assistance. Both groups were tasked with executing three specific activities: dye testing, material cleaning, and emulsion homogenization.

This study aimed to assess the efficacy of enabling SMEs to embed their expertise directly into the physical work environment. The findings of this research are encouraging, demonstrating a notable difference in task completion times between the two groups, independent of task complexity. To further evaluate the system's usability, participants in Group B were provided with a Likert-scale questionnaire. The responses indicated a positive perception of the system, underscoring its potential benefits.

1.3 Structure of the thesis

This thesis is structured as follows:

Chapter 1 This chapter introduces the thesis, along with its motivation and goals.

Chapters 2, 3 and 4. These chapters present the selection of research articles that support this thesis, specifically:

Paper 1. Izquierdo-Domenech, J., Linares-Pellicer, J., & Orta-Lopez, J. (2023). Towards achieving a high degree of situational awareness and multimodal interaction with AR and semantic AI in industrial applications. Multimedia Tools and Applications, 82(10), 15875-15901. 10.1007/s11042-022-13803-1

Paper 2. Izquierdo-Domenech, J., Linares-Pellicer, J., & Ferri-Molla, I. (2023). Environment awareness, multimodal interaction, and intelligent assistance in industrial augmented reality solutions with deep learning. Multimedia Tools and Applications, 1-28. 10.1007/s11042-023-17516-x

Paper 3. Izquierdo-Domenech, J., Linares-Pellicer, J., & Ferri-Molla, I. (2023). Large Language Models for in Situ Knowledge Documentation and Access With Augmented Reality. International Journal of Interactive Multimedia and Artificial Intelligence, 10.9781/ijimai.2023.09.002

Chapter 5 This chapter discusses the thesis results, derived publications, and future research.

Chapter 2

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

Izquierdo-Domenech, J., Linares-Pellicer, J., & Orta-Lopez, J. (2023). Towards achieving a high degree of situational awareness and multimodal interaction with AR and semantic AI in industrial applications. Multimedia Tools and Applications, 82(10), 15875-15901. 10.1007/s11042-022-13803-1

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.

2.1 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 (Caudell and Mizell 1992). 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 (Radkowski, Herrema, and Oliver 2015), (Makris et al. 2016), as a step-by-step guide (Scurati et al. 2018) or maintenance tasks (Garza et al. 2013), (Benbelkacem et al. 2013). 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 (Mourtzis, Siatras, and Angelopoulos 2020). 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 (Endsley 1995). 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 (Kamat and Sugandhi 2020) and (Zonta et al. 2020).

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 (Lai et al. 2020). 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), ((Vaswani et al. 2017) and (Devlin et al. 2018)), 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 (Coli et al. 2020)) based on MultiWordNet (Pianta, Bentivogli, and Girardi 2002), and Yu et al. uses natural language to retrieve answers based on previous questions to the system (Yu et al. 2020).

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 2.2 gives a detailed explanation of the proposed architecture, section 2.3 explains the followed approach for validating the operators' actions, section 2.4 describes how a chatbot can help the operator in retrieving industrial data, section 2.5 focuses on the problem of asking questions on technical documentation, section 2.6 proposes the usage of AR for indicating the operator the position of an anomaly. Finally, section 2.7 explains the application developed to test the proposed architecture, section 2.8 evaluates the system in an industrial environment, and section 2.9 presents the conclusions.

2.2 Architecture overview

Since the concept of Industry 4.0 appeared in 2013 (Kagermann et al. 2013), 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 (Guerreiro et al. 2018). 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 (Gorecky et al. 2014). Therefore, based on the three key elements that make up a Cyber-physical system (CPS) (Drath and Horch 2014):

- 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 figure 2.1, 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.

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.

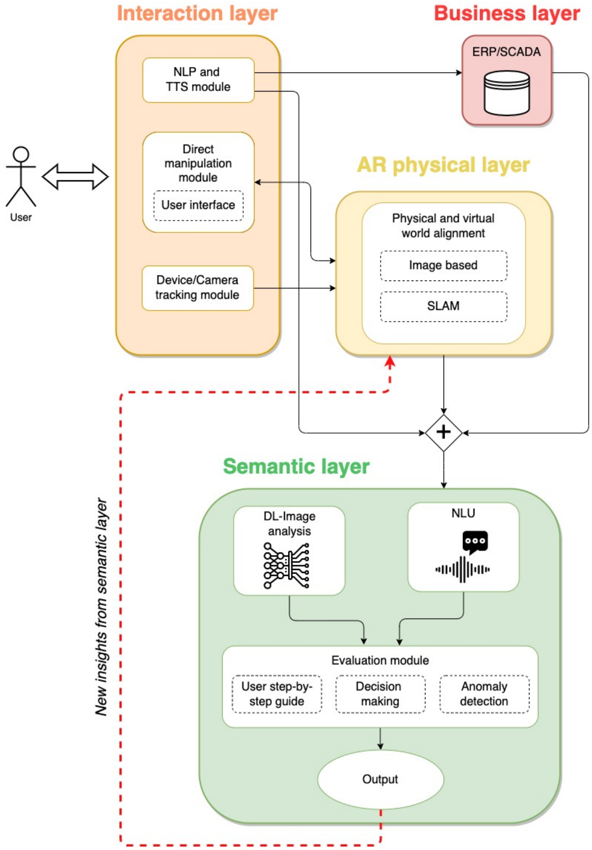


Figure 2.1: General architecture overview

- 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

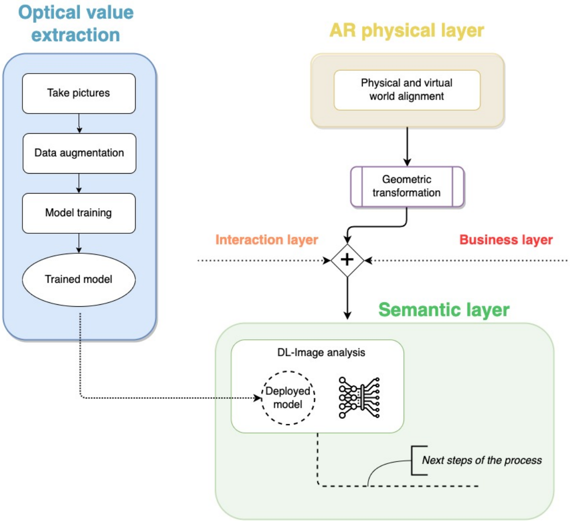


Figure 2.2: Detailed diagram of the physical and semantic layers

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.

2.2.1 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 (Gonzalez et al. 2019).

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

2.2.2 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 (Elia, Gnoni, and Lanzilotto 2016).

In this layer, it is possible to use any solution based on image tracking (Tsai 1987), surface tracking (Simon, Fitzgibbon, and Zisserman 2000), or even Simultaneous Localization and Mapping (SLAM) techniques, which allow the device to discover its position in an unknown environment, and in real-time (Jinyu et al. 2019). 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 figure 2 as the last image adaptation before the semantic layer.

Also, figure 2.2 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 figure 2.2, 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) 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.

2.2.3 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 figure 2.8.

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 2.1. Furthermore, it is also possible to ask specific questions about technical documentation, as seen in table 2.1. This synergy with visuals, sensory, and natural language interaction will be described in further detail later.

2.2.4 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.

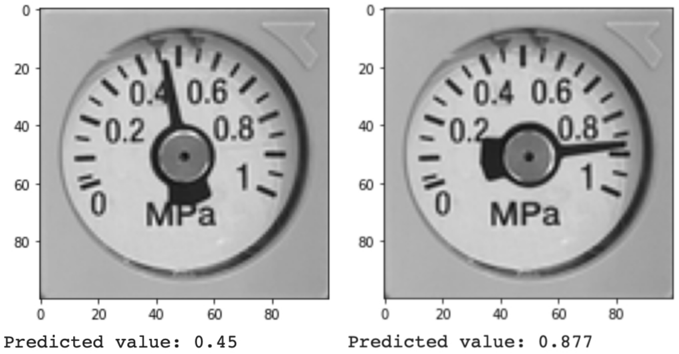
One of the most significant benefits of this approach is the relief of the operator from having to learn specific commands or actions in complex menus. This common way of interacting with SCADA or ERP systems requires essential training time; otherwise, they are only within reach of experts. In the proposed 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.

2.3 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 figure 2.2, 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 figure 2.3, 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.

As has already been mentioned, AR can be used for many tasks such as product design (Ong and Shen 2009), process control (Yuan, Ong, and Nee 2008), and maintenance and training tasks (Garza et al. 2013). 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 (Bottani and Vignali 2019). In this way, it is possible to develop a virtual expert able to help the operators.



```
Predicted value: 0.45
```

```
Predicted value: 0.877
```

Figure 2.3: Using a CNN with regression to interpret the values of a pressure gauge

In the experiments carried out, different CNN architectures for classification and regression tasks based on the images captured by the device are used. In figure 2.4, 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 figure 2.5, the system can interpret the value through an analog control that uses a pointer to indicate the current pressure value. In the case of figure 2.5, 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 figure 2.12 in appendix 2.9.1, a simplified CNN based on (Alexeev et al. 2020) that gives great precision in estimating 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 figure 2.12 in appendix 2.9.1 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.

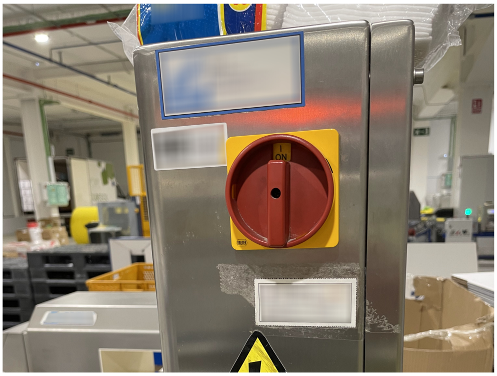


Figure 2.4: Classification example

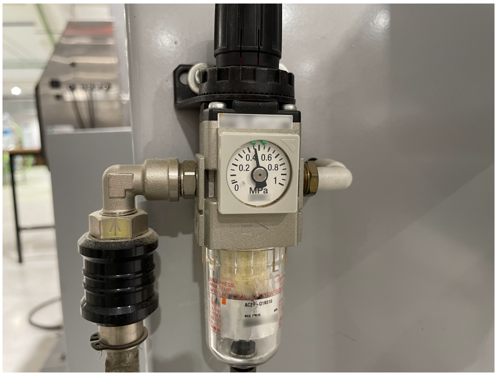


Figure 2.5: Regression example

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 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.

2.4 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 (Coli et al. 2020). 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 2.1 with the AR solution facing the example in figure 2.6. The flexibility of this approach enables the possibility of making the same query/intent for different elements/entities.

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 (Baldauf et al. 2018). 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.

```
{
  "text": "Rollers in stock?",
  "intents": [
    {
      "id": "1606940483084759",
      "name": "get_stock",
      "confidence": 0.9984
    }
  ],
  "entities": {
    "element:element": [
      {
        "id": "1087430018514134",
```

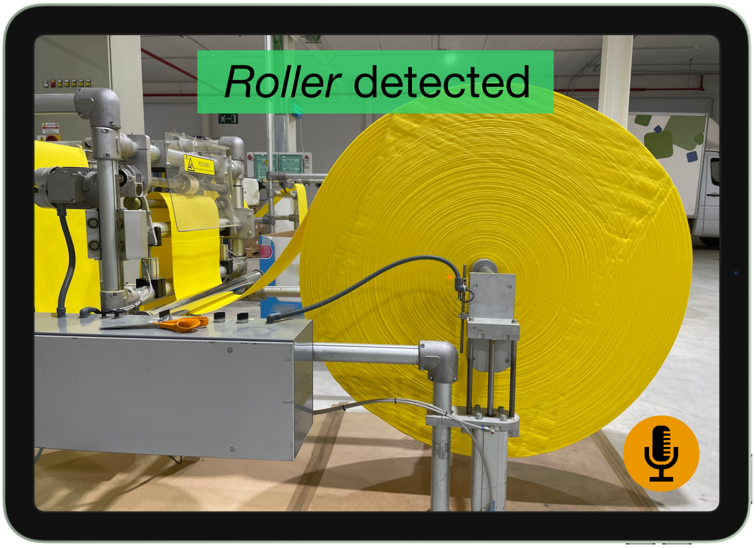


Figure 2.6: The operator can ask questions in natural language about this machine. The AR system gives information regarding what element is the operator in front of, so the question is complemented with the required context

```
        "name": "element",
        "role": "element",
        "start": 0,
        "end": 7,
        "body": "Rollers",
        "confidence": 0.9995,
        "entities": [],
        "value": "rollers",
        "type": "value"
      }
    ]
  }
}
```

Listing 2.1: 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

2.5 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 (Barakonyi, Psik, and Schmalstieg 2004).

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 2.1 and 2.4.

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 (Liu et al. 2019), DistilBERT (Sanh et al. 2019), and Google's T5 (Raffel et al. 2020). 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) (Rajpurkar et al. 2016).

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 (Zafrir et al. 2021) 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 2.1.

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 2.7 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.

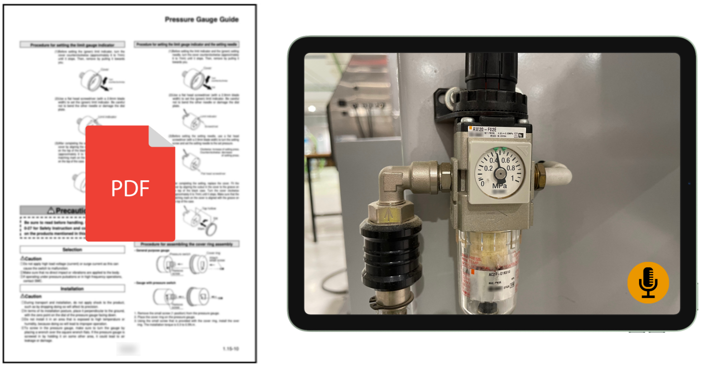
The final model's accuracy metrics, capable of answering questions from the operator in front of a technical document describing different processes related to a device or machine, can only be evaluated in a specific context. If there are some manuals, a set of questions, and the answers obtained by the model, the only way to evaluate the model's adequateness is by comparing its responses to the ones given by humans (Rajpurkar et al. 2016).

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 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

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

Table 2.1: Using a transformer to ask questions in NL in technical documentation



- 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

...

Figure 2.7: Some examples of real questions using a manual of a pressure gauge

perform preliminary training on a chatbot architecture, as discussed in section 2.4.

2.6 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) (Chalapathy and Chawla 2019). 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 figure 2.8, different unsupervised classification algorithms have been tested for anomaly detection. Some examples have been Isolation Forest (Tony Liu, Ming Ting, and Zhou 2008) or K-Means (Ball and Hall 1965), 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.

2.7 General multimodal AR approach

As a consequence of the elements proposed in the suggested architecture from figure 2.1, 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.

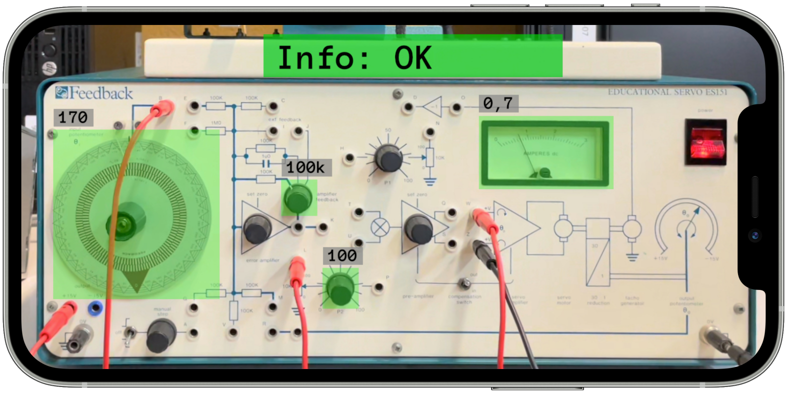


Figure 2.8: The combination of several values can be seen as standard or as an anomaly, and visual cues are possible in AR

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 2.8 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 2.9 and 2.10 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 figure 2.9. 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 figure 2.10. 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.

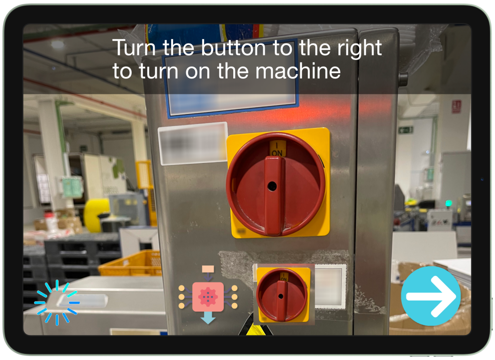
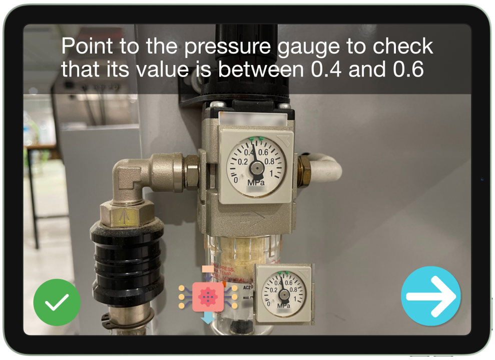

Turn the button to the right to turn on the machine

Figure 2.9: When the machine is switched on, the app lets the operator move to the next step



Point to the pressure gauge to check that its value is between 0.4 and 0.6

Figure 2.10: Automatic value extraction from a pressure gauge

2.8 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:

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) (Brooke 1996) 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 ((Bottani and Vignali 2019), (Fraga-Lamas et al. 2018), and (X. Wang, Ong, and Nee 2016)). 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 figures 2.6, 2.8, 2.9, and 2.10. This process consists of the following set of tasks:

Task 1 *Machine activation* (figure 2.9). 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* (figure 2.10). 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* (figure 2.10). 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* (figure 2.6). 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* (figure 2.8). 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 2.2.

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

Table 2.3 and figure 2.11 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.

	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

Table 2.2: Task completion times without and with semantic layer

*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.

	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)		

Table 2.3: Descriptive and statistical contrasts

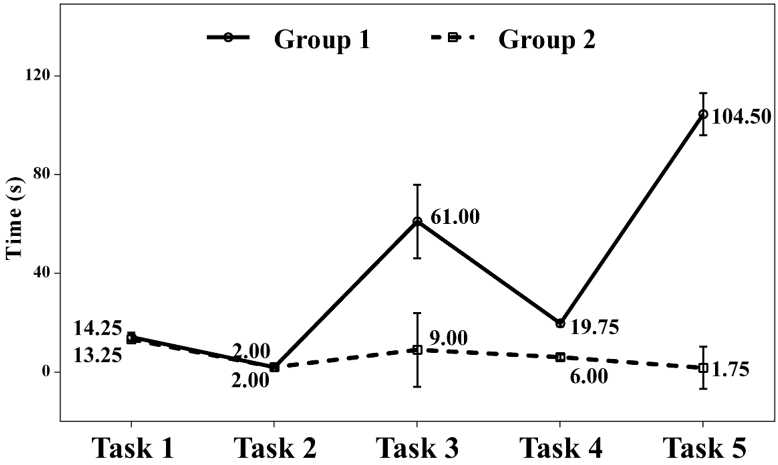


Figure 2.11: Tasks comparison box plot

2.9 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 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.

Declarations

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

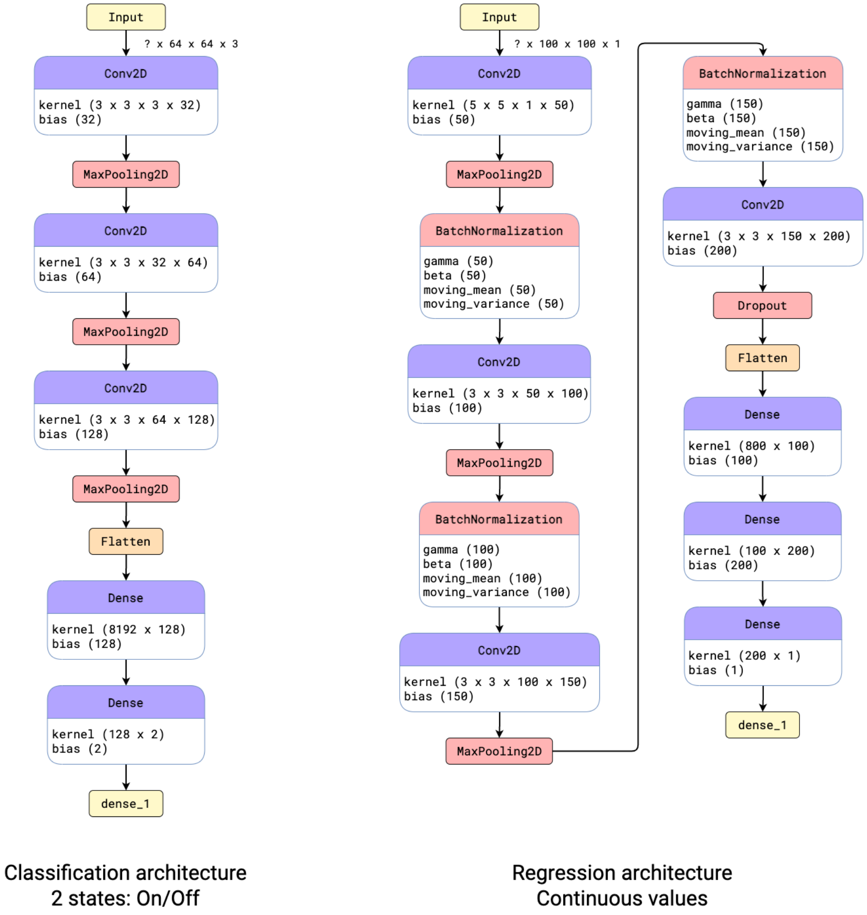
Data Availability

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

Appendices

2.9.1 CNN architectures

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



Classification architecture
2 states: On/Off

Regression architecture
Continuous values

Figure 2.12: Classification and regression CNN architectures

Chapter 3

Environment awareness, multimodal interaction, and intelligent assistance in industrial augmented reality solutions with deep learning

Izquierdo-Domenech, J., Linares-Pellicer, J., & Ferri-Molla, I. (2023). Environment awareness, multimodal interaction, and intelligent assistance in industrial augmented reality solutions with deep learning. Multimedia Tools and Applications, 1-28. 10.1007/s11042-023-17516-x

Augmented reality is increasingly used in various fields, especially industrial applications. Although augmented reality devices' characteristics and technological benefits are still evolving, augmented reality's clear advantages in facilitating mechanical tasks and improving operator performance have made it popular. In industrial settings, the human factor remains irreplaceable, but the evolution of artificial intelligence has allowed any activity on the shop floor to be given new semantic possibilities. Through a semantic layer, it is possible to interpret and validate the environment, provide multimodal interaction, and analyze and evaluate information to detect anomalies or risky situations. Deep learning has opened up new possibilities for existing augmented reality solutions, such as visual interpretation of the environment, natural language understanding for problem-solving, or automatic anomaly detection. This new intelligent layer minimizes unnecessary interactions with the environment, validates the operator's actions, and increases comfort, safety, and focus, making them more efficient in high cognitive level tasks. This work presents a general architecture based on a Semantic layer that relies on augmented reality systems and validates its advantages in a real industrial setting. Overall, integrating artificial intelligence and augmented reality solutions in industrial settings offers significant potential for improving productivity, safety, and worker satisfaction.

