ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

SCIENZA E CULTURA DEL BENESSERE E DEGLI STILI DI VITA

Ciclo 36

Settore Concorsuale: 01/B1 - INFORMATICA

Settore Scientifico Disciplinare: INF/01 - INFORMATICA

ENTANGLING ARTIFICIAL INTELLIGENCE WITH EXTENDED REALITY PARADIGMS FOR HUMAN ACTIVITY SUPPORT

Presentata da: Lorenzo Stacchio

Coordinatore Dottorato

Laura Bragonzoni

Supervisore

Gustavo Marfia

Co-supervisore

Giuseppe Lisanti

Esame finale anno 2024

"*The advance of technology is based on making it fit in so that you don't really even notice it, so it's part of everyday life.*"

Bill Gates - Co-founder of Microsoft

"*Our adventure in the Digital World might be over for now, but that gate won't stay closed forever. I have a feeling that this won't be the last time we see our pals. You wait and see...One day that portal will open again and we'll return to the Digital World. I know I'll never forget Agumon or the rest of the Digimon. None of us will!*"

Tai Kamiya, Digidestined

Declaration

I here declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is entirely based on my work and contains nothing that is the outcome of work done in collaboration with others, except as specifically stated. The contributions reported in this thesis have been partially published in [362, 363, 357, 360, 364, 368, 366, 365, 358, 367, 38, 251, 370, 369, 335, 14, 10, 361, 93, 371, 359, 372].

January 2024

Abstract

Despite the "Metaverse" being a fairly recent inclusion in scholars' vocabulary, technologies needed to allow its creation and enable its use cases are now approaching their maturity. In the Metaverse, eXtended Reality (XR) covers the fundamental role of providing natural and ergonomic ways to visualize and manipulate the elements in such a digital layer. However, the rapid development of Metaverse use cases is also driven by advancements in other disciplines, such as Artificial Intelligence (AI), blockchains, and the Internet of Things. In particular, XR and AI are jointly defining a new body of research at their intersection, here named "Extended Artificial Intelligence", that has potentially far-reaching effects on sectors like industrial production, health care, fitness, archival, creativity, and cultural heritage. With the "Extended Artificial Intelligence", we could envision an entire digital world, that could define new ways of living, by supporting humans in their everyday activities. Hence, it becomes fundamental to study how XR and AI should be orchestrated and composed, from both an academic and industrial perspective, to support humans in every aspect of their lives. In such a research space, this thesis focuses on studying those dynamics in different use cases regarding three fields of study: Cultural Heritage, Creative industries, and Industrial production.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we offer an overview of the thesis's background and its contextual framework. Additionally, we present the research inquiries that guided the investigation. To conclude, we provide an outline of the thesis.

1.1 Background

The term "Metaverse" is a recent inclusion in scholars' vocabulary but was introduced in 1992 through Neal Stephenson's novel, "Snow Crash", where the "Metaverse" was portrayed as a virtual reality (VR) environment that leverages the Internet and augmented reality (AR) using avatars and software agents [106, 3].

A generalization of such a concept was embraced by Mark Zuckerberg when he launched Meta, envisioning the Metaverse as a unified and immersive ecosystem in which the divisions between the digital and physical realms are invisible to users. Within this vision, the Metaverse enables immersive features, while including realistic avatars and holograms for work, interaction, and socialization through simulated shared experiences [454]. An academic definition of the Metaverse came in [84] where the authors defined it as the layer between a subject and reality, a 3D virtual shared world where all activities can be carried out with the help of AR and VR paradigms.

Despite the latter definition referring to AR and VR, we can create a generalization of such definition, replacing them with eXtended Reality (XR). The term XR refers to all real-and-virtual combined environments and human-machine interactions generated by computer technology and wearables, where the 'X' represents a variable for any current or future spatial computing technologies, that allow blending or extending the whole set of the possible realities [106, 324, 176]. In other words, it is an umbrella term encapsulating AR, Mixed Reality (MR), Augmented Virtuality (AV), VR, and all the other possible degrees of reality.

More in detail, AR refers to a real-time direct or indirect view of the physical world environment that has been enhanced/augmented by adding virtual computer-generated information to it [242, 59]. Increasing the level of immersiveness, MR allows users to place virtual objects in a context-aware fashion making them indistinguishable from real ones. At the same time, MR paradigms provide natural interaction logic to manipulate virtual objects, which facilitates visualization and manipulation of 3D models and scenes. There is however still a debate about an objective definition of MR [105, 356]. Augmented Virtuality (AV) refers instead to an immersive VR experience that includes elements from the real world, including physical objects, environments, or people into a virtual space, enhancing the virtual experience with aspects of the real world. Finally, VR refers to all those experiences that let users be fully immersed in a three-dimensional computer-generated simulation that can be explored and interactable. It is worth noticing that, when we refer to XR we also refer to haptic controls, holograms, and, in general, all-immersive techniques, which can either enhance or deceive our natural senses or both [113]. Different XR applications require different spatial devices and technological stacks based on the use case and the considered reality(ies). It is worth noticing that, an emerging paradigm called Cross-Reality allows users to live the same experience through different degrees of reality with several benefits from both cost and performance perspectives [332].

Considering the integration of XR technologies, we can now define the

Metaverse as a *3D digital and shared layer over the physical reality, where all activities are enhanced with the support of XR paradigms* [106, 324, 226]. In such a definition, XR paradigms cover a fundamental role: providing natural and ergonomic ways to visualize and manipulate the elements in such a digital layer [242]. Given the recent advancements, and exhibited flexibility of XR paradigms, systems built on top of these technologies are no longer limited to the gaming sphere: their possible range of applications has rapidly grown [168, 259, 353], driving a transformation that could potentially have far-reaching effects on sectors like industrial production, education, tourism, health care, archival, creativity, cultural heritage, and research [168, 259, 353, 105, 226, 266, 87, 431]. In the future, XR will provide an effortless shift between the physical and virtual realms, improving our experiences and interactions, and unlocking a boundless realm of opportunities, that will support researchers, consumers, and professionals in their everyday activities, unlocking the full potential of the Metaverse.

1.2 Extended Artificial Intelligence

The rapid evolution of Metaverse-related applications is not only driven by XR, but also by advancements in other disciplines, such as Artificial Intelligence (AI), Computer Hardware Architecture Design, Networking and Wireless technologies (e.g., 5G), the Internet of Things (IoT), and Blockchains [307, 325, 134, 106, 162, 345, 324, 176, 3, 226, 178, 266, 135]. Those technologies enable different functionalities and use cases, including but not limited to object and action recognition to enable AR/MR experiences, immersive analytics/simulations for research, and collaborative experiences for scenarios such as immersive e-commerce. Holistically, those allow the definition of new paradigms, such as the recently introduced Digital Twins (DTs), which allow for a virtual replicate of physical world entities, with a bidirectional binding. DT's main use cases particularly benefit from XR, AI, and IoT, providing unique ways to make predictions, data generation, simulations,

and visualizations of past, current, and future DT states [26, 431]. With the aligned usage of the mentioned technology, we could envision the Metaverse as a novel form of Internet application, that exploits XR to provide immersive experiences on different degrees of reality while establishing an economic framework relying on blockchain technology and seamlessly exploits AI, IoT, and DTs to define a new way of living our everyday life [266].

In particular, AI is a field of study that attempts to understand and build intelligent entities, taking inspiration from the underlying mechanisms of our brain: how we can perceive, understand, predict, and manipulate a world far larger and more complicated than itself [318]. AI encompasses a huge variety of subfields, ranging from constraint and logic programming to Machine and Deep Learning, for solving various tasks such as planning, classification, object detection, and data generation. Despite the general attention to the joint usage of all the mentioned technologies, AI was the one that was most applied in combination with XR. The combination of those has defined a new body of research at their intersection, here named "Extended Artificial Intelligence", that will cover a fundamental role in building the Metaverse [420, 162, 226].

As described in [420, 162], XR researchers employed AI methods to solve problems like foveated rendering, object tracking, immersive content generation, diminished reality, improving natural interaction systems, virtual agent communications, and situated predictions. AI researchers, instead, adopted XR technologies to address issues such as understandability and explainability, by, for example, visualizing neural networks (NN) in VR, or visualizing predictions of NN for different use cases (e.g., medical) and also for the generation of synthetic dataset [307, 325, 106, 345, 324, 176, 3, 178]. In such a context, an increasing body of research is orchestrating, composing, and pipelining XR and AI paradigms for the benefit of several fields of study, from both an academic and industrial perspective [420, 162].

This novel cross-research field stimulated contributions in multidisciplinary contexts, including cultural heritage, industrial production, and creative in-

dustries, which will cover fundamental roles within the Metaverse [286, 268, 162, 267]. The "Extended Artificial Intelligence" will cover a fundamental role in building a digital layer over-imposed on our physical reality to support humans in everyday activities. Hence, it becomes fundamental to study its use cases, paradigms, and applications, from both an academic and industrial perspective [420, 162].

1.3 Contributions to the Extended Artificial Intelligence Ecosystem

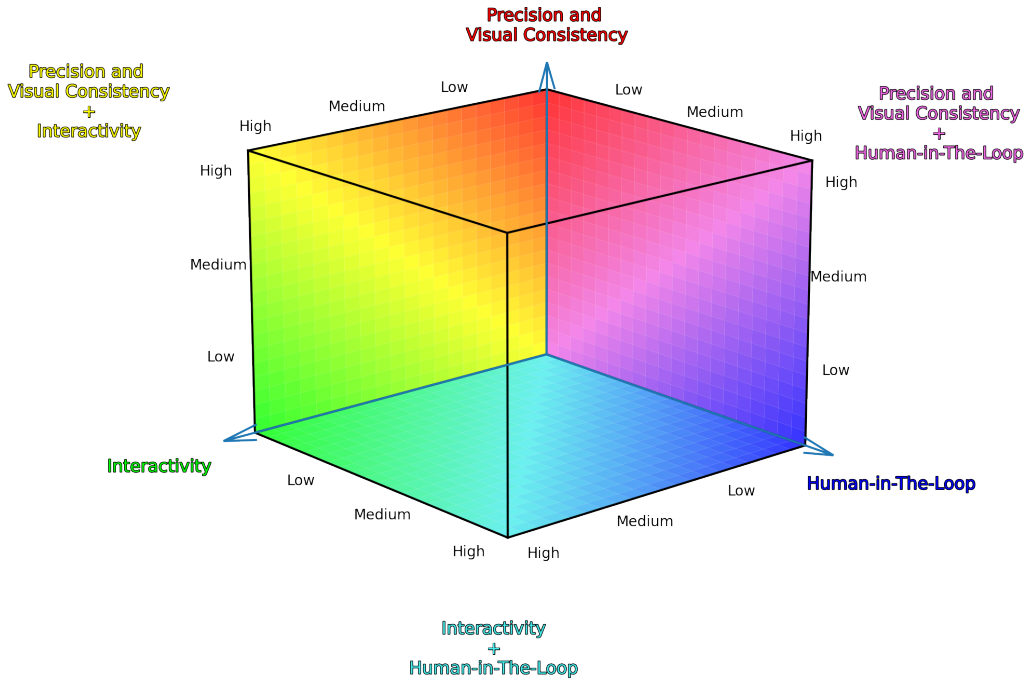


Figure 1.1: 3D Taxonomy defining the domain space for our contributions in the XR and AI domain.

We will now contextualize the domain space of this thesis and its contri-

bution regarding "Extended Artificial Intelligence", highlighting the major factors of analysis and projecting them as axes in a representative research space. Taking inspiration from the taxonomy introduced in [121], we here introduce an adaption of it (depicted in Figure 1.1): the three considered factors (axis) here examined, regarding the combination of XR and AI, amounts to *Interactivity*, *Human-In-The-Loop*, and *Precision and Visual Consistency*.

Interactivity is a well-explored conceptual construct [121] that garnered multiple definitions over time, but we here consider the one that interprets the degree of interactivity in visualization as *its ability to contextually provide users with the opportunity to engage in interactive behavior synchronously with real-time updates* [121].

The second regards the *Precision and Visual Consistency* continuum [121]. With the term precision, we refer to all the elements that can contribute to enhancing the reliability of what is visualized. Visual consistency refers instead to the ability of a visualization to display photo-realistic qualities, making it easier for the viewer to suspend disbelief.

The third one corresponds to *Human-in-the-loop* (HITL) which involves human supervision and decision-making in automated or semi-automated systems (driven by AI), providing the ability to remotely monitor and control them, with customization functionalities [228, 282, 426]. We provide this axis instead of the original Automation one detailed in [121], considering the strong role of Humans in XR research, and also considering the use cases here analyzed.

Each of the mentioned dimensions was categorized based on their degree of contribution as *Low*, *Medium*, or *High*. Those can also be cross-projected defining compositional axes (e.g., HITL + Interactivity as reported in Figure 1.1).

Considering such a domain space, in the following section, we contextualize the different research directions, applications, and use cases analyzed in this thesis. As mentioned, Cultural Heritage, Creative Industries, and Industrial Production & Retail where XR and AI, together considered, have

already provided significant contributions [84, 324, 162]. We so focused on such application domains, considering new and different perspectives, tasks, and contexts for each one. Such contexts of use, however, were also selected because they are predicted to be among the first where Extended Artificial Intelligence will be pervasively adopted [8, 234, 342, 243]. This also means that, as often happens in technology, the paradigms developed in such contexts would also be adopted in many others [306].

It is worth noticing how the research that has been carried out delves into a spectrum ranging from purely academic contexts to industrial applications: Cultural Heritage contributions are almost entirely related to academic settings, while Creative Industries often fall at the intersection between industry and academia.

1.3.1 Cultural Heritage

In Cultural Heritage (CH), XR and AI were combined for the conservation, cataloging, visualization, and restoration of tangible and intangible items [156, 18, 402, 162]. We focused on the definition of an "Extended Artificial Intelligence" pipeline to solve a specific task in CH, which was not previously considered in the literature: historical family photo album cataloging and preservation. We employed Deep Learning for Computer Vision paradigms to identify and classify those pictures. Then we exploited the representation learned by those models to contribute to both historical mixed qualitative-quantitative and cross-cultural influences analysis. We then deployed such models to augment and automatically catalog the photos utilizing AR paradigms, performing a user experience analysis, to evaluate and validate the proposed interface [84, 324]. The path taken involves automation through AI paradigms, which however includes a strong HITL component, while carefully providing a way to revive those materials through XR paradigms.

1.3.2 Creative Industries

Creative industries and research, such as fashion, music, and arts, are being driven by AI and XR in several ways: intelligent assistance, virtual avatar synthesis, 3D item visualization and manipulation, and synthetic generation [302, 115, 350, 145, 162]. It is worth noticing that creative fields are considered synergic but yet different concerning CH [81, 271]. The former is related to traditional forms of preservation and creation of human heritage, while the latter includes the applied practices and innovations for generating tangible or intangible items that could also be related to profit and creation of jobs by creating intellectual property [81, 271]. In this thesis, we approached such a field, considering in particular the fashion arena. We analyzed which kind of technological paradigms, involving both XR and AI, could be applied to enable commerce-related experiences in the Metaverse, concentrating our efforts on VR, intelligent vocal assistants, Human DT, and aesthetic and emotional perception. For all the considered technologies, we performed a user study, to evaluate the proposed frameworks, and applications, analyzing their possible utility and adoption, at the intersection between HITL and Precision and Visual Consistency [84, 324].

1.3.3 Industrial Applications

Different industrial applications adopted XR and AI paradigms, often integrating DT ones, in different fields and with different missions, such as work processes optimizations, retail, food, and training [379, 310, 58, 62, 345, 324, 176, 33, 226, 162]. In this thesis, we analyzed what is the role of humans in such pipelines (HITL), introducing a DT-based framework injected with human collaborative knowledge and experience. This framework exploits both XR and AI paradigms to respectively provide accessible visualization in any degree of reality and predictive analytics to support human actions, related to a specific DT. After its design, we applied it in two different use cases: family album knowledge preservation (CH) and electrical

board assembly information manipulation, validating them with a user assessment. Considering its flexibility, we also applied such a framework in the food industry, where we defined an AR wine recognition application. It is worth highlighting that this app applies a DL-based OCR and a custom textual-driven search algorithm to overcome the limits posed by image retrieval approaches. This research direction considered mostly the intersection between HITL, AI, and Precision and Visual Consistency, considering that human decisions are strongly based on information flowing from predictive models and that virtual elements should be visualized in a contextual, clear, and non-obstructive fashion.

1.3.4 A visual representation of our contributions

The three directional vectors, representing the contribution of each of the three main factors of analysis in each of the considered fields, are reported in Figure 1.2.

The Cultural Heritage directional vector points towards the *Human-in-The-Loop* component with a high magnitude concerning *Interactivity* while providing a low force to *Precision and Visual Consistency*. This is given by the fact, that the main topic analyzed in such a field involves an automatic XR and AI pipeline related to socio-historical photography, where the human has a pivotal role, with a low degree of manipulation. Compared with it, the Industrial Applications directional vector has a higher magnitude in both the *Human-in-The-Loop* and *Interactivity* dimensions, while also slightly contributing to *Precision and Visual Consistency*. In fact, we analyzed how humans could exploit XR and AI in industrial automatic processes playing an active role in it, where the visualization of reliable information covers a fundamental aspect. Finally, the Creative Industries vector points towards *Precision and Visual Consistency* with the highest magnitude, also strongly pointing to *Human-in-The-Loop* and *Interactivity*. This is because we analyzed those use cases that involve the usage of XR and AI technologies to support humans in creative activities, where visual and aesthetic perception

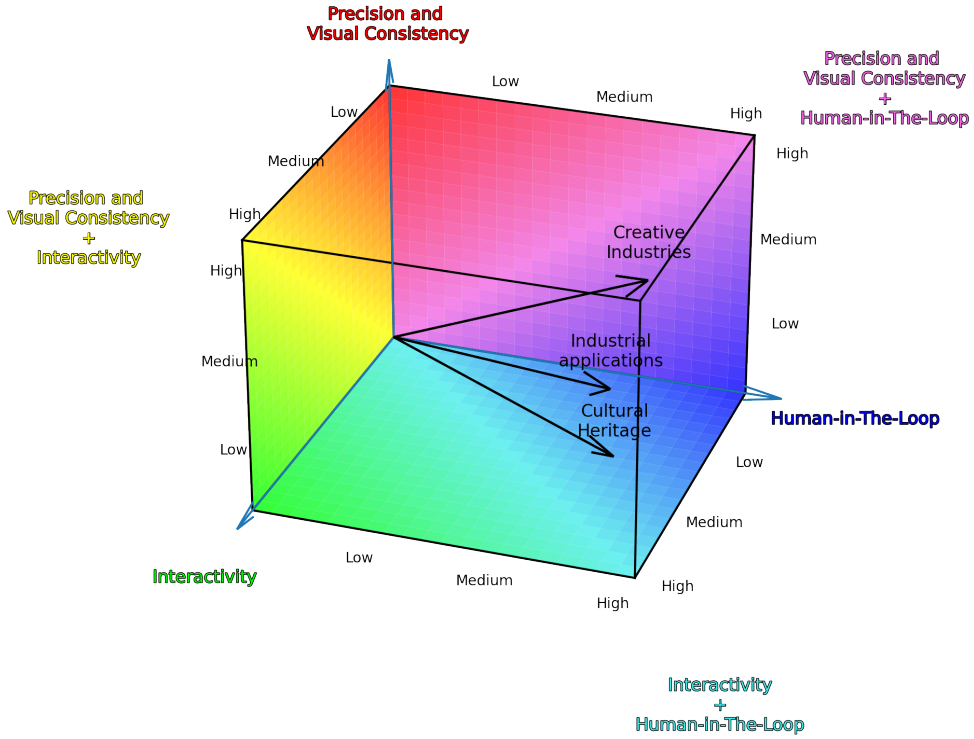


Figure 1.2: View of the directional vectors related to the main topics analyzed in this thesis: Cultural Heritage, Creative Industries, and Industrial applications.

have a fundamental role.