3.1 Introduction

Industry 4.0 aims to enhance industrial production efficiency, speed, quality, optimization, and resilience by adopting new technologies (Kagermann et al. 2013; L. D. Xu, E. L. Xu, and Li 2018). The Industrial Internet of Things (IIoT), additive manufacturing, Artificial Intelligence (AI), cybersecurity, and Augmented Reality (AR) play vital roles in this transformation, with AR particularly emphasizing the human factor (C. H. Chu et al. 2021).

Despite significant automation efforts in Industry 4.0, some tasks on the shop floor still require a human-centered approach and cannot be fully automated (Guerreiro et al. 2018; Runji, Y.-J. Lee, and C.-H. Chu 2022). To address this, adopting a Lean philosophy is recommended, where small changes are introduced and evaluated (Womack, Jones, and Roos 1992). AR has gained popularity in the industrial field due to its potential to improve various aspects, including assembly, maintenance, training, and waste reduction (X. Wang, Ong, and Nee 2016; Runji, Y.-J. Lee, and C.-H. Chu 2022; Zonta et al. 2020; Palmarini, Erkoyuncu, et al. 2018; Jaschke 2014; Huenerfauth 2014). Since AR strongly focuses on the human factor in the industry, its implementation can benefit operators and task performance.

Accessing documentation can be complex on the shop floor, especially in non- or partially automated industries where technical documents may exist in paper format (Kollatsch and Klimant 2021). The lack of documentation standardization further complicates information retrieval, increasing the mental load on operators, especially in critical situations (Gattullo et al. 2019). Operators may sometimes require support from Subject Matter Experts (SMEs), which may only be feasible occasionally due to cost or availability (Gilchrist 2016). Operator attentiveness can also lead to task resolution errors (Backs and Seljos 1994). Traditional problem-solving approaches may not be the most effective strategy for learning new concepts or procedures; instead, reducing cognitive load is crucial for learning (Sweller 1988). Stress can impact cognition and knowledge acquisition, but controlled exposure can facilitate cognitive function (Sandi 2013). This study aims to develop a system to reduce cognitive stress for operators facing unfamiliar tasks in a shop floor setting.

Furthermore, the concept of Operator 4.0 aims to establish reliable and interactive relationships between humans and machines, empowering "smart operators" with cutting-edge gadgets and novel skills to exploit Industry 4.0 technologies (Romero, Stahre, and Taisch 2020; Romero, Stahre, Wuest, et al. 2016). Peruzzini et al. highlight these technologies' potential to alleviate the cognitive burden on operators (Peruzzini, Grandi, and Pellicciari 2020). Additionally, the emerging notion of Operator 5.0 aims to create more intuitive, symbiotic, human-centered, and cognitively supportive computing environments to enhance human adaptation capabilities, productivity, and mental well-being (Zambiasi et al. 2022). The integration of technologies like AR and AI is driving the emergence of "softbots" as virtual systems in computing environments to automate tasks, offer conversation-like interactions, exhibit system intelligence, autonomy, proactivity, and process automation (Romero and Stahre 2021; Rabelo, Romero, and Popov Zambiasi 2018).

AI tools such as Machine Learning (ML) and Deep Learning (DL) can complement existing systems to reduce cognitive load, improve context information, and enhance shop floor safety. In cases where older machines lack machine-generated data, visual information becomes crucial. Techniques like Convolutional Neural Networks (CNNs) enable the interpretation of visual cues, such as reading pressure gauge values. Algorithms like K-means (Ball and Hall 1965) and Support Vector Machine (SVM) (Cortes, Vapnik, and Saitta 1995) can check for anomalies and highlight them through AR. Transformers (Vaswani et al. 2017) enable operators to ask questions in Natural Language (NL) and receive responses linked to AR systems through visual cues.

The main objective of this study is to enhance current AR solutions by adding cognitive capabilities through DL and foundation models (Bommasani et al. 2021). Recent industry interest has been observed in AR solutions with cognitive capabilities, aiming to extend current AR solutions with a semantic layer, demonstrated in studies (Rasmussen et al. 2022; Eversberg et al. 2022; Z. Wang et al. 2021; Zhang et al. 2022). These studies illustrate the benefits of integrating environment awareness, multimodal interaction, and cognitive assistance into AR solutions, crucial for extracting information in fast-paced and evolving production environments (Sheu 2010). This work develops and evaluates a system allowing users to interact multimodally, including NL Question Answering (QA) and contextualized visual indications in the AR environment, enhancing operator comfort and reducing the risk of errors. Integrating AR with AI technologies also impacts aspects such as comfort, security, focus, and knowledge acquisition (Sahu, Young, and Rai 2021).

The main contributions of this research are:

- To enhance the capabilities of existing AR systems by adding semantic skills to comprehend and interpret the environment,
- To make a profit from ML models to gain insights, such as anomaly detection or information retrieval with transformers, in industrial settings,
- To reduce the cognitive load from operators, improve task performance, and foster technology adoption.

This article is structured as follows: In section 3.2, the state-of-the-art research related to this work is presented. Then, in section 3.3, the different components of the proposed system are explained in detail. After that, in section 3.4, the technologies and their characteristics for implementing the evaluated system are explained. In section 3.5, a specific case study is performed to evaluate the system's validity, and the results are discussed in section 3.6. Finally, the conclusions are drawn in section 3.8.

3.2 Related work

3.2.1 AI in the Industrial Field

AI, especially ML, has significantly improved various aspects of the industrial field, including waste reduction, error prevention, enhanced quality, risk prevention, and faster learning (Bertolini et al. 2021). In complex scenarios where traditional analytical methods struggle to relate variables and outcomes, ML proves to be a critical tool (Esen et al. 2009). For instance, the "You Only Look Once" (YOLO) algorithm has been utilized by Zamora et al. to detect real-time assembly task actions (Zamora-Hernández et al. 2021). ML and DL techniques excel in anomaly detection, outperforming traditional methods (Javaid et al. 2015). Architectures such as CNNs, RNNs, LSTMs, and AEs are popular for detecting anomalies in unstructured data like images, videos, audio, and time series (Chalapathy and Chawla 2019). Various studies have applied these techniques to detect anomalies in videos and sequential data (Ionescu et al. 2019; Lu et al. 2017), as well as appearance-based anomalies in videos using pre-trained ResNet-50 models (Pang et al. 2020). Additionally, GANs have been used to detect anomalies in image datasets and intrusion networks (H. Zenati et al. 2018). Traditional ML algorithms, especially unsupervised training models, have also shown strong performance in anomaly detection (Škvára, Pevný, and Šmídl 2018). ML and DL are increasingly essential in monitoring and controlling industrial processes, demonstrating their relevance in AR industrial environments (Song et al. 2022; Gopaluni et al. 2020).

3.2.2 AR in Industrial Tasks

AR technology has demonstrated great potential in the industrial field by simplifying complex tasks for operators. However, the complexity of the task and the visual cues utilized should be considered to avoid overloading the operator and diluting their focus of attention (Radkowski, Herrema, and Oliver 2015). Several authors have studied the development and evaluation of AR applications in various industrial tasks, including step-by-step guides (Scurati et al. 2018), product design Luh et al. 2013; Shen, Ong, and Nee 2010; Ong and Shen 2009, process control Ong and Z. B. Wang 2011; Yuan, Ong, and Nee 2008, maintenance and security Mourtzis, Siatras, and Angelopoulos 2020; Tatić and Tešić 2017; Espíndola et al. 2013; Garza et al. 2013; Benbelkacem et al. 2013; Ziaei et al. 2011; N. Zenati, Zerhouni, and Achour 2004; Barakonyi, Psik, and Schmalstieg 2004, and operator training Monroy Reyes et al. 2016; Webel et al. 2013; De Crescenzio et al. 2011. A review of the current literature conducted by Bottani et al. in 2019 indicates that the most researched fields regarding AR in the industry are assembly, maintenance, and training Bottani and Vignali 2019. It is essential to consider not only the complexity of the task but also the complexity added by AR to the interface.

3.2.3 Voice-directed Interfaces and NL Interaction

Voice-directed interfaces are gaining popularity in industrial settings due to their hands-free control and benefits for operators, particularly those with functional diversity (Baldauf et al. 2018). In scenarios where complex machinery requires information from multiple sources, including paper documents, the ability for operators to interact through NL questions proves advantageous (Coli et al. 2020). Examples of chatbots being used to train operators and improve usability in industrial settings, as well as their application in Maintenance, Repair, and Overhaul (MRO) tasks, have been documented in the literature (Casillo et al. 2020; Mleczko 2021).

3.2.4 Cloud Computing Systems for Remote Assistance

Current investments in communication and Cloud Computing systems have spurred the development of remote assistance systems, such as TeamViewer Assist AR (TeamViewer 2021) and Vuforia Chalk (PTC 2017), facilitating remote communication between operators and experts. While these AR technologies have been applied to maintenance tasks (Mourtzis, Siatras, and Angelopoulos 2020), challenges persist, notably the cost and availability of online Subject Matter Experts (SMEs).

In conclusion, AR, voice-directed interfaces, AI, and Cloud Computing have greatly enhanced various industrial aspects, simplifying tasks, reducing errors, and expediting learning. However, it's crucial to consider task complexity and the potential for AR interfaces to overwhelm operators. ML and DL techniques, especially in anomaly detection, outperform traditional methods. Despite challenges, AR, voice interfaces, and AI hold substantial potential in the industrial field, necessitating further research and development to unlock their capabilities. Our research aims to leverage modern DL techniques and AI foundation models to enhance information access in complex environments.

3.3 System architecture overview

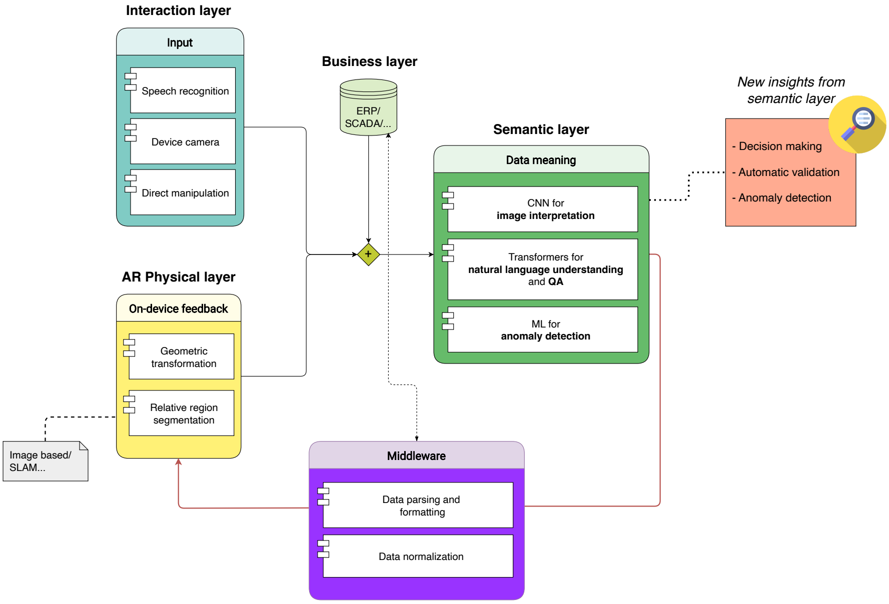


Figure 3.1: System architecture layers.

This study proposes an architecture combining AR, ML, and DL techniques to support operators' tasks on a production shop floor. The architecture is designed to enable the creation of applications that offer intelligent assistance to the operator through multimodal interaction, potentially replacing or enhancing the need for an SME, even when using teleoperated AR systems.

The system consists of four communicating layers, as shown in Fig. 3.1. The central axis is the AR layer, facilitating integration and creating synergy beyond individual components. Further details about each layer are provided below.

3.3.1 Interaction layer

The Interaction layer provides various modes of operator interaction. Three modules enable multimodal interaction, as depicted in Fig. 3.1:

- **Speech recognition**: The system recognizes and responds to the operator's NL queries, providing prompt answers based on available technical information, reducing query time significantly, and avoiding the need for consulting SMEs or searching technical documentation.

- **Device camera**: Using a camera is a fundamental element in current AR solutions, available on mobile devices and AR Head-mounted displays (HMD). It supplies the possibility of adding visual cues to the operator and facilitates the validation of performed tasks using AI-based elements described later.
- **Direct manipulation**: This interaction style relates real-world actions to device-based actions (Shneiderman et al. 2016). The operator interacts through touch screens of mobile devices and hand gestures in specific AR HMDs. The operator can touch the digital representation of environment elements, such as buttons or AR indicators, to retrieve additional information or perform specific tasks.

The system's multimodality allows users to interact with it in three ways and provides feedback through AR cues or textual information, as explained later.

3.3.2 AR Physical layer

The AR Physical layer performs crucial tasks, including displaying synthetic content overlaid on the device's camera feedback, determining the operator's location and relative position in the industrial environment, and extracting areas of interest for later analysis. It locates analog controls, like pressure gauges, in the camera's captured image (Relative region segmentation in Fig. 3.1). Geometric transformation is applied to remove the perspective between the camera position and the control (Geometric transformation in Fig. 3.1), reducing input image variability for more effective CNN training (i.e., the machine view remains orthogonal to the device camera). Reference marks, such as printed images on machines or specific places on the shop floor, can guide the operator in the real environment. Moreover, the layer provides indications of possible anomalies, and answers operator queries about technical documentation.

Simultaneous Localization and Mapping (SLAM) creates a 3D map of the environment to determine the user's 3D location, allowing real-time combination of virtual objects and indoor tracking. The traditional A^* algorithm can guide the operator on the shop floor, while marker-based tracking and SLAM are employed to determine the operator's position.

The evaluated application utilizes AR technology to enable operators to perform complex tasks. Various tools and technologies are available for developing AR applications, including ARCore for Android (Google 2018), ARKit for iOS (Apple 2017), and AR Foundation (Unity 2018). Libraries and SDKs like ARToolkit (Kato and M. Billinghurst 1999) or Vuforia streamline the integration of AR functionality, reducing programming complexity and supporting cross-platform development. Vuforia SDK is used and tested on an iPad device for this research.

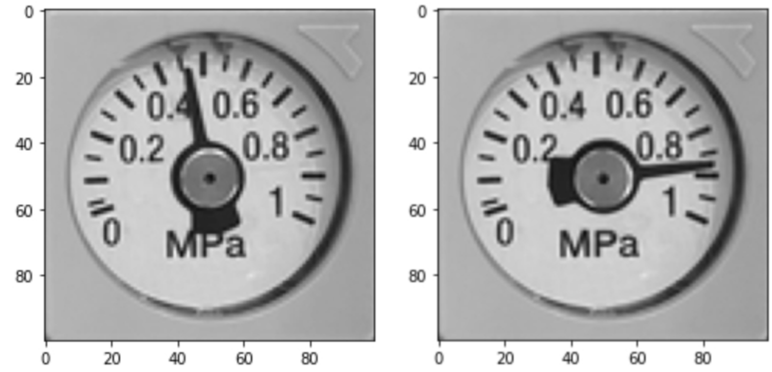
3.3.3 Business layer

The Business layer plays a pivotal role in our architecture by seamlessly integrating shop floor operational systems with higher-level decision-making processes. It leverages Enterprise Resource Planning (ERP) and Supervisory Control And Data Acquisition (SCADA) systems, enabling efficient resource management and process supervision. Combining data from ERP, SCADA, and the semantic layer through CNNs enhances our system's capabilities, particularly in tasks like anomaly detection.

This layer acts as a bridge between operational and business components, utilizing ERP and SCADA data alongside DL from the semantic layer to provide comprehensive insights for informed decision-making. Our architecture promotes a holistic understanding of the industrial environment, improving operational efficiency and supporting intelligent decisions. The intermediate middleware layer serves as an interface, facilitating seamless communication between the AR Physical layer and the Business layer. Its primary function is to abstract and enable the retrieval of information from sensorized machines, enabling system scalability and flexibility.

3.3.4 Semantic layer

DL has enabled the development of CNN architectures for image input, utilizing convolutional layers to extract significant features and perform classifications or regressions, such as in Fig. 3.2, where a CNN can predict the pressure value from the image of a pressure gauge. Analyzing that image through CNNs is crucial when a machine is not sensorized (i.e., it is not connected to a SCADA system or does not have a PLC), thus being the only source of information the system can extract data from. Nevertheless, it's important to note that this visual data should not be perceived as a substitute for the structured data derived from PLCs and SCADA systems, as these systems excel in tasks related to automation, control, and historical data analysis. Instead, the integration



```
Predicted value: 0.45          Predicted value: 0.877
```

Figure 3.2: Reading the pressure value with a CNN in non-sensorized machines.

of both structured and visual data sources emerges as a strategic approach, affording a comprehensive perspective of industrial operations that, in turn, enhances decision-making capabilities and facilitates process optimization.

On the other hand, ML allows the use of supervised models like SVM or Isolation Forest, and unsupervised models like K-means, to identify anomalies or outliers in feature vectors, such as values or states from machines or production lines. Supervised models require normal and anomalous feature vectors, while unsupervised models detect anomalies based on deviations from primary clusters in the training data.

The Semantic layer extracts meaning from the operator/device's environment and interactions, integrating three input types (See Fig. 3.1) to enhance system accuracy.

- CNNs classify discrete and analog controls to predict their values,
- ML algorithms detect anomalies in combinations of interpreted values,
- Transformers are used to answer NL questions from the operator in the context of a specific machine, process, or task.

Regarding the last point, the transformer-based answers significantly improve problem resolution for the operator. Combining SMEs' descriptions with existing technical documentation determines the transformer's context. The responses are transmitted to the middleware layer, where they are stored for subsequent retrieval by the AR Physical layer. This retrieval process allows for the presentation of new information within the operator's field of view. The proposed architecture considers displaying information such as the next step or anomalous information next to a control. Fig. 3.4 shows an example of this functionality.

Transformers for QA

In complex industrial tasks, AR systems provide visual cues to assist operators. By adding a Semantic layer (explained in Sec. 3.3.4) that utilizes CNNs to extract information and context, operators can validate actions, detect anomalies, and simplify interactions. The Semantic layer also enables automatic progression to the next step based on confirmed actions.

During MRO tasks, operators may encounter questions that require answers found in technical documentation and procedure manuals. These inquiries should be resolved quickly to avoid risks or costs resulting from delays. The operators may not have the necessary knowledge, leading them to consult either the technical documentation or a remote expert. One of the most promising neural network architectures, transformers, can address this issue. Transformers can help interpret technical documentation and procedure manuals, even when written in NL, enabling operators to find the answers they need promptly.

Transformers outperform recurrent networks by incorporating attention mechanisms and acquiring language semantics through extensive text training. They possess high-level capabilities in NL processing, including text classification, chatbots, text summarization, and language translation. Popular transformer architectures include RoBERTa (Liu et al. 2019), DistilBERT (Sanh et al. 2019), and Google's T5 (Raffel et al. 2020). In QA tasks, transformers can be specialized to provide context-based answers to NL questions. Training can be performed on large datasets like SQuAD (Stanford Question Answering Dataset) (Rajpurkar et al. 2016), or pre-trained transformers can be used for their ability to understand various question formulations in NL. Fine-tuning, as detailed in section 3.5, further enhances their accuracy, particularly when considering real industrial contexts.

The need for knowledge may arise during MRO tasks, with or without AR guidance. This study evaluates the availability of tools such as QA based on technical documentation and SME knowledge to analyze whether it improves task efficiency in complex environments such as shop floors.

Connecting transformers with AR for multimodal interaction

Transformers have proven to be an effective method for answering technical documentation questions in NL, offering valuable benefits such as streamlining the search process for specific problems and reducing cognitive load. In the evaluated application, the transformer's responses are integrated with AR technology, linking NL answers to physical locations and specific elements in the spatial mapping (see Fig. 3.4). This integration aims to provide precise indications of the elements of interest in the operator's physical environment alongside the NL response to a query (Fig. 3.3).

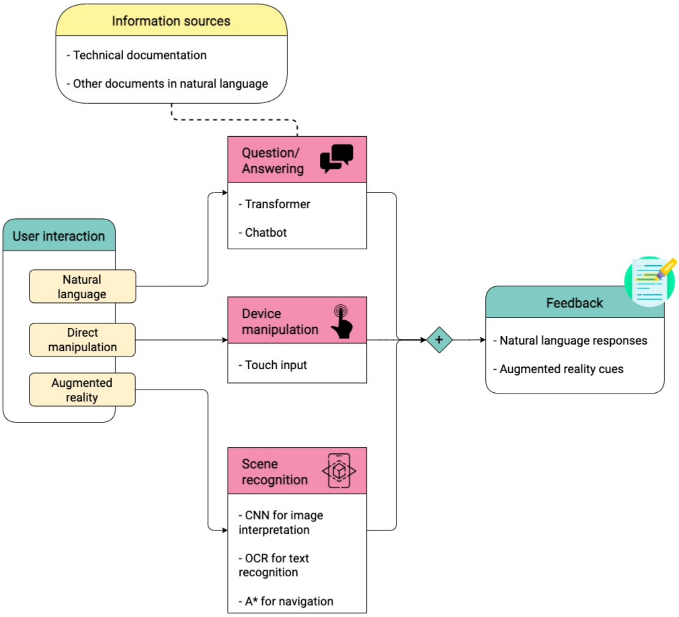


Figure 3.3: Users can interact multimodally to obtain AR and NL feedback.

The proposed method for linking questions and AR responses involves preprocessing the technical documentation and utilizing existing 3D scanning technology to map the environment of interest. To achieve this, we add extra text to the technical documentation to identify elements of interest in the 3D scanned mesh unequivocally. The relevant elements in the industrial environment, such as panels, controls, indicators, and actuators, are identified and assigned unique identifiers in the 3D mesh of the shop floor area.

The connection between the identified elements and the technical documentation is established through two steps:

- Additional sentences are incorporated into the context, indicating the specific identifiers assigned to the elements of interest in the physical space. For instance, Fig. 3.6 illustrates four marked nozzles (0001, 0002, 0003, and 0004) of an Extruder machine, and their corresponding identifiers are listed in lines 26 to 28 of Listing 3.1.
- In cases where the text does not clearly describe the element in question, the previous sentences are extended to provide the necessary context, enabling the transformer to obtain the required answer when asking a question.

The interaction between the operator, the transformer, and the AR system is depicted as follows (Fig. 3.4): The operator asks a question in NL, and the system forwards it to the transformer, which predicts and returns the answer. However, in certain cases, additional information is necessary to pinpoint the element referred to by the operator in the AR space. In such situations, a second internal question is sent to the transformer to obtain the identifier of the element of interest. Therefore, the application receives a response comprising two elements: (1) the answer provided by the transformer and (2) the ID of the element of interest associated with a specific location in the AR environment.

Explicit statements are added to identify specific controls in subsequent questions (see Listing 3.1, lines 26-28) to ensure that the transformer can accurately relate the response to a physical control in the real environment.

It is worth noting that the transformer used in this study belongs to the *Extractive Open QA* family. The answers are generated solely from the context and do not introduce new content. This approach is recommended when precise answers are required, based solely on the available technical documentation.

3.4 System implementation

The proposed architecture, its main elements, multimodal capabilities, and synergies are described, along with a real-world implementation and evaluation. This study analyzes the usage of two machines, an Extruder, and an Injector, for creating filaments and molds from a polymer mixture. The process requires various machine parameters, such as nozzle temperatures, which vary based on the polymer type. The client-server architecture for implement-

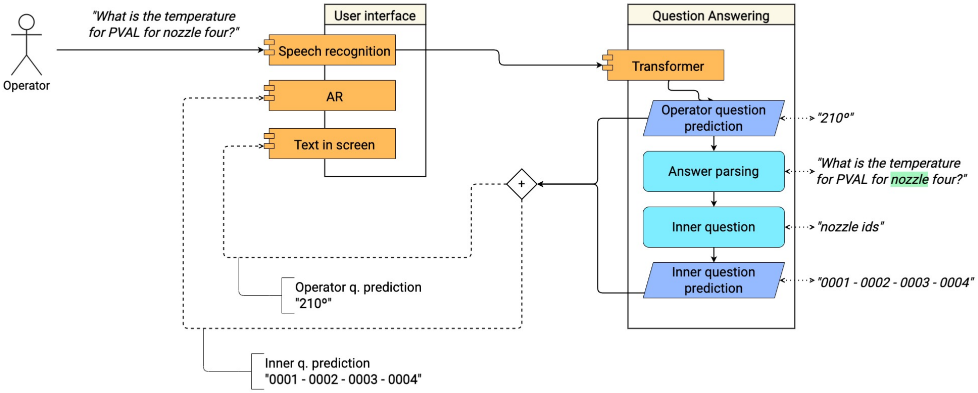


Figure 3.4: Linking answers in NL to AR cues.

ing, validating, and verifying the proposed system is detailed in section 3.4.1 and 3.4.2, respectively.

3.4.1 Client-side application

The primary application that the operator interacts with has been developed using the Unity engine, allowing for the development of interactive, cross-platform apps for desktop and mobile devices. The app was tested on an iPad Air (4th generation). Since the system relies on AR for spatial location and operator guidance, a tool that facilitates AR application development in Unity was required. Vuforia was chosen for environment scanning since it offers prior 3D spatial mapping of the workspace and allows AR elements placement during application runtime. The initial scanning procedure is necessary as the system's primary objective is to guide the user during various tasks and processes. The result of the 3D scanned environment using an iPhone 13 with a LiDAR sensor can be seen in Fig. 3.5. While alternative methods like 3D laser scans exist for environment scanning, their high cost led us to choose Vuforia Area targets as the preferred technology. After the scan process, this tool allows for environment detection and tracking, so placing AR elements on the environment is possible. Additionally, a marker placed close to the control of interest aids in interpreting the CNN by eliminating the perspective of the image taken from the control. Regarding operator guidance on the shop floor, the Unity native's implementation of the A^* algorithm is utilized.

To distribute and optimize the Semantic layer functionality between the client and the server, the client performs local inferences from CNNs trained using the Open Neural Network Exchange standard (ONNX) (The Linux Foundation 2017). The proposed system uses the open-source Keras library (Chollet 2015) to develop the architecture and train the model, which is subsequently exported to ONNX for inference on the mobile device, enabling the device to perform classification and regression tasks, which interpret images semantically.

As explained in Sec. 3.3.4, labeling the elements of interest (e.g., buttons, machines, nozzles) is required for the system to interpret the transformer's responses and identify the relevant elements in the environment. Fig. 3.6 displays the labels used for the temperature nozzles of the Extruder.

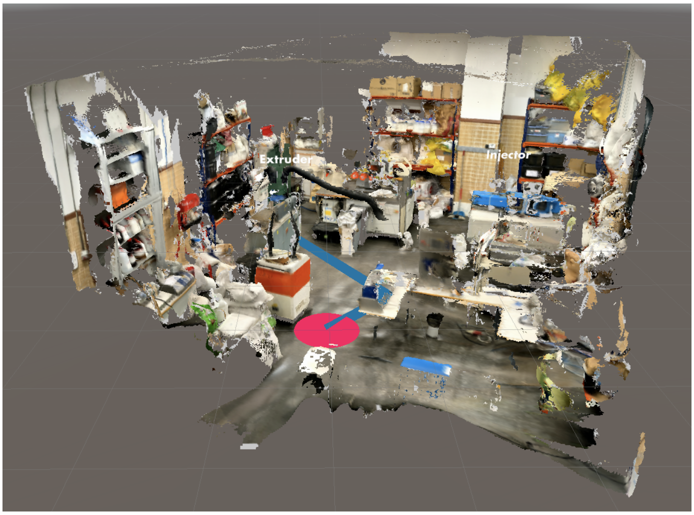


Figure 3.5: Scanned mesh (Unity).

3.4.2 Server-side application

To improve the system's overall efficiency, tasks involving complex DL models, such as anomaly detection models or transformers for QA, are distributed and solved on a server instead of the operator's device. This approach offloads the computational burden, leading to faster response times and reduced battery consumption.

For the evaluated implementation, the FastAPI framework for Python was used to run a transformer on the server. Specifically, the transformer was 80% 1x4 Block Sparse BERT-Large (uncased) Fine Tuned on SQuADv1.1 (Zafrir et al. 2021) due to its high prediction accuracy for the questions asked. Fur-

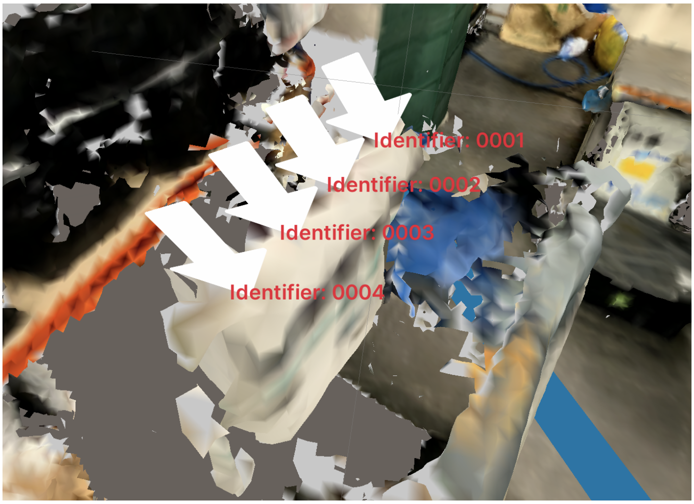


Figure 3.6: Link between a physical element and transformer identifier (Unity).

ther insights into the decision-making process and analysis of the results are described in section 3.6.

Transformer selection

To establish a relationship between the transformer's response (Semantic layer) and the visual AR cue (AR Physical layer), controls are linked with IDs, such as 0001 or 0002, as shown in Listing 3.1. This linking allows triggering a second question to the transformer if a keyword like "nozzle" or "funnel" is detected in the operator's query. The result from the second question returns the control identifier and activates the cue in AR, as depicted in Fig. 3.4.

Four state-of-the-art QA transformers were evaluated to provide NL interaction, all available at Hugging Face.

1. Intel/bert-large-uncased-squadv1.1-sparse-80-1x4-block-pruneofa
2. bert-large-uncased-whole-word-masking-finetuned-squad
3. csarron/bert-base-uncased-squad-v1
4. csarron/mobilebert-uncased-squad-v2

Fourteen questions related to the context were performed to select the QA transformer for this study. The average scores and their results were calculated (refer to Table 3.1 and Table 3.2). The chosen model was Intel/bert-large-uncased-squadv1.1-sparse-80-1x4-block-pruneofa. It is important to note that the validity of the answers depends on the technical documentation and specific question formulation. The score value alone should not be the only factor in choosing a transformer. The possibility of using an Ensemble model to choose the highest-scoring model for a specific question should also be considered. It is worth mentioning that the study's focus is not to compare transformer models but instead to properly integrate their capabilities into an AR environment.

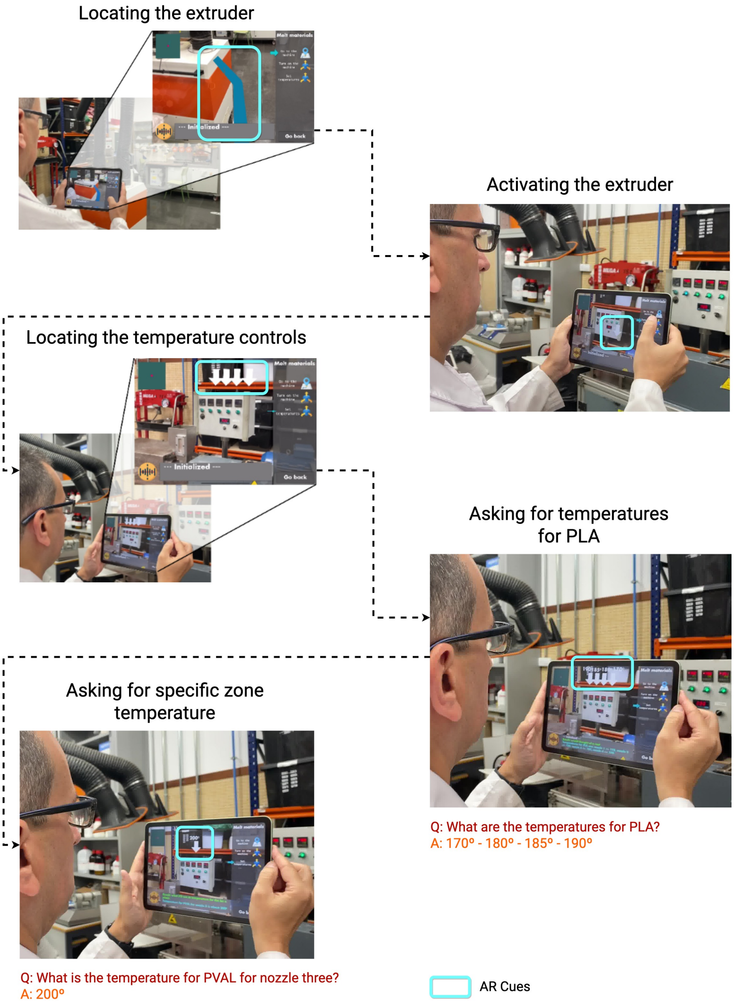
3.4.3 System operation

The presented architecture's benefits over a traditional approach have been evaluated using two applications. The first application utilizes AR, Transformers for NLP, and CNN for action validation, while the second only provides AR cues for operator guidance. The developed applications allow preparing new materials using different polymers from a list of possible processes. The AR Physical layer detects the current environment to ensure the operator is in the work area. The AR system guides the operator to the first machine using an AR route and a mini-map. Once in front of the machine, a sequence of AR cues shows the operator the actions to be performed if the process includes handling different elements. An example of the flow of use of the developed application is shown in Fig. 3.7. Again, the complete system, implemented in the first application, allows for asking questions in NL, thus generating responses through AR cues and textual information, and is also capable of interpreting analog values, such as pressure gauges.

3.4.4 Implementation guidance

To generalize the benefits of the proposed architecture and facilitate its incorporation into various industrial environments, this research evaluates a specific implementation in a particular environment. The following common aspects should be considered:

1. **Technical documentation compilation** Collecting technical documentation from various sources is necessary for NL questions. The data is usually in PDF files or paper documents, but SMEs may have additional information acquired from experience. Specific tools are required to extract this data efficiently.



Locating the extruder

Activating the extruder

Locating the temperature controls

Asking for temperatures for PLA

Q: What are the temperatures for PLA?
A: 170º - 180º - 185º - 190º

Asking for specific zone temperature

Q: What is the temperature for PVAL for nozzle three?
A: 200º

Figure 3.7: Extruder workflow using the application, setting the nozzle temperatures. Visual validation, AR guiding, and voice interaction.

2. **Operatable elements** Define the machines the architecture needs to understand and operate. Automating the process of determining operable machines and controls would streamline development.

3. **AI model definition** The architectures need to be trained with several models for each control and its variations to achieve better results since

elements such as gauges or buttons can significantly vary, even in the same machine. Simple CNN architectures have shown excellent results, especially with the preliminary perspective removal stage, requiring only a few pictures.

4. **Environment scanning** Enable the operator to move freely within the shop floor, necessitating 3D environment tracking. While 3D scanning and recognition tools are recommended, 2D markers from specific machines could serve as an alternative. Current AR systems generally offer these possibilities with specific authoring tools.

5. **Task definition** To highlight the AR cues linked to specific tasks, a good knowledge of what processes and tasks the operator will perform is necessary. The application should be flexible enough to manage these abstractions and highlight related elements to specific operator questions.

6. **Context and environment linkage** The architecture's advantage is the strategy used to link responses from the transformer and AR elements, as described in section 3.3.4. However, the implementation may differ for different cases.

3.5 System evaluation

A suitable environment was selected to evaluate the proposed system, focusing on tasks involving creating new materials using polymers mixed with various materials at different temperatures. These tasks require the use of multiple machines located in different areas. The system evaluation was conducted using the following elements:

- Three groups of people were evaluated (see Sec. 3.5.1), a group without AR, a group with traditional AR, and a group with the proposed system (i.e., AR and Semantic layer).

- Three work-related questions were posed to operators to measure the resolution time. Two groups of operators were tested, one with physical documentation and the other with the transformer integrated into the proposed AR system.

- A Likert questionnaire was used for qualitative analysis to evaluate two groups of operators, one with traditional AR and the other with the proposed semantic AR.

The chosen evaluation criteria follow similar studies in the literature to obtain qualitative and quantitative measures for assessing the suitability of the proposed AR system.

3.5.1 Participants

The proposed system was evaluated through the developed application by three groups of operators, each consisting of six operators with experience on the shop floor. Although the operators were familiar with industrial processes, they were not acquainted with the tasks evaluated, allowing for assessing potential application improvements.

Group A: The first group used an application that incorporated semantic AR through CNNs and transformers to complete the designated tasks. Additionally, this group could ask questions in NL when performing a specific task, providing text and AR cues (if possible).

Group B: The second group had access to an application that only used AR to display visual cues, which is the current standard in AR applications used in the industry. They also had access to technical documentation in a traditional electronic format (i.e., PDF files).

Group C: The third group did not have access to the application or any AR support. They were given access to the same source of technical documentation as group B and a list of tasks to be performed.

3.5.2 Task description

The system evaluation involved the operators performing eight tasks, comprising a complete process that required using two machines, the Extruder, and the Injector, described in section 3.4.

Task 1, Extruder Location: This task required locating the machine responsible for dissolving the selected material within the manufacturing plant. While groups A and B successfully used AR guidance, group C faced challenges due to limited information in the technical documentation.

Task 2, Turn on the Extruder: Upon locating the Extruder, the operators had to power it on. This step proved to be straightforward for all groups, involving a standard control and clear location visibility. Notably, group A benefited from the assistance of a CNN for automatic switch validation.

Task 3, Nozzles temperatures setup: Operators were tasked with establishing the temperatures of the four nozzles, which vary depending on the melting material. Group A utilized AR cues (e.g., arrows pointing at the nozzles) and NL interaction for easier execution (e.g., texts indicating the temperatures of the nozzles, as additional AR cues), while groups B and C faced difficulties relying solely on technical documentation.

Task 4, Extruder material insertion: In this step, operators introduced the compound into the Extruder's top funnel. AR cues provided additional assistance to groups A and B.

Task 5, Injector Location: Similar to Task 1, this step required locating the Injector machine. Groups A and B used AR cues, while group C encountered challenges without this assistance.

Task 6, Turn on the Injector: Activating the Injector involved using a standardized control, similar to task 2. Groups A and B used AR cues to find the control, while group C had minor difficulties finding the control location due to its concealed location.

Task 7, Injector material insertion: Operators needed to locate the Injector's funnel to insert the new mixture. No significant differences were observed among the three groups.

Task 8, Purge and injection: For this task, the operator had to use the Injector's Human Machine Interface (HMI) to perform purging and injection. Groups A and B received AR assistance on the HMI, while group C relied on technical documentation and encountered some challenges using the HMI menus.

To evaluate the possibility of asking in NL, operators from group A had to ask three questions to see the response from the system. These questions were:

Question 1: Asking for the temperature range of the nozzles for the Extruder for a specific material different from PLA.

Question 2: Query about a machine's location on the shop floor.

Question 3: Inquiry regarding the range of values on the pressure during the injection process.

In addition, a six-question questionnaire was administered to operators from groups A and B using a 5-point Likert scale (see Table 3.6). The questionnaire aims to compare and evaluate the use of standard AR applications versus

Index	Question	Selected transformer prediction
0	*What is PLA?*	polylactide
1	*What are all the temperatures for melting PLA?*	170º - 180º - 185º - 190º
2	*What are the temperatures for PLA?*	170º - 180º - 185º - 190º
3	*What is the range of temperatures for melting PLA?*	170º - 180º - 185º - 190º
4	*What is the temperature for the first nozzle for PLA?*	170º
5	*What are the temperatures for PVAL?*	190º - 195º - 200º - 210º
6	*What is the temperature for PVAL for nozzle four?*	210º
7	*How do I start using the extruder?*	setting the nozzle controls to the desired temperature
8	*What is the next step after setting the temperatures?*	pour the materials into the funnel
9	*Where is the extruder's funnel?*	top of the machine
10	*Nozzle id's*	0001 - 0002 - 0003 - 0004
11	*nozzle id's*	0001 - 0002 - 0003 - 0004
12	*Activation button id*	0005
13	*funnel ID*	0006

Table 3.1: Transformer predictions to questions (Intel/bert-large-uncased-squadv1.1-sparse-80-1x4-block-pruneofa).

Model index	Mean score
1	±68.46
2	±30
3	±53.38
4	±28

Table 3.2: Scores for the 4 QA transformers tested.

adding a semantic layer on top of it with the capability of asking questions in NL. The results of the questionnaire can be found in section 3.6.

3.6 Analysis and results

Table 3.7 presents the task completion times for the three groups. Table 3.8 compiles the time required to answer three specific questions for groups A and C. As a reminder, operators from group A could use NL to ask questions to the selected transformer. Group B and group C operators had to search for answers in technical documentation.

A two-factor ANOVA test was conducted, with one factor using repeated measures, to assess whether the group impacts the task execution time. Table 3.3 displays the means and standard deviations of task execution times for each group (refer to Fig. 3.8 in appendix 3.9 for a comparative view). The test results indicate that there is a statistically significant difference between the groups, regardless of the task performed ($F(7, 105) = 288.39$; $p < 0.001$; $\eta^2 = 0.951$). However, the interaction between group and task was also significant, suggesting that the execution time of a task depends on the group performing it ($F(14, 105)$ =8.30; $p < 0.001$; η^2 =0.525).

Tasks 2, 4, 6, and 7 show no statistically significant differences in execution times between the operator groups. Conversely, in tasks 1, 5, and 8, the execution times of operators in group C, who only used the documentation, were significantly longer than those of operators in groups A and B, who used the applications; no significant differences were found between groups A and B. For task 3, the execution time of operators in group C was significantly longer than that of operators in group A; no significant differences were observed between groups A and B or B and C.

Table 3.4 presents the p-values of the pairwise comparisons with Bonferroni correction to adjust the significance level of each pairwise comparison between the groups and tasks.

	Task							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	Mean (Sd)	*Mean (Sd)*	*Mean (Sd)*	*Mean (Sd)*	*Mean (Sd)*	*Mean (Sd)*	*Mean (Sd)*	*Mean (Sd)*
Group								
Group A	18,67 (3,56)	5,50 (2,26)	50,67 (15,95)	23,00 (6,81)	19,67 (4,93)	12,00 (3,29)	6,83 (3,82)	89,83 (8,91)
Group B	18,50 (2,26)	4,67 (1,37)	64,17 (11,58)	15,83 (4,71)	18,83 (6,31)	10,50 (2,59)	6,83 (2,64)	85,50 (11,52)
Group C	37,83 (8,04)	5,67 (3,01)	77,83 (14,58)	16,83 (5,49)	53,17 (10,07)	14,83 (7,91)	7,67 (2,25)	126,00 (23,38)

Table 3.3: Descriptive execution times of tasks and statistical contrasts.

	Task							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Group A Vs. Group B	1	1	0,358	0,141	1	1	1	1
Group A Vs. Group C	< **0,001**	1	**0,014**	0,247	< **0,001**	1	1	**0,004**
Group B Vs. Group C	< **0,001**	1	0,345	1	< **0,001**	0,501	1	**0,002**

Table 3.4: Two to two comparisons P-values (Bonferroni correction).

The response times to the questions posed were analyzed, and the results are presented in table 3.5 (see also Fig. 3.9 in appendix 3.9). The analysis indicates a statistically significant difference between groups A and C concerning the three proposed questions. The response times of operators who worked with the application were significantly lower than those of operators who performed tasks with documentation for all three questions ($p < 0,001$, $p < 0,001$, and

$p < 0,001$, respectively). However, the interaction between the group and the task was not significant.

	Question			**Tests of within-subjects effects**	
	1	**2**	**3**	**Group**	**Group*Question**
	Mean (Sd)	*Mean (Sd)*	*Mean (Sd)*	**F(g.l.);** $p - valor(\eta^2)$	**F(g.l.);** $p - valor(\eta^2)$
Group				$F(2;20) = 3,93;$ $p = 0,036(0,274)$	$F(2;20) = 0,57;$ $p = 0,572(0,054)$
Group A	19,67 (8,89)	10,67 (5,79)	15,17 (5,08)		
Group C	87,33 (12,31)	70,83 (19,45)	87,83 (21,74)		

Table 3.5: Descriptive answer times and statistical contrasts.

Table 3.6 describes and compares the scores given to each statement based on the group of workers. The results indicate that operators from group A showed more significant agreement than those from group B for the questions *"It helped me solve technical questions seamlessly"* and *"I can access additional information effortlessly"*. These answers suggest that workers in group A found the system more helpful and easier to use than group B workers.