In summary, we here analyzed several systems, and solutions for different use cases in all of the three mentioned research fields and contexts with a particular focus on (a) XR & AI for Cultural Heritage research (i.e., family album photos analysis); (b) XR & AI for Creative Industries (i.e., metaverse for fashion); (c) XR & AI to support humans in industrial-related tasks (i.e., industrial DT and XR for production optimization and retail).

1.4 Research Questions

In this broad XR and AI spectrum, we focused on those paradigms that could be effectively composed to solve real-world problems, then abstracting theoretical methodology from them. Such a study should produce some high-level reflections that can contribute to the framing of how the XR and AI composition has changed, predicting its future role, and also highlighting its limit (i.e., an "Extended Artificial Intelligence" analysis).

- RQ-1 - Can Artificial Intelligence and eXtended Reality be combined to respond to unanswered cultural heritage questions?

  To answer this research question, we delved into exploring the multi-faceted ways in which the fusion of AI and XR technologies can contribute to a particular field of Cultural Heritage, namely socio-historical research, which lacks tools to speed up and optimize processes such as cataloging, qualitative and cultural analysis. In such a context, considering historical pictures, we attempted to analyze how Computer Vision and Augmented Reality paradigms could be adopted to fill the mentioned gaps. In particular, this RQ aims at examining a particular kind of picture, i.e., family album photos, which amounts to a visual media that has been used pervasively to capture the world throughout the 20th century defining a unique source of knowledge to study and learn from the past [51, 245, 248]

- RQ-2 - Can Artificial Intelligence improve user interactions and task completion efficacy in eXtended Reality systems?

  XR and AI were jointly adopted to support both consumers and workers in their everyday tasks [420, 162]. This team-up was investigated on dimensions with a complementary perspective, indicating how XR and AI could fulfill some of their reciprocal theoretical gaps [162]. With this RQ, we aim to study and analyze different XR and AI combinations that aim instead at solving real-world problems, with a particular focus

on how AI could improve user interaction and task completion assistance with XR paradigms. We applied such a perspective by exploiting: (i) Computer Vision and Augmented Reality in Cultural Heritage scenarios; (ii) Vocal Assistance, Digital Twins, and Virtual Reality in Creative Industry applications; (iii) Computer Vision, Digital Twins, and Augmented and Virtual Reality paradigms in Industrial contexts [310].

- RQ-3 - Which kind of eXtended Reality and Artificial Intelligence technologies could be applied in the creative industry?

  Building upon the exploration of XR and AI synergy in the creative domain, this RQ seeks to understand how AI can elevate user experience efficacy within XR systems. The current literature highlights the adoption of XR and AI in diverse creative domains like fashion, music, and art. Despite these conceptualizations, the practical application of XR and AI technologies in such fields is still in its early stages [178, 324]. This RQ aims to explore the uncertainties surrounding factors that contribute to the immersiveness and usability of these systems, in particular in the context of fashion, through technologies such as Intelligent Vocal Assistants (VA) and Digital Twins (DT), and the potential aesthetic impact of immersive (generated) content, emphasizing the need for a robust human-in-the-loop role in their evaluation, where XR paradigms are foundational [288, 311].

- RQ-4 - Can industrial workers take advantage of eXtended Reality and Artificial Intelligence in their everyday activities?

  Separately, both XR and AI were applied to support workers in their everyday activities, while their combination was just recently analyzed for industrial use cases, mainly concerning visualization [420, 171, 162]. In such a context, Digital Twins (DT) are adopted to model real-world objects and processes, providing insights about the current and future state of the DT itself [26]. An interesting question amounts to how the information flowing through both the physical and virtual worlds

should be visualized and manipulated, and the role of XR in supporting humans in the interaction with such a DT flow [370, 162]. Those interactions should speed up both the access and the comprehension of DT-related information that was automatically or manually generated. With this perspective, this RQ aims to explore methods and paradigms to connect all these components, to create seamless XR access to knowledge linked to DTs (including AI and collaborative human knowledge), contextualizing it to cultural heritage, industrial production, and retail scenarios [310].

1.5 Outline

Given the background concepts exposed above, a detailed overview of the work presented in this thesis is provided in the following.

Chapter 1 The first chapter introduces the research background of the thesis, presents its analyzed research questions, and summarizes its content.

Chapter 2 This chapter explores the integration of AI and XR paradigms to assist cultural heritage studies, particularly in the photography domain, applied to social history. Emphasizing the significance of socio-historical studies, the chapter delves into the role of Computer Vision and XR in creating automated tools to aid socio-historians. It introduces the IMAGO dataset and a deep learning tool for classifying historical photos based on shooting date and socio-historical context. The chapter advocates for applying mixed qualitative-quantitative methods to uncover cross-cultural influences in family album photos. Additionally, it presents an AR application for Hololens 2, utilizing the introduced models and a fine-tuned version of the Yolo V5 architecture, to automatically cropping and cataloging physical family album photos. In summary, the chapter introduces a comprehensive framework that combines AI and XR to support socio-historical research, addressing research

questions related to efficacy and efficiency in automatic family album photo cataloging for both experts and non-experts (RQ-1, RQ-2).

Chapter 3 This chapter explores the potential synergy between XR and AI paradigms in supporting creative research and industry. Beginning with a literature review, the examination focuses on the adoption of these technologies across diverse creative domains, with a focus on fashion. In this creative field, numerous works have conceptualized systems leveraging XR and AI to develop different tools, such as intelligent assistants, and realistic fashion model avatars. Despite these efforts, the application of such technologies and the exploration of their potential impacts are still in their early stages, particularly in the e-commerce arena. Uncertainty surrounds the factors enhancing the usability of immersive fashion commerce environments. Simultaneously, there is a lack of studies regarding the social acceptance of self-human avatars, to enable both consumers and professionals to perform actions in the Metaverse without directly controlling them. To address these aspects, the chapter analyzes the usage of two technologies, that employ AI, applied to XR fashion environments: Intelligent Vocal Assistants and Human Digital Twins (RQ-2, RQ-3). Within the same context, the rapid advancements of Generative AI are impacting immersive digital fashion commerce. However, a notable gap exists in the literature regarding works that define an aesthetic human evaluation framework for such data, which also integrates proper visualization hardware. This chapter aims to partially address this gap, validating the usage of XR paradigms to define a human evaluation framework for aesthetic immersive items (RQ-3).

Chapter 4 This chapter examines the pivotal role of XR and AI technologies in advancing Industry 4.0 across diverse contexts, highlighting how they contribute to real-time visualization, decision-making, and adaptive automation. In this scenario, the integration of DTs bridges the gap between physical and digital realms, facilitating direct interaction across remote locations and paving the way for smart human–machine collaborations. However, a

gap was highlighted regarding the injection of human collaborative knowledge into DT information flow, along with its smart visualization. On this line, we introduced a novel framework incorporating XR paradigms and Human Collaborative Intelligence in DTs to generate and easily access human knowledge from any degree of reality [139]. This framework was then applied in different scenarios and DTs to demonstrate its flexibility: (i) memory preservation for family album photographs [358]; (ii) ergonomic manipulation of information related to electrical systems to optimize industrial production [370]; (iii) a wine bottle recognition system for accelerate and improve wine information retrieval and annotation [365, 10]. In all these use cases, we exploited XR-based Multimedia Information Retrieval (X-MIR), to retrieve previously stored information, exploiting Computer Vision paradigms (RQ-2, RQ-4). In particular, for the wine use case, we introduced a novel textual-based wine typology recognition algorithm, which exploits deep learning-based Optical Character Recognition, and classical search algorithm paradigms.

Chapter 5 This chapter ends the thesis by summarizing the here introduced contributions and paving the way for future ones.

Chapter 2

Artificial Intelligence and eXtended Reality for Cultural Heritage: supporting socio-historical studies

In this chapter, we will describe how Artificial Intelligence (AI) and eXtended Reality (XR) paradigms can be composed for the benefit of Cultural Heritage (CH), in particular to tangible one (TCH) [286, 267], which refers to physical, touchable elements of a culture that have historical, artistic, scientific, or cultural significance [104]. Preserving TCH is crucial for maintaining a connection to the past [104] and AI & XR paradigms were often employed to this aim, for example, by digitizing, analyzing, and visualizing them [342].

A lot has been done on this line, however, some fields of study intersecting with TCH preservation and analysis, are still overlooked. In particular, socio-historical studies and one of their most important subject of study: family album photographs [51].

This kind of picture represents an example of vernacular photography that has drawn the attention of researchers and public institutions, due to

their ability to reveal sociological and historical insights regarding specific cultures in space and times [326, 53] that *represent a reference point for the conservation, transmission, and development of a community Social Heritage* [53]. Despite their significance, conducting a comprehensive analysis of such photo collections is frequently impractical due to the absence of well-organized and extensive digital repositories, Moreover, the manual analysis of a hundred pictures is excessively time-consuming and also requires a strong domain knowledge. [326, 53]. Despite their importance and those evident limits in their manual analysis, there is a lack of automatic and easy-to-use tools to analyze and preserve them [367].

To this date, we here review the scientific background related to family album photography, and how Computer Vision (CV) and XR paradigms could be put to good use to define automatic tools to support socio-historians (and non) in their everyday activities. Then, we introduce, to the best of our knowledge, the first family album photos dataset, called IMAGO, and a novel Deep Learning (DL) based tool, optimized to learn how to classify those pictures for two socio-historical labels: the shooting date and socio-historical context (RQ-1). Such a tool is then exploited to debate how (i) mixed qualitative-quantitative methods could be applied in the research field of family album photos and (ii) how they could be applied to discover cross-cultural influences by visual cues (RQ-1). Finally, an AR application for the Hololens 2, which exploits the mentioned models and a fine-tuned version of Yolo V5 [392] will be detailed, showing how it could be used to automatically crop and catalog physical family album photos, to also revive the social phenomenon of visualizing together family album photos (RQ-2).

The rest of this chapter is organized as follows. Section 2.2 sets the theoretical background elucidating the contribution of this chapter. Section 2.3 elucidates the socio-historical background that motivates our work while Section 2.4 describes the most similar works in the context of historical picture analysis. Section 2.5 describes the novel-introduced IMAGO dataset which was used to train the DL architecture described in both Section 2.6 and Sec-

tion 2.7. The obtained results and their analysis are described in Section 2.8. Sections 2.9, 2.10, 2.11 describe three different applications developed for the benefit of socio-historical studies, built on top of the obtained DL models and results. Finally, Section 2.12 provides an overall perspective on these contributions while fostering future extensions and applications of our work.

I here declare that the content of this chapter is entirely based on my work and contains nothing that is the outcome of work done in collaboration with others and that all the contributions reported here have been partially published in [362, 363, 360, 364, 366, 367, 369].

2.1 Research Questions

Considering the different aspects analyzed in the previous Sections, this chapter aims to answer the following research questions:

RQ-1 - Can Artificial Intelligence and eXtended Reality be combined to respond to unanswered cultural heritage questions?

RQ-2 - Can Artificial Intelligence improve user interactions and task completion efficacy in eXtended Reality systems?

2.2 Introduction

Following Kodak's invention of the first-megapixel sensor in 1986, digital photography has slowly grown to substitute its analog predecessor, playing a key role in the early 21st-century digital revolution and social transformation [280, 339]. As a relevant example, photography has modified the way mobile phones are used, as their integration of digital cameras has at once fostered an exponential growth of the photos that are shot and uploaded to the Internet every year, as well as a paradigm shift in mobile communications, which today rely on high-quality multimedia [435, 64, 39, 298]. These

phenomena have proven to be game-changers for both how people communicate and the bloom of new fields of research, as both academia and industry, have exploited such plethora of visual data to develop and apply, as an example, CV models to a variety of different problems (e.g., face recognition, autonomous driving) [446, 200, 207, 66, 396, 312].

CV is a field of AI that enables machines to interpret visual information from the world captured through a camera lens and projected into the pixel space [318, 406]. Nowadays, CV is mostly driven by DL, which is a subset of Machine Learning (ML) that utilizes neural networks with multiple layers (deep neural networks) to learn from a huge amount of data using optimization algorithms [318, 406].

Now, while a wealth of research is being devoted to applying CV and DL paradigms for the processing and analysis of digital images, much has to be done regarding analog ones, mainly because printed images representing places, people, and/or objects at a given time may be: (i) scattered in numerous public and private collections, (ii) of variable quality, and, (iii) damaged due to hard or continued use or exposure. Furthermore, any analysis utilizing image processing and computer vision algorithms is contingent upon the initial digitization step, which has the potential to degrade the overall image quality.

Despite the complexities and difficulties posed by analog photographs, they serve as an unparalleled source of information about the recent past: in fact, no other visual media has been used as pervasively to capture the world throughout the 20th century, as the availability of consumer-grade photo cameras supported the spread and popularity of vernacular photography practices [245, 248]. Vernacular photography refers to the creation and use of photographs by everyday people, as opposed to professional photographers. In the context of photography, the term "vernacular", refers to all those pictures that emphasize the ordinary, everyday aspects of life captured by amateurs (e.g., family snapshots, travels, friends and classes, workers) [245, 248]

In particular, pictures stored in Family photo albums, represent an example of vernacular photography that has drawn the attention of researchers and public institutions, due to their ability to reveal sociological and historical insights regarding specific cultures in space and times [326, 53]. In fact, a recent work defines family photo albums *a globally circulating form that not only takes locally specific forms but also "produces localities" that create and negotiate individual stories* [326]. Also [92] stated that: *"as people struggled with this broadening of their family album, other narratives began to emerge within those already established of colonialism, imperialism, migration, and dispossession"*. Along the same line, [53] highlight how family albums *represent a reference point for the conservation, transmission, and development of a community Social Heritage* [53]. Despite their significance, conducting a comprehensive analysis of such photo collections is frequently impractical due to the absence of well-organized and extensive digital repositories. The manual verification of the characteristics of several hundred pictures is excessively burdensome, especially given that, in numerous instances, there are no accompanying descriptions. As a result, contributions in this field typically rely on the examination of small corpora of photos. [326, 53].

In this chapter, we address such a problem, taking as a case study the socio-historical analysis of a collection of family album photographs: we here present the design and implementation of a multimedia application that, resorting to DL models, implements their classification for cataloging purposes. To verify the validity of such an approach, the application is exploited to classify a novel dataset for socio-historical image classification studies, IMAGO, collected and maintained at the University of Bologna [53, 367]. This tool allows the classification and cataloging of family album pictures according to their shooting date and socio-historical context (that will be further discussed in Section 2.5) which are relevant categories to discriminate and understand what happened in the past. Exceeding our initial expectations, such an approach has revealed its merit in terms of performance, but also in terms of the foreseeable implications for the benefit of different socio-historical research.

We so adopted this DL-based tool to further investigate its role in three crucial aspects of socio-historical studies: (a) comparing classical qualitative approaches for classification quantitative ones (i.e., our tool) [364], (b) analyzing automatical approaches for searching cross-cultural relationships in different places and times [366] and (c) automate physical family album photos cataloging with usable and portable tools (in our case, a MR Device) [360, 363].

In fact, (a) the relations between quantitative and qualitative analyses, their potentials, and limits represent open questions within different research communities [36, 241, 386, 29]. Due to the growth of digital and digitized data, qualitative analyses are becoming more and more expensive and difficult to apply to massive datasets, and quantitative methods seem to be the only way to deal with them [29]. Eventually, the results obtained adopting quantitative methods could converge to those returned by qualitative ones, as demonstrated by [29] that compared a qualitative approach, from interpretive social science, and a quantitative one, from natural language processing on textual data. However, some criticism emerges also for quantitative methods, which may improperly apply the definition of measurement, simply matching tasks, objects, and events to numbers, according to specific rules [241]. In addition, they may be misleading due to insufficient care in data collection, feature definition and processing, and adherence to the domain of interest [36, 70]. Despite such critics, recent research in computer science has been focusing on how quantitative methods may be able to support qualitative analyses [29, 296, 129, 329]. Authors of [296] highlighted that, although a variety of issues have emerged with the use of machine learning models in social data analyses, the intersection of machine learning and the social sciences has provided critical new insights. In particular, [129] observed the similarities between creating human-labeled datasets and content analysis, underlining the importance of utilizing high-quality training data, also labeled with qualitative methods.

In this chapter, we want to contribute to such debate, showing how quan-

titative and qualitative methods can coexist to carry out integrated analyses to evaluate more data than those usually examined in a qualitative process while adopting a well-defined theoretical foundation. For this reason, we built our analyses on the developed classification tool, to support socio-historians in drawing their conclusions using both quantitative and qualitative analysis approaches [326, 53, 289]. The adoption of qualitative methods has been so far justified by the small number of items socio-historians have at their disposal and by a general skepticism around the adoption of quantitative methods. Therefore, we verify whether quantitative techniques, built resorting to results obtained from qualitative processes, may be employed to perform specific qualitative analyses, yielding a mixed qualitative-quantitative one. To start, we focused on existing socio-historical categories derived from previous sociological and historical qualitative studies [51] that were used to label the IMAGO dataset and to train our quantitative classification tool. This dataset allowed us the mentioned deep learning classification tool, capable of recognizing salient features of socio-historical interest, opening the possibility of exploiting quantitative methods with qualitative ones to improve cataloging processes. To demonstrate its efficacy, we compared the results obtained with our approach to the ones obtained by a socio-historian who manually assessed the photos. The results acquired with this test confirmed that quantitative approaches may integrate qualitative ones to benefit the overall performance and speed of socio-historical analyses.

On top of this mixed quantitative-qualitative contribution, we tested whether our models could be employed in another crucial socio-historical investigation: analyzing cross-cultural influences (b). Family album pictures contain, for example, the different clothes that people wear, their haircut styles, the tools and machinery, the natural landscape, the overall environment, etc., which may exhibit the culture of a given time and place [247, 285]. In particular, analyzing the time dimensions allows us to search for relationships in human habits among different places and bound possible intercultural influences through time itself. For example, by analyzing changes

in fashion, technology, and other visual cues over time, we can gain insights into how cultural practices and social norms have evolved and also identify patterns to connect different communities and their reciprocal influence. In such a context, having automatic methods that could learn meaningful visual cues that discriminate over the time dimension, could be exploited for such kind of analysis. This method can be especially valuable when other sources of information, such as written records. In this chapter, we so investigate such a problem, by applying our developed DL classifiers, to discover possible intercultural influences (i.e., the adoption of different customs and habits in different epochs and countries) by analyzing the differences in dating, resulting from a cross-dataset experiment [136, 321] while providing qualitative cross-dataset qualitative visualizations.

Finally, (c) we considered the lack of digital technologies tools apt to individuate, digitize, and share elements of human cultural heritage in an easy and portable way [315, 363]. In fact, to the best of our knowledge, no works designed or implemented similar solutions concerning family album photos, exploiting eXtended Reality (XR) paradigms. For this reason, we put to good the previously defined classification tool and a novel fine-tuned version of a known object detector, namely YOLO [305] to define an AR application support socio-historians, but also non-academics, to automatically crop, catalog and augment physical family album photos, just wearing an AR headset [363, 360]. This novel tool has been validated through an assessment model, asking a group of ten people to provide their comments regarding the use of our prototype.

To summarize, the contributions of this chapter amount to [360, 363, 367, 366, 364, 369]:

- A deep learning-based multimedia application to assist socio-historians in their cataloging work which consists of identifying the socio-historical information of an image, i.e., its shooting year and socio-historical context, according to the definitions provided in [53, 367]. While the dating task has been so far considered in literature [136, 321, 258], the estima-

tion of the socio-historical context has not been yet investigated and will be further defined and explored in Section 2.3;

- The introduction of a family photo album collection, namely IMAGO, comprising over 80,000 analog photos taken between 1845 and 2009, belonging to approximately 1,500 families, primarily from the Emilia-Romagna and immediately neighboring regions in Italy.

- A thorough evaluation of the performance obtained by Convolutional Neural Network (CNN) models [187, 380, 169] trained on the IMAGO dataset for both the dating and the estimation of the socio-historical context.