Question	**Group**, *median (IR)*		**Mann-Whitney U test**	
	Group A	**Group B**	**U**	p**-value**
I think AR is helpful	5 (5 - 5)	5 (4 - 5)	15	0,523
I feel I can finish the tasks faster	4,5 (4 - 5)	4 (4 - 5)	13,5	0,423
I felt more confident and safe using the application	5 (5 - 5)	5 (5 - 5)	15	0,317
It helped me solve technical questions seamlessly	5 (4 - 5)	2,5 (2 - 3)	1	**0,005**
I can focus better on the task	4,5 (4 - 5)	4 (4 - 5)	15	0,575
I can access additional information effortlessly	5 (5 - 5)	2 (2 - 2)	0	**0,002**

IR: interquartile range

Table 3.6: Descriptive and comparative scores to the questions raised about the performance of the tasks.

The results show that the proposed semantic layer reduces the time required for tasks of a more cognitive nature and provides greater comfort and security in the qualitative assessment of results. Consequently, group A shows better results and higher levels of satisfaction. The evaluation of the proposed solution confirms the significant advantages that can be obtained by complementing current AR systems in industrial environments with the proposed semantic and multimodal layers. These findings align with similar studies such as Rasmussen et al. on workspace awareness using a multi-camera solution (Rasmussen et al. 2022), Eversberg et al. on cognitively assisted AR through digital twins

(Eversberg et al. 2022), Wang et al. on spatial cognition (Z. Wang et al. 2021), and Zhang et al. on the benefits of multimodal interaction (Zhang et al. 2022). It is important to acknowledge that further investigation and analysis are necessary to address the open questions surrounding the semantic layer and to refine the system's capabilities in future research.

3.7 Discussion

The present study investigated the effectiveness of a proposed semantic layer in augmenting industrial workers' AR tasks. The analysis of the study's data revealed several significant findings. Firstly, the two-factor ANOVA test demonstrated a statistically significant difference in task execution times between the three groups (A, B, and C). Operators in group C, who relied solely on technical documentation, exhibited significantly longer task execution times than those in groups A and B, who used the proposed application-based semantic layer. Moreover, task completion times varied significantly across different tasks, suggesting that the semantic layer's impact depends on the task's nature. The response times to specific questions showed that operators in group A, using the semantic layer, had significantly lower response times than those in group C, who used only technical documentation. This finding indicates that the semantic layer can streamline obtaining information and answering queries, potentially improving workers' efficiency. The likert-scale evaluation indicated that operators in group A expressed higher levels of agreement with statements related to the proposed system's helpfulness, confidence, and ease of use. These results suggest that the semantic layer contributes positively to the overall user experience and satisfaction.

These outcomes imply that the semantic layer has the potential to enhance workers' cognitive capabilities and enable them to perform complex tasks more effectively. The advantages observed in tasks requiring more cognitive processing further emphasize the importance of addressing human-computer interaction challenges in AR systems. The seamless integration of contextual information and multimodal interaction in the proposed solution potentially alleviates cognitive load, improving task performance.

We acknowledge several limitations in this study:

1. The sample size for each group was relatively small, which may limit the generalizability of the results. Future studies with more extensive and diverse samples are necessary to validate and extend our findings.

2. While various tasks were considered, the chosen tasks may only partially represent some possible scenarios in industrial settings. Further investigation across a broader range of tasks would enhance the study's comprehensiveness.

3. The study was conducted in a controlled environment, and real-world industrial conditions may introduce additional complexities not fully accounted for in this research.

3.8 Conclusions

In the Industry 4.0 era, the operator plays a crucial role in incorporating the mechanisms that benefit from this revolution. This is becoming a fundamental element of the new Industry 5.0 definition that has, as one essential component, the human-centric approach and the value-added work of future operators. AR is an essential technology to enable this evolution toward an automated, efficient, and user-centric environment. As tasks performed by operators on the shop floor become increasingly complex, there is a growing need for qualified labor. However, remote assistance or access to an SME may only sometimes be available. In critical situations, the operator's cognitive load can increase stress levels, affect problem resolution, and compromise safety.

This study proposes a system consisting of four layers to assist operators in the Industry 4.0 era. The AR Physical layer is the primary interface between the operator and the shop floor. Multimodal interaction using NL, AR, and direct manipulation helps operators learn the system more quickly and interact with it more naturally. By incorporating ML and DL models, the system further assists operators in understanding the environment (e.g., using CNNs to get states and values of non-sensorized machines), enabling them to interact in NL and receive verbal and visual AR feedback.

The proposed system was implemented and evaluated with three groups of operators, and its real-life use cases have demonstrated significant benefits:

- **Enhanced productivity and reduced errors**: The first group used the proposed system with a semantic layer, and they experienced a substantial increase in task completion efficiency and a notable reduction in errors. This improvement showcases how AR technology, when integrated with semantic layers, can significantly enhance operator performance in complex industrial tasks.

- **Comparison with limited application**: The second group, using a limited application version without the semantic layer, struggled to match the efficiency and accuracy of the first group. This comparison underscores the importance of incorporating semantic layers in AR solutions to improve cognitive tools and environmental awareness.
- **Importance of semantic layers**: In contrast, the third group, which relied solely on technical documentation, faced longer task completion times and a higher error rate. These findings further highlight the importance of integrating AR solutions with semantic layers, as they outperformed traditional documentation in assisting operators.

These findings highlight the importance of including semantic layers in AR solutions to enhance cognitive tools and environmental awareness, and are aligned with recent studies, such as the one by Eswaran et al. on augmented opportunities in the industry (Eswaran et al. 2023).

The proposed system effectively reduces task completion time and minimizes errors, ultimately contributing to improved operational efficiency and operator performance. As the Industry 4.0 revolution progresses, it is crucial to continue exploring and refining such AR systems to empower operators, mitigate cognitive load, and foster a safe and productive industrial setting. As Industry 4.0 and 5.0 advance, this research paves the way for future investigations in AR systems. There is a need for further refinement and exploration of novel interactions to enhance AR system efficiency for operator assistance. Integrating AI-driven decision support can empower operators to handle complex tasks effectively. Additionally, extending semantic layers within AR solutions shows promise in augmenting operators' cognitive tools and environmental awareness. These future research directions can significantly advance AR systems, creating a more user-centric, efficient, and safe industrial environment. These advancements hold the potential to revolutionize various industries, from manufacturing and logistics to healthcare and maintenance, providing tangible benefits to operators and organizations alike.

Declarations

Conflict of interest

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

Funding

The authors did not receive support from any organization for the submitted work.

Data Availability

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

Appendices

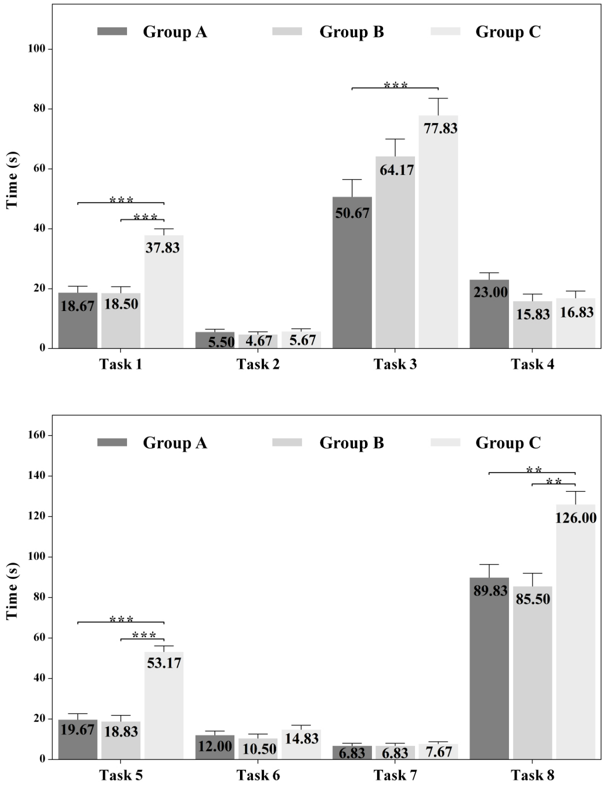
3.8.1 Transformer context

```
context = """
Extruder information:
- PLA is Polylactide.
- For melting PLA, 170° - 180° - 185° - 190°.
- PLA nozzle one/first: 170°.
- PLA nozzle two/second: 180°.
- PLA nozzle three/third: 185°.
- PLA nozzle four/fourth: 190°.

- PVAL is Polivinylalcohol.
- For melting PVAL, 190° - 195° - 200° - 210°.
- PVAL nozzle one/first: 190°.
- PVAL nozzle two/second: 195°.
- PVAL nozzle three/third: 200°.
- PVAL nozzle four/fourth: 210°.

Extruder tasks:
1° Turn on the machine turning the activation button to the
   left.
2° Heat up the machine by using setting the nozzle controls to
    the desired temperature.
3° When the machine is heated, pour the materials into the
   funnel, found at the top of the machine.
4° The material will appear through the fourth nozzle.

-------------------------------------

```

```
Extruder IDs:
- Nozzle ID ( 0001 - 0002 - 0003 - 0004 ).
- Activation button ID ( 0005 ).
- Funnel ID ( 0006 ).
"""
```

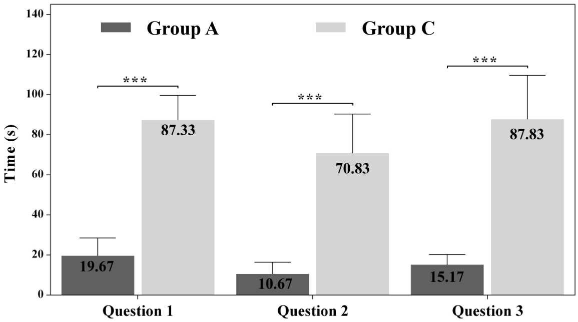
Listing 3.1: Transformer context based on technical documentation (Python language).

3.9 System evaluation



$**p < 0.01, ***p < 0.001$

Figure 3.8: Comparative of tasks, group A (Semantic AR) Vs. group B (Only AR) Vs. group C (Traditional).



$***p < 0.001$

Figure 3.9: Comparative of answers (Group A Vs. Group C).

	Group A						Group B						Group C						Description
	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	Op. 6	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	Op. 6	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	Op. 6	
Task 1	16s	21s	18s	24s	19s	14s	20s	17s	21s	15s	18s	20s	34s	45s	28s	32s	49s	39s	*Extruder locating*
Task 2	4s	4s	3s	9s	7s	6s	4s	7s	3s	5s	4s	5s	9s	3s	8s	2s	4s	8s	*Extruder activation*
Task 3	52s *	42s	81s *	38s	51s	40s	72s	64s	82s	57s	61s	49s	87s	98s	67s	57s	76s	82s	*Temperature setup*
Task 4	23s	21s	22s	28s	12s	32s	15s	19s	23s	10s	16s	12s	24s	14s	9s	21s	14s	19s	*Material insertion*
Task 5	20s	17s	17s	22s	28s	14s	24s	13s	17s	16s	29s	14s	54s	64s	43s	43s	66s	49s	*Injector locating*
Task 6	14s	8s	17s	12s	9s	12s	10s	7s	12s	13s	8s	13s	19s	7s	13s	17s	6s	27s	*Injector activation*
Task 7	6s	4s	14s	4s	5s	8s	5s	6s	5s	12s	7s	6s	7s	8s	6s	6s	7s	12s	*Material insertion*
Task 8	78s	85s	97s	87s	89s	103s	91s	101s	83s	92s	69s	77s	110s	143s	89s	131s	129s	154s	*Purge and injection*

Table 3.7: Task time results (Group A Vs. Group B Vs. Group C).

	Group A						Group C						Question
	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	Op. 6	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	Op. 6	
Question 1	10s	35s *	14s	21s	23s	15s	98s	76s	89s	103s	87s	71s	*Temperature range*
Question 2	6s	9s	16s *	10s	4s	19s *	68s	45s	70s	59s	102s	81s	*Machine location*
Question 3	13s	17s	8s	13s	17s	23s *	56s	113s	83s	91s	110s	74s	*Pressure values*

* Ask again to transformer (bad answer to a first question)

Table 3.8: Question time results (Group A Vs. Group C).

Chapter 4

Large Language Models for in situ knowledge documentation and access with Augmented Reality

Izquierdo-Domenech, J., Linares-Pellicer, J., & Ferri-Molla, I. (2023). Large Language Models for in Situ Knowledge Documentation and Access With Augmented Reality. 10.9781/ijimai.2023.09.002

Augmented reality (AR) has become a powerful tool for assisting operators in complex environments, such as shop floors, laboratories, and industrial settings. By displaying synthetic visual elements anchored in real environments and providing information for specific tasks, AR helps to improve efficiency and accuracy. However, a common bottleneck in these environments is introducing all necessary information, which often requires predefined structured formats and needs more ability for multimodal and Natural Language (NL) interaction. This work proposes a new method for dynamically documenting complex environments using AR in a multimodal, non-structured, and interactive manner. Our method employs Large Language Models (LLMs) to allow experts to describe elements from the real environment in NL and select corresponding AR elements in a dynamic and iterative process. This enables a more natural and flexible way of introducing information, allowing experts to describe the environment in their own words rather than being constrained by a predetermined structure. Any operator can then ask about any aspect of the environment in NL to receive a response and visual guidance from the AR system, thus allowing for a more natural and flexible way of introducing and retrieving information. These capabilities ultimately improve the effectiveness and efficiency of tasks in complex environments.

4.1 Introduction

Augmented Reality (AR) and its capability for superimposing synthetic elements on top of real environments has been, indeed, a key factor in the rise of Industry 4.0 (Kagermann et al. 2013). There are numerous definitions of AR, with one of the most well-known being the one proposed by Azuma: *"AR is a system that supplements the real world with virtual (computer-generated) objects that appear to coexist in the same space as the real world"*(R. Azuma et al. 2001). Billinghurst also defines AR as an interactive experience in which real-world objects are enhanced by computer-generated perceptual information (Mark Billinghurst, Clark, and G. Lee 2015); nonetheless, AR had been applied in the industry field even before such definitions (Caudell and Mizell 1992). The ability to enhance environments with this technology has been utilized in various industrial applications, including product design, process design and control, maintenance processes, and learning. Its benefits have been widely demonstrated. Examples include using AR to visualize and manipulate 3D models during the design process, to provide real-time guidance and instructions for Maintenance, Repair, and Overhaul (MRO) tasks, and to enhance training and education programs through interactive simulations and visualizations. Industry 4.0 lays on several pillars, such as the Industrial Internet of Things (IIoT), cloud computing, additive manufacturing, and AR; however, the latter is unique in that it focuses on the human factor (C. H. Chu et al. 2021). On the other hand, shop floor operators have seen how their roles and required knowledge have been transformed to match completely different profiles, leading to a need for more skilled operators with advanced education in the use of technologies (Jaschke 2014; Marino et al. 2021; Gattullo et al. 2019; Masood and Egger 2019). AR can serve as an assistive technology to support shop floor operators in these environments.

The evolution of Artificial Intelligence (AI) and its integration into the industry is one the most critical components behind what it has been defined as Industry 4.0. It can lead to a shift in the role of workers towards more value-added tasks, which can increase job satisfaction and improve overall productivity. By incorporating these technologies, manufacturers can create a more efficient and flexible workforce, ultimately leading to a better future of work in the manufacturing industry (X. Xu et al. 2021). The current proposal is a step forward in achieving these objectives.

The proper training of operators is always the first challenge to be met to guarantee their subsequent ability to work effectively and efficiently. Apart from the emergence of new possibilities in this training, such as multimedia tools, Virtual Reality (VR) and AR, 'one-to-one' training is still very beneficial. Direct interaction is still a precious element in training in complex contexts, such as laboratories, control centers, and shop floors in the industry. However, after training, operators need immediate access to documentation that can solve new doubts or problems that may arise at any time. On these occasions, the presence of a specialized expert is very rare or impossible. In contrast to initial training, it is unlikely that the expert or Subject Matter Expert (SME) will be available for the operator's day-to-day work (Garza et al. 2013).

It is, therefore, essential to develop solutions that enable the operator to access information quickly and efficiently in case of need. In this context, there is a need to know what means and interrogation mechanisms are available. An ideal solution would offer multiple interaction options, including operators' ability to ask questions and receive answers in Natural Language (NL), as discussed in (Izquierdo-Domenech, Linares-Pellicer, and Orta-Lopez 2022). It is necessary to consider technical documentation and expert knowledge to provide adequate answers. It is very interesting to offer not only the possibility to give answers in NL to the operator based on the technical documentation, but also the information provided by the experts; however, technical documentation and expert information are typically unstructured, presenting a significant challenge for creating operator assistance systems in complex environments. This often leads to the creation of time-consuming, *ad hoc* solutions for different environments, which can be overwhelming and cause an excessive workload, specially in environments with a high degree of diversity or work volume. Therefore, it is necessary to address the issue of unstructured technical documentation and expert information to create operator assistance systems that are efficient, effective, and sustainable in complex environments.

One of the latest technologies based on Deep Learning (DL), Large Language Models (LLMs), can aid in NL interaction and information retrieval by operators. The proposed system enables experts to train operators on the job, allowing for the system to serve as a knowledge source for subsequent consultations. This type of learning, known as *in situ* learning or Scenario Based Training (SBT), has been demonstrated to enhance knowledge acquisition and retention according to prior research (Lave and Wenger 1991).

One-on-one training, primarily SBT, is still the best way to provide knowledge of complex systems. Combining SBT with an automatic acquisition of information, adding multimodal elements, avoiding the need to structure or post-process the information, and making this knowledge available to the operators, is an element of great interest, and the main interest of this work.

Another issue when discussing complex environments, such as shop floors, is documentation access. The complexity of these environments tends to increase exponentially, as does the specialized knowledge and technology required by operators. Documentation about the different machines spread over a shop floor is critical for making them work and learning and maintenance processes. However, traditional forms of documentation, such as paper manuals, can be cumbersome and difficult to use due to their lack of portability, the potential for inaccuracies, and interpretation issues. As Ventura highlights, these issues can make it difficult to effectively utilize this type of information (Ventura 2000). To address these challenges, several alternatives using AR technology have been proposed. For example, AR can be used to display machine-specific documentation on a user device in real-time as they work, allowing for easier reference and reducing the risk of errors due to outdated or incomplete information (Quint and Loch 2015; Gattullo et al. 2019; Kollatsch and Klimant 2021). In the context of Industry 4.0, the need for such accessible and reliable information becomes even more pressing as the demands for decentralized, accurate, modular, and fast access to information become increasingly important (Hermann, Pentek, and Otto 2016). By utilizing AR technology, it may be possible to improve the accessibility and usability of documentation in complex environments such as shop floors, ultimately leading to increased efficiency and productivity.

AI and its subsets, such as Machine Learning (ML) and DL, also play an indispensable role in the industry 4.0 field. The capabilities to develop solutions that range from Computer Vision (CV), Natural Language Processing (NLP), and finding patterns hidden in vast amounts of data are being applied in tasks such as predictive maintenance (Carvalho et al. 2019), process automation (Ribeiro et al. 2021), or security enhancement (Bécue, Praça, and Gama 2021). Some architectures that enable developing applications that tackle these kinds of tasks are Convolutional Neural Networks (CNNs) for CV and Transformers for NLP. CNNs are a type of neural network architecture typically used for image classification, while Transformers have revolutionized NLP tasks by allowing for attention-based mechanisms to capture semantic dependencies between words. Transformers are behind the current LLMs and are mainly used for tasks such as language translation, text summarisation, sentiment analysis, question answering (QA), and language modeling (Vaswani et al. 2017). AI tools are used to analyze large amounts of data, automate repetitive tasks, and improve decision-making processes, leading to increased efficiency, cost savings, and improved customer experiences. AI-powered systems are also being used to monitor, predict, and prevent potential equipment failures and downtime, reducing maintenance costs and increasing overall productivity.

This study aims to evaluate the effectiveness of using multimodal interaction and AR to enrich complex environments with additional information from a variety of sources (e.g., technical documentation, experience-based knowledge) in an unstructured format and to assess the feasibility of novice workshop operators accessing this information multimodally by anchoring it in the environment through AR.

The main contributions of this work are:

1. The ability to incorporate the knowledge and experience of SMEs in a flexible format and to continually update it through an iterative process,
2. The collection of information in multiple formats, anchored in the physical space and using NL, to reduce the need for access to technical documentation and SMEs,

3. Reducing the time between the emergence of a doubt and receiving a response.

This paper is structured as follows: In section 4.2, numerous examples of AR and AI being applied to industrial settings are enumerated and described to emphasize this research's novelties. Then, in section 4.3, the principal user roles and technology used for this research are explained. Section 4.4 presents and evaluates the results from the experiment. Finally, in section 4.5, the conclusions are described, highlighting the significance of the findings.

4.2 Background and context

Technological advancements have led to the integration of AR and AI into various industries in recent years. These technologies have the potential to revolutionize the way shop floor operators and SMEs interact with and perceive the environment around them. This section will explore some of the current state-of-the-art applications of AR and AI in the industry, highlighting their potential impact and future possibilities.

4.2.1 AR in industry

As a rapidly emerging technology, AR is increasingly being adopted across various industrial sectors, providing plenty of potential applications that can improve efficiency, productivity, and safety. There is a considerable amount of AR applications in the industry, and some examples are step-by-step guides (Scurati et al. 2018), manufacturing (Caudell and Mizell 1992; Sääski et al. 2008; Salonen et al. 2009), design (M. Fiorentino, Monno, and Uva 2009; Michele Fiorentino et al. 2013) and evaluation (Hou, Xiangyu Wang, Bernold, et al. 2013; Hou, Xiangyu Wang, and Truijens 2015). Wang et al. (X. Wang, Ong, and Nee 2016) highlights in their literature review the need for research in several aspects when applying AR to the industry field, such as knowledge representation and contextual awareness. AR can provide many benefits in product design, allowing for faster and more collaborative tasks (Baroroh, C. H. Chu, and L. Wang 2021; P. Wang et al. 2021; Sereno et al. 2022; Marques, Silva, Joao Alves, et al. 2022; Marques, Silva, João Alves, et al. 2022). Process design and control is another field of interest, as indicated by Elia et al. (Elia, Gnoni, and Lanzilotto 2016), and several applications and systems have been developed (Yuan, Ong, and Nee 2008; Ong and Z. B. Wang 2011), bringing to attention the benefits of using AR in this field of application. Regarding maintenance, much interest has been put into solving challenging problems, such as reducing the Mean Time To Repair (MTTR) (Mourtzis, Siatras, and Angelopoulos 2020), reducing the cost of having SMEs on site (Gilchrist 2016), guiding in bad viewing conditions (Ziaei et al. 2011), or focusing on the operators' safety (Tatić and Tešić 2017). There has also been research about using AR for accessing information, such as the ARES framework, to adapt the information shown to the operator depending on several conditions, such as the time to perform a task (Syberfeldt et al. 2016). This highlights the importance of the different roles and experiences on the shop floor and how the interface must show more or less information depending on these characteristics. Additionally, the findings indicate that using authoring tools by SMEs makes creating instructions more efficient and user-friendly. For example, Palmarini et al. developed the FARA authoring tool, which facilitates the creation of step-by-step AR animations for various procedures, such as maintenance, repair, and overhaul (MRO) (Palmarini, Fernández, et al. 2022).