- A comparison between the performance of the adopted CNN-based approach and a novel visual backbone (i.e., Vision Transformer) is performed, discussing the pros and cons of both approaches for the considered classification tasks.

- A comparison of the performance obtained by a socio-historical scholar, employing a qualitative analysis of the photos, with the performance of the CNN and Transformer-based DL models. This provides an exemplary case integrating quantitative and qualitative approaches to speed and increase the amount of processed and cataloged data.

- A cross-dataset experiment, based on the IMAGO dataset and the datasets from [136, 321], to verify possible intercultural influences (i.e., the adoption of different customs and habits in different epochs and countries) by analyzing the differences in dating errors and providing qualitative visualizations exploiting the Uniform Manifold Approximation and Projection (UMAP) algorithm.

- The definition and assessment of an AR application built for the Hololens 2, integrating a fine-tuned version of the YOLO object detector and the

mentioned classification model, to support socio-historians by automatically cropping and cataloging pictures from physical family photo albums. The developed application was validated through a user study, which highlighted positive adoption intentions.

2.3 Socio-historical Background on Family Photo Albums

In this Section, we sketch the socio-historical background required to set the stage for this work. No classification tool can be defined without first clarifying which classification categories are and how they are constructed. This review offers essential insights for comprehending the emergence of contexts and categories in socio-historical studies. To achieve this, we initially outline the key distinctions between conventional and social history. Subsequently, we elucidate the inclusion of family photo albums in the purview of this field of study, ultimately introducing the methodology employed by socio-historians in organizing a dataset.

In the words of Cabrera, traditional history and social history differ as follows: *Traditional history, especially classical political history, was based on the concept of the subject: the subjectivity of historical agents was rational and autonomous; the subject a preconstituted center; and, therefore, actions were caused, and fully explained, by the intentions that motivated them. Social history, on the other hand, was based on the concept of society. For social historians, subjectivity and culture are not rational creations but representations or expressions of the social context in which the causes of actions were to be found* [48]. Such social contexts, with their historical logic, represent the ground on which categories are constructed, to grasp the meaning and organize social reality [47]: the categories represent a complex relational network whose nature is neither subjective nor objective but the result of a specific historical phenomenon with its own behavior. Therefore, the categories do not constitute a simple means for transmitting social reality

but are an active part in its definition and are called *socio-historical contexts*.

Now, the starting point of a socio-historical analysis is the space in which the interweaving between individual initiative and social coercion takes place. An attempt is usually made to explain how society works on different theoretical bases resorting to traditional oppositions: public/private, subjective/objective, ideal/material, visible/invisible, body/conscience. Further analyses are then introduced turning to the concept of social imaginary, defined as "The way in which ordinary people imagine their social contexts which, often, does not translate into a theoretical formulation but is conveyed in images, stories and legends" [51]. In essence, any socio-historical context introduced in such analyses should describe the evolution of social history and therefore the change of sociality and of peoples' behavior in a defined space/time. To this aim, socio-historical categories are first identified by studying historical archival documents from different topics (e.g., economics, traditions, wars), including multimedia sources, which nowadays cover a fundamental role for socio-historical analysis [75]. Out of the many multimedia sources today available, photography emerges as the one capable of covering the greatest time span so far, even if was exploited as a primary source of information just in the last few decades [352].

For this work, socio-historical categories have been obtained relying on the study of family album photos. This particular kind of picture originates from and at the same time represents a fundamental component of social structures, also a well-known socio-historical abstraction, the Family [42]. The Family is, indeed, a fundamental construct in Social History studies, since it embodies at once the public and the private spheres [352]. The photos contained in the family albums can be read, on the one hand, as private visual memories of one's history, destined to remain hidden from society, and, on the other hand, as traces and signs of the collective social imaginary of a given historical period. So, family album photographs depict the daily existence of their time, not considering them solely as memories but also as a network of signs, traces, and documents that may be used to

interpret the past [352].

Although socio-historical contexts may emerge from the study of archival documents and family album photographs, the specific context of a specific photo may remain hard to tell. This is because without knowing when a picture was taken and what the people there portrayed were doing, it may be impossible to associate any accurate information with the picture. This highlights the pivotal role of the picture shooting date, usually reported in their backside, which in any case requires additional analysis to be confirmed. To this date, the most accurate source of information may be obtained only by resorting to the knowledge of the subjects represented in the photo, for both the shooting date and the social category. For this reason, socio-historians rely on the knowledge of the main source, if available, which may be one of the subjects or the actual owner of the photograph.

However, such information could be impossible to find: when studying and cataloging a corpus of photos, no reliable source of information may be available. This problem is common for socio-historical scholars, in such cases, they resort to other approaches, which may include classifying data based on a visual inspection and implementing onerous processes to reduce as much as possible the misclassifications of socio-historical features. As a relevant example, authors of [108] collected, analyzed, and classified 355 photos related to women involved in agriculture learning activities using visual analysis only. Considering such an aspect, we hypothesize that it would be possible to classify family album pictures, by resorting to AI and CV paradigms.

In the following, we present the most relevant works related to the area of research so far described.

2.4 Related Works

In this Section, we analyze the works that fall closest to ours in terms of datasets, classification methods, and applications related to socio-historical studies. Concerning datasets and methods to classify pictures from a histori-

cal perspective, only a few works have analyzed analog collections of vernacular photographs [136, 321, 258]. For example, [136] employed a deep learning approach to analyze and date 37,921 historical frontal-facing American high school yearbook photos taken from 1928 to 2010 [136]. Here, a Convolutional Neural Network (CNN) architecture was trained to analyze people's faces and predict the year in which a photo was taken. Along the same line [321] presented a dataset containing images from high school yearbooks, covering in this case the 1950 to 2014 time span (considering 1,400 photos per year). They resorted to CNNs to estimate the precise image shooting year. To assess the characteristics that allow them to correctly classify a picture, they considered both color and gray-scaled images containing: (i) faces, (ii) torsos (i.e., upper bodies including people's faces), and, (iii) random regions from the images. The best performance was obtained considering color images portraying the torso of people. Their results confirmed that human appearance is strongly related to time. Authors of [258] instead analyzed the dating task through the lenses of vernacular and landscape photos belonging to years 1930 through 1999, amounting to at most 25,000 pictures per year. The authors proposed different baselines relying on deep CNNs, considering the dating as both a regression and a classification task.

Reference	**Type(s) of photography**	**Type(s) of camera**	**Theme**	**Cardinality**	**Period**
[136]	Portrait	Digital and analog	Frontal face from High school yearbook	168,055	1905 - 2013
[321]	Portrait	Digital and analog	High school yearbook	ca 600,000	1912 - 2014
[258]	Vernacular and landscape	Digital and analog	No specific theme	1,029,710	1930 - 1999
IMAGO	**Vernacular**	**Analog**	**Family albums**	**ca 80,000**	**1845 - 2009**

Table 2.1: Characteristics of existing datasets and IMAGO.

In Table 2.1 we summarize the characteristics of the archives employed in the works described so far. In most cases, only specific subsets of such

archives have been analyzed employing computer vision techniques. To provide a comfortable comparison, the same information regarding the IMAGO collection is provided in the last row.

Going beyond the dating task, other works have already investigated the digital cataloging of historical photos [72, 25]. For example, [213] developed a prototype to find duplicates and tag photos depicting similar scenes in the Carnegie Mellon University Archives' General Photograph Collection. In [381], authors based on semiotics and visual cultural studies, developing a framework called *distant viewing*, to individuate larger patterns within a corpus that may be difficult to discern by closely studying only a small set of objects (e.g., narrative arcs in American sitcoms). One of the works that fall closest in scope was published by [418], where the CHRONIC and the SIAMESET datasets were introduced to study the transition from illustrations to photographs in the history of Dutch newspapers.

However, in all the aforementioned research works, no pre-defined socio-historical categories were utilized as means of analysis but only years were used to distinguish pictures from a historical perspective. In addition, none of the considered works analyzed family album pictures: to the best of our knowledge, this is the first contribution to investigate their classification according to the socio-historical context definitions and background.

Focusing now on the possibility of supporting socio-historical photography research by exploiting deep learning-based classifiers, a lot of work is yet to be done, particularly in: (a) comparing and mixing qualitative and quantitative analysis, (b) analyzing, with an automatic approach, visual cues to discover historical cross-cultural influences and (c) implementing automatic tools to catalog physical pictures.

In the case of (a), to the best of our knowledge, no works have provided meaningful contributions, although many already debated about qualitative vs quantitative analysis [241, 36, 70, 29, 296, 129, 329].

On the same line, we did not find works that specifically address the automatic analysis of historical cross-cultural influences by analyzing errors

in machine learning predictions (b). Finally, to the best of our knowledge, no AR system allows a user to automatically crop and catalog pictures from a physical family album photo using a combination of object detection and classification (c) [406, 21].

2.5 IMAGO: A New Dataset Of Family Album Pictures To Support Socio-historical Studies

The IMAGO project was started in 2004 by socio-historical scholars to study the evolution of Social History through the lens of family album photographs. This produced a digitized collection, namely IMAGO [1], of analog family album photos gathered year by year and conserved by the Department of the Arts of the University of Bologna. The collection comprises ca 80,000 photos, taken between 1845 and 2009, belonging to about 1,500 Italian family albums, offering the opportunity to study the evolution of Italian society during the twentieth century. Among these, 16,642 images have been labeled by the bachelor students in the Fashion Cultures and Practices course, under the supervision of the socio-historical faculty following a strict and sound labeling protocol.

2.5.1 Annotation Process

The annotation process followed a simple but strict protocol, involving the following steps:

1. During the first lecture the socio-historical background, the IMAGO dataset construction project, and the different classification categories were presented and explained;

[1] The IMAGO resources are available upon request.

2. During the second lecture, the annotation problem has been covered in more detail. In particular, the lecture focused on the importance of the reliability and authenticity of sources of socio-historical materials (including the shooting year). This meant explaining that the original owner of the photo should be interviewed whenever possible. In case such person(s) were not available (e.g., the photo is very old), one could find a secondhand informed party (e.g., anyone who might be aware of the context of the given photo). Alternatively, an attempt to infer the socio-historical context and the shooting year (if possible) could be made by analyzing any written annotations scripted behind the photo. In case none of such solutions were possible, no annotation would be added.

Hence, the information provided by a photograph's owner amounts to the ground truth from a socio-historical point of view. This assumption in the labeling process is what injects the social component along with the historical one in the dataset. Such an approach is not new to the computer vision community either, other works in literature have considered as image metadata the information provided by their owners [275, 27]. These elements highlight the uniqueness of such datasets: since only the owner (or a directly connected party such as a relative or a friend) holds the ground truth, it is not possible to resort to just any standard labeling services (e.g., Amazon SageMaker Ground Truth or the Google AI Platform Data Labeling Service [6, 142]). This annotation process generated two socio-historical metadata per each photo: (i) the socio-historical context and (ii) the shooting year [53].

2.5.2 Socio-Historical Context

We here explain how the classes employed to analyze IMAGO have been defined from a socio-historical point of view. To this aim, we here report on the rationale behind the use of two exemplar ones, "Motorization" and "Affectivity", while a more in-depth analysis of all classes may be found in [50, 352, 51, 52].

The "Motorization" class is meant to mark an important change in people's lifestyles. We can take as an example the boom of sales for motorcycles. Such phenomena not only changed the production trend and its related economic ecosystem but also changed the social behavior of people in the area in which such a boom took place. It affected society's idea of mobility and how people gathered together. In these terms, the motorization aspect becomes therefore fundamental for the study of Social History.

On a completely different plane, instead, the "Affectivity" class regards personal feelings. Such a class wants to represent the changes that occurred between the affective and family relationships. For example, in the first decades of the twentieth century, family emotional relationships were considered estrangement. This phenomenon is also reflected in the photographs that depict wife and husband, parents and children, brothers and sisters. Although all members of the same family, they all posed without any affectional gestures (e.g., hugs). From the second post-war onward, things changed starting with younger people who changed poses in terms of distances, contacts, hugs, etc. In the following, we provide the socio-historical categories individuated in the IMAGO dataset [352], along with a brief explanation:

- *Work*, photos belonging to this class are mostly characterized by people sitting and/or standing in workplaces and wearing work clothes and/or gear;

- *Free-time*, includes scenes of leisure time, reconstructing, wherever possible, generational and gender differences. It also includes images representing people visiting far off landmarks, expanding social relationships and interacting with nature;

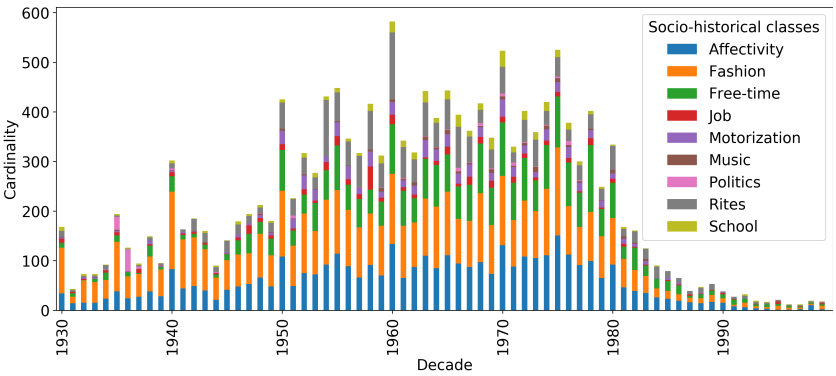
- *Motorization*, although often closely related to the *Free-time* category, this class has been distinguished as it includes symbolic objects such as cars and motorcycles, which represent a social and historical landmark;

- *Music*, as for the *Motorization* one, this class may also include scenes from leisure time, characterized in this case by the appearance of musical instruments or events;
- *Fashion*, as clothing represents a mirror of the articulated intertwining of socio-economic, political, and cultural phenomena. This class is characterized by the presence of symbolic objects and clothes, such as suits, trousers, skirts, and coats;
- *Affectivity*, characterized by the presence of people (e.g., couples, friends, families, or colleagues) bound by inter-personal relationships;
- *Rites*, portraits of sacred and/or celebratory events from family lives;
- *School*, this class includes all the photos that represent schools, often characterized by symbolic objects (e.g., desk, blackboard) or groups of students;
- *Politics*, this class contains photos related to political gatherings, demonstrations, and events.

These aforementioned categories amount to the ones that from now on will be used to implement the socio-historical classification task.

2.5.3 Exploratory Analysis

In Figure 2.1a we show the number of labeled images available per year in the 1930 to 1999 time frame, out of such time interval, the number of available images is too little to be visually represented. Such a figure also exhibits the distribution of the socio-historical information (i.e., shooting year and socio-historical context) over the entire dataset. From such a plot, it is evident the unbalance that exists in terms of the number of photos both per year and socio-historical context. Figure 2.1b shows four exemplar images from the IMAGO dataset, which belong to different decades and represent different socio-historical contexts. These images are representative of the



(a) Classes distribution

(b) Sample images

Figure 2.1: IMAGO characteristics.

different characteristics that may be found in each photo (e.g., number of people, clothing, colors, and location).

2.6 A Deep Learning-based Socio-historical Cataloging Tool For Family Photo Albums

Socio-historical analyses include dealing with various sources of information, systematically examining their soundness, exemplarity, and meaning, and seeking for inter and intra-correlations and relationships that may help to understand what happened in the past [41]. Sources are in general not objective but *shaped by the politics, practices, and events that selectively doc-*

ument protest [71]. In summary, the procedure of historical inquiry implies the following steps: (i) identification and selection of sources, (ii) registration and classification for further investigation, and, (iii) a critical inquiry of the collection. From here, a socio-historian's work can then proceed in multiple directions. A sound socio-historical study may hence require the inspection and classification of hundreds or even thousands of documents and images [31, 331, 108].

This amounts to burdensome work which often seeks for the big picture provided by large corpora of data, rather than the specific information returned by a single document or image. Such type of process opens to the use of automatic tools, capable of classifying great amounts of data in short amounts of time. This has already been discussed over two decades ago, for example, in [122], where the author illustrated linguistic and statistical tools that could be profitably used by historians and social historians in the study of events. Now, much more can be expected thanks to the development of computing tools, capable of handling growing amounts of multimedia data originating from heterogeneous sources. This would require a holistic approach taking care of source(s): (i) digitization, (ii) accessibility through standard interfaces, and, (iii) analysis with models capable of translating socio-historical tasks into computing ones.

Now, a typical socio-historical task amounts to inferring from and subsequently applying categorical models to large corpora of data (Section 2.3). We apply such an idea to the case of family photos, proposing a multimedia tool capable of processing and cataloging such types of pictures. To this aim, in Figure 2.2 we show the components of the proposed application. The core is the Socio-Historical Module (SHM), which is composed of one or more classifiers, depending on socio-historical tasks of interest. For this work, such tasks have been defined on top of family album photos, originating from the IMAGO dataset (details regarding its socio-historical value are discussed in Section 2.5). Such a dataset offered the opportunity to predict two socio-historical information: the context and the shooting year. In brief, the SHM

amounts to a tool that may automatically label photos with the obtained predictions giving, in addition, the opportunity to confirm or correct such estimates, when necessary, during cataloging procedures.

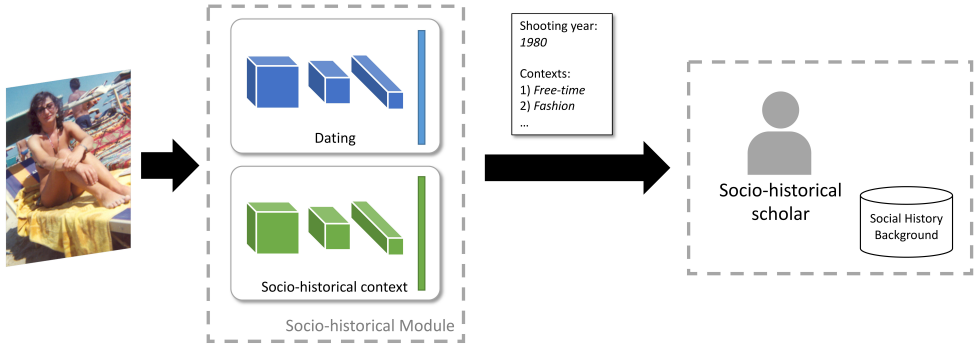


Figure 2.2: Schema of the multimedia support application for socio-historians.

The classifiers that compose the SHM could be defined by exploiting DL approaches, which have generally provided higher accuracies [336], considering the analysis of historical picture datasets [213, 381], in particular for the dating task [136, 321, 258]. Inspired by the work of [321], we trained several DL classifiers considering different image regions, belonging to the same picture, selected using different criteria. To this aim, we considered the whole image and the crops enclosing the faces and the full figures of the people there portrayed (further details in Section 2.7.1). Such patches are always present when dealing with family album photos since those always include at least one person. To effectively estimate the value provided by such patches in terms of prediction performance, we also considered random ones. Hence, for the whole image and each of the aforementioned regions, we trained two specific single-input classifiers, one per each of the two socio-historical tasks of interest.

Such classifiers are named following the analyzed patches: full-image, faces, people, and random patches. The single-input architecture utilizes either a CNN or a Vision Transformer-based as the backbone and an additional

fully connected layer for the final classification. It is worth noticing that the results of such classifiers may not be comparable, as the amount of data utilized to perform a prediction varies depending on the fact that the full image is used during testing, or parts of it (patches). This fact required establishing a different evaluation method, considering not a single face/person/random-patch but introducing a layer that merged all of such activations into a single one per picture. In practice, the activation vectors returned by a single-input classifier (e.g., the face classifier) for each face region were averaged per image to compute the most probable class. This process was applied also to the people classifier and the random-patches classifier.