4.2.2 AI in industry

Equally important, AI is being applied in industrial fields rapidly, alongside AR in various domains. MRO, diagnosis, and predictive maintenance are among the fields where AI has found widespread applications. Predicting a possible error in the system before it happens leverages AI to foresee potential system failures before they occur, thus enabling proactive maintenance and increased operational efficiency. Both Carvalho et al. (Carvalho et al. 2019), and Zonta et al. (Zonta et al. 2020) perform a systematic literature review where several ML and DL models are being applied for predictive maintenance, demonstrating the increasing research interest in this field. Other applications focus on customer support, where chatbots and recommendation systems can help companies provide faster and more personalized support to clients. Casillo et al. develop a chatbot framework for real-time assistance and efficient and personalized training (Casillo et al. 2020). Pattern recognition and prediction are, by nature, key applications of ML and DL algorithms. Detecting patterns in vast amounts of data might help data scientists obtain valuable insights, for example, to predict changes in product demand. Moroff et al. evaluate several ML and DL models such as Random Forest, XGBoost, Long-term short-term memory (LSTM) networks, and a multilayer perceptron (MLP), among others, as forecasting models (Moroff, Kurt, and Kamphues 2021). Finally, AI models are being increasingly applied in the field of automation. Operators' previous tasks can be automatized intelligently, such as optimizing production processes and improving customer support. Maschler and Weyrich highlight, in their literature review, several studies in fields such as anomaly detection, time series prediction, fault diagnosis and prognostics, quality management, and computer vision (Maschler and Weyrich 2021).

4.2.3 Synergy between AR and AI in industry

As a complementary element to AR, AI opens new synergy possibilities. In the field of information access, Chidambaram develops a solution utilizing AR and the YOLO foundation model (Redmon and Farhadi 2018) to generate instructions that differentiate between novice users and SMEs (Chidambaram et al. 2021). As described by Standford, a foundation model means to *"Train one model on a huge amount of data and adapt it to many applications."*, or in other words, it is a model that has been pre-trained and provides various features that can be utilized for transfer learning or fine-tuning to fit specific requirements (Bommasani et al. 2021). Examples of foundation models are YOLO for object detection, Stable Diffusion for image generation (Rombach et al. 2022) or GPT for text generation (Radford, Narasimhan, et al. 2018). Our previous research has focused on developing tools that guide and enhance the safety of shop-floor operators using AR and AI (Izquierdo-Domenech, Linares-Pellicer, and Orta-Lopez 2022); however, the present proposal in this work takes a closer look at the other side of the equation, the SMEs and how to use AR and AI to enrich the environment with information in a comfortable manner. Little research has been done regarding the use of these two technologies in enhancing unstructured information management and access, and this work proposes an approach to fill this gap.

4.2.4 Documentation management and access

With its human-centric view, the advent of Industry 5.0 (Rožanec et al. 2022; Akundi et al. 2022) brings about significant challenges in the realm of documentation and information access, owing to various factors such as decentralization, virtualization, and modularity (Gattullo et al. 2019). This highlights the need for more effective methods for managing documentation. In light of the need for information to be easily accessible, updatable, and translatable, paper-based documentation is becoming obsolete. Further research must be conducted in this area, as several authors have emphasized (Ventura 2000; Quint and Loch 2015; Kollatsch and Klimant 2021). This study seeks to address a key challenge in shop floor operations by exploring novel strategies to enhance information accessibility. The ultimate aim of this proposal is to leverage the knowledge and expertise of SMEs to create dynamic environments, enhancing them with knowledge anchors into spatial 3D real environments to improve efficiency and profitability.

4.2.5 Information retrieval and mental decay

In accordance with the discussion presented in section 4.1, the most optimal way to gain expertise in an industrial setting is to perform SBT and personalized tutelage with an SME; however, one of the most critical issues associated with this process is the maintenance of the acquired knowledge, particularly its tendency to deteriorate over time.

Mental decay, also known as knowledge decay, is a passive process in which the knowledge and skills of a person gradually decline over time if not actively reinforced. Studies have shown that mental decay can occur even when an individual is exposed to new information, with decay increasing as the time between exposure and retrieval increases (Hardt, Nader, and Nadel 2013). Numerous studies have been conducted on knowledge retention and information retrieval in industrial settings in recent years. One such study by Adesope et al. found that repeated exposure to information leads to better knowledge retention compared to solitary exposure (Adesope, Trevisan, and Sundararajan 2017). This finding is supported by other studies, such as the work of Karpicke and Roediger, who showed that retrieval practice can enhance long-term retention of information (Karpicke and Roediger 2008). In addition, research has also been conducted on the impact of aging on knowledge retention and retrieval. For example, Bissig and Lustig found that older adults experience greater difficulty retrieving information from long-term memory compared to younger adults (Bissig and Lustig 2007). This finding has important implications for industrial settings, as the aging workforce is becoming increasingly prevalent and might be a focus of interest in future research (Wolf et al. 2018).

In industrial environments, knowledge about machines and elements on the shop floor is often distributed through multiple documents and SMEs. Hence, having a reliable source for accessing and retrieving this information is essential. Information access is of paramount importance in industrial settings as it plays a critical role in ensuring the efficient performance of operations. Understanding how the knowledge provided to operators fades over time becomes increasingly important. It is essential to note that the operators involved in the experiments were only given the task to perform with prior knowledge. As explained in section 4.4, operators were subjected to repeated exposures of the same information because this can lead to better knowledge retention compared to a solitary exposure (Cepeda et al. 2006; Carpenter et al. 2012).

The proposed tool aims to fulfill this gap of mental decay, thus providing access to technical and SME information at all times.

4.3 Proposed system

One of the significant challenges faced by shop floor operators in industrial settings is knowledge retrieval from the environment, as previously discussed in section 4.1 and 4.2. It has been discussed that having an SME and utilizing SBT may be the ideal solution, but not always feasible in practice. Furthermore, mental decay adds to the difficulty as the shop floor operator may not always retain all of the information taught. Although AR applications have been proposed as a means of providing additional information in a context-aware system boosted by AI systems; accessing information naturally when technical documentation and SMEs are the only sources of information remains a challenge. This section presents a detailed description of the proposed system, highlighting the key roles of the SME, the shop floor operator, and their interactions with the system. This research aims to address the gap in the literature and justify the main contributions outlined in section 4.1.

4.3.1 SME: context enrichment with information

The SME is an expert who has acquired extensive knowledge in a particular field or topic; however, disseminating their knowledge and its contribution to the field remains challenging. While the possibility to ask the SME in case there is doubt exists, it may only sometimes be feasible, as the constant presence of an SME in the work environment may not be practical (Garza et al. 2013; Gilchrist 2016). Indeed, AR systems can reduce the cost of having SMEs on-site, but the challenge remains in effectively transferring the SME knowledge to the worksite. To bridge the gap in knowledge transfer to the site, this study proposes an architecture that considers the SMEs roles as a *"Knowledge Transfer Experts"*.

The SMEs are responsible for utilizing the system to introduce "pills" of knowledge across the environment. In this research paper, a "pill" refers to a small unit of knowledge that can be added to a system. The term is chosen for its memorable connotation and aligns with the concept of intentional knowledge management. It is important to highlight that the presented architecture implementation relies on the fact that the environment needs to be previously scanned, a common feature in current SLAM-based AR solutions. Upon entering the environment, the SME can interact with their surroundings using touch interaction. This way, the SME can add specific "pills" of information to any element they find interesting to enrich, regardless of whether they are machines, control panels, or any other element of interest in complex environments. This information "pills" will be used by the system with two purposes:

1. To retrieve a specific "pill" linked to a specific position in the environment *as-is*,
2. For obtaining answers to specific questions.

The present study depicts a specialized tool that supports SMEs in contributing to a digitized environment. The tool facilitates data input through the means of either voice recognition or written text. Using ray-casting techniques, alongside touch interaction in AR, enables SMEs to pinpoint and enrich specific features of the 3D scanned mesh from the virtual environment. The interaction in the system is performed using touch input that is implemented differently depending on the final device used. Specific AR devices can trigger the touch action with hand-specific controls or even by using hands, while for mobile devices like tablets, touching the screen at the desired object triggers the touch action. In both cases, the interaction is implemented using ray-casting, which calculates a line or ray from the touch 2D coordinates and with the direction derived from the camera frustum. Then, the ray intersections are checked, and the object selected is the one that is closest to the user in 3D coordinates. The AR library maintains a congruent mapping of spatial coordinates between the physical and virtual worlds, resulting in accurately identifying elements in the 3D virtual space. A visualization of the SMEs task in the environment is displayed in Fig. 4.1.

4.3.2 Shop floor operator: information retrieval

Once the site has been enriched with anchored "pills" of knowledge, it is time for the shop floor operator to utilize these resources, reducing the frequency of their need to seek clarification from the SME and technical documentation.

As depicted in Fig. 4.2, the presented application offers two alternatives for accessing such information in a multimodal manner:

Touch & Area Selection Since many anchors might be disseminated around the site, the shop floor operator has the option to select a rectangular area and retrieve all the "pills" within that selection, as shown in Lane C in Fig. 4.3. Applying the ray-casting methods as introduced in section 4.3.4.3.1, the system can obtain the 3D virtual coordinates of a chosen

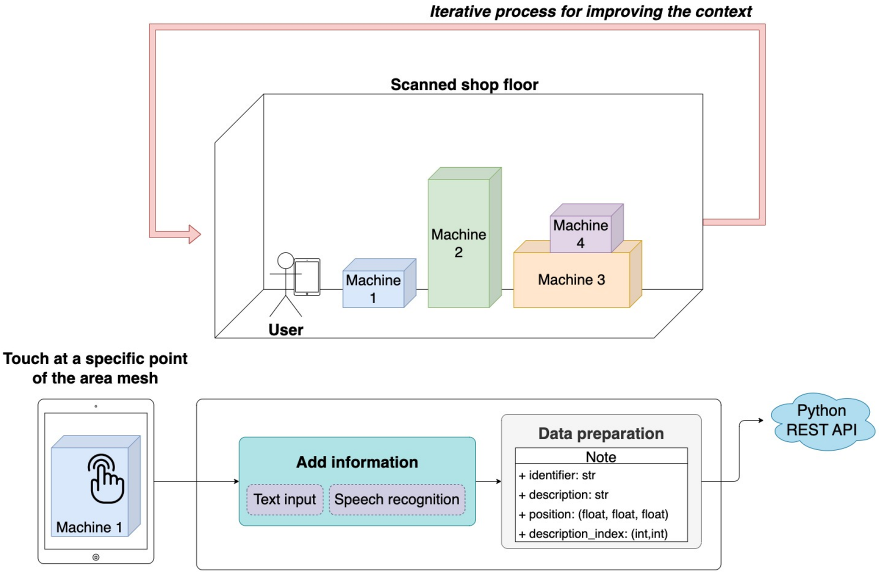


Figure 4.1: SMEs annotate scanned environments.

location on the shop floor by utilizing touch interaction. The process involves the shop floor drawing an area of interest from which anchored "pills" can be retrieved. The approach leverages the same ray-casting techniques employed by SMEs in the aforementioned section, demonstrating the tool's versatility across multiple domains.

Speech Recognition for NL Queries While the *touch and area selection* method is proactive, the system also includes a more reactive approach to obtaining information. The shop floor operator can ask any query in NL, such as "What temperature does glycol evaporate at?". The system will then process the query and provide the answer (e.g., "360º") as well as the location within the work environment where the SME anchored the corresponding "pill" of knowledge. This is also shown in Lane B in Fig. 4.3.

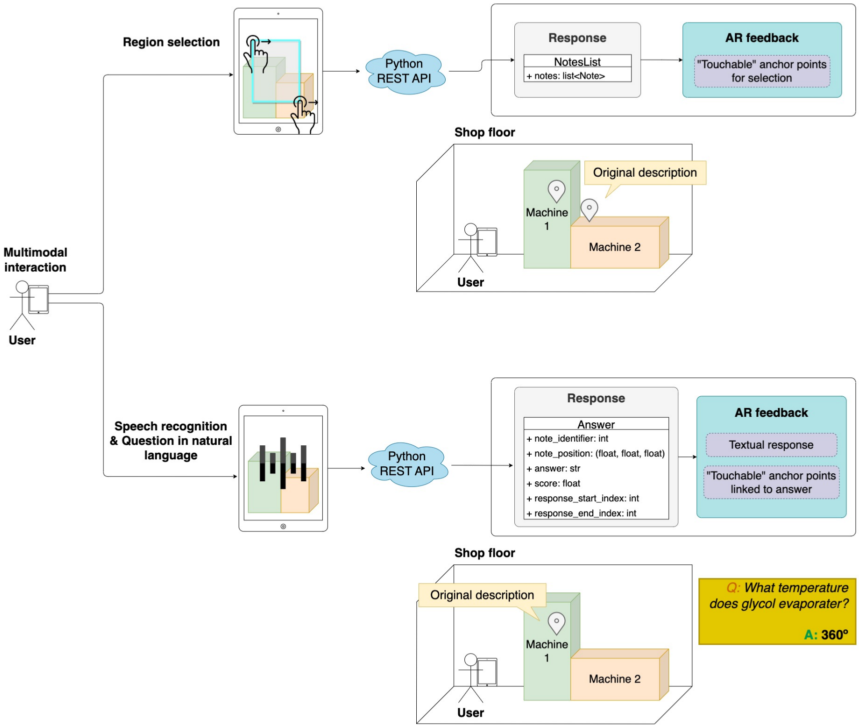


Figure 4.2: Shop floor operator retrieves anchored information via NL query or area selection.

4.3.3 System implementation

For evaluating the proposed system, a mobile application has been developed using the Unity platform, which allows for developing applications for both Android and iOS devices. AR through the Vuforia library has been integrated for scene recognition (i.e., the scanned laboratory). For Speech Recognition on device, the Vosk toolkit has been used, which allows for real-time voice recognition in diverse languages. Regarding the server side, since the information must persist between application uses, the Python FastAPI framework has been utilized for generating the different endpoints using a RESTful API.

Fig. 4.3 briefly summarises the application's key features. Lane A highlights the capabilities of SMEs within the app. After detecting the environment, the SME can add pieces of information, either handwritten or by voice, at any point, using touch interaction. Lanes B and C outline the interaction from the

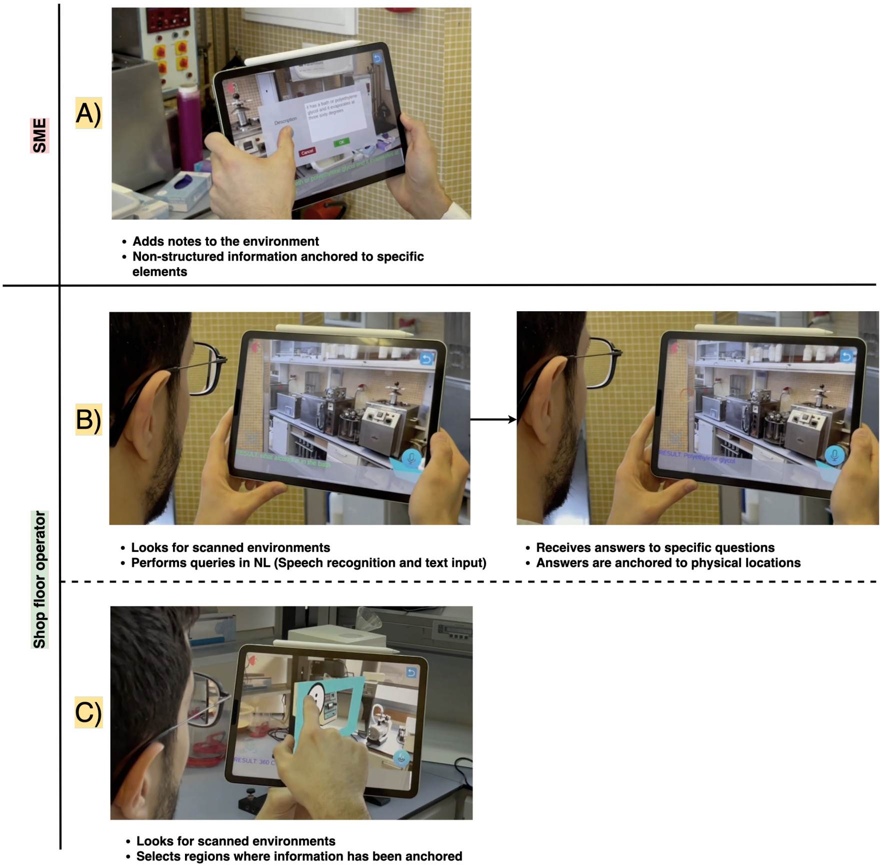

SME
A)
• Adds notes to the environment
• Non-structured information anchored to specific elements
Shop floor operator
B)
• Looks for scanned environments
• Performs queries in NL (Speech recognition and text input)
• Receives answers to specific questions
• Answers are anchored to physical locations
C)
• Looks for scanned environments
• Selects regions where information has been anchored

Figure 4.3: Two shop floor roles: SMEs add information, operator retrieves it via NL/AR.

shop floor perspective. In Lane B, the operator can use the application to ask, in NL, about anything regarding contextual information. This query is then sent to the transformer for processing. The specific transformer architecture is explained in detail in subsection 4.3.4. The result is displayed not only as a text answer but also as a highlighted location in the environment where the SME added the information. Finally, Lane C briefly shows that the operator can retrieve any information in any environment by drawing a rectangle around the area of interest using touch interaction. The application will display all notes anchored by the SME within that area.

4.3.4 Consistency of the information

Ensuring the consistency of textual information units when integrating them into a system is a crucial factor to consider. It is important to estimate whether a new block of textual information is contradictory, redundant, or provides new information content, particularly in a proposal where unstructured information is the fundamental input. The topic of consistency in LLMs is a recurring subject of study, as noted by Elazar et al. in their work on measuring LLMs (Elazar et al. 2021). In this context, consistency is treated systematically and comprehensively, with a focus on ensuring the correctness of the answers provided by the model prior to paraphrasing the input questions.

In this study, the main focus is on consistency, which involves ensuring that there are no redundancies or contradictions when introducing a new block of information related to a particular aspect of the environment. Redundant information can reduce the efficiency of the model, while contradictory information is even more critical. It is crucial to prevent SMEs from providing conflicting information about the same element, as this can cause problems when correcting subsequent information queries. Such conflicts may arise not only from an SME's error but also from speech-to-text conversion, for instance. To solve this problem, LLMs allow a fine-tuning process after pre-training to improve their behavior when faced with more specific problems. For this specific problem we use a Transformer architecture.

Transformer architecture is a neural network model composed of two key components: an encoder and a decoder. The encoder comprises multiple layers comprising two sub-layers: an attention mechanism and a fully connected feed-forward neural network. The decoder, on the other hand, follows a similar structure. However, in this case, each sub-layer comprises three sub-layers: the attention sub-layer, the feed-forward one, and a third sub-layer in which multi-head attention is applied to the output of the encoder. Using this mechanism makes it possible to have better results than with recurrent networks since it is possible to take into account the semantics of the input sentence more efficiently. In addition, the training process can be unsupervised; however, it is necessary to use significant amounts of unlabelled text. Due to the cost of training a transformer from scratch, not only in terms of time but also in terms of computational units and the amount of data needed to obtain a good performance, it is common to use pre-trained models. That is, using architectures that have already been trained with vast amounts of data. This allows the pre-trained transformers to have already learned most of the semantics of NL, so they can process and answer most of the questions or suggestions that the user asks in NL in a very flexible way. However, it is important to carefully select the dataset to fine-tune the model, as it is of balanced and representative data.

It is possible to find a wide variety of transformers with a wide range of capabilities in NL processing. Such as machine translation between language pairs, text summarisation, text-relevant QA, or conversational systems. Among the most commonly used transformer-type pre-training systems currently in use are DistilBERT (Sanh et al. 2019), RoBERTa (Liu et al. 2019), Google's T5 (Raffel et al. 2020), BLOOM (Scao et al. 2022), or GPT-3 (Radford, Narasimhan, et al. 2018). Arroni et al. (Arroni et al. 2023) provide a compelling example of the effective use of LLMs in their work on semantic classification.

In the case of this project, the GPT-JT model was used as resulting of the fine-tuning of the GPT-J model with UL2 training objectives (Tay, Dehghani, et al. 2022; Tay, Wei, et al. 2022), achieving results similar to models of 175B parameters in many tasks, such as InstructGPT davinci v2, but with only 6B parameters (Together 2022). GPT-JT has been trained in a decentralized way and allows its free download and use, as well as its installation and local use. The final LLM used in the final solution can be adapted to specific final requirements of the solution and available options. GPT-JT was a good commitment in evaluating the proposed solution, allowing good results in QA on technical documentation and a high degree of flexibility on few-short learning. Few-shot learning means that the model can be presented with one or several examples of the task to be solved to achieve higher degrees of accuracy in its responses.

With this purpose, the GPT-JT model has been tested to detect if a new piece of textual information is redundant or in contradiction with the previous information assigned to a specific element of the scene of the AR environment.

Although the main aim of this research was not to compare various models on the same instruct-based tasks, we set specific criteria for selecting the most appropriate model. Based on its Open Source availability, superior performance against instruct-based queries (as per the Hugging Face ranking), and ease of installation on local servers (6B parameters only), the GPT-JT model was chosen. The preliminary findings indicated that the chosen model classified the information consistently.

Table 4.1 shows a concrete example used to discriminate between "new", "contradictory", or "redundant" with few-shot learning, in this case, a single training example. Using only one example, it has been possible to verify the correctness of the classification of the new information block in all the tested examples. An exhaustive evaluation of this approach transcends the objectives of the present proposal, more typical of computational linguistics, requiring the creation of comprehensive and specific evaluation datasets which, in the best of cases, cannot guarantee their results in the face of domain problems.

The proposal in this paper focuses on using the few-shot learning consistency test of the transformer to detect redundancy or contradiction problems better and visually notify the SME of this possibility. When the SME is warned of a possible redundancy or contradiction, it can examine the text entered associated with an element and repeat in case of a possible error or reconfirm the correction of the new information element.