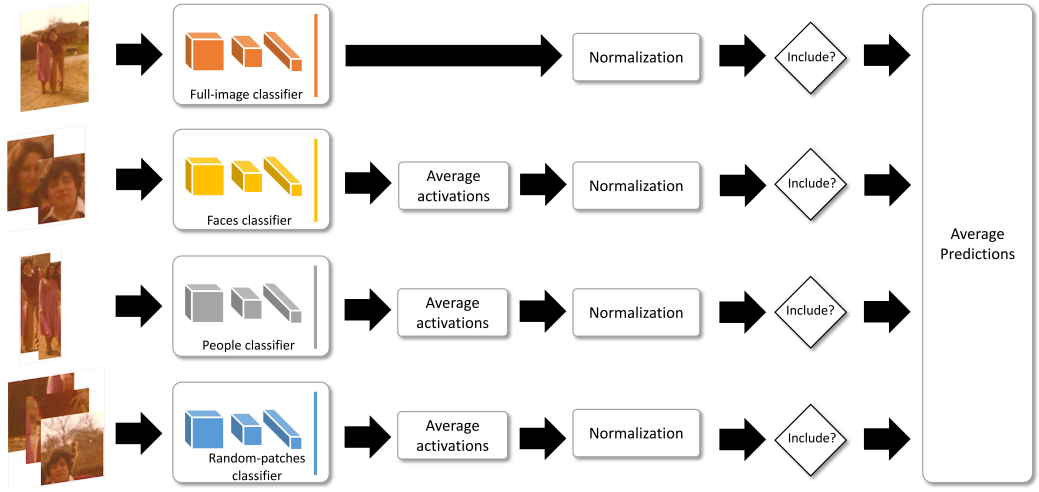


Figure 2.3: Ensemble the different models trained on the proposed datasets. Depending on the information exploited to obtain the final prediction the activations from a model may be included or not.

Finally, we also exploited the ensemble of these models (Figure 2.3). We resorted to such an approach as it has been successfully applied in literature [292] and did not require any additional training and tuning of hyper-parameters. This kind of approach was employed, not only to exploit the averaging effect [35] but also because it helps identify which type of classifier and data provide valid contributions at inference time. However, since we are considering activations coming from a single image or obtained averag-

ing across multiple regions, these may contain values at different scales. For this reason, we $l2$-normalized the different inputs of the ensemble, to support the combination of the activation vectors coming from the full-image, faces, people, and random-patches classifiers. In particular, the final prediction is obtained by averaging the outputs described above and then computing the most probable class. As also represented in Figure 2.3, the final system results in a modular approach, supporting the selection of the single-input classifiers.

2.7 Methods

We here provide details about the dataset pre-processing, deep learning models adopted architecture and training process. For the sake of clarity, we highlight that the adopted subset of the IMAGO dataset is composed of 16,642 labeled photos, spanning over the 1845-2009 time period. While it was entirely used during the analysis of the socio-historical context classification task, we consider only 15,673 pictures for the dating task, covering the 1930 to 1999 temporal interval, have been employed to avoid those years with a very limited number of samples, as already shown in Figure 2.1a.

2.7.1 Data Pre-processing

We performed a data pre-processing phase aimed at (i) isolating the regions of interest from each photo, and (ii) improving the quality of the images composing the dataset, resorting to different techniques.

As reported in Section 2.6, both faces and people represent regions of interest to be exploited for the dating analysis [321, 136]. Following such insight, we created the IMAGO-FACES and the IMAGO-PEOPLE datasets, comprising over 60,000 samples each: the first composed of individual faces, the second of a single person's full figure images. These have been obtained by processing each image of the IMAGO dataset using the open-source implementations of YOLO-FACE and YOLO available at [382, 179], respectively.

Figure 2.4: Sample of different patches: (a) IMAGO-FACES, (b) IMAGO-PEOPLE, and, (c) IMAGO-RANDOM samples.

The IMAGO-FACES dataset has been constructed accounting for the number of people portrayed in a photo. In fact, by adopting a fixed-size bounding box it may be possible to lose relevant details (e.g., hairstyle) or to include pixels related to the faces of other people. To avoid such a problem, an adaptive strategy has been adopted: the size of the bounding box used to crop a face depends on the number of people portrayed in a photo, the greater the number of people, the smaller the bounding box. In this way, it was possible to extract the shoulders and the full head of a single person even when a picture portrayed tens of people. Figure 2.4a shows some sample images taken from the IMAGO-FACES dataset considering different decades and different socio-historical contexts. The construction of the IMAGO-PEOPLE dataset follows the same criteria employed for IMAGO-FACES, though, images can present different aspect ratios (i.e., people may be standing or sitting in photos). Figure 2.4b shows exemplar images from IMAGO-PEOPLE. It is possible to appreciate that IMAGO-PEOPLE includes details that are not present in IMAGO-FACES (e.g., the clothing of a person).

We then verified the utility of performing denoising and super-resolution operations, as all the images considered in this work derive from scans of the analog prints. For denoising, we tested the neural network model from [443] and the Bilateral Filter [319]. For super-resolution, we used an open-source implementation of the ESRGAN model [415] within the Image Restoration Toolbox [442]. The overall improvement obtained from adopting such strate-

gies was revealed to be negligible, we hence opted for an analysis based on the original scans of analog photos. The IMAGO-FACES and IMAGO-PEOPLE were defined only to fine-tune the deep learning models for the socio-historical tasks introduced with the IMAGO dataset. So, we will not release such datasets, since their creation is technology-dependent. Indeed, in the future, algorithms or models providing more accurate bounding boxes for faces and people regions could be introduced.

Finally, to study the possible use of non-human features within a family album photo dataset, we also created a dataset called IMAGO-RANDOM, comprising 8 randomly cropped regions, of 128×128 pixels, from each image in the IMAGO dataset (some samples are reported in Figure 2.4c). Other window sizes were also tested but returned a lower performance.

2.7.2 Deep Learning Models Architecture

As a first approach, we considered as single-input classifiers the most adopted DL-based vision backbone: Convolutional Neural Network (CNN). CNN is a type of feedforward neural network designed to extract features from data using convolutional structures (i.e., kernels) which are learned during training [208]. Different CNN-based architectures were trained using the ImageNet [89] which is one of the most adopted datasets to train and test the generalization capabilities of vision classifier-based architectures, such as the ResNet50 [187]. Considering such models, it is possible to adjust their learned parameter, and so transfer the learned knowledge to a task that differs from the initial one with, for example, a fine-tuning approach [294].

We proceeded by following such a vastly adopted approach, modifying the considered pre-trained model and replacing the top-level classifier layer with a new one, whose structure depends on the socio-historical task (i.e., the number of output classes) and whose weights have been randomly initialized [141]. The pre-trained convolutional layers have been specifically fine-tuned for the given input data and task. To verify the independence of our dataset from the specific CNN architecture, we have also considered

other two well-known ones: InceptionV3 [380] and DenseNet121 [169]. We here anticipate that considering the similar performance among the cited architectures, we decided to choose the ResNet50 as the main backbone for our further analysis since it represents a good trade-off between performance and the number of parameters (details in Section 2.8) [80].

Additionally, we took into consideration a recently introduced vision backbone in the DL panorama: Vision Transformers (ViT). The Transformer is a DL architecture that relies entirely on the self-attention mechanism to draw global dependencies between input and output [399]. Recent works have shown that such an approach can achieve comparable or even superior performance than CNNs [100, 390, 150]. In particular, the ViT architecture, proposed by [100], has achieved state-of-the-art performance on several computer vision benchmarks. Considering the capacity of ViT to correlate features within different image patches, we hypothesize that it could improve the performance of the Single Input CNNs somehow mimicking our ensembling approach in an implicit fashion [100]. For these reasons, we decided also to employ the ViT architecture in the development of our socio-historical tool. To this aim, we proceeded to fine-tune different ViT configurations (Tiny, Small, Base, and Large), varying the size of the input images (i.e., 224×224 or 384×384) and considering patches of 16×16 pixels.

2.7.3 Data Partition & Training setting

All the datasets introduced in the previous sections have been partitioned as follows: 80% for training and 20% for testing; in addition, 10% of the training images are used as the validation set for hyperparameters tuning. For each image in the train set of IMAGO, the faces, and the people there are portrayed, and the random patches are extracted and added to the corresponding dataset subset. This process is repeated for the validation and test sets, as it guarantees that none of the training samples may end in the validation and test sets.

During the training phase, we applied data augmentation techniques (e.g.,

random crop and horizontal flip) to make the model less prone to overfitting. Each model has been fine-tuned using a weighted cross entropy loss to counter the unbalance in our dataset [283]. For training the CNN-based architectures, we employed the Adam optimizer with a learning rate of 1e-4 and a weight decay of 5e-4. We set the batch size to 32 for the training of the full-image classifier and to 64 for the faces, people, and random-patches models

For the training of ViT instead, we followed the procedure reported in [100], while adopting again weighted cross-entropy loss to counter the dataset unbalance [283] while preserving the same subdivision in training, validation, and test sets used in our previous experiments. These training settings were adopted for both the socio-historical context classification and dating tasks.

2.8 Results and Analysis

In the following Section, we proceed to report on the performance obtained with single-input classifiers and with the ensemble model for both Socio-historical classification and dating tasks using CNN-based architectures. We then provide a comparison of the obtained performance with ViTs.

2.8.1 Socio-historical Classification Results

This Section reports on the results obtained for the socio-historical context classification using the different deep learning models described so far on the IMAGO dataset.

2.8.1.1 Single-input and Ensemble Classifiers

We first analyze the performance of each of the CNN-based architectures considering the top-k classification accuracy (i.e., Single-input classifier) on the full-image IMAGO dataset. The top-k metric considers not only the

top-1 prediction but the first k ones, and the prediction is considered correct if the correct label is present within those top-k predictions, otherwise, it is counted as incorrect. As reported in Table 2.2, the model based on ResNet-50 achieves, in most cases, the best performance among all the other CNN backbones. For this reason, from now on, we will adopt the ResNet-50 architecture for all the experiments that follow.

	Model		
	CNN		
Architecture	**DenseNet121**	**ResNet-50**	**InceptionV3**
input dim	256	256	299
#params (K)	6,963	23,526	25,130
Top-1	63.72	**64.35**	64.08
Top-2	83.38	**85.00**	83.83
Top-3	92.37	**92.85**	92.28
Top-4	96.54	96.66	**96.75**
Top-5	98.47	98.35	**98.53**

Table 2.2: Socio-historical model accuracies for an increasing Top-k classification (k ranging from 1 to 5).

	Single-input classifiers			
Top-k	**full-image**	**faces**	**people**	**random-patches**
Top-1	**64.35**	41.30	56.54	37.35
Top-2	**85.00**	65.55	78.48	62.40
Top-3	**92.85**	82.75	89.90	80.31
Top-4	**96.66**	90.86	94.74	90.42
Top-5	**98.35**	94.98	97.42	95.35

Table 2.3: Accuracy for the socio-historical single-input classifiers considering the Top-k predicted classes (k ranging from 1 to 5).

Now, considering the Resnet-50 as our reference backbone, we could analyze the results regarding the training of the picked backbone with different

image patches, which results are in Table 2.3. It is possible to appreciate that the full-image classifier exhibits a higher accuracy compared to the other single-input classifiers. To further investigate the reasons behind such results we report in Table 2.4 a comparison between the accuracy of each class considering the different single-input classifiers.

	Single-input classifiers			
Socio-historical Context	**full-image**	**faces**	**people**	**random-patches**
Affectivity	**64.54**	28.25	43.15	29.58
Work	**30.00**	24.00	22.00	29.00
Fashion	65.79	55.87	**67.60**	38.80
Motorization	**88.44**	17.01	51.02	29.66
Music	**40.62**	15.62	25.00	12.50
Politics	65.52	24.14	48.28	**66.67**
Rites	**71.50**	42.50	66.50	39.59
School	**60.58**	22.12	48.08	14.42
Free-time	58.09	46.09	**58.78**	51.94

Table 2.4: Single class accuracy for each socio-historical context classifier.

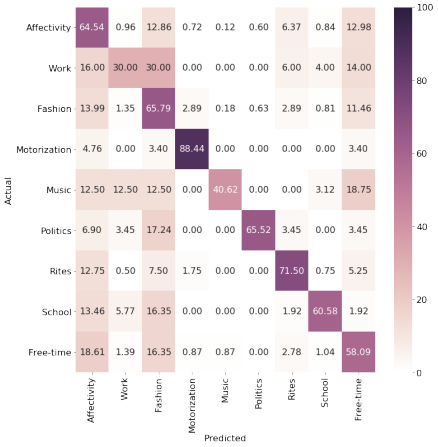


Figure 2.5: Confusion matrix for the full-image classifier.

As it is possible to observe, the model trained on IMAGO provides the best performance for the *Motorization*, *Rites*, *Music*, *School*, *Affectivity* and *Work* classes. This may be due to the presence of specific objects that drive the performance of the model, also considering that the model was initialized with the ImageNet pre-trained weights [89], which contains classes such as race car and car wheel. Indeed, from a socio-historical point of view, images from the classes *Rites* and *Music* could contain physical objects and/or symbols that are representative of that class (e.g., formal attires, musical instruments). Nevertheless, such objects only acquire meaning when people deal with them. However, the fact that the full-image classifier reached the highest accuracy for the *School*, *Affectivity* and *Work* classes mean that the network has also learned to recognize the presence of groups of people (e.g., school classes, friends standing in front of a monument, mother hugging her child) and specific clothing. Despite this classifier performing best, some

peculiar results have to be discussed. For example, the people classifier performs slightly better for the *Fashion* and *Free-Time* socio-historical contexts. This is probably because the network may be focusing on people's clothing details and poses instead of exploiting specific objects and/or backgrounds that are not present in the people's crops. Exemplar areas on which the models focus to classify its images are reported in the next paragraph of this section. Finally, the *Politics* class amounts to the only one for which, in terms of performance, the random-patches classifier is comparable to the full-image one.

We also evaluated different ensemble classifiers obtained from the combinations of the single-input classifiers. However, such combinations did not provide any significant improvement with respect to just considering the full-image model. For this reason, from now on, we consider the full-image classifier for the analysis that will follow and as the socio-historical context classifier in our application (check Figure 2.2).

Figure 2.5 shows the confusion matrix obtained with the full-image classifier. It is possible to observe that the classes responsible for the largest share of misclassifications are *Fashion*, *Affectivity*, and *Free-time*. This may be due to different causes. Firstly, some classes share visual elements. For example, pictures labeled with *Work* class often depict people in uniform in workspaces. These could mistakenly be classified as belonging to the *Fashion* class, as pictures in this class are characterized by people in a pose wearing some particular cloth items. Another example involves the *Music* and *Free-time* classes. Indeed, the *Music* category is characterized by photos portraying people playing some instruments or taking part in some musical event. The latter, however, could be easily associated with *Free-time* photos, since they also often portray a group of people in similar environments and poses. Secondly, the IMAGO dataset is unbalanced, as reported in Figure 2.1a (Section 2.5). Indeed, the most misclassified classes are also those which contain fewer samples.

2.8.1.2 Grad-Cam Analysis

We here report a qualitative analysis that aims at highlighting which visual cues led the classifier to associate a specific socio-historical context to a picture. To do so, we exploited the Grad-Cam algorithm [337], which delimits the areas driving the predictions performed by a deep learning model.

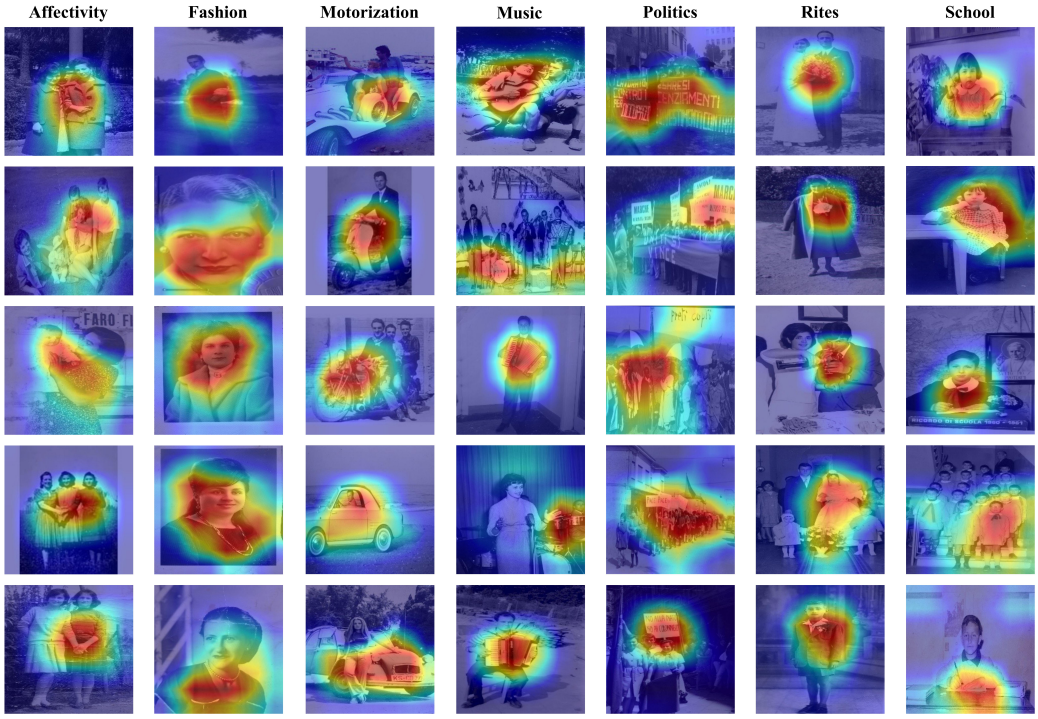


Figure 2.6: Grad-Cam analysis of socio-historical contexts of pictures within IMAGO.

Figure 2.6 depicts samples of correctly classified IMAGO images processed by the Grad-Cam algorithm. Each column, starting from the left, shows five exemplary images belonging to the *Affectivity*, *Fashion*, *Motorization*, *Music*, *Politics*, *Rites* and *School* classes, respectively. Such images are representative of the regions exploited by the full-image classifier. More in detail, people in certain poses close to each other (e.g., hugs, holding a baby, handshakes), as shown in the first column of Figure 2.6, are characteristic of the *Affectivity* class. Specific objects like earrings, necklaces, and lapels

but also particular hairstyles, are used to classify a picture as belonging to the *Fashion* class (second column of the Figure). All kinds of vehicles, as well as musical instruments, are used to recognize a given picture as a member of the *Motorization* or the *Music* classes, shown in the third and fourth columns, respectively. The presence of a political banner is typical of pictures in the *Politics* class (fifth column). The model also appears to individuate the objects that characterize the *Rites* class (e.g., white dress, flowers, pour a drink, cheers), as shown in the sixth column of Figure 2.6. Finally, children wearing school uniforms, as well as school gear (e.g., books, pens, desks) are used to recognize pictures in the *School* class (last column). It is not surprising that the model was able to correctly classify pictures belonging to the *Motorization* and *Music* classes, as these are clearly characterized by specific objects but, more importantly, already part of the model pre-trained on ImageNet [89]. However, also for the majority of the other classes (not studied so far in the literature, to the best of our knowledge), the model seems to be able to isolate and focus on the details that distinguish them.

Figure 2.7 shows instead some failure cases for the full-image classifier. From the leftmost picture and its probability histogram it is possible to see that a photo containing a car was classified as belonging to the *Motorization* class but the ground truth label assigned to the picture was *Affectivity* (two people standing close to each other in a specific pose). Instead, the rightmost picture and its corresponding probability histogram show that a picture depicting a school class was classified as belonging to *School*, while the actual one was *Work* (a teacher is standing in the rightmost part of the picture).

Such misclassifications may be traced back to the fact that the IMAGO dataset has been labeled by the owners of the pictures. The pictures thus convey such specific points of view, which may not be correctly predicted by the network. On the other hand, however, the point of view of the photo owner amounts to the ground truth, according to the methods adopted in socio-historical studies. In fact, the leftmost picture presented in Figure 2.7 was classified as *Affectivity* since the owner of the photograph was the child

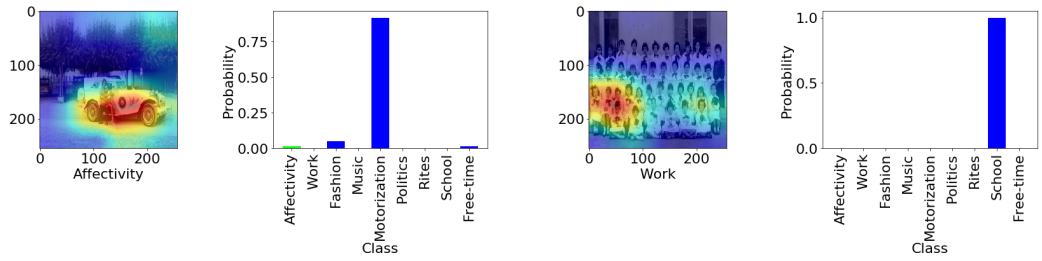


Figure 2.7: Grad-Cam examples of failure cases: *Affectivity* recognized as *Motorization* and *Work* recognized as *School*.

of the couple there portrayed. The same phenomenon happens in the rightmost one since the one who labeled the photo was a teacher of those students. This proves the intrinsic challenge that the socio-historical classification task poses, since any classifier, including an expert socio-historian, may be subject to such kind of errors. For such reason, we further investigate such phenomenon in Section 2.9, analyzing the differences between the predictions obtained with the deep learning model and the choices made by a socio-historian.

2.8.2 Dating

The dating amounts to a task where the DL models should predict the shooting year of a picture. In our experiments, we evaluated the models of such models using the same framework adopted in the previous section, but using as the reference metric the time distances, as in [136, 321]. The time distance defines the tolerance accepted in predictions concerning the actual year. For example, if a photo was labeled with the year 1942 and the model returned 1937 (or even 1947) this would be considered correct if the time distance is set to be equal or greater than 5, otherwise, it represents an error. In this work, model accuracies were computed considering temporal distances of 0, 5, and 10 years. The results are reported in Table 2.5. It is possible to appreciate that different baseline models (i.e., ResNet-50, InceptionV3, DenseNet121) return similar accuracies.

	Single-input classifier		
	ResNet-50	**InceptionV3**	**DenseNet121**
time distance	***full-image***		
d = 0	**11.31**	10.45	10.68
d = 5	**62.56**	61.38	60.77
d = 10	82.54	**82.82**	82.47
time distance	***faces***		
d = 0	**15.01**	14.60	12.91
d = 5	**58.09**	56.95	57.81
d = 10	78.39	78.46	**79.70**
time distance	***people***		
d = 0	**15.77**	12.56	13.99
d = 5	**62.40**	60.04	59.69
d = 10	**82.47**	81.39	81.42

Table 2.5: Model accuracies on different IMAGO patches and different time distances (d = 0, d = 5, d = 10).

Picking the Resnet50 as the best backbone, we focus on the comparison of the single-input classifiers in Table 2.6, where we also included a comparison with the performance of the model trained on random-patches, to provide a complete ablation of the role and the importance of the selected patches. The models fine-tuned on faces and people regions achieved a higher accuracy compared to the full-image classifier when considering a time distance equal to 0. This is also true for the random-patches classifier, which however performed worse with larger time distances. These results could be explained by model averaging, as using more data allows controlling uncertainty and reducing the prediction error rate [35]. Nevertheless, considering the gap among random-patches and faces, people classifiers, this increase in performance may also be because the faces and people classifiers learned specific visual features characteristic of given time slices [136, 321].

To verify whether such improvement was due to the averaging effect, we designed a specific experiment. We considered a test subset composed of all those images containing at least $n = 8$ faces or people crops (as in the case of random crops, see Section 2.7.1). To weigh the role of the number of

	Single-input classifiers			
Time distance	**full-image**	**faces**	**people**	**random-patches**
d = 0	11.31	15.01	15.77	11.64
d = 5	62.56	58.09	62.40	54.26
d = 10	82.54	78.39	82.47	76.12

Table 2.6: Comparison of single-input classifiers dating performance. The accuracy is reported for different time distances (d = 0, d = 5, d = 10).

faces/people, the accuracy values were computed considering k faces/people, with k growing from 1 to n. To ensure the completeness and fairness of this experiment, 1,000 random trials per each k faces/people/random-patches were considered. Results have been grouped by k and reported in Table 2.7. From these results, we can observe that averaging across multiple inputs, in general, results in a higher performance, which increases when considering the faces and people regions.

	Single-input classifiers		
# of crops	**faces**	**people**	**random-patches**
1	11.70 (1.47)	11.74 (1.56)	6.35 (1.27)
2	12.88 (1.39)	14.32 (1.46)	6.97 (1.23)
3	13.46 (1.27)	15.09 (1.44)	8.01 (1.20)
4	13.87 (1.25)	15.47 (1.26)	8.15 (1.14)
5	14.19 (1.19)	15.71 (1.14)	8.16 (1.07)
6	14.40 (1.10)	15.89 (1.06)	8.42 (0.95)
7	14.58 (1.06)	16.07 (1.04)	8.47 (0.86)
8	14.82 (0.95)	15.93 (0.91)	9.00 (0.00)

Table 2.7: Single-input classifiers averaging accuracies, along with their standard deviation, considering an increasing number of patches and a time distance d = 0.

Differently from the socio-historical context classification, an ensemble of different classifiers provides positive results for the dating task. Following

the flow described in Figure 2.3, we proceeded to evaluate different ensemble combinations, exploiting the full-image, faces, people, and random-patches classifiers. Since no significative improvements were observed employing the random-patches classifier, for the sake of clarity, Table 2.8 only includes the results which involve the full-image (T), faces (F), and people (P) classifiers. It is possible to observe that the best overall performance is obtained with the ensemble combination of all these classifiers. This shows that the model may benefit from averaging across different classifiers, as well as across multiple regions [35]. From now on, we consider, for all the following experiments, the model that reached the best performance which is the ensemble of the full-image, faces, and people classifiers.

	Ensemble classifiers			
Time distance	**T + F**	**T + P**	**F + P**	**T + F + P**
d = 0	17.14	16.79	17.91	**18.51**
d = 5	66.51	66.44	64.02	**67.53**
d = 10	85.66	84.80	83.75	**86.17**

Table 2.8: Ensemble model considering different combinations of full-image (T), faces (F), and people (P) classifiers. The accuracy is reported for different time distances (d = 0, d = 5, d = 10).

We now proceed to analyze which years were better modeled by our best classifier, through the confusion matrix, calculated with a time distance equal to 0, reported in Figure 2.8.

The diagonal structure demonstrates that the confusion mostly occurs between neighboring years, except for the initial and the final decades (this has been observed also in other works, as in [136]). The confusion created within the first 20 years may be caused by the low quality of the images and the limited number of samples representing those years. The confusion created within the last 20 years, instead, may be related to the fact that the number of images for these years is limited (as shown in Figure 2.1a).

Nevertheless, it is interesting to observe the information provided in Fig-

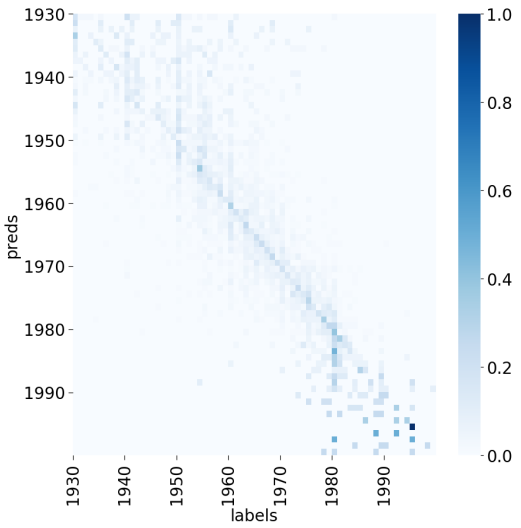


Figure 2.8: Confusion matrix for the dating task considering a time distance d = 0.

ure 2.9, where the model accuracy and the number of samples per decade are reported. This Figure confirms the finding exhibited by the confusion matrix, that is, the model accuracy improves after the 50s.

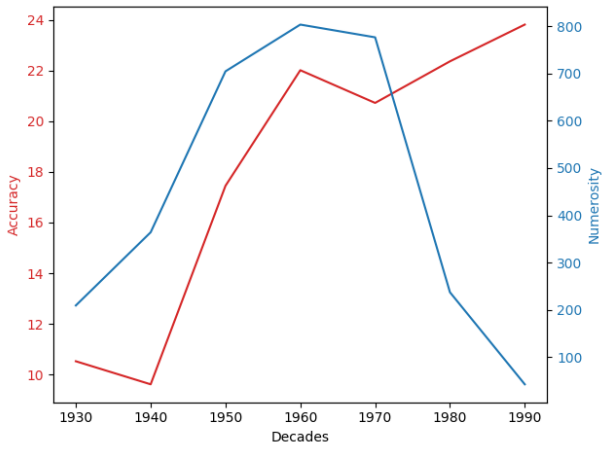


Figure 2.9: Model accuracy (red line) and number of samples (blue line) by decade for a time distance d = 0

Figure 2.9 also shows that, despite a reduction in terms of available sam-

ples per decade after the 80s, the performance of the model does not decrease. The accuracy generally improves after the 50's (also when the number of samples drops), again, this could be related to the fact that the images are of better quality than the previous decades. Differently from the socio-historical task we did not carry out a qualitative analysis for the dating task, as such type of analysis may already be found in literature [136, 321].

2.8.3 CNNs vs ViTs

As previously mentioned, we wanted to verify whether ViT models could improve the performance concerning our CNN baseline (i.e., Resnet50). To this aim, we fine-tuned different ViT configurations (Tiny, Small, Base, and Large), varying the size of the input images (i.e., 224×224 or 384×384) and considering patches of 16×16 pixels. For the training, we followed the procedure reported in [100], while adopting a weighted cross-entropy loss to counter the dataset unbalance [283] and preserving the subdivision in training, validation, and test sets used in our previous experiments. This process was adopted for both the socio-historical context classification and dating for all of the proposed datasets (i.e., IMAGO, IMAGO-FACES, IMAGO-PEOPLE, and IMAGO-RANDOM).

The results obtained with ViT for the socio-historical classification tasks are reported in Table 2.9, Table 2.10 and Figure 2.10. These should be contrasted with those previously presented in Sections 2.8.1. For the sake of clarity, the metrics there reported followed the same evaluation protocol followed for the CNN, considering Single-input classifiers, as detailed in Section 2.6.

On one hand, considering the socio-historical classification from the results reported in Table 2.9, it is possible to observe that in most cases either ViT-Base or ViT-Large outperforms the ResNet50 while requiring a much higher number of parameters and thus increasing the complexity of the model. When instead a similar number of parameters is used (e.g., ViT-Small with input size 224×224), ViTs exhibit a slightly lower performance. Neverthe-

Single-input classifiers									
	CNN	Vision Tranformer							
Architecture	**ResNet50**	**ViT-Tiny**	**ViT-Small**	**ViT-Base**	**ViT-Large**	**ViT-Tiny**	**ViT-Small**	**ViT-Base**	**ViT-Large**
#params (K)	23,526	5,526	22,669	85,806	303,311	5,599	21,815	86,097	303,700
input dim	256	224				384			
full-image									
Top-1	64.35	53.62	60.96	**66.24**	**67.87**	57.43	**65.13**	**68.53**	**69.19**
Top-5	98.35	96.63	97.72	**98.71**	**98.74**	97.14	97.84	**99.01**	**99.10**
faces									
Top-1	41.30	35.58	41.23	**42.98**	**43.13**	35.61	37.21	40.64	39.43
Top-5	**94.98**	89.87	93.54	92.03	93.84	89.84	91.67	93.90	93.21
people									
Top-1	56.54	48.42	53.23	56.08	**59.21**	46.58	51.99	**60.35**	**62.51**
Top-5	97.42	93.78	96.15	97.32	**97.45**	93.36	95.85	**98.02**	**97.69**

Table 2.9: Comparison of single-input classifiers for socio-historical context classification, considering both ResNet50 and ViT models. The accuracy is reported considering the Top-1 and Top-5 predicted classes.

	Single-input classifiers		
Socio-historical Context	**full-image**	**faces**	**people**
Affectivity	**52.76**	30.41	32.21
Work	**55.00**	4.00	26.00
Fashion	54.24	55.32	**58.75**
Motorization	**95.24**	37.41	93.20
Music	53.12	12.50	**56.25**
Politics	**79.31**	51.72	58.62
Rites	**72.25**	44.75	62.75
School	**80.77**	49.04	63.46
Free-time	**66.09**	34.43	58.61

Table 2.10: Single class accuracy for each socio-historical context classifier based on ViT-Small.

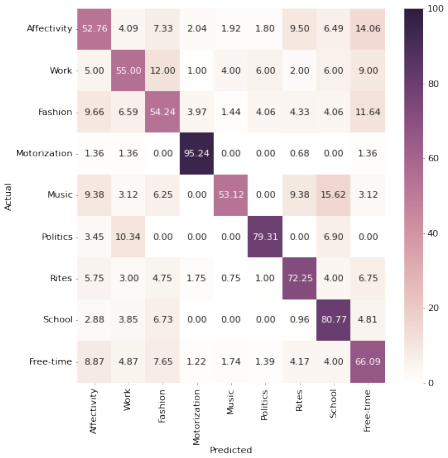


Figure 2.10: Confusion matrix for the ViT-Small full-image classifier.

less, comparing the results shown in Table 2.10 and Figure 2.10 with those reported in Table 2.4 and Figure 2.5, it is worth noticing that ViT-Small obtains a more balanced per-class accuracy.

Single-input classifiers									
	CNN	Vision Tranformer							
Architecture	**ResNet50**	**ViT-Tiny**	**ViT-Small**	**ViT-Base**	**ViT-Large**	**ViT-Tiny**	**ViT-Small**	**ViT-Base**	**ViT-Large**
#params (K)	23,651	5,538	21,693	85,852	303,373	5,611	21,839	86,144	303,763
input dim	256	224				384			
full-image									
d = 0	**11.31**	5.16	7.27	9.47	10.26	4.62	7.11	10.17	9.97
d = 5	**62.56**	38.85	43.40	50.72	53.44	35.21	46.37	54.11	55.74
d = 10	**82.54**	58.38	62.84	71.92	73.68	54.46	66.19	74.98	75.21
faces									
d = 0	**15.01**	3.47	5.10	6.66	7.43	4.11	4.78	6.72	7.46
d = 5	**58.09**	31.39	39.42	46.46	46.59	34.58	38.34	45.73	49.24
d = 10	**78.39**	51.21	60.77	68.51	68.58	55.32	59.85	68.01	71.70
people									
d = 0	**15.77**	4.05	4.40	7.11	7.87	4.14	4.68	7.33	7.97
d = 5	**62.40**	33.65	34.51	46.88	48.69	32.22	38.91	46.18	49.17
d = 10	**82.47**	54.21	51.56	68.90	70.24	52.42	59.40	67.81	70.04

Table 2.11: Comparison of single-input classifiers for the dating, considering both ResNet50 and ViT models. The accuracy is reported for different time distances (d = 0, d = 5, d = 10).

Considering the dating task, the results in Table 2.11, which should be contrasted with the ones introduced in Section 2.8.2, show that the ResNet50 outperforms all single-input ViT configurations for dating. We also considered different ensemble combinations but no relevant improvements were detected and for this reason, are not here reported.

Concluding, the ViT approach exhibits divergent behaviors when applied to dating and socio-historical context classification. Why this occurred may be explained by resorting to [297], where the authors highlighted how ViT: (a) incorporates more global information than ResNet at lower layers, leading to different features, and, (b) strongly preserves spatial information adopting class tokens. Indeed, the inclusion of more global information at lower layers and the strong preservation of spatial information could be the reason why socio-historical context classification obtained a better accuracy than dating. This is qualitatively represented by a few GradCam examples reported in Figure 2.11: more accurate activations are obtained when compared to the corresponding examples for ResNet50, reported in Figure 2.6. On the contrary, dating often requires focusing on specific local visual cues rather

than on global ones, as also highlighted by [136].

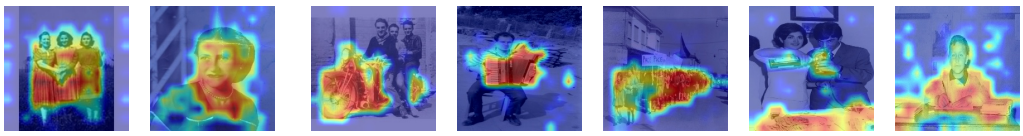


Figure 2.11: Grad-Cam analysis of socio-historical contexts using ViT-Small.

2.9 Qualitative vs Quantitative analysis in socio-historical studies: Human vs Machine assessment

As introduced in Section 2.2, we aimed to show how mixing qualitative and quantitative methods may overcome classical difficulties in socio-historical research. Therefore, we wanted to verify whether quantitative techniques, built resorting to results obtained from qualitative processes, may be employed to perform specific qualitative analyses, yielding a mixed qualitative-quantitative one.

More in detail, we exhibit the process that follows: (a) we took the DL models previously trained on family album pictures for the socio-historical and date classification tasks which categories derived from previous sociological and historical qualitative studies described in Section 2.5; (b) then we adopt those models, which are capable of recognizing salient features of socio-historical interest, and compared their performance to the ones obtained by a socio-historian who manually assessed the photos employing qualitative paradigms.

In particular, to assess how deep learning models performed in comparison to a scholar, we designed a specific experiment where a socio-historian was asked to categorize all the 3,327 images belonging to the IMAGO test set, providing for each picture its socio-historical context and the shooting

date. It is worth noticing that the socio-historian employed a qualitative approach to classifying each image, exploiting his/her sociological and historical background, and not only based on the contents of the picture.

2.9.1 Human vs Machine: Socio-historical Classification

For this experiment, we select the best deep learning model trained on socio-historical classes with a comparable number of parameters (i.e., Resnet50) that provides the best top-k accuracy. In contrast, the socio-historian could freely select multiple categories per photo. Although unrestricted to use as many labels as desired, it is interesting to note that no more than three have been considered at once (one class was chosen for 2,147 photos, two classes for 1,131, and three classes for 49 pictures).

To make a fair comparison, we considered the k most probable classes chosen by the SHM model and compared them with the k classes selected by the socio-historian. We then proceeded to compute the accuracy of the socio-historian and of the model, as follows. For example, if the ground truth for a photo was "*Affectivity*", the predictions provided by the application and the selections made by the socio-historian would be considered positive if both contained "*Affectivity*". Since the scholar could choose the number of categories to assign, we computed such scores cumulatively.

		Top-k Accuracy	
Cumulative k	**Cardinality**	**Socio-historical context module**	**Socio-historian**
1	2,147	64.88	54.82
1-2	3,278	72.02	66.53
1-2-3	3,327	72.24	66.93

Table 2.12: Human vs Machine: Accuracy comparison for increasing values of k (k indicates the number of selections made by the socio-historical scholar and the most probable classes returned by the model).

In particular, in correspondence of **Cumulative k**, in Table 2.12, with

$k = 1$, a prediction is counted as positive in case it matches the ground truth. It follows that, if $k > 1$, a positive match is recorded in case one of the k predictions matches the ground truth. The results are reported in Table 2.12. The first, simple observation, is that the proposed application obtained accuracy levels that surpassed those obtained by the socio-historical scholar. For example, when we consider those pictures that were tagged with only one category by the socio-historical scholar, an accuracy of 54.82% was obtained vs an accuracy of 64.89% of the application. This occurred also when considering those pictures for which the socio-historian chose one or more classes, the application was still able to obtain a higher performance.

In Figure 2.12 we show a representative example of a case where the model predicts the correct label, unlike the socio-historian. In fact, the socio-historian fails to recognize a particular detail that only the owner could have known (the subject of the photo is posing wearing a particular outfit); on the contrary, the model correctly classified this image.