4.4 Evaluation and results

4.4.1 Experimental setup

Experiments were conducted in a textile laboratory at Universitat Politècnica de València, Campus d'Alcoi. The section of the laboratory that was scanned for subsequent identification is a representative sample of an overall facility. It was selected because it contains a variety of equipment commonly used in textile manufacturing and provides a suitable environment for testing the developed system. The equipment used in the experiments includes machines for fabric dye testing, material cleaning, and emulsion homogenization, all essential for producing high-quality textile products. The laboratory's wide range of equipment and diverse capabilities make it an ideal environment for simulating potential scenarios, providing a representative setting for testing and evaluating various approaches and solutions. To evaluate the developed application, we selected an iPad Air device because of its compatibility with the system and development tools and its ease of use for the operators.

4.4.2 Participants

As far as participants are concerned, a textile master teacher served as the SME for explaining and adding information. Two groups of participants were selected to evaluate the system. 30 participants were selected and distributed between groups A and B.

The 30 participants recruited for the study were all master's students in engineering, ranging in age from 22 to 28 years old, with an equal distribution of male and female participants. While all participants had previous experience

Context	**Input**	**Output**
Few-shot learning		
To turn on the machine switch on the red button. To turn off the machine switch off the red button.	*To turn on the machine it is necessary to switch on the red button.*	"redundant"
Same context.	*To turn on the machine it is necessary to switch on the blue button.*	"contradictory"
Same context.	*To pause the machine it is necessary to switch on the blue button.*	"new"
After few-shot learning...		
The emulsion is homogenized with an agitator. One field of use is microcapsule emulsions or cosmetic creams. At the top of the panel is the button to raise and lower the agitator. In the central part of the panel are the buttons to turn on and off, and a wheel to control the number of Revolutions Per Minute (RPM). At the bottom, we find the motor and ignition indicators.	*To lower the agitator, use the button on the top of the panel.*	"redundant"
Same context.	*The machine is called homogenizer.*	"new"
Same context.	*We can find the buttons to turn off the machine at the bottom.*	"contradictory"

Table 4.1: GPT-JT tests label information as "new", "redundant", or "contradictory" based on context.

with similar machines, none were familiar with the specific machine used in this study, making it a novel task for all participants.

While both groups were exposed to the explanation of the SME, during the system evaluation phase, group A had access to the documentation about the machines to be utilized and the SME. In contrast, group B had access to the developed application. The only restriction applied to group B was that they were instructed to use the application first for any questions about the machines. If the answer from the application was incorrect or lacked enough information, they were allowed to search in the technical documentation and ask the SME.

Both groups, A and B, were exposed to information about the environment and the machines they were going to use during the experiment. The information exposure took place simultaneously a week before the experiment for both groups. The information was presented by a SME, who, at the same time, introduced the information into the system. Three days before the experiment, both groups were again presented with the same information to boost retention.

4.4.3 Tasks

The participants were instructed to perform several interactions with three types of machines; fabric dye testing (Task 1), material cleaning with an ultrasonic machine (Task 2), and homogenizing emulsions (Task 3). Fig. 4.4 compares the completion time between groups A and B while performing the same tasks in seconds. To ensure objective and consistent measurements, the data was collected by a single external observer who followed standardized procedures throughout the data collection process.

The following standardized procedures were employed by the observer:

1. **Training and Familiarization**: The observer underwent comprehensive training to become familiar with the research goals, the tasks to be performed, and the specific machines involved. This training aimed to ensure that the observer had a thorough understanding of the procedures and requirements for accurate data collection.

2. **Non-Intrusive Observation**: The observer adopted a non-intrusive approach to minimize any potential influence on the participants' behavior or performance. The observer focused on discreetly observing the participants without interfering with their interactions or affecting the natural flow of the tasks. This approach aimed to ensure that the participants' actions were representative of their usual behavior.

3. **Data Recording**: The observer used the same data collection sheet to record relevant information during the observation process. This included capturing the start and end times of each task, any relevant observations

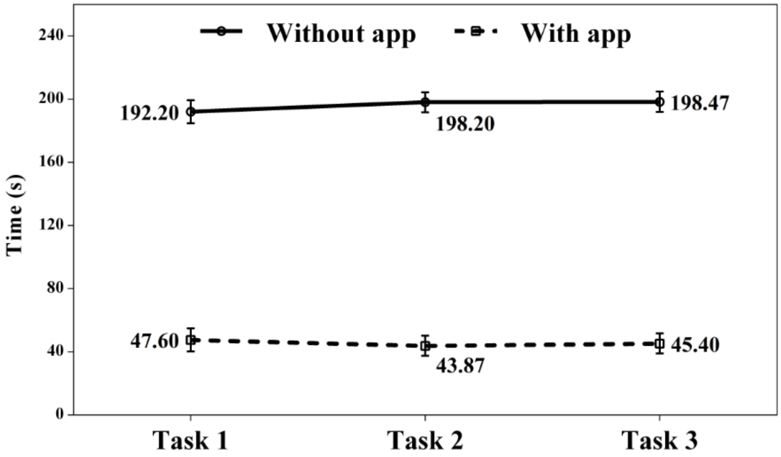


Figure 4.4: Time comparison between groups A (No app) and B (With app).

or notes, and any additional contextual information that could be important for later analysis.

This approach allowed for precise time measurement and reduced the potential for biases or errors that could arise from multiple observers or inconsistent methods. It is clear that there is a significant reduction in time when specific information needs to be retrieved, thereby supporting the second and third main contributions of this research: anchoring information in the environment reduces the need for consulting technical documentation and SME, and the reduction in time between the emergence of a doubt and receiving a response.

4.4.4 Results

Table 4.2 shows the results of the two-factor ANOVA with repeated measures on one of them performed to determine if the effect of the group influences the task execution time. The results show that there is a statistically significant difference between the groups, regardless of the task ($p < 0{,}001$) such that the task execution time for workers using the app (45,62 seconds) was significantly lower than for workers not using the app (196,29 seconds). No differences were observed between tasks ($p = 0.964$), and the group and task interaction were not significant ($p = 0{,}781$). The term "group and task interaction" refers to the relationship between the different groups of participants (Groups A and

B) and the specific tasks they performed. In this context, a non-significant interaction suggests that the effect of the group on task execution time was consistent across all tasks. In other words, the app usage had a similar effect regardless of the task performed.

Additionally, a Likert questionnaire of 5 points was delivered to the participants of group B (see table 4.3) to measure their perceptions towards the combination of AR and anchored information retrieval in a multimodal manner and the NL interaction complement in the AR system. The questionnaire had ten questions, with values that ranged between 1 (Strongly disagree) and 5 (Strongly agree). The results support the notion that the participants had a favorable view towards integrating anchored information and its retrieval by AR systems. This indicates that this approach could potentially lead to improved outcomes in future studies and practical applications.

4.5 Conclusions

The roles of SMEs and shop floor operators are essential in Industry 4.0, but even more in the future advent of Industry 5.0. AR and AI techniques are being applied to improve the efficiency and effectiveness of these roles. However, while the use of AR and AI techniques is receiving much attention, only some studies have investigated the value of SMEs as a source of information. The unstructured nature of this information makes it challenging to manage and integrate with technical documentation. In this study, we developed and evaluated a system to extract information from SMEs that can be integrated with technical documentation. We used state-of-the-art AI architectures such as Transformers and LLMs to perform useful tasks such as QA and multimodal interaction on AR systems. Our results demonstrate the potential of integrating SME information with technical documentation to reduce the time it takes for operators to access relevant information. It is worth noting that Industry 5.0 is a human-centric approach to the industry that emphasizes the value added by people in the manufacturing process. While Industry 4.0 focuses on using advanced technology to automate and optimize production, Industry 5.0 recognizes the importance of operator comfort and satisfaction in the workplace. This approach considers the physical and emotional well-being of the workforce, as well as their creativity, problem-solving abilities, and interpersonal skills. By combining the strengths of both human workers and technology, Industry 5.0 aims to create a harmonious and efficient work environment that benefits all stakeholders.

	Tasks			**Effects tests**		
	1	**2**	**3**	**Group**	**Tasks**	**Group*Task**
	Mean *(Sd)*	*Mean* *(Sd)*	*Mean* *(Sd)*	***F*(g.l.);** ***p*-value (η^2)**	***F*(g.l.);** ***p*-value (η^2)**	***F*(g.l.);** ***p*-value (η^2)**
Time (seconds)				F(1;28) = 1.545,22; ***p* <0,001** (0,982)	F(2;56) = 0,04; p = 0,964 (0,001)	F(2;56) = 0,25; p = 0,781 (0,009)
Group A	$192,20$ $(34,32)$	$198,20$ $(31,26)$	$198,47$ $(32,45)$			
Group B	$47,60$ $(20,28)$	$43,87$ $(15,24)$	$45,40$ $(13,62)$			

Table 4.2: Comparison of time between groups A and B, along with the corresponding p-values

	Min-Max	Mean (Sd)
I found the AR app easy to use for accessing information about the machines.	3-5	4 (0,85)
The information provided by the AR app was accurate and reliable.	2-5	3,27 (0,8)
The AR app helped me perform my tasks more efficiently.	3-5	4,6 (0,63)
The information provided by the AR app was useful and relevant to my needs.	4-5	4,8 (0,41)
Information was provided by the AR app quickly when I requested it.	3-5	4,07 (0,88)
I did not encounter errors or issues while using the AR app.	2-5	3,47 (0,74)
I did not need to refer to technical documentation/technical operator in addition to the AR app to find the information I needed.	3-5	3,93 (0,7)
The information provided by the AR app was helpful in completing my tasks.	4-5	4,73 (0,46)
The AR app made it easier to access information compared to other methods I have used in the past.	4-6	4,47 (0,64)
I would be willing to use the AR app on a regular basis as part of my workday.	4-5	4,33 (0,72)

Table 4.3: Likert questionnaire

This study highlights the importance of SMEs' knowledge for improving shop floor operations; however, there is still room for improvement in automating the extraction process and maintaining the accuracy and relevance of the information. Future research could focus on developing more advanced NLP techniques to better extract and organize SMEs' knowledge while ensuring that the information remains up-to-date and reliable.

Although the Vosk toolkit was used for the system implementation for speech recognition, in the presence of noisy or industrial environments, speech-to-text accuracy can be significantly improved by employing the latest Open Source models, such as Whisper (Radford, Kim, et al. 2022).

Another area for future research is integrating SME knowledge with technical documentation. It would be beneficial to investigate how different types of information can be presented in a way that is easy to access and use for operators. Additionally, there is potential for integrating AR and AI techniques with SME knowledge to further enhance the efficiency and effectiveness of shop floor operations. For example, by implementing automatic information retrieval methods using object detection models, thus allowing operators to access relevant information while exploring the shop floor quickly.

As Industry 5.0 emphasizes the human-centric nature of manufacturing, it is essential to explore ways to improve operator comfort and satisfaction in the workplace. Future studies could investigate using AR and VR technologies to create more engaging and interactive training materials or wearable technologies to monitor and improve operator well-being.

4.6 Appendix

This appendix illustrates the different requests the application can make to the server, shown in Fig. 4.5. It outlines the possible requests the SME and the shop floor operator can make to the server.

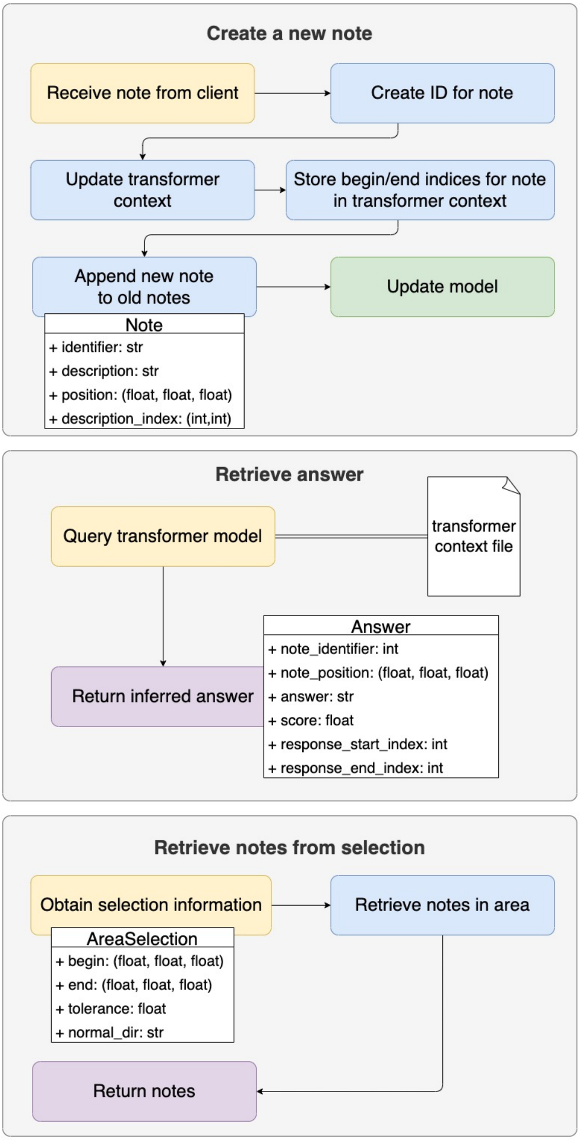

Create a new note
Receive note from client
Create ID for note
Update transformer context
Store begin/end indices for note in transformer context
Append new note to old notes
Update model
Note
+ identifier: str
+ description: str
+ position: (float, float, float)
+ description_index: (int,int)
Retrieve answer
Query transformer model
transformer context file
Answer
+ note_identifier: int
+ note_position: (float, float, float)
+ answer: str
+ score: float
+ response_start_index: int
+ response_end_index: int
Return inferred answer
Retrieve notes from selection
Obtain selection information
Retrieve notes in area
AreaSelection
+ begin: (float, float, float)
+ end: (float, float, float)
+ tolerance: float
+ normal_dir: str
Return notes

Figure 4.5: Rest API calls.

Chapter 5

Conclusions

This chapter presents a summary of the key contributions of this thesis and suggests directions for future research.

5.1 Conclusions

In this thesis, the integration of AI with AR has significantly enhanced the capabilities of AR systems. This synergy has facilitated advancements ranging from automated validation to improved learning experiences, introducing diverse multimodal interaction methods with the systems and effectively disseminating SME knowledge. A comprehensive architecture was devised and evaluated across multiple applications to assess the viability and advantages of this AI-AR fusion, yielding promising outcomes. Moreover, the process enabling SMEs to impart their knowledge for subsequent retrieval has been examined, demonstrating substantial success.

This thesis encompasses three studies to evaluate the practicality of integrating AR and AI via a semantic layer alongside the dissemination of SME knowledge. Each study benchmarked the system's performance against conventional information dissemination methods, such as paper and PDF documents. Furthermore, in the second study, the effectiveness of the developed AR-AI application was also compared with that of a traditional AR application. Additionally, in the second and third studies, participants were asked to complete a Likert-scale questionnaire designed to assess the usability and overall appeal of the proposed system.

The following outlines the conclusions drawn from each of the three studies:

Study 1: Towards achieving a high degree of situational awareness and multimodal interaction with AR and semantic AI in industrial applications

- The introduction of a semantic layer in the AR system facilitated automatic transitions to subsequent tasks, reducing completion time for specific tasks. This was particularly evident in Task 1, where Group 2, utilizing the semantic layer, could automatically proceed to the next task without manual intervention, unlike Group 1.
- In Task 2, the semantic layer provided additional security by automatically validating and warning operators when a specific threshold was exceeded in the analog control. This feature not only reduced the cognitive load on operators but also improved their response times and accuracy in task execution
- Task 3 highlighted the significant time efficiency achieved by querying through natural language instead of consulting technical documentation. The use of NLP and voice responses in the AR system considerably sped up information retrieval despite occasional needs for question repetition due to errors in transformer interpretation.
- The addition of the semantic layer showed a notable impact on the execution time of tasks, especially those that were more complex. Tasks 3, 4, and 5 demonstrated significant time savings for Group 2 (using the semantic layer) compared to Group 1. This finding suggests that the semantic layer is particularly beneficial in tasks where operators require more guidance and where the complexity is higher, such as anomaly detection.
- Overall, it has been demonstrated that automatic validation and guidance are beneficial, even for tasks that do not require high cognitive effort.

Study 2: Environment awareness, multimodal interaction, and intelligent assistance in industrial augmented reality solutions with deep learning

- Integrating a semantic layer in AR significantly improved task performance. Operators in Group A, who used semantic AR, demonstrated more efficient completion of tasks compared to those in Group C, who relied solely on traditional documentation. This was particularly noticeable in tasks that required more complex cognitive processes, underscoring the semantic layer's role in enhancing cognitive capabilities and operational efficiency.
- The study revealed that operators using the semantic AR system (Group A) experienced considerably lower response times when asking questions than those using only technical documentation (Group C), indicating that the semantic layer effectively streamlines information retrieval, enabling faster decision-making and problem-solving.
- Operators utilizing the semantic AR system reported higher satisfaction levels, as evidenced by their responses to the Likert-scale questionnaire. They found the system to be more helpful and easier to use, suggesting that the semantic layer improves task performance and enhances the overall user experience.
- The analysis showed a statistically significant difference in task execution times between the groups, regardless of the task performed. This highlights the effectiveness of the semantic AR system in reducing the time required to complete tasks, particularly for Group A operators.
- For tasks that were less cognitively demanding (such as Task 2, 4, 6, and 7), no significant differences in execution times were observed among the operator groups. This suggests that the benefits of the semantic AR system are more pronounced in tasks that require higher cognitive input and complex decision-making.
- Both groups that utilized AR systems (Groups A and B) concurred on the advantages of implementing AR in shop floor operations. Notably, participants from Group A (app with semantic capabilities) perceived the system as significantly more beneficial, especially for information retrieval purposes.

Study 3: Large Language Models for in situ knowledge documentation and access with Augmented Reality

- The integration of Large Language Models (LLMs) with AR in the textile laboratory setting significantly enhanced task efficiency. This technology facilitated more rapid completion of tasks related to fabric dye testing, material cleaning, and emulsion homogenization compared to traditional methods.

- The experimental setup, involving various of equipment used in textile manufacturing, proved to be a suitable environment for assessing the application of AR and LLMs. The diversity and complexity of the equipment provided a realistic and challenging environment for testing the system's capabilities.

- Participants in the study, comprising master's students in engineering, could effectively use the developed AR application despite unfamiliar with the specific machines used. This suggests the system is intuitive and user-friendly, even for those new to the specific industrial environment.

- The findings demonstrated a distinct benefit in utilizing the AR application for information retrieval in a shop floor setting. Compared with conventional approaches, like referring to technical documentation or consulting an SME, who cannot be available at any time, the AR application offered a more rapid and effective means of accessing essential information.

Based on the aforementioned specific conclusions, the following general conclusions are drawn:

- The integration of semantic layers and LLMs with AR technology significantly enhanced task efficiency. This was evident in tasks that required higher cognitive input and complex decision-making, where AR systems facilitated more rapid completion and increased accuracy.

- While the advantages of semantic AR systems were more pronounced in tasks requiring higher cognitive input, they also proved beneficial in less demanding tasks by providing automatic validation and guidance, enhancing overall efficiency.

- The use of AI-enhanced AR systems, and more specifically with NLP capabilities, resulted in considerably faster information retrieval and more effective decision-making compared to traditional methods like consulting technical documentation or seeking SME assistance.

- Operators using AR systems with semantic layers reported higher satisfaction levels and found the systems more helpful and easier to use.
- The diverse and challenging environments used for the studies, such as a textile laboratory or a shop floor, demonstrated the suitability of AR systems with semantic layers for a wide range of industrial applications.
- Even participants without prior experience with specific machinery could use the developed AR applications effectively, indicating the systems' user-friendliness.

5.2 Future work

This thesis underscores the advantages of augmenting traditional methods and AR applications with cutting-edge AI techniques. While significant progress has been made, there remain several avenues for further research:

- In Studies 1 and 2, we conducted a general evaluation of the semantic layer without specifically focusing on gender and age diversity. Future research should involve a more diverse participant pool to broaden the applicability and robustness of our conclusions.
- The tasks involving visual validation, such as the operation of pressure valves or activation buttons in Study 1, necessitated specialized CNN model training. This approach, while effective, may not be ideal for complex environments with numerous visual elements that belong to non-sensorized machines (e.g., not connected to a SCADA system). Future work should explore the potential of Large Multimodal Models (LMMs) like Flamingo, CLIP, or GPT-4 with Vision to enhance the capacity for interpreting and interacting with a wide array of visual controls, thus minimizing the necessity for training one model per each control.
- The tasks selected for system evaluation were specific and relevant to our studies. However, the diversity of tasks on the shop floor extends far beyond these. Subsequent research should aim to evaluate the effectiveness of these systems across a broader spectrum of tasks, thereby affirming their versatility and applicability in various industrial contexts.
- In our current setup, the SME manually added and updated information through the AR interface, a method that might not be scalable in larger, more complex settings. Future studies should focus on developing more efficient and scalable techniques for inputting and updating expert

knowledge to enhance the practicality and adaptability of the system in diverse industrial environments.

5.3 Scientific contributions

The following publications have emerged from this thesis:

5.3.1 Papers in journals indexed in JCR

- Izquierdo-Domenech, J., Linares-Pellicer, J., & Orta-Lopez, J. (2023). Towards achieving a high degree of situational awareness and multimodal interaction with AR and semantic AI in industrial applications. Multimedia Tools and Applications, 82(10), 15875-15901.
- Izquierdo-Domenech, J., Linares-Pellicer, J., & Ferri-Molla, I. (2023). Environment awareness, multimodal interaction, and intelligent assistance in industrial augmented reality solutions with deep learning. Multimedia Tools and Applications, 1-28.
- Izquierdo-Domenech, J., Linares-Pellicer, J., & Ferri-Molla, I. (2023). Large Language Models for in Situ Knowledge Documentation and Access With Augmented Reality. International Journal of Interactive Multimedia and Artificial Intelligence.

5.3.2 Conferences

- Izquierdo-Domenech, J., Linares-Pellicer, J., & Orta-Lopez, J. (2020, December). Supporting interaction in augmented reality assisted industrial processes using a CNN-based semantic layer. In 2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR) (pp. 27-32). IEEE.
- Izquierdo-Domenech, J., Linares-Pellicer, J., & Orta-Lopez, J. (2021, November). Semantic Computing Enhancement of Industrial Augmented Reality Solutions with Machine Learning. In Proceedings of the 2021 3rd International Conference on Video, Signal and Image Processing (pp. 131-138).