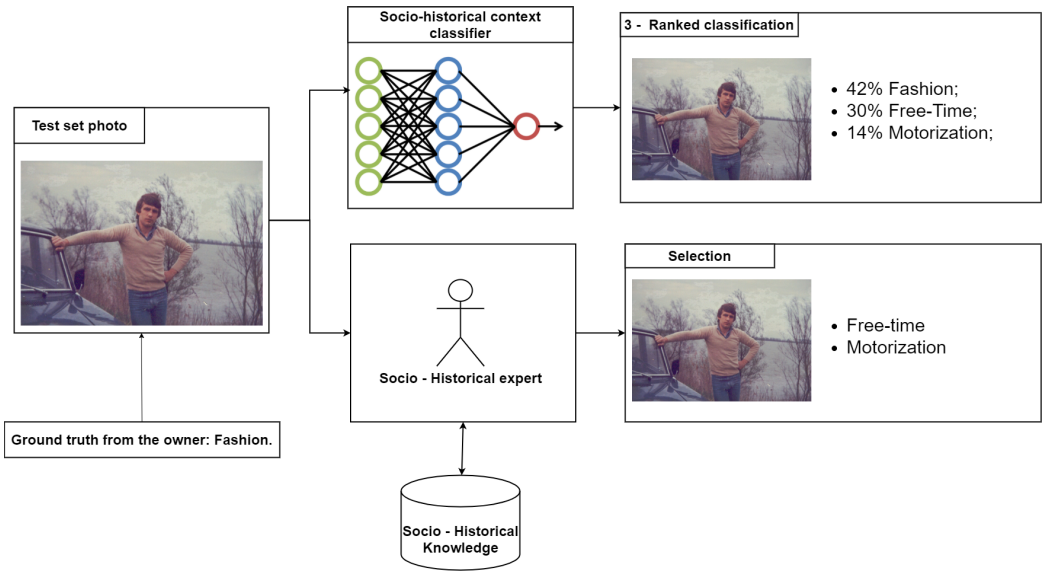


Figure 2.12: Human vs Machine: experiment diagram.

2.9.2 Human vs Machine: Dating Classification

To compare dating performance instead, the socio-historian labeled all the pictures belonging to the test set assigning a year in the [1930,1999] range. The obtained results are reported in Table 2.13. The dating module (Ensemble classifier 2.8) performed better than the socio-historian considering the specific picture shooting year (+12.58%). The difference in performance decreases when a higher time distance is considered, arriving at 3.64% when the time distance equals 10.

	Accuracy	
Time distance	**Dating module**	**Socio-historian**
d = 0	18.51	5.93
d = 5	67.53	56.36
d = 10	86.17	82.53

Table 2.13: Human vs Machine: accuracy reported for different time distances (d = 0, d = 5, d = 10).

2.10 Searching for Cultural Relationships With Cross-Dataset Experiments

Considering the existence of the USA-Italy cross-cultural influence on visual appearances between individuals, throughout the second half of 1900 [147, 56] we carried out an analysis to verify whether this effect could be also quantified using DL. To achieve such a goal, we adopted a cross-dataset approach considering the American-people datasets provided by [136, 321] (described in Section 2.4) and IMAGO as the Italian counterpart.

In particular, among all the relatable datasets [136, 321, 258, 247] no one includes family album photos (each picture contain at least one person). However, [136, 321] share some common traits with IMAGO: they analyzed American datasets comprising people's faces and torsos, where subjects are

often in pose and dressed for a specific occasion. This means that it is possible to extract what characterizes all of them: people's faces and torsos. Considering such features, the cross-dataset experiment will consider along with such datasets the pictures in the IMAGO one that are comparable to them (i.e., IMAGO-FACES and IMAGO-PEOPLE). Finally, all the images within the selected datasets (IMAGO, [136], [321]) were shot during the 20th century.

2.10.1 Cross Dataset Experiments Methodology

The trained models from [136, 321] should be adopted to perform cross-dataset experiments. However, those models weren't available for the deep learning framework used in such work to both train and evaluate the IMAGO models. So, we proceeded by mimicking the training procedure listed in the respective works [136, 321] to define different deep learning-based models that could be adopted to perform the target analysis. It is worth noticing that, the dataset provided by [136] contained only face-patches, making the comparison with our trained people-classifier model impossible. To achieve such a goal, we first fine-tuned the VGG16 and AlexNet architectures, respectively used in [136, 321], following the procedures described by the authors. In all the cases, an 80%-20% training-test split was considered. All the information is reported in Table 2.14. Important to highlight that the dataset introduced in [136] considers only people's faces, while the one introduced in [321] offers both people's faces and torsos.

Original dataset	**Architecture**	**Train cardinality (%)**	**Test cardinality (%)**
[136]	VGG16	28,554 (80.0%)	8,716 (20.0%)
[321]	AlexNet	72,800 (80.0%)	18,200 (20.0%)
IMAGO collection	ResNet50	11,252 (80.0%)	4,421 (20.0%)

Table 2.14: Models settings and accuracies of existing solutions and IMAGO considering the dating task.

We then evaluated these models on the IMAGO dataset. Vice versa,

the *faces* and *people* classifiers, presented in this work, have been evaluated on the corresponding regions offered in the datasets from [136, 321]. For a fair evaluation, the experiments were carried out on the 1930-1999 time-span for the [136] vs. IMAGO comparison, while considering 1950-1999 for the [321] vs. IMAGO one, respectively. The results of such evaluation are reported in Tables 2.15, 2.16 and 2.17. As expected, the final performance is really poor in both directions, i.e., the models fine-tuned on our dataset and evaluated on the test set of the related works and vice versa. This may be due to the domain-shift effect (these datasets have been acquired from multiple locations, using different cameras) [293]. However, another reason for such poor performance could be addressed as the intercultural influence that changes the visual appearance of people of different ages.

	Faces classifier cross-dataset comparison with [136] – range 1930-1999	
time distance	**Our faces classifier tested on [136]**	**Model from [136] tested on IMAGO-FACES**
d = 0	2.50	1.02
d = 5	24.49	12.4
d = 10	41.33	25.68

Table 2.15: Comparison of our faces classifier evaluated on the test set of [136] with the model from [136] evaluated on the IMAGO-FACES test set. We considered the common time slice 1930-1999.

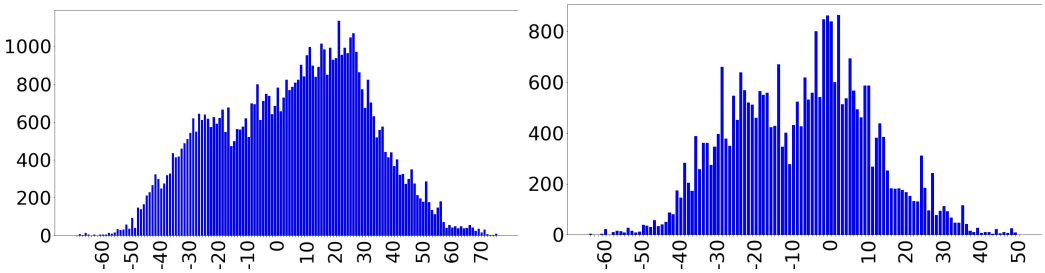
	Faces classifier cross-dataset comparison with [321] – range 1950-1999	
time distance	**Our faces classifier tested on [321]**	**Model from [321] tested on IMAGO-FACES**
d = 0	1.45	2.46
d = 5	14.02	25.09
d = 10	26.20	46.13

Table 2.16: Comparison of our faces classifier evaluated on the test set of [321] with the model from [321] evaluated on the IMAGO-FACES test set. We considered the common time slice 1950-1999.

To explore such possible influence quantitatively, we collected the error between the predicted and the actual year per each picture.

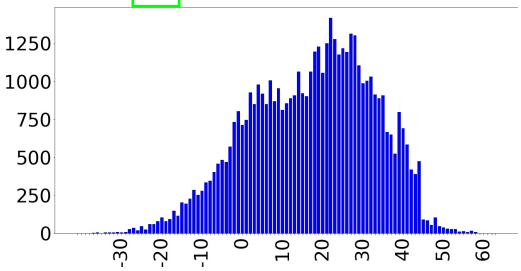
	People classifier cross-dataset comparison with [321] – range 1950-1999	
time distance	**Our people classifier tested on [321]**	**Model from [321] tested on IMAGO-PEOPLE**
d = 0	1.49	1.74
d = 5	18.08	18.22
d = 10	35.21	35.43

Table 2.17: Comparison of our people classifier evaluated on the test set of [321] with the model from [321] evaluated on the IMAGO-PEOPLE test set. We considered the common time slice 1950-1999.

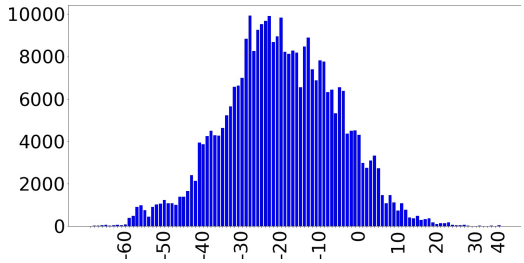


(a) [136] on IMAGO-FACES

(b) IMAGO-FACES on [136]



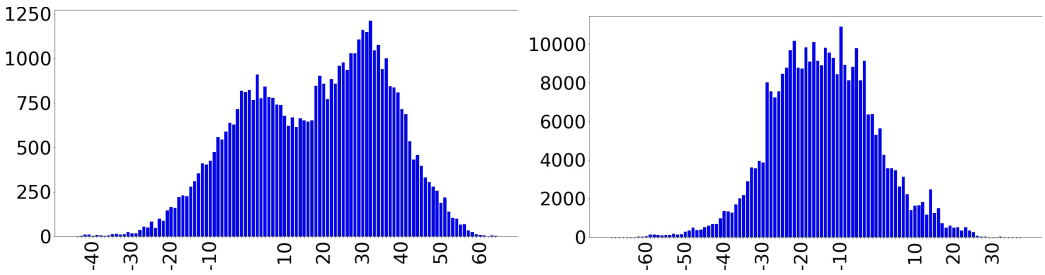
(c) [321] on IMAGO-FACES



(d) IMAGO-FACES on [321]

Figure 2.13: Dating error distributions for faces.

The error distributions are reported in Figure 2.13 and Figure 2.14 for the cross-dataset experiments involving faces and people images. In particular, Figures 2.13a and 2.13c depicts that the date estimation error distributions are shifted towards positive values, while, in Figs. 2.13b and 2.13d towards negative ones. The models built on top of American datasets [136, 321] applied to IMAGO-FACES tend to overestimate the image shooting year while the opposite phenomenon (underestimation) occurs when the model presented in this work is applied to [136] and [321]. The same phenomenon



(a) [321] on IMAGO-PEOPLE (b) IMAGO-PEOPLE on [321]

Figure 2.14: Dating error distributions for people.

appeared considering people's torsos. Nevertheless, we were able to analyze such phenomena only for [321] which provides pictures of full-figure instead of only faces. The obtained results are reported in Figure 2.14. To further investigate whether the errors were statistically significant, we performed a data analysis process. Firstly, we measured the normality of the error distributions by adopting a normality test that combines skew and kurtosis to produce an omnibus test [94, 83]. The normality test was adopted to discriminate between parametric and non-parametric statistical tests. In our experimental sessions, none of the considered distributions passed the normality test (p-value < 0.001, the null hypothesis test that a sample comes from a normal distribution). For this reason, we proceeded by adopting non-parametric tests. In particular, we evaluated whether the difference between the ground truth and model prediction pairs (i.e., error distributions) were statistically significant when performing the Wilcoxon signed-rank test. The Wilcoxon signed rank is a non-parametric test where the null hypothesis states: "two related paired samples come from the same distribution". In particular, it tests whether the distribution of the differences is symmetric about zero [73]. Also, in this case, the null hypothesis was rejected for all the conditions (p-value < 0.001), indicating that the considered differences exhibit different distributions. Finally, we verified whether the shift between two cross-dataset (e.g.,) settings came effectively from two different distributions with the Mann-Whitney U test [225], providing some clues

about the significance of the overestimation/underestimation effect. The non-parametric Mann-Whitney U rank test hypothesizes two independent samples and tests the null hypothesis that the distribution underlying the first sample is the same as the distribution underlying the second sample. Even for the Whitney U test the null hypothesis was rejected for all the conditions (p-value < 0.001), indicating that the considered cross-shift differences came from different distributions. These results motivated us to perform an additional visual analysis to qualitatively explore the possible time-shift phenomenon in a cross-dataset setting.

2.10.2 Qualitative Analysis Of Visual Intercultural With UMAP

Considering the results reported in the previous Section, we decided to visually explore the images that were most shifted, from a dating perspective, while evaluating the models described in Section 2.10.1 on the IMAGO datasets, and the IMAGO models on [136, 321]. In practice, we exploited the CNN extracted features (embeddings) in such a cross-dataset setting. However, for the considered models (ResNet50, VGG, AlexNet), the embedding vectors lie in a latent space of cardinality $2,048$ or $4,096$. To project both of them in a space with the same cardinality which can also be visualized, we adopted dimensionality reduction, which aims to preserve as much of the significant structure of the high-dimensional data as possible in a low-dimensional map (i.e., 2 or 3 dimensions).

In particular, we adopted one of the most used data dimensionality reduction algorithms: UMAP [236]. When the data presents a non-linear structure (as in the case of a CNN latent space), UMAP and the t-distributed stochastic neighbor embedding (t-SNE) represent a valid method to reduce them due to their non-linear nature [236, 397]. However, UMAP is faster and scales better for both dataset dimensionality and cardinality while better preserving the global structure of the data [236]. In particular, t-SNE has been observed to distort distances between clusters in the original high-

dimensional space, while UMAP preserves them more accurately [274, 86]. In other words, this technique produces high-quality visualizations by reducing the high-dimensional data revealing structures in them also considering large data sets [236]. In our analysis, we employed the official implementation of the UMAP algorithm [237]. To carry out a cross-dataset analysis, we picked as target datasets the one introduced in [321] and IMAGO which includes people's torsos. This choice was mainly driven by the fact that these datasets possess a greater, more detailed, and more varied number of pictures concerning the one introduced in [136].

Figure 2.15: UMAP applied to the embeddings of the model trained with [321] (indicated as A) on the IMAGO-PEOPLE dataset. The model correctly predicted the selected images within a decade of confidence.

With such a setting, we first analyzed the clusters extracted by the UMAP algorithm while being applied to the embeddings extracted by inferring date with the model trained with [321] on IMAGO-PEOPLE. In Figure 2.15 we reported a sample of images that were correctly predicted for each of the

considered decades in the common dataset time-span. In Figure 2.16, instead, we report a sample of images that were wrongly predicted with a shift of 30 years, which is the most occurrent shift reported in Figure 2.14 (Section 2.10.1). It is worth noticing that in Figure 2.15 the UMAP algorithm was able to highlight clusters for different decades that however possess intersection with clusters of adjacent decades (e.g. some pictures from 1950 are mixed with the ones of 1960). In Figure 2.16 instead, it is interesting to note that many samples that were labeled with a 30-year shift are not colored: this could mean that the model exploited other cues apart from the colors to date those images (e.g., the style of men in lower pictures in Figure 2.16 possess very similar fashion style).

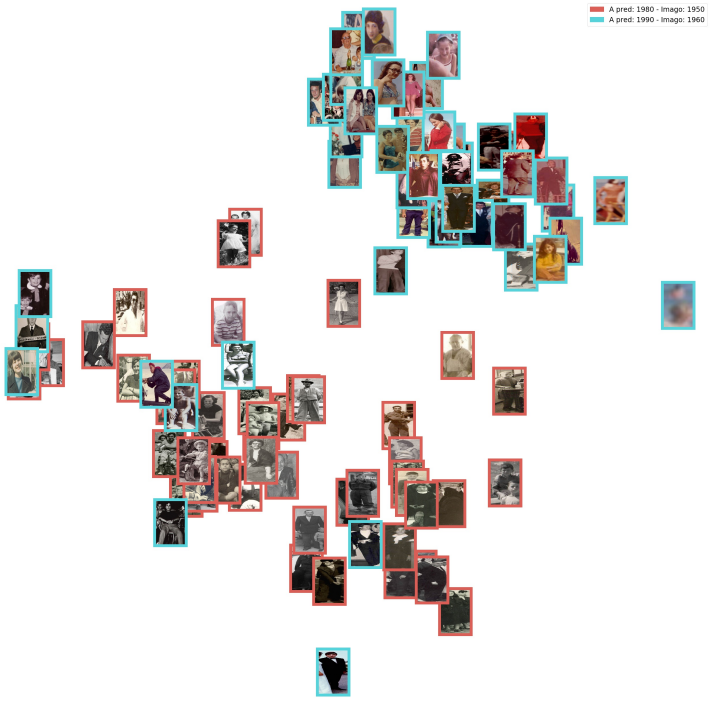


Figure 2.16: UMAP applied to the embeddings of the model trained with [321] (indicated as A) on the IMAGO-PEOPLE dataset. The selected images were wrongly predicted to be 30 years forward the real shooting date.

Secondly, we explored the output of the UMAP algorithm while evaluating the embeddings extracted by inferring the date on [321] with the model

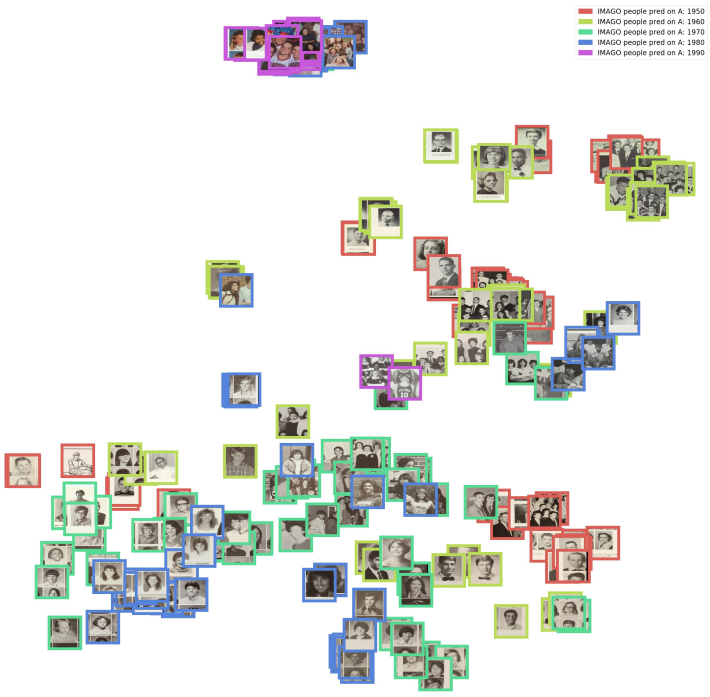


Figure 2.17: UMAP applied to the embeddings of the model trained with IMAGO-PEOPLE on [321] (indicated as A). The model correctly predicted the selected images within a decade of confidence.

trained with IMAGO-PEOPLE. In Figure 2.17 we report a sample of images that were correctly predicted for each of the considered decades in the common dataset time-span. In Figure 2.18 instead, we report a sample of images that were wrongly predicted with a shift of -20 years, which is the majority shift reported in Figure 2.14 (Section 2.10.1). Also in this case, the UMAP algorithm was able to highlight clusters for different decades that however possess intersection with clusters of adjacent decades. It is worth noticing that the majority of samples that were labeled with a -20 shift are in black-white: the model may exploited other cues apart from the colors to date those images (e.g. similar female hairstyles in the left-lower cluster).

We here highlight that these results were obtained in a qualitative analysis setting, considering the cardinality of the considered samples, and so they cannot be generalized [37]. However, the adoption of dimensionality reduc-

tion and visualization algorithms, such as the UMAP, to visualize neighbor images in the latent space eases and speeds up the classical analysis approach that would be done in museums or academia for searching relationships with visual cues. If correctly automatized, this approach could be a valuable tool for socio-historical researchers, as it allows for a deeper understanding of complex phenomena, such as cross-cultural influences, just by analyzing visual learned cues.

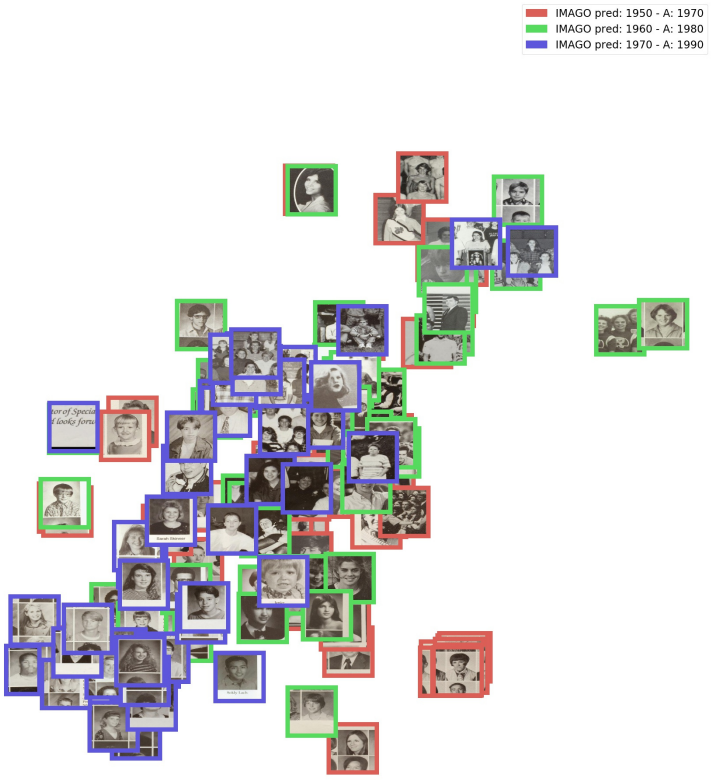


Figure 2.18: UMAP applied to the embeddings of the model trained with IMAGO-PEOPLE on [321] (indicated as A). The selected images were wrongly predicted to be -20 years forward the real shooting date.

2.11 Composing eXtended Reality With Deep Learning To Support Socio-historical Research

As mentioned in Section 2.2, taking into consideration the limited availability of digital tools capable of identifying, digitizing, and easily sharing components of human cultural heritage, in particular, referred to family photo albums, we explored the possibility of using XR paradigms. In particular, we exploited the best DL classification models defined in Section 2.6 to develop an AR system for the digitization and cataloging of collections of analog family album photographs. To this date we adopted: (a) the HoloLens 2 [394] as an AR wearable device to visualize and catalog family album photos, (b) a well-known object detector, namely YOLO [305], to automatically crop images in the user's view, (c) the DL models trained on the IMAGO dataset (Section 2.8) to classify cropped images and (d) providing the chance to share with remote users the HoloLens 2 scene view. The produced tool interface was validated through an assessment model, asking a group of ten people to provide their comments regarding the use of our prototype.

2.11.1 System Architecture

We first designed an AR system architecture that pipelines: (a) the HoloLens 2 interface, (b) Deep learning processing, and, (c) a sharing layer. As depicted in Figure 2.19, we envisioned a simple application for the HoloLens 2 device. The application sends all of the frames within the user's view to the DL models which, in case one or more historical pictures are detected, provides the bounding box(es) and the label(s) that can be then visualized in AR. Such information is utilized to augment the visualization of the family photographs by resorting to the HoloLens 2 interface. In addition, the application supports the sharing of the augmented HoloLens user's view to other devices (e.g., smartphone, tablet, PC).

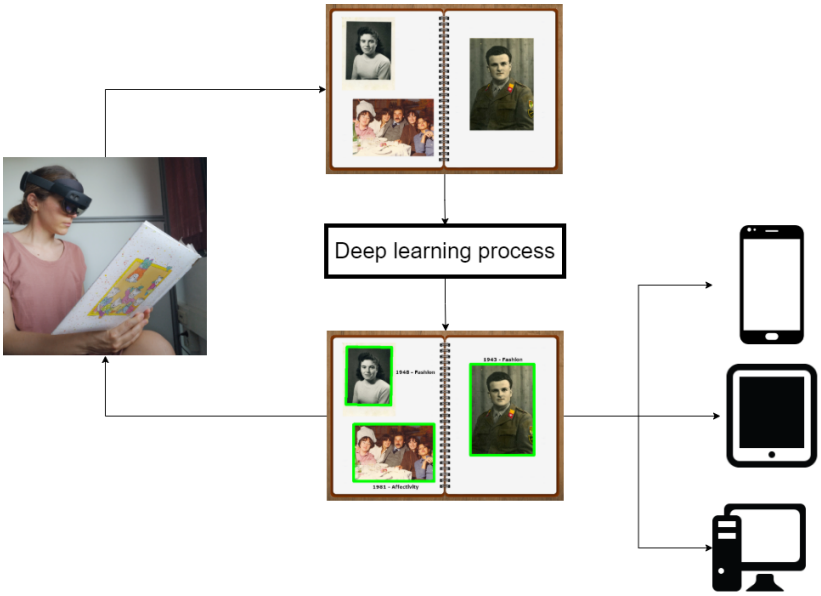


Figure 2.19: Hololens 2 interface architecture.

Figure 2.20 depicts instead the deep learning pipeline adopted in our AR app. As a first approach, we resorted to a classical computer vision object detection pipeline to recognize the area enclosing family album photos, stacking respectively: pre-processing image (i.e., bilateral filtering), edge-detection (i.e., canny edge detector) and contour-detection algorithms (i.e., Sobel). However, given the poor performances obtained in a dynamical lighting experimental setting, we leveraged DL-based object detectors, as they can be generalized in more varied and complex visual environments [155]. Within the DL object detectors zoo, the YOLO architecture has emerged [392], considering its vast adoption and deployability on XR applications [2]. In particular, we resorted to YOLOv5s, because it represents a good trade-off between accuracy and time/memory complexity, making it a good candidate to deploy it on the HoloLens 2, which has limited hardware resources([164]).

Despite the huge performance of YOLO models, they were not trained to recognize pictures within a family album photo, since they were trained to recognize and classify objects within the ImageNet dataset [317], which also

[2] https://github.com/ultralytics/yolov5

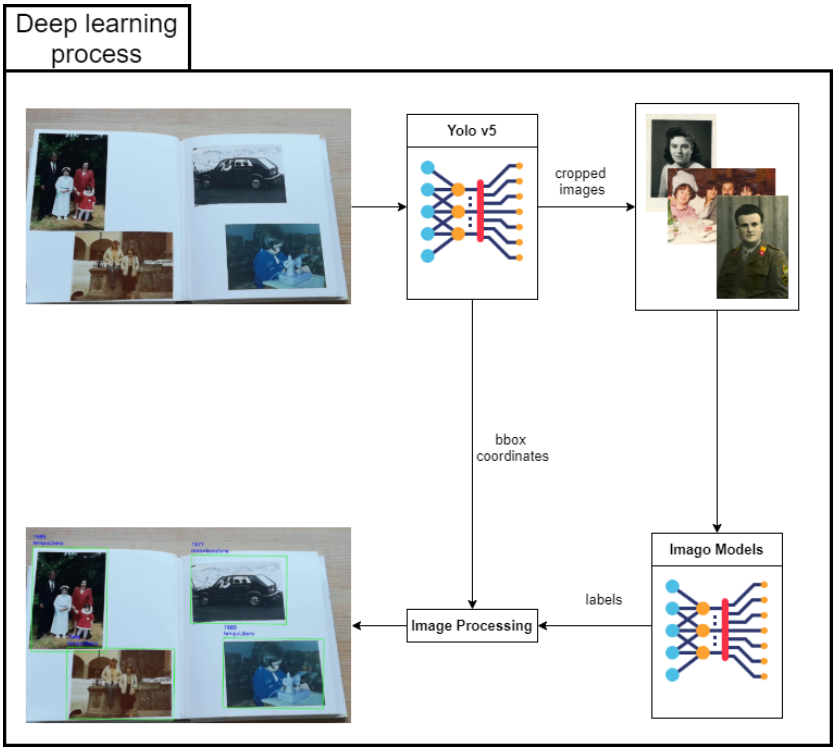


Figure 2.20: Deep learning processing architecture.

furnish object bounding boxes for each picture [3]. To let so this class of models able to solve the analyzed task, we first synthesized a new dataset by random pasting, on random backgrounds, n pictures (with n ranging between 0 and 4, which is the usual range in family album pictures [53]), casually picked from the IMAGO dataset (e.g., paper, wall backgrounds). Images might also partially overlap to increase the detection robustness (examples are given in Figure 2.21). With this process, 9,006 images were obtained, and partitioned into training (7,372) and test (1,634) sets.

We then proceeded to fine-tune the YOLOv5s model, exploiting data augmentation techniques (e.g., random brightness, horizontal and vertical flipping) for 10 epochs, with a batch size of 32 using the Adam optimizer. We set the learning rate as 1e-3 with a weight decay equal to 5e-4. A sample of the evaluation of such a trained model on our test set is depicted in Figure 2.22.

[3] https://image-net.org/download-bboxes.php

Figure 2.21: Images from the synthetic dataset.

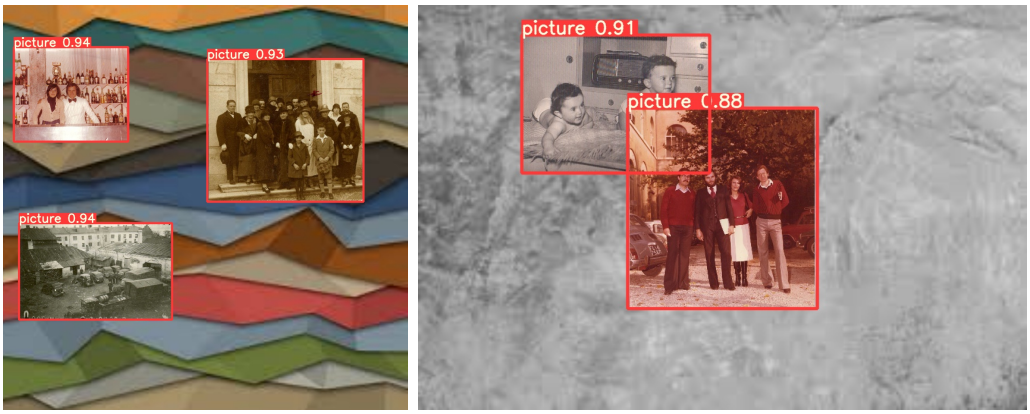


Figure 2.22: Images classification obtained from the synthetic dataset test set with YOLOv5.

The result of this stage is a DL model capable of cropping historical pictures appearing in family albums. Finally, the IMAGO DL models are exploited to predict the date and the socio-historical context of each cropped picture. As specified in Section 2.7, the model is capable of dealing with pictures whose date falls within the 1930-1999 interval and whose socio-historical context belongs to one of the following {Work, Free-time, Motorization, Music, Fashion, Affectivity, Rites, School, Politics}, according to their definition. Such labels, along with the ones provided by YOLO, are then sent to the HoloLens 2 to augment the view of the photographs with such information

At this point, we already have a functioning and portable tool to catalog family album pictures just by watching them with the headset.

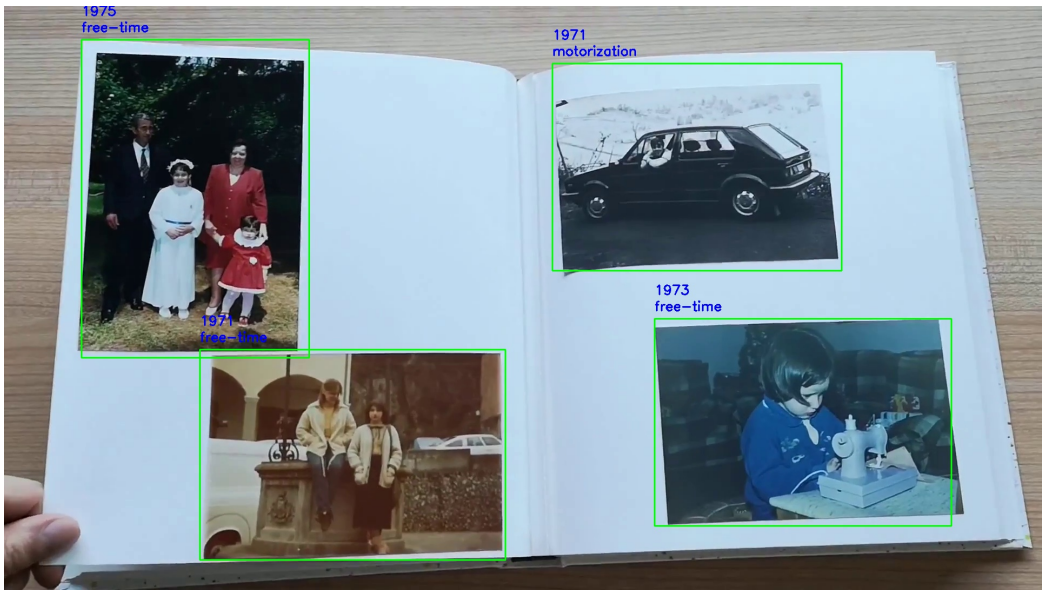


Figure 2.23: Real-world example of Augmented HoloLens 2 view.

The captured stream, along with the overlayed labels over each picture, could be leveraged as a piece of information that may be shared following a collaborative style and hence sent to the interfaces of those users who are viewing photo albums from a remote location. To this aim, we built a simple HTTP server to continuously stream, to any kind of device (e.g., smartphone, tablet, PC), the augmented view of the HoloLens 2. In brief, the server processes the video stream captured by the HoloLens 2 and adds to each of its frames the labels returned by the YOLO and IMAGO DL models. The use of HTTP is a design choice meant to support easy access to the stream, from any type of device. A real-world example of the augmented view, as seen from the HoloLens 2, is provided in Figure 2.23.

2.11.2 System Evaluation & Assessment Model

To evaluate the system detailed in Section 2.11, we asked a group of ten participants to answer a survey regarding their experience. This group had

an average age of 26 years and was composed of 3 females and 7 males. The number of participants has been chosen as a trade-off between the necessity of acquiring sufficient feedback data from a population and the time spent for the evaluation phase: ten participants have repeatedly proven to be a sufficient population to discover over 80% existing interface design problems [117, 323].

Written consent to participate in this experimental study was collected from each subject. The experimental session was possible thanks to the full compliance with the COVID-19 sanitary protocol adopted by the University of Bologna at the time of the experimentation.

The survey was designed to assess four constructs, namely, Perceived Ease and Enjoyment of Use (PEEU), Deep Learning Gain (DLG), HoloLens Perspective (HLP), and Receiver Perspective (RP).

The first chunks aimed to investigate the PEEU constructs and the DLG ones (both evaluated through a 5-point Likert scale). For simplicity, a general overview of these constructs is reported in Table 2.18.

Construct	Question
PEEU	(A1) I found the new interface easy to understand
	(A2) I would prefer watching an Augmented Family Photo Album respect to a normal one
	(A3) I enjoyed the overall experience
DLG	(B1) I appreciated the automatic identification of pictures
	(B2) I appreciated the automatic estimate of pictures' date
	(B3) I appreciated the automatic estimate of pictures' socio-historical context

Table 2.18: Items and questions used in the survey to assess the Perceived Ease and Enjoyment of Use (PEEU) and the Deep Learning Gain (DLG) constructs. All the questions here reported were evaluated on a 5-Point Likert Scale.

Individuals' satisfaction and acceptance of a technological innovation, such as an AR application, may be analyzed through different theoretical approaches. The Technology Acceptance Model (TAM) [88] amounts to one of the most popular assessment approaches, as it allows to measure of user

intentions in terms of their attitudes, subjective norms, perceived usefulness, perceived ease of use, and related variables. In our case, we want to concentrate on the perceived usefulness and ease of use. Perceived usefulness is defined as the degree to which individuals believe that adopting one particular technology will improve an aspect of their life, whereas perceived ease of use is the degree to which an individual thinks that adopting a particular technology will be easy to use. Starting from these definitions we composed the PEEU construct:

A1. I found the new interface easy to understand;

A2. I would prefer watching an Augmented Family Photo Album with respect to a normal one;

A3. I enjoyed the overall experience.

The A1 sentence immediately gets to the point; item A2 has been introduced as a further investigation, to understand if the users prefer to live an augmented experience concerning a classical one. Through A3, we ask for a broad evaluation of the experience. Following this path, we also want to evaluate the usefulness of the three DL models that have been developed to carry out the three different computer vision tasks present in this work: family album photo recognition, date, and socio-historical context estimations. For such reason, we also designed the chunk of question items defined as Deep Learning Gain (DLG), which is thought to measure the utility of our DL models:

B1. I appreciated the automatic identification of the pictures;

B2. I appreciated the automatic estimation of the pictures' dates;

B3. I appreciated the automatic estimation of the pictures' socio-historical context.

Construct	Question	Evaluation Method
HLP	(C1) Would you use the HoloLens application to share your memories?	Yes/No question
	(C2) Nowadays, would you use the HoloLens application to share your photo family album with a distant affection?	Yes/No question
	(C3) Nowadays, would you prefer to share your memories with the HoloLens rather than sharing them in presence?	Yes/No question
	(D1) Would you use the HoloLens application to share with anyone your photo family album?	5-point Likert scale
	(D2) Do you think this HoloLens application would push you to contact more your affections?	5-point Likert scale
	(D3) Do you think this HoloLens application would push you spend more time visualizing your photo family album?	5-point Likert scale
RP	(C4) Would you use this application to revive memories with a distant affection?	Yes/No question
	(C5) Do you think this application would push you to contact more your affections?	Yes/No question
	(D4) Would you use this application to visualize photo family albums of strangers?	5-point Likert scale
	(D5) Nowadays, do you think this application could foster the creation of bonds between strangers?	5-point Likert scale

Table 2.19: Items and questions used in the survey to assess the HoloLens Perspective (HLP) and the Receiver Perspective (RP) construct.

The second chunk of question items regards the HLP and RP constructs are defined in Table 2.19. This additional set of questions was defined to explore the different perspectives of users enjoying our application (i.e., the one of the HoloLens 2 wearer and the remote one). In particular, they are based on the concept of Behavioural Intention, that is the individual intention to use a particular technology, that was designed and adapted from the ones reported in [309]. However, different from the first constructs which were meant to exclusively measure the usefulness of our system, these questions aim at inspecting more intimate aspects of the users' intentions (i.e., the use they would make of this application and its impact on their daily lives). In particular, both constructs were investigated by exploiting two groups of questions: the C and D groups. The C group is formed by Yes/No question scale questions, to avoid neutral scores:

C1. Would you use the HoloLens application to share your memories?

C2. Nowadays, would you use the HoloLens application to share your family photo album with a distant friend or relative?

C3. Nowadays, would you prefer to share your memories with the HoloLens,

rather than sharing them in presence?

C4. Would you use this application to revive memories with a distant affection?

C5. Do you think this application would push you to contact more your affections?

This group of items appears sufficient to answer and evaluate our constructs. Indeed, they face the problem of sharing memories from different perspectives: C1, C2, and C3 regard the intentions of the HoloLens 2 user while C4 and C5 are about the remote user ones. Nevertheless, we also wanted to explore deeper aspects of Behavioural Intentions. For this reason, we also introduced the D group, evaluated through a Likert scale ranging from 1 to 5, to capture all the nuances of the user's intentions. The following questions form it:

D1. Would you use the HoloLens 2 application to share with anyone your family photo album?

D2. Do you think this HoloLens 2 application would push you to contact more your affections?

D3. Do you think this HoloLens 2 application would push you to spend more time visualizing your photo family album?

D4. Would you use this application to visualize family photo albums of strangers?

D5. Nowadays, may this application help creating bonds between strangers?

The D-items group formed by D1, D2, and D3 reinforces the opinion regarding the role of our AR application in the revival of the family photo albums' cultural phenomena. The second group, which is composed of D4 and D5, regards instead the possible role that our design could have in socialization, inspecting the possibility of sharing such intimate material with strangers.