Bibliography

Abramovici, Michael, Andreas Krebs, and Thomas Schindler (2013). “Design for Usability by Ubiquitous Product Documentation”. In: *Lecture Notes in Production Engineering*. Vol. Part F1158. Springer Nature, pp. 633–641. DOI: 10.1007/978-3-642-30817-8{_}62 (cit. on p. 6).

Adesope, Olusola O., Dominic A. Trevisan, and Narayankripa Sundararajan (June 2017). “Rethinking the Use of Tests: A Meta-Analysis of Practice Testing”. In: *Review of Educational Research* 87.3, pp. 659–701. ISSN: 19351046. DOI: 10.3102/0034654316689306 (cit. on p. 85).

Akundi, Aditya et al. (Feb. 2022). “State of Industry 5.0—Analysis and Identification of Current Research Trends”. In: *Applied System Innovation* 5.1. ISSN: 25715577. DOI: 10.3390/asi5010027 (cit. on p. 84).

Alexeev, Alexey et al. (2020). “A highly efficient neural network solution for automated detection of pointer meters with different analog scales operating in different conditions”. In: *Mathematics* 8.7, pp. 1–12. ISSN: 22277390. DOI: 10.3390/math8071104 (cit. on p. 24).

Apple (2017). *ARKit* (cit. on pp. 2, 53).

Arroni, Sergio et al. (2023). “Sentiment Analysis and Classification of Hotel Opinions in Twitter With the Transformer Architecture”. In: *International*

Journal of Interactive Multimedia and Artificial Intelligence 8.1, p. 53. ISSN: 1989-1660. DOI: 10.9781/ijimai.2023.02.005 (cit. on p. 92).

Azuma, Ronald et al. (2001). “Recent advances in augmented reality”. In: *IEEE Computer Graphics and Applications* 21.6, pp. 34–47. ISSN: 02721716. DOI: 10.1109/38.963459 (cit. on pp. 1, 78).

Azuma, Ronald T (1997). “A Survey of Augmented Reality”. In: *Presence: Teleoperators and Virtual Environments* 6.4, pp. 355–385. DOI: 10.1162/pres.1997.6.4.355 (cit. on p. 1).

Backs, Richard W and Kimberle A Seljos (1994). “Metabolic and cardiorespiratory measures of mental effort: the effects of level of difficulty in a working memory task”. In: *International Journal of Psychophysiology* 16, pp. 57–68 (cit. on p. 47).

Baldauf, Matthias et al. (Sept. 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.* Association for Computing Machinery, Inc, pp. 119–126. ISBN: 9781450359412. DOI: 10.1145/3236112.3236128 (cit. on pp. 26, 50).

Ball, Geoffrey H and David J Hall (1965). *ISODATA, a novel method of data analysis and pattern classification.* Tech. rep. Stanford research inst Menlo Park CA (cit. on pp. 32, 47).

Barakonyi, István, Thomas Psik, and Dieter Schmalstieg (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. ISBN: 0769521916. DOI: 10.1109/ISMAR.2004.11 (cit. on pp. 28, 49).

Baroroh, Dawi Karomati, Chih Hsing Chu, and Lihui Wang (Oct. 2021). “Systematic literature review on augmented reality in smart manufacturing: Collaboration between human and computational intelligence”. In: *Journal of Manufacturing Systems* 61, pp. 696–711. ISSN: 02786125. DOI: 10.1016/j.jmsy.2020.10.017 (cit. on p. 82).

Bécue, Adrien, Isabel Praça, and João Gama (June 2021). "Artificial intelligence, cyber-threats and Industry 4.0: challenges and opportunities". In: *Artificial Intelligence Review* 54.5, pp. 3849–3886. ISSN: 15737462. DOI: `10.1007/s10462-020-09942-2` (cit. on p. 81).

Benbelkacem, Samir et al. (July 2013). "Augmented reality for photovoltaic pumping systems maintenance tasks". In: *Renewable Energy* 55, pp. 428–437. ISSN: 09601481. DOI: `10.1016/j.renene.2012.12.043` (cit. on pp. 14, 49).

Bertolini, Massimo et al. (Aug. 2021). "Machine Learning for industrial applications: A comprehensive literature review". In: *Expert Systems with Applications* 175, p. 114820. ISSN: 0957-4174. DOI: `10.1016/J.ESWA.2021.114820` (cit. on p. 49).

Billinghurst, Mark, Adrian Clark, and Gun Lee (2015). "A survey of augmented reality". In: *Foundations and Trends in Human-Computer Interaction* 8.2-3, pp. 73–272. ISSN: 15513963. DOI: `10.1561/1100000049` (cit. on pp. 2, 4, 78).

Bissig, David and Cindy Lustig (Aug. 2007). "Who benefits from memory training?" In: *Psychological Science* 18.8, pp. 720–726. ISSN: 09567976. DOI: `10.1111/j.1467-9280.2007.01966.x` (cit. on p. 85).

Bommasani, Rishi et al. (Aug. 2021). "On the Opportunities and Risks of Foundation Models". In: *arXiv preprint arXiv:2108.07258*. DOI: `10.48550/arXiv.2108.07258` (cit. on pp. 48, 84).

Bottani, Eleonora and Giuseppe Vignali (Mar. 2019). "Augmented reality technology in the manufacturing industry: A review of the last decade". In: *IISE Transactions* 51.3, pp. 284–310. ISSN: 24725862. DOI: `10.1080/24725854.2018.1493244` (cit. on pp. 23, 35, 50).

Brooke, John (1996). "SUS - A quick and dirty usability scale". In: *Usability evaluation in industry* 3.3 (cit. on p. 35).

Carpenter, Shana K. et al. (Sept. 2012). "Using Spacing to Enhance Diverse Forms of Learning: Review of Recent Research and Implications for Instruction". In: *Educational Psychology Review* 24.3, pp. 369–378. ISSN: 1040726X. DOI: `10.1007/s10648-012-9205-z` (cit. on p. 85).

Carvalho, Thyago P. et al. (Nov. 2019). “A systematic literature review of machine learning methods applied to predictive maintenance”. In: *Computers and Industrial Engineering* 137. ISSN: 03608352. DOI: `10.1016/j.cie.2019.106024` (cit. on pp. 81, 83).

Casillo, Mario et al. (Dec. 2020). “Chatbot in industry 4.0: An approach for training new employees”. In: *Proceedings of 2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering, TALE 2020.* Institute of Electrical and Electronics Engineers Inc., pp. 371–376. ISBN: 9781728169422. DOI: `10.1109/TALE48869.2020.9368339` (cit. on pp. 50, 83).

Caudell, T.P. and D.W. Mizell (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. ISBN: 0-8186-2420-5. DOI: `10.1109/HICSS.1992.183317` (cit. on pp. 1, 2, 14, 78, 82).

Cepeda, Nicholas J. et al. (May 2006). “Distributed practice in verbal recall tasks: A review and quantitative synthesis”. In: *Psychological Bulletin* 132.3, pp. 354–380. ISSN: 00332909. DOI: `10.1037/0033-2909.132.3.354` (cit. on p. 85).

Chalapathy, Raghavendra and Sanjay Chawla (2019). “Deep Learning for Anomaly Detection: A Survey”. In: *arXiv preprint arXiv:1901.03407*, pp. 1–50. DOI: `10.48550/arXiv.1901.03407` (cit. on pp. 32, 49).

Chidambaram, Subramanian et al. (June 2021). “ProcessAR: An augmented reality-based tool to create in-situ procedural 2D/3D AR Instructions”. In: *DIS 2021 - Proceedings of the 2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere.* Association for Computing Machinery, Inc, pp. 234–249. ISBN: 9781450384766. DOI: `10.1145/3461778.3462126` (cit. on p. 84).

Chollet, Francois (2015). *Keras* (cit. on p. 59).

Chu, Chih Hsing et al. (Oct. 2021). “Augmented reality in smart manufacturing: Enabling collaboration between humans and artificial intelligence”. In: *Journal of Manufacturing Systems* 61, pp. 658–659. ISSN: 02786125. DOI: `10.1016/j.jmsy.2021.05.006` (cit. on pp. 46, 79).

Coli, Elena et al. (May 2020). "Towards Automatic building of Human-Machine Conversational System to support Maintenance Processes". In: *arXiv preprint arXiv:2005.06517*. DOI: `10.48550/arXiv.2005.06517` (cit. on pp. 16, 26, 50).

Cortes, Corinna, Vladimir Vapnik, and Lorenza Saitta (1995). "Support-Vector Networks Editor". In: *Machine Learning* 20, pp. 273–297 (cit. on p. 48).

De Crescenzio, Francesca et al. (2011). "Augmented reality for aircraft maintenance training and operations support". In: *IEEE Computer Graphics and Applications* 31.1, pp. 96–101. ISSN: 02721716. DOI: `10.1109/MCG.2011.4` (cit. on p. 49).

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805*. DOI: `10.48550/arXiv.1810.04805` (cit. on pp. 6, 15).

Drath, Rainer and Alexander Horch (2014). "Industrie 4.0: Hit or hype?" In: *IEEE Industrial Electronics Magazine* 8.2, pp. 56–58. ISSN: 19324529. DOI: `10.1109/MIE.2014.2312079` (cit. on pp. 7, 17).

Elazar, Yanai et al. (Dec. 2021). "Measuring and Improving Consistency in Pretrained Language Models". In: *Transactions of the Association for Computational Linguistics* 9, pp. 1012–1031. ISSN: 2307-387X. DOI: `10.1162/tacl{\_}a{\_}00410` (cit. on p. 91).

Elia, Valerio, Maria Grazia Gnoni, and Alessandra Lanzilotto (Nov. 2016). "Evaluating the application of augmented reality devices in manufacturing from a process point of view: An AHP based model". In: *Expert Systems with Applications* 63, pp. 187–197. ISSN: 09574174. DOI: `10.1016/j.eswa.2016.07.006` (cit. on pp. 20, 82).

Endsley, M. R. (1995). "Toward a theory of situation awareness in dynamic systems". In: *Human Factors* 37.1, pp. 32–64. ISSN: 00187208. DOI: `10.1518/001872095779049543` (cit. on p. 15).

Esen, Hikmet et al. (Oct. 2009). "Artificial neural network and wavelet neural network approaches for modelling of a solar air heater". In: *Expert Systems with Applications* 36.8, pp. 11240–11248. ISSN: 0957-4174. DOI: `10.1016/J.ESWA.2009.02.073` (cit. on p. 49).

Espíndola, Danúbia Bueno et al. (May 2013). "A model-based approach for data integration to improve maintenance management by mixed reality". In: *Computers in Industry* 64.4, pp. 376–391. ISSN: 01663615. DOI: `10.1016/j.compind.2013.01.002` (cit. on p. 49).

Eswaran, M. et al. (Mar. 2023). "Augmented reality-based guidance in product assembly and maintenance/repair perspective: A state of the art review on challenges and opportunities". In: *Expert Systems with Applications* 213, p. 118983. ISSN: 0957-4174. DOI: `10.1016/J.ESWA.2022.118983` (cit. on p. 71).

Eversberg, Leon et al. (Oct. 2022). "A cognitive assistance system with augmented reality for manual repair tasks with high variability based on the digital twin". In: *Manufacturing Letters* 34, pp. 49–52. ISSN: 2213-8463. DOI: `10.1016/J.MFGLET.2022.09.003` (cit. on pp. 48, 69).

Fiorentino, M., G. Monno, and A. E. Uva (2009). "Tangible digital master for product lifecycle management in augmented reality". In: *International Journal on Interactive Design and Manufacturing* 3.2, pp. 121–129. ISSN: 19552513. DOI: `10.1007/s12008-009-0062-z` (cit. on p. 82).

Fiorentino, Michele et al. (Nov. 2013). "Design review of CAD assemblies using bimanual natural interface". In: *International Journal on Interactive Design and Manufacturing* 7.4, pp. 249–260. ISSN: 19552505. DOI: `10.1007/s12008-012-0179-3` (cit. on p. 82).

Fraga-Lamas, Paula et al. (2018). "A Review on Industrial Augmented Reality Systems for the Industry 4.0 Shipyard". In: *IEEE Access* 6, pp. 13358–13375. ISSN: 21693536. DOI: `10.1109/ACCESS.2018.2808326` (cit. on p. 35).

Garza, Luis Eduardo et al. (2013). "Augmented reality application for the maintenance of a flapper valve of a fuller-kynion type m pump". In: *Procedia Computer Science* 25, pp. 154–160. ISSN: 18770509. DOI: `10.1016/j.procs.2013.11.019` (cit. on pp. 6, 14, 23, 49, 79, 86).

Gatara Munyua, John, Geoffrey Mariga Wambugu, and Stephen Thiiru Njenga (2021). *A Survey of Deep Learning Solutions for Anomaly Detection in Surveillance Videos*. Tech. rep. 5, pp. 2279–0764 (cit. on p. 7).

Gattullo, Michele et al. (2019). "Towards augmented reality manuals for industry 4.0: A methodology". In: *Robotics and Computer-Integrated Manufacturing* 56.March 2018, pp. 276–286. ISSN: 07365845. DOI: `10.1016/j.rcim.2018.10.001` (cit. on pp. 6, 47, 79, 80, 84).

Gilchrist, Alasdair (2016). "Introducing Industry 4.0". In: *Industry 4.0*. Springer. Chap. 13, pp. 195–215. ISBN: 978-1-4842-2046-7. DOI: `10.1007/978-1-4842-2047-4` (cit. on pp. 47, 83, 86).

Gonzalez, Andres Vargas et al. (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. ISBN: 9781728109879. DOI: `10.1109/ISMAR.2019.00032` (cit. on p. 19).

Google (2018). *ARCore* (cit. on pp. 2, 53).

Gopaluni, R. Bhushan et al. (Jan. 2020). "Modern Machine Learning Tools for Monitoring and Control of Industrial Processes: A Survey". In: *IFAC-PapersOnLine* 53.2, pp. 218–229. ISSN: 2405-8963. DOI: `10.1016/J.IFACOL.2020.12.126` (cit. on p. 49).

Gorecky, Dominic et al. (2014). "Human-machine-interaction in the industry 4.0 era". In: *2014 12th IEEE International Conference on Industrial Informatics (INDIN)*, pp. 289–294. ISBN: 9781479949052. DOI: `10.1109/INDIN.2014.6945523` (cit. on p. 17).

Griffor, Edward R et al. (June 2017). *Framework for cyber-physical systems: volume 1, overview*. Tech. rep. Gaithersburg, MD: National Institute of Standards and Technology. DOI: `10.6028/NIST.SP.1500-201` (cit. on p. 7).

Guerreiro, Bruno V. et al. (2018). "Definition of smart retrofitting: First steps for a company to deploy aspects of industry 4.0". In: *Advances in Manufacturing*. Vol. 0. Springer Heidelberg, pp. 161–170. ISBN: 9783319666969. DOI: `10.1007/978-3-319-68619-6{\_}16` (cit. on pp. 5, 17, 46).

Hardt, Oliver, Karim Nader, and Lynn Nadel (Mar. 2013). "Decay happens: The role of active forgetting in memory". In: *Trends in Cognitive Sciences*

17.3, pp. 111–120. ISSN: 13646613. DOI: `10.1016/j.tics.2013.01.001` (cit. on p. 85).

Hermann, Mario, Tobias Pentek, and Boris Otto (Mar. 2016). “Design principles for industrie 4.0 scenarios”. In: *Proceedings of the Annual Hawaii International Conference on System Sciences.* Vol. 2016-March. IEEE Computer Society, pp. 3928–3937. ISBN: 9780769556703. DOI: `10.1109/HICSS.2016.488` (cit. on p. 81).

Hou, Lei, Xiangyu Wang, Leonhard Bernold, et al. (Sept. 2013). “Using Animated Augmented Reality to Cognitively Guide Assembly”. In: *Journal of Computing in Civil Engineering* 27.5, pp. 439–451. ISSN: 0887-3801. DOI: `10.1061/(asce)cp.1943-5487.0000184` (cit. on p. 82).

Hou, Lei, Xiangyu Wang, and Martijn Truijens (Jan. 2015). “Using Augmented Reality to Facilitate Piping Assembly: An Experiment-Based Evaluation”. In: *Journal of Computing in Civil Engineering* 29.1. ISSN: 0887-3801. DOI: `10.1061/(ASCE)CP.1943-5487.0000344` (cit. on p. 82).

Huenerfauth, A. M. (2014). “Mobile technology applications for manufacturing, reduction of muda (waste) and the effect on manufacturing economy and efficiency”. In: *International Journal of Interactive Mobile Technologies* 8.4, pp. 20–23. ISSN: 18657923. DOI: `10.3991/ijim.v8i4.3797` (cit. on p. 47).

Ing Tay, Shu et al. (2019). *An Overview of the Rising Challenges in Implementing Industry 4.0.* Tech. rep. 6 (cit. on p. 5).

Ionescu, Radu Tudor et al. (2019). “Object-centric Auto-encoders and Dummy Anomalies for Abnormal Event Detection in Video”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7842–7851 (cit. on p. 49).

Izquierdo-Domenech, Juan, Jordi Linares-Pellicer, and Jorge Orta-Lopez (Sept. 2022). “Towards achieving a high degree of situational awareness and multimodal interaction with AR and semantic AI in industrial applications”. In: *Multimedia Tools and Applications.* ISSN: 1380-7501. DOI: `10.1007/s11042-022-13803-1` (cit. on pp. 79, 84).

Jaschke, Steffen (Jan. 2014). "Mobile learning applications for technical vocational and engineering education: The use of competence snippets in laboratory courses and industry 4.0". In: *Proceedings of 2014 International Conference on Interactive Collaborative Learning, ICL 2014*. Institute of Electrical and Electronics Engineers Inc., pp. 605–608. ISBN: 9781479944378. DOI: `10.1109/ICL.2014.7017840` (cit. on pp. 47, 79).

Javaid, Ahmad Y. et al. (2015). "A deep learning approach for network intrusion detection system". In: *EAI International Conference on Bio-inspired Information and Communications Technologies (BICT)*. ISSN: 24116777. DOI: `10.4108/eai.3-12-2015.2262516` (cit. on p. 49).

Jinyu, Li et al. (2019). "Survey and evaluation of monocular visual-inertial SLAM algorithms for augmented reality". In: *Virtual Reality & Intelligent Hardware* 1.4, pp. 386–410. ISSN: 20965796. DOI: `10.1016/j.vrih.2019.07.002` (cit. on p. 20).

Kadir, Bzhwen A (2020). "DESIGNING NEW WAYS OF WORKING IN INDUSTRY 4.0 Aligning humans, technology, and organization in the transition to Industry 4.0 Bzhwen A Kadir". PhD thesis (cit. on p. 4).

Kagermann, Henning et al. (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 (cit. on pp. 3, 17, 46, 78).

Kamat, Pooja and Rekha Sugandhi (2020). "Anomaly detection for predictive maintenance in industry 4.0-A survey". In: *E3S Web of Conferences*. Vol. 170, pp. 1–8. DOI: `10.1051/e3sconf/202017002007` (cit. on p. 15).

Karpicke, Jeffrey D. and Henry L. Roediger (Feb. 2008). "The critical importance of retrieval for learning". In: *Science* 319.5865, pp. 966–968. ISSN: 00368075. DOI: `10.1126/science.1152408` (cit. on p. 85).

Kato, H. and M. Billinghurst (1999). "Marker tracking and HMD calibration for a video-based augmented reality conferencing system". In: *Proceedings - 2nd IEEE and ACM International Workshop on Augmented Reality, IWAR 1999*. Institute of Electrical and Electronics Engineers Inc., pp. 85–94. ISBN: 0769503594. DOI: `10.1109/IWAR.1999.803809` (cit. on p. 53).

Kollatsch, Christian and Philipp Klimant (June 2021). “Efficient integration process of production data into Augmented Reality based maintenance of machine tools”. In: *Production Engineering* 15.3-4, pp. 311–319. ISSN: 18637353. DOI: `10.1007/s11740-021-01026-6` (cit. on pp. 6, 47, 80, 84).

Lai, Ze Hao et al. (2020). “Smart augmented reality instructional system for mechanical assembly towards worker-centered intelligent manufacturing”. In: *Journal of Manufacturing Systems* 55, pp. 69–81. ISSN: 02786125. DOI: `10.1016/j.jmsy.2020.02.010` (cit. on p. 15).

Lave, Jean and Etienne Wenger (1991). *Situated learning: Legitimate peripheral participation*. Cambridge university press. DOI: `10.1017/CBO9780511815355` (cit. on p. 80).

Liu, Yinhan et al. (July 2019). “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *arXiv preprint arXiv:1907.11692*. DOI: `10.48550/arXiv.1907.11692` (cit. on pp. 28, 55, 92).

Lu, Weining et al. (Sept. 2017). “Unsupervised Sequential Outlier Detection with Deep Architectures”. In: *IEEE Transactions on Image Processing* 26.9, pp. 4321–4330. ISSN: 10577149. DOI: `10.1109/TIP.2017.2713048` (cit. on p. 49).

Luh, Yuan Ping et al. (Oct. 2013). “Augmented reality-based design customization of footwear for children”. In: *Journal of Intelligent Manufacturing* 24.5, pp. 905–917. ISSN: 09565515. DOI: `10.1007/s10845-012-0642-9` (cit. on p. 49).

Makris, Sotiris et al. (2016). “Augmented reality system for operator support in human–robot collaborative assembly”. In: *CIRP Annals - Manufacturing Technology* 65.1, pp. 61–64. ISSN: 17260604. DOI: `10.1016/j.cirp.2016.04.038` (cit. on p. 14).

Malaiya, Ritesh K. et al. (2019). “An Empirical Evaluation of Deep Learning for Network Anomaly Detection”. In: *IEEE Access* 7, pp. 140806–140817. ISSN: 21693536. DOI: `10.1109/ACCESS.2019.2943249` (cit. on p. 7).

Marino, Emanuele et al. (May 2021). “An Augmented Reality inspection tool to support workers in Industry 4.0 environments”. In: *Computers in Industry*

127. ISSN: 01663615. DOI: 10.1016/j.compind.2021.103412 (cit. on p. 79).

Marques, Bernardo, Samuel Silva, Joao Alves, et al. (Dec. 2022). "A Conceptual Model and Taxonomy for Collaborative Augmented Reality". In: *IEEE Transactions on Visualization and Computer Graphics* 28.12, pp. 5113–5133. ISSN: 19410506. DOI: 10.1109/TVCG.2021.3101545 (cit. on p. 82).

Marques, Bernardo, Samuel Silva, João Alves, et al. (Mar. 2022). "Remote collaboration in maintenance contexts using augmented reality: insights from a participatory process". In: *International Journal on Interactive Design and Manufacturing* 16.1, pp. 419–438. ISSN: 19552505. DOI: 10.1007/s12008-021-00798-6 (cit. on p. 82).