2.11.3 Results

We here report the results obtained for the system and the method defined in Section 2.11. As mentioned, we asked a group of ten participants to answer some additional questions regarding their experience, with an average age of 26 years and composed of 3 females and 7 males. All the collected data have undergone a reliability check to test their internal consistency and validate our research, through the widely used Cronbach's alpha index. However, Cronbach's alpha may result in low values for constructs when the tested population is equal to or less than ten items [276]. Therefore, we have also analyzed the Mean Inter-Item Correlation (MIIC), which is appropriate for our case [284]. In a range from 0 to 1, the MIIC confidence interval is 0.15 to 0.50, whereas higher values denote the item's overlap.

Construct	**PEEU**	**DLG**	**HLP**	**RP**	**HLP**
Items	**A1-A3**	**B1-B3**	**D1-D3**	**D4-D5**	**C1-C3**
α	0.73	0.81	0.69	0.56	0.71
MIIC	0.46	0.58	0.43	0.39	0.45

Table 2.20: Cronbach's α index and MIIC for the considered constructs.

As reported in Table 2.20, all scales demonstrate to be reliable concerning the MIIC measure (all MIICs > 0.15). As you can see, our analysis doesn't take into consideration the group C4-C5. This is because such questions concern very different aspects. The first one regards the application we are proposing, while the second involves family and personal aspects which are beyond the scope of this research.

In Figure 2.24 are reported the survey results about the Perceived Ease and Enjoyment of Use (PEEU) and the Deep Learning Gain (DLG) construct items. In particular, we have detailed the mean and the standard deviation for each of them. From such responses, it is evident that there is a strong agreement about the usefulness and ease of use of our application. Indeed, only the A2 item highlights a mean lower than 4. This is because some of the

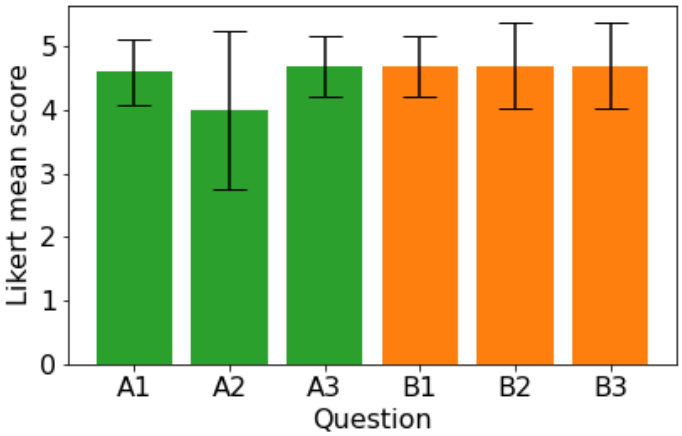


Figure 2.24: Histogram comparison of 5-point Likert questionnaire results related to the Perceived Ease and Enjoyment of Use (PEEU) and the Deep Learning Gain (DLG), reporting mean scores with error bars per question item.

respondents continue to prefer reviving their old memories physically with their affections. Surprisingly, all the questions regards the DLG construct have a mean of 4.5. This outcome was not so obvious, since the respondents are suggesting their preference for the use of modern technologies in the given application scenario.

Figure 2.25 and Figure 2.26 report the survey results for the HoloLens Perspective (HLP) and the Receiver Perspective (RP). In particular, Figure 2.25 depicts the percentage of agreement concerning the C-x items of the two groups. In contrast, the second describes the likelihood concerning the D-x ones, evaluated with the mean and the standard deviation of Likert scores.

Given the percentage of agreement on the C-x items, reported in Figure 2.25, we can infer that the considered population, from both the HoloLens 2 user and the remote perspectives, would use our AR application to contact their affections and revive together their memories, when physically distant. This is of great importance since our work could be useful to bring back to

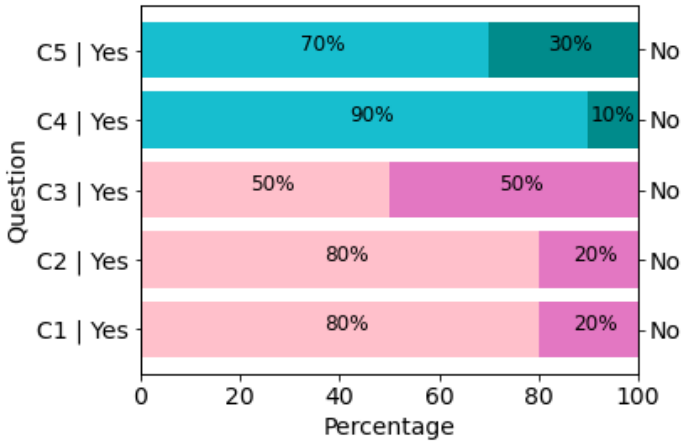


Figure 2.25: Yes/No answer percentages for the C-x items. C-x items are those related to the HoloLens Perspective (HLP) and Remote Perspective (RP), respectively colored in pink and light blue.

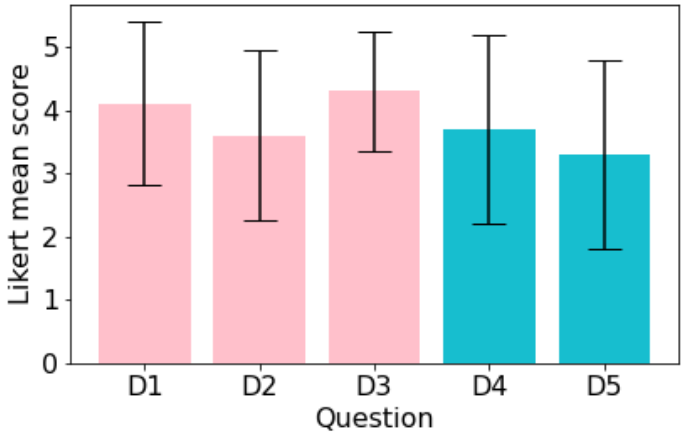


Figure 2.26: Histogram comparison of 5-point Likert questionnaire results related to D-x items which are relative to the HoloLens Perspective (HLP) and Remote Perspective (RP). In pink and light blue we report the mean scores obtained by the HLP and RP respectively, along with their standard deviations.

life the tradition of family reunions in front of family albums, even when a family is geographically spread. However, we can notice from the answer to C3, in line with the results reported for PEEU and DLG, that our respondents were equally divided when asked whether they'd prefer to live such a moment physically or virtually.

The results described in Figure 2.26 follow this trend, nevertheless, the D2 answer highlights that the system may not be sufficiently convincing to contact an affection, more than usual (this could also be linked to the participant's habits, not evaluated in our survey). Moreover, D4 and D5 scores underline that a large part of our respondents are not so comfortable regarding the sharing of such intimate materials with anyone who wants to appreciate it. Nevertheless, these answers may provide additional inspiration for future works.

2.12 Discussion and Conclusions

In this chapter, we first proposed a multimedia application to assist socio-historians in cataloging family album photos and support socio-historical research. For this purpose, the IMAGO dataset, composed of photos belonging to family albums, a source of socio-historical knowledge, was introduced. The dataset amounts to 16,642 pictures, each labeled with its socio-historical metadata: shooting year and context [53]. We then trained and tested single-input and ensemble deep-learning models to carry out those tasks, considering both CNN and ViT-based backbones. To the best of our knowledge, this is the first work addressing the socio-historical context classification. This consists of identifying the sociological and historical context of a picture, according to the definitions provided by socio-historical scholars [53].

We also debate how the integration of quantitative and qualitative methods should occur, experimenting with how quantitative methods may support qualitative ones. To this aim, we used the best-trained DL model comparing their performance with the one of a socio-historical scholar. The results

of such assessment proved that quantitative methods could not only speed up the cataloging processes but also support socio-historians in carrying out qualitative analyses of complex or large catalogs of visual information. This is only one step in the direction of exploiting quantitative models to support qualitative analysis, which may take into account all the processes involved in the complex socio-historical domain.

To so analyze how such paths could be followed, we used the developed DL models to verify the possibility of automatically discovering visual cues of intercultural influences through cross-dataset experiments, in this case considering only the dating task due to the lack of socio-historical classes for the compared datasets [136, 321]. Results from this experimental session outlined an interesting asymmetry in dating predictions, that could lead to intercultural visual influences, reflected in the latent representations. This motivated us to perform a qualitative UMAP analysis, that further supports the discovery of such visual cues (first quantitatively explored). Despite those interesting results, our cross-cultural visual cues analysis framework has a limitation: the observed domain shift effect [293] may be due to the different distribution of the considered datasets, even if such effect is partially alleviated considering that the models share similar classification tasks and that the datasets share some common visual features such as people's hairstyles, clothing, and earrings which amount to useful cues to individuate the date of an image [136, 321, 436, 367, 262]. To uncover which kind of visual features most influenced dating errors we could proceed by systematically compare the photos of the datasets based on their actual and predicted date. In other words, it could be possible to apply the IMAGO model, for example, to the [136] one, collect all the photos misclassified within a decade, and compare those photos to the ones within the IMAGO dataset that have been correctly classified within the same decade, as long as their distribution doesn't differ too much. This approach may automate the comparison of different styles across different countries at different times and be supported by the use of well-known visualization tools such as UMAP or t-SNE.

Other approaches could also be employed, such as using object detectors to identify particular objects in both the pictures that are present in the misclassified images from a dataset and in the correctly classified images of the other dataset (e.g., particular dresses, haircuts, face features, physical objects) [363, 453]. This may lead to the creation of a further layer of knowledge including those objects which most frequently appear in the presence of cross-dataset misclassifications and within-dataset correct classifications. At the same time, we could try advanced restoration deep learning models, such as the one introduced in [411], to reduce noise, picture imperfections, and non-useful cues that could improve the classification performance of the models.

All of these contributions only represent a step in the direction of creating a holistic tool that could support the socio-historic cataloging process, as many are the involved processes and sources of information. For example, in our specific case, the DL models were trained to utilize an unbalanced dataset and consider image regions that often included non-relevant information for classification purposes (e.g., background). In addition, when focusing on the socio-historical classification or dating, scholars perform analyses that resort at once to different sources of information (e.g., newspapers, magazines, archival documents), as well as to traces belonging to the same historical period. These represent three of the most relevant limits for this work. Further investigations in this domain, hence, may consider (i) larger amounts and more balanced sets of data, (ii) a better segmentation of the relevant areas of the images, and, (iii) the implementation of a multi-modal approach, capable of including also other sources of information and data formats [295]. Regarding the first point, the availability of larger datasets could improve the model's discriminative power, reducing possible unbalanced problems [35]. Regarding the second, the use of segmentation models may benefit the individuation of more relevant regions [194]. Finally, multi-modal learning appears as the approach that may best replicate the comprehensive approach normally adopted by socio-historians during cataloging processes. Indeed,

exploiting knowledge from historical archival documents (and other sources) could improve the general cataloging and analysis effort. This could also define novel methods to discover new latent representations, to cluster social history from a spatiotemporally perspective. For example, knowing how people dressed during a specific period might improve the classification for both the socio-historical context and dating. Such a path, although complex, may not be impossible to follow. Indeed, recent natural language processing solutions can provide discriminative features that could be exploited in our models to improve the overall performance [91, 143].

Despite those mentioned future works, the general performance of our trained models motivated us to define an AR system to catalog family album photos, bridging the revival of the biggest family traditions, i.e., family photo album exploration. The AR system exploits those models to catalog and classify pictures in the user's view by date and socio-historical contexts. Those classifications are then used to augment a given HoloLens 2 user's view, which is streamed on the web, allowing users to access shared photo albums from any kind of device. The system has been assessed with the interview of 10 users who found the interface easy to use and who provided enthusiastic feedback regarding the proposed experience. Based also on the users' comments, we could individuate possible future directions of work. In particular, we aim to include an active collaboration between HoloLens 2 users and remote ones, letting them synchronously manipulate the augmented and shared view, through any kind of non-AR device (e.g., smartphone, PC) and AR devices (e.g., HoloLens 2). Such kind of manipulation amounts to provide: (a) multimedia data annotation capabilities, and, (b) affine transformations such as moving, flipping, rotating, etc.

In conclusion, this chapter introduced four different frameworks that exploit both AI and XR to support socio-historical research (RQ-1), focusing on how the combination of those technologies could be used to increase both the efficacy and efficiency of family album photo cataloging and information retrieval tasks for both experts and non-experts (RQ-2).

Chapter 3

eXtended Reality and Artificial Intelligence in the Creative Industries

This chapter aims to analyze how paradigms from both XR and AI could support creative research and industry [368, 251, 361]. To mitigate the potential for misinterpretations, we here define the difference between CH and Creative Industries and Research (CIR). Despite CH and CIR often being used together, the former is related to traditional forms of preservation and creation of human heritage, while the latter includes the applied practices and innovations for generating a tangible or intangible item, that can be also related to profit and subjected to intellectual property [81, 271, 57, 300]. It is worth noticing how those two have however a strong connection, enabling reciprocal benefits [152].

Considering CIR, starting from a literature review, we examined how XR and AI technologies are adopted in a particular field of application: fashion [106, 288, 254]. Many works envisioned and designed systems to jointly them for different fashion tasks, such as immersive fashion shopping, gar-

ment retrieval, collaborative garment design, virtual-try ons, and intelligent fashion assistant [68, 288, 162, 254]. However, the systems using such technologies and the works that analyzed their possible adoption and effects are still in their early stage [288, 311]. Despite the expectations, one of the most unexplored and challenging fashion use-cases amounts to immersive and intelligent retail experiences [254, 288, 311, 85]. Into the Metaverse, consumers could dive into experiences that break the classical bounds of a fashion shop, both spatially and temporally [254, 288, 311, 85]. It is, however, yet unclear which factors and technologies could improve the immersiveness and usability of such systems [190] (RQ-2). In such a context, some related works stated that Intelligent Vocal Assistants (VAs) could play a pivotal role: the definition of easier-to-use systems while guaranteeing interaction that resembles the one provided by a fashion shop clerk [355, 254, 288]. On such a line, another interesting research question involves the usage of Human Digital Twins (HDT), which could model human behaviors in the Metaverse, defining a new way to perform actions and collaborative interactions (RQ-2). Considering such aspects, this chapter contributes to analyzing the acceptance of such technologies and their role in what we defined as x-commerce: e-commerce projected in an immersive environment, through XR paradigms.

Following the line of x-commerce scenarios, another impactful technology amounts to Generative AI [254]. Despite the fast-paced advancement in this field, there is a lack of works that include a robust Human evaluation framework for those tools, considering two crucial factors in visual industries like fashion: emotional and aesthetic [348]. This gap in the literature is even greater considering immersive content, like 3D objects or 360° videos. For this reason, we here resort to XR paradigms to scratch a new evaluation pipeline starting from a pivotal case study in the creative research for emotional and aesthetics evaluation: an immersive musical concert [265, 82]. This item was chosen because it is one of the few experiences that can be lived and experimented with in different degrees of reality (RQ-3). It also has a strong connection with immersive fashion catwalks, which are approaching maturity

in showcasing and selling automatically designed fashion garments. [153, 254, 417].

The rest of the chapter is organized as follows. Section 3.2 and 3.3 set the theoretical background for the considered composition of XR and AI in the creative industry. Section 3.4 provides the analysis, implementation, and results of injecting an intelligent vocal assistant in an x-commerce experience. Following the results obtained in this analysis, Section 3.5 explores the possible adoption of human digital twins in commerce environments. Section 3.6 introduces a new framework to inject human aesthetic and emotional evaluation of immersive content through XR paradigms, which possible adoption is validated with a user study. Finally, Section 3.7 concludes this chapter, providing new directions for future works.

I here declare that the content of this chapter is entirely based on my work and contains nothing that is the outcome of work done in collaboration with others and that all the contributions reported here have been partially published in [368, 251, 335, 361].

3.1 Research Questions

Considering the different aspects analyzed in the previous Sections, this chapter aims to answer the following research questions:

RQ-2 - Can Artificial Intelligence improve user interactions and task completion efficacy in eXtended Reality systems?

RQ-3 - Which kind of eXtended Reality and Artificial Intelligence technologies could be applied in the creative industry?

3.2 Introduction

The perception that XR applications are limited to the gaming sphere is disappearing, as their possible range of applications is rapidly expanding

[168, 259, 353]. From a managerial point of view, it is important to highlight that the amount of funding poured into such technologies is booming: the consumer and industrial market spending is estimated to grow by +69% and +134% in the next three years, respectively. While their cost is falling, their performance is rising: a widespread adoption is expected soon, along with the new habits and routines they will shape in people's lives [410].

As such technologies mature, it is also interesting to observe that the traditional separation imposed by hardware devices between AR and VR is fading. Indeed, more and more head-mounted displays, starting from those based on mobile phones up to those engineered for cutting-edge applications such as the Varjo XR-3 [177], are converging towards tools capable of implementing all types of XR scenarios, opening the path to an increasingly mixed world. This phenomenon was intensively applied in the fashion world, where also AI paradigms play a crucial role [281, 246, 288, 254].

In the fashion industry, the digital transformation process is pushing online branding and the adoption of e-commerce platforms [281]. To this date is worth 332 billion USD, representing 28% of all e-commerce transactions. Such value will likely increase [95]: mobile commerce (m-commerce) is sustaining such growth in the US for a share that exceeds 50% of the retail clothing market. It is also worth noticing that a large increase in investments dedicated to XR and AI technologies in the world of fashion is expected by 2030 [115]. It follows that we could hypothesize a future vast joint application of XR and AI paradigms for the benefit of e-commerce [229, 219, 172, 254].

However, despite AI technologies being already put to good use in this context for several applications (e.g., visual search, and chatbot assistance) [219, 254], the use of XR technologies in fashion scenarios is currently limited to exclusive and sporadic shows, aiming at achieving a "wow-effect" capable of attracting customers and/or enhancing brand reputation [254, 288, 85]. The reason is a combination of (a) lack of maturity of XR technologies, (b) low adoption among consumers, (c) low understanding of their use by

non-specialists, be they consumers or businesses, and (d) lack of intelligent methods injected in this experiences to improve user experience [250, 85]. Not too long ago, in the early 2000s, a similar situation took place: the e-commerce market was, in fact, in its primordial stage [250, 251]. At that time, authors of [400], investigated by which means it could grow with the formulation of the following research questions:

1. "What do consumers want?";
2. "Would consumers adopt it?";
3. "How would consumers behave?";
4. "What capabilities are required to make e-commerce viable?".

Two decades later, the same guidelines may be adopted when approaching a next-generation digital commerce system.

For question #1, let us isolate one of the aspects that may influence the shopping experience of a customer, making him/her feel at ease. The scientific literature has found in the relationship built with a professional salesperson an aspect which may improve the loyalty and the satisfaction of a customer entering a retail clothing shop [138, 233]. In terms of interaction, such a relationship entails the exchange of visual and verbal information. In terms of products, a novel selling form is defined by highly customizable fashion items by Generative Content Models (GCMs) like Generative Adversarial Network (GAN) and Diffusion Models (DM) in the form of NFTs [447, 313, 127, 361, 412, 254, 154, 17].

Building upon this foundation, addressing question #2 first involves supplying the visual and auditory interactions essential for an advanced digital commerce system. This can be achieved by leveraging emerging technologies like VR and AI-driven Voice Assistants (VAs), essentially encompassing x-commerce systems. Recent projections suggest that in the United States, by the conclusion of 2021, approximately 57.4 million individuals will engage

with VR, and nearly 120 million will be utilizing VAs [395, 405]. Simultaneously, consumers express a willingness to invest in products generated by GCMs if the functional and emotional value closely mirrors that of real products, which are however hard to measure [348, 181]. On this line [348] highlights how emotional factors play a pivotal role in users' intention to buy GAN-generated items in e-commerce environments. However, no standardized scale for the aim has been validated, to the best of our knowledge.

We can now concentrate on how consumers may accept x-commerce (question #3). Interestingly, preliminary works show promising results when integrating VR and voice in a simple gaming environment [416]. Nevertheless, the potential of embedding an off-the-shelf voice assistant in a more complex retail x-commerce scenario has not been investigated to this date, to the best of our knowledge. At the same time, investigating the usage of Human Digital Twins (HDT) paradigms to empower such intelligent assistants or even virtual users, is a pathway that can be further explored for x-commerce [389, 368]. The concept of Digital