Martin, Juhas, Juhasova Bohuslava, and Halenar Igor (Nov. 2018). "Augmented reality in education 4.0". In: *International Scientific and Technical Conference on Computer Sciences and Information Technologies*. Vol. 1, pp. 231–236. ISBN: 9781538664636. DOI: 10.1109/STC-CSIT.2018.8526676 (cit. on p. 6).

Maschler, Benjamin and Michael Weyrich (June 2021). "Deep Transfer Learning for Industrial Automation: A Review and Discussion of New Techniques for Data-Driven Machine Learning". In: *IEEE Industrial Electronics Magazine* 15.2, pp. 65–75. ISSN: 19410115. DOI: 10.1109/MIE.2020.3034884 (cit. on p. 84).

Masood, Tariq and Johannes Egger (Aug. 2019). "Augmented reality in support of Industry 4.0—Implementation challenges and success factors". In: *Robotics and Computer-Integrated Manufacturing* 58, pp. 181–195. ISSN: 07365845. DOI: 10.1016/j.rcim.2019.02.003 (cit. on p. 79).

Milgram, Paul et al. (1995). "Augmented reality: a class of displays on the reality-virtuality continuum". In: *Telemanipulator and Telepresence Technologies* 2351, pp. 282–292. ISSN: 0277786X. DOI: 10.1117/12.197321 (cit. on pp. 1, 2).

Mleczko, Katarzyna (Sept. 2021). "Chatbot as a Tool for Knowledge Sharing in the Maintenance and Repair Processes". In: *Multidisciplinary Aspects of Production Engineering* 4.1, pp. 499–508. DOI: 10.2478/mape-2021-0045 (cit. on p. 50).

Monroy Reyes, Alejandro et al. (Nov. 2016). “A mobile augmented reality system to support machinery operations in scholar environments”. In: *Computer Applications in Engineering Education* 24.6, pp. 967–981. ISSN: 10990542. DOI: `10.1002/cae.21772` (cit. on p. 49).

Moroff, Nikolas Ulrich, Ersin Kurt, and Josef Kamphues (2021). “Machine Learning and Statistics: A Study for assessing innovative Demand Forecasting Models”. In: *Procedia Computer Science* 180, pp. 40–49. ISSN: 18770509. DOI: `10.1016/j.procs.2021.01.127` (cit. on p. 83).

Mourtzis, Dimitris, Vasileios Siatras, and John Angelopoulos (Mar. 2020). “Real-time remote maintenance support based on augmented reality (AR)”. In: *Applied Sciences (Switzerland)* 10.5. ISSN: 20763417. DOI: `10.3390/app10051855` (cit. on pp. 4, 15, 49, 50, 83).

Ng, L. X. et al. (June 2011). “GARDE: A gesture-based augmented reality design evaluation system”. In: *International Journal on Interactive Design and Manufacturing* 5.2, pp. 85–94. ISSN: 19552513. DOI: `10.1007/s12008-011-0117-9` (cit. on p. 4).

Ong, S. K. and Y. Shen (2009). “A mixed reality environment for collaborative product design and development”. In: *CIRP Annals - Manufacturing Technology* 58.1, pp. 139–142. ISSN: 00078506. DOI: `10.1016/j.cirp.2009.03.020` (cit. on pp. 23, 49).

Ong, S. K. and Z. B. Wang (2011). “Augmented assembly technologies based on 3D bare-hand interaction”. In: *CIRP Annals - Manufacturing Technology* 60.1, pp. 1–4. ISSN: 00078506. DOI: `10.1016/j.cirp.2011.03.001` (cit. on pp. 4, 49, 82).

P, Karrupusamy (Jan. 2021). “Machine Learning Approach to Predictive Maintenance in Manufacturing Industry - A Comparative Study”. In: *Journal of Soft Computing Paradigm* 2.4, pp. 246–255. DOI: `10.36548/jscp.2020.4.006` (cit. on p. 7).

Palmarini, Riccardo, John Ahmet Erkoyuncu, et al. (2018). “A systematic review of augmented reality applications in maintenance”. In: *Robotics and Computer-Integrated Manufacturing* 49, pp. 215–228. ISSN: 07365845. DOI: `10.1016/j.rcim.2017.06.002` (cit. on p. 47).

Palmarini, Riccardo, Iñigo Fernández, et al. (2022). “Fast Augmented Reality Authoring: Fast Creation of AR step-by-step Procedures for Maintenance Operations”. In: *IEEE Access*. DOI: `10.1109/ACCESS.2017.DOI` (cit. on pp. 4, 83).

Pang, Guansong et al. (2020). “Self-trained Deep Ordinal Regression for End-to-End Video Anomaly Detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12173–12182 (cit. on p. 49).

Paolanti, Marina et al. (Aug. 2018). “Machine Learning approach for Predictive Maintenance in Industry 4.0”. In: *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications, MESA 2018*. Institute of Electrical and Electronics Engineers Inc. ISBN: 9781538646434. DOI: `10.1109/MESA.2018.8449150` (cit. on p. 7).

Pejsa, Tomislav et al. (Feb. 2016). “Room2Room: Enabling Life-Size telepresence in a projected augmented reality environment”. In: *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. Vol. 27. Association for Computing Machinery, pp. 1716–1725. DOI: `10.1145/2818048.2819965` (cit. on pp. 2, 3).

Peres, Ricardo Silva et al. (2020). “Industrial Artificial Intelligence in Industry 4.0 -Systematic Review, Challenges and Outlook”. In: *IEEE Access*. ISSN: 21693536. DOI: `10.1109/ACCESS.2020.3042874` (cit. on p. 7).

Peruzzini, Margherita, Fabio Grandi, and Marcello Pellicciari (Jan. 2020). “Exploring the potential of Operator 4.0 interface and monitoring”. In: *Computers and Industrial Engineering* 139. ISSN: 03608352. DOI: `10.1016/j.cie.2018.12.047` (cit. on p. 47).

Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi (2002). “MultiWordNet: developing an aligned multilingual database”. In: *First international conference on global WordNet*, pp. 293–302 (cit. on p. 16).

Pierdicca, Roberto et al. (2017). “The use of augmented reality glasses for the application in industry 4.0”. In: *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*. Vol. 10324 LNCS, pp. 389–401. ISBN: 9783319609218. DOI: `10.1007/978-3-319-60922-5{\_}30` (cit. on p. 6).

Prause, Martin (Oct. 2019). "Challenges of Industry 4.0 technology adoption for SMEs: The case of Japan". In: *Sustainability (Switzerland)* 11.20. ISSN: 20711050. DOI: 10.3390/su11205807 (cit. on p. 5).

PTC (2017). *Vuforia Chalk* (cit. on p. 50).

Quint, Fabian and Frieder Loch (2015). "Using smart glasses to document maintenance processes". In: *Mensch und Computer 2015–Workshopband*, pp. 203–208. ISSN: 3110443902. DOI: 10.1515/9783110443905-030 (cit. on pp. 6, 80, 84).

Rabelo, Ricardo J, David Romero, and Saulo Popov Zambiasi (2018). "Softbots Supporting the Operator 4.0 at Smart Factory Environments". In: *IFIP International Conference on Advances in Production Management Systems (APMS)*, pp. 456–464. DOI: 10.1007/978-3-319-99707-0{_}57 (cit. on p. 47).

Radford, Alec, Jong Wook Kim, et al. (Dec. 2022). "Robust Speech Recognition via Large-Scale Weak Supervision". In: *arXiv preprint arXiv:2212.04356* (cit. on p. 100).

Radford, Alec, Karthik Narasimhan, et al. (2018). "Improving Language Understanding by Generative Pre-Training". In: *OpenAI* (cit. on pp. 84, 92).

Radkowski, Rafael, Jordan Herrema, and James Oliver (May 2015). "Augmented Reality-Based Manual Assembly Support With Visual Features for Different Degrees of Difficulty". In: *International Journal of Human-Computer Interaction* 31.5, pp. 337–349. ISSN: 15327590. DOI: 10.1080/10447318.2014.994194 (cit. on pp. 14, 49).

Raffel, Colin et al. (Oct. 2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *The Journal of Machine Learning Research* 21.1, pp. 5485–5551. DOI: 10.48550/arXiv.1910.10683 (cit. on pp. 6, 28, 55, 92).

Rai, Rahul et al. (2021). *Machine learning in manufacturing and industry 4.0 applications.* DOI: 10.1080/00207543.2021.1956675 (cit. on p. 7).

Rajpurkar, Pranav et al. (June 2016). "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *arXiv preprint arXiv:1606.05250.* DOI: 10.48550/arXiv.1606.05250 (cit. on pp. 29, 55).

Rasmussen, Troels et al. (Oct. 2022). "Supporting workspace awareness in remote assistance through a flexible multi-camera system and Augmented Reality awareness cues". In: *Journal of Visual Communication and Image Representation*, p. 103655. ISSN: 1047-3203. DOI: 10.1016/J.JVCIR.2022.103655 (cit. on pp. 48, 68).

Redmon, Joseph and Ali Farhadi (2018). "YOLOv3: An Incremental Improvement". In: *arXiv preprint arXiv:1804.02767.* DOI: 10.48550/arXiv.1804.02767 (cit. on pp. 7, 84).

Rekimoto, Jun (1995). "The World through the Computer: Computer Augmented Interaction with Real World Environments". In: *Proceedings of the 8th annual ACM symposium on User interface and software technology*, pp. 29–36 (cit. on pp. 4, 5).

Ribeiro, Jorge et al. (2021). "Robotic Process Automation and Artificial Intelligence in Industry 4.0 - A Literature review". In: *Procedia Computer Science.* Vol. 181. Elsevier B.V., pp. 51–58. DOI: 10.1016/j.procs.2021.01.104 (cit. on p. 81).

Rombach, Robin et al. (2022). "High-Resolution Image Synthesis with Latent Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695. DOI: 10.48550/arXiv.2112.10752 (cit. on p. 84).

Romero, David and Johan Stahre (2021). "Towards the Resilient Operator 5.0: The Future of Work in Smart Resilient Manufacturing Systems". In: *Procedia CIRP* 104, pp. 1089–1094. ISSN: 22128271. DOI: 10.1016/j.procir.2021.11.183 (cit. on p. 47).

Romero, David, Johan Stahre, and Marco Taisch (Jan. 2020). *The Operator 4.0: Towards socially sustainable factories of the future.* DOI: 10.1016/j.cie.2019.106128 (cit. on p. 47).

Romero, David, Johan Stahre, Thorsten Wuest, et al. (2016). "Towards an Operator 4.0 Typology: A Human-Centric Perspective on the Fourth Industrial Revolution Technologies". In: *Proceedings of the international conference on computers and industrial engineering (CIE46)*. Tianjin, China, pp. 29–31 (cit. on p. 47).

Rožanec, Jože M. et al. (Nov. 2022). "Human-centric artificial intelligence architecture for industry 5.0 applications". In: *International Journal of Production Research*, pp. 1–26. ISSN: 0020-7543. DOI: `10.1080/00207543.2022.2138611` (cit. on p. 84).

Runji, Joel Murithi, Yun-Ju Lee, and Chih-Hsing Chu (June 2022). "Systematic Literature Review on Augmented Reality-Based Maintenance Applications in Manufacturing Centered on Operator Needs". In: *International Journal of Precision Engineering and Manufacturing-Green Technology.* ISSN: 2288-6206. DOI: `10.1007/s40684-022-00444-w` (cit. on pp. 46, 47).

Rushe, Ellen and Brian Mac Namee (2019). "Anomaly detection in raw audio using deep autoregressive networks". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3597–3601. ISBN: 9781538646588 (cit. on p. 7).

Sääski, Juha et al. (2008). "Integration of design and assembly using augmented reality". In: *Micro-Assembly Technologies and Applications: IFIP TC5 WG5. 5 Fourth International Precision Assembly Seminar (IPAS'2008).* Chamonix, France: Springer, pp. 295–404. DOI: `10.1007/978-0-387-77405-3{\_}39` (cit. on p. 82).

Sahu, Chandan K., Crystal Young, and Rahul Rai (2021). "Artificial intelligence (AI) in augmented reality (AR)-assisted manufacturing applications: a review". In: *International Journal of Production Research* 59.16, pp. 4903–4959. ISSN: 1366588X. DOI: `10.1080/00207543.2020.1859636` (cit. on p. 48).

Salonen, Tapio et al. (2009). "Data pipeline from CAD to AR based assembly instructions". In: *Proceedings of the ASME/AFM World Conference on Innovative Virtual Reality 2009, WINVR2009*, pp. 165–168. ISBN: 9780791843376. DOI: `10.1115/WINVR2009-705` (cit. on p. 82).

Sandi, Carmen (May 2013). "Stress and cognition". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 4.3, pp. 245–261. ISSN: 19395078. DOI: `10.1002/wcs.1222` (cit. on p. 47).

Sanh, Victor et al. (Oct. 2019). “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: *arXiv preprint arXiv:1910.01108*. DOI: `10.48550/arXiv.1910.01108` (cit. on pp. 28, 55, 92).

Scao, Teven Le et al. (Nov. 2022). “BLOOM: A 176B-Parameter Open-Access Multilingual Language Model”. In: *arXiv preprint arXiv:2211.05100*. DOI: `10.48550/arXiv.2211.05100` (cit. on p. 92).

Scurati, Giulia Wally et al. (June 2018). “Converting maintenance actions into standard symbols for Augmented Reality applications in Industry 4.0”. In: *Computers in Industry* 98, pp. 68–79. ISSN: 01663615. DOI: `10.1016/j.compind.2018.02.001` (cit. on pp. 14, 49, 82).

Sereno, Mickael et al. (June 2022). “Collaborative Work in Augmented Reality: A Survey”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.6, pp. 2530–2549. ISSN: 19410506. DOI: `10.1109/TVCG.2020.3032761` (cit. on p. 82).

Sevinç, Ali, Şeyda Gür, and Tamer Eren (Dec. 2018). “Analysis of the difficulties of SMEs in industry 4.0 applications by analytical hierarchy process and analytical network process”. In: *Processes* 6.12. ISSN: 22279717. DOI: `10.3390/pr6120264` (cit. on p. 5).

Shen, Y., S. K. Ong, and A. Y.C. Nee (Mar. 2010). “Augmented reality for collaborative product design and development”. In: *Design Studies* 31.2, pp. 118–145. ISSN: 0142694X. DOI: `10.1016/j.destud.2009.11.001` (cit. on pp. 4, 49).

Sheu, Phillip C-y (2010). *Semantic computing*. Wiley Online Library. ISBN: 9780470464953 (cit. on p. 48).

Shneiderman, Ben et al. (2016). *Designing the user interface: strategies for effective human-computer interaction*. Pearson (cit. on p. 52).

Simon, G., A. W. Fitzgibbon, and A. Zisserman (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. ISBN: 0769508464. DOI: `10.1109/ISAR.2000.880935` (cit. on p. 20).

Škvára, Vít, Tomáš Pevný, and Václav Šmídl (2018). "Are generative deep models for novelty detection truly better?" In: *arXiv preprint arXiv:1807.05027*. DOI: `10.48550/arXiv.1807.05027` (cit. on p. 49).

Song, Xiaona et al. (June 2022). "Event-driven NN adaptive fixed-time control for nonlinear systems with guaranteed performance". In: *Journal of the Franklin Institute* 359.9, pp. 4138–4159. ISSN: 0016-0032. DOI: `10.1016/J.JFRANKLIN.2022.04.003` (cit. on p. 49).

Sweller, John (Apr. 1988). "Cognitive Load During Problem Solving: Effects on Learning". In: *Cognitive Science* 12.2, pp. 257–285. DOI: `10.1207/s15516709cog1202{\_}4` (cit. on p. 47).

Syberfeldt, Anna et al. (2016). "Dynamic Operator Instructions Based on Augmented Reality and Rule-based Expert Systems". In: *Procedia CIRP* 41, pp. 346–351. ISSN: 22128271. DOI: `10.1016/j.procir.2015.12.113` (cit. on p. 83).

Tatić, Dušan and Bojan Tešić (Feb. 2017). "The application of augmented reality technologies for the improvement of occupational safety in an industrial environment". In: *Computers in Industry* 85, pp. 1–10. ISSN: 01663615. DOI: `10.1016/j.compind.2016.11.004` (cit. on pp. 49, 83).

Tay, Yi, Mostafa Dehghani, et al. (May 2022). "UL2: Unifying Language Learning Paradigms". In: *arXiv preprint arXiv:2205.05131*. DOI: `10.48550/arXiv.2205.05131` (cit. on p. 92).

Tay, Yi, Jason Wei, et al. (Oct. 2022). "Transcending Scaling Laws with 0.1% Extra Compute". In: *arXiv preprint arXiv:2210.11399*. DOI: `10.48550/arXiv.2210.11399` (cit. on p. 92).

TeamViewer (2021). *TeamViewer Assist AR* (cit. on p. 50).

The Linux Foundation (2017). *ONNX: Open Neural Network Exchange* (cit. on p. 59).

Together (2022). *GPT-JT* (cit. on p. 92).

Tony Liu, Fei, Kai Ming Ting, and Zhi-Hua Zhou (2008). "Isolation Forest". In: *Eighth IEEE International Conference on Data Mining*. IEEE, pp. 413–422 (cit. on p. 32).

Tsai, Roger Y. (1987). "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses". In: *IEEE Journal on Robotics and Automation* 3.4, pp. 323–344. ISSN: 08824967. DOI: 10.1109/JRA.1987.1087109 (cit. on p. 20).

Unity (2018). *AR Foundation* (cit. on p. 53).

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in neural information processing systems.* Vol. 30, pp. 6000–6010. DOI: 10.48550/arXiv.1706.03762 (cit. on pp. 6, 15, 48, 81).

Ventura, Cheryl A (2000). "Why switch from paper to electronic manuals?" In: *Proceedings of the ACM conference on Document processing systems*, pp. 111–116. DOI: 10.1145/62506.62525 (cit. on pp. 6, 80, 84).

Wang, Peng et al. (Dec. 2021). "AR/MR Remote Collaboration on Physical Tasks: A Review". In: *Robotics and Computer-Integrated Manufacturing* 72. ISSN: 07365845. DOI: 10.1016/j.rcim.2020.102071 (cit. on p. 82).

Wang, X., S. K. Ong, and A. Y.C. Nee (Mar. 2016). "A comprehensive survey of augmented reality assembly research". In: *Advances in Manufacturing* 4.1, pp. 1–22. ISSN: 21953597. DOI: 10.1007/s40436-015-0131-4 (cit. on pp. 35, 47, 82).

Wang, Zhuo et al. (Jan. 2021). "The role of user-centered AR instruction in improving novice spatial cognition in a high-precision procedural task". In: *Advanced Engineering Informatics* 47. ISSN: 14740346. DOI: 10.1016/j.aei.2021.101250 (cit. on pp. 48, 69).

Webel, Sabine et al. (Apr. 2013). "An augmented reality training platform for assembly and maintenance skills". In: *Robotics and Autonomous Systems* 61.4, pp. 398–403. ISSN: 09218890. DOI: 10.1016/j.robot.2012.09.013 (cit. on p. 49).

Wolf, Matthias et al. (2018). “Current and future industrial challenges: demographic change and measures for elderly workers in industry 4.0”. In: *Annals of the Faculty of Engineering Hunedoara* 16.1, pp. 67–76 (cit. on p. 85).

Womack, James P, Daniel T Jones, and Daniel Roos (1992). “A máquina que mudou o mundo”. In: *Campus: Rio de Janeiro* (cit. on pp. 5, 46).

Xu, Li Da, Eric L. Xu, and Ling Li (2018). “Industry 4.0: State of the art and future trends”. In: *International Journal of Production Research* 56.8, pp. 2941–2962. ISSN: 1366588X. DOI: 10.1080/00207543.2018.1444806 (cit. on p. 46).

Xu, Xun et al. (Oct. 2021). “Industry 4.0 and Industry 5.0—Inception, conception and perception”. In: *Journal of Manufacturing Systems* 61, pp. 530–535. ISSN: 02786125. DOI: 10.1016/j.jmsy.2021.10.006 (cit. on p. 79).

Yu, Wenhao et al. (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 (cit. on p. 16).

Yuan, M. L., S. K. Ong, and A. Y.C. Nee (Apr. 2008). “Augmented reality for assembly guidance using a virtual interactive tool”. In: *International Journal of Production Research* 46.7, pp. 1745–1767. ISSN: 00207543. DOI: 10.1080/00207540600972935 (cit. on pp. 4, 23, 49, 82).

Zafrir, Ofir et al. (Nov. 2021). “Prune Once for All: Sparse Pre-Trained Language Models”. In: *arXiv preprint arXiv:2111.05754*. DOI: 10.48550/arXiv.2111.05754 (cit. on pp. 29, 59).

Zambiasi, Lara P et al. (2022). “Supporting Resilient Operator 5.0: An Augmented Softbot Approach”. In: *IFIP International Conference on Advances in Production Management Systems*, pp. 494–502 (cit. on p. 47).

Zamora-Hernández, Mauricio Andrés et al. (2021). “Deep learning-based visual control assistant for assembly in Industry 4.0”. In: *Computers in Industry* 131. ISSN: 01663615. DOI: 10.1016/j.compind.2021.103485 (cit. on pp. 7, 49).

Zenati, Houssam et al. (Feb. 2018). “Efficient GAN-Based Anomaly Detection”. In: *arXiv preprint arXiv:1802.06222*. DOI: 10.48550/arXiv.1802.06222 (cit. on p. 49).

Zenati, Nadia, Noureddine Zerhouni, and Karim Achour (2004). “Assistance to maintenance in industrial process using an augmented reality system”. In: *Proceedings of the IEEE International Conference on Industrial Technology*. Vol. 2, pp. 848–852. DOI: 10.1109/icit.2004.1490185 (cit. on p. 49).

Zhang, Jie et al. (Oct. 2022). “Projected augmented reality assembly assistance system supporting multi-modal interaction”. In: *The International Journal of Advanced Manufacturing Technology 2022 123:3* 123.3, pp. 1353–1367. ISSN: 1433-3015. DOI: 10.1007/S00170-022-10113-6 (cit. on pp. 48, 69).

Ziaei, Z. et al. (Oct. 2011). “Real-time markerless Augmented Reality for Remote Handling system in bad viewing conditions”. In: *Fusion Engineering and Design* 86.9-11, pp. 2033–2038. ISSN: 09203796. DOI: 10.1016/j.fusengdes.2010.12.082 (cit. on pp. 49, 83).

Zonta, Tiago et al. (Dec. 2020). “Predictive maintenance in the Industry 4.0: A systematic literature review”. In: *Computers and Industrial Engineering* 150. ISSN: 03608352. DOI: 10.1016/j.cie.2020.106889 (cit. on pp. 15, 47, 83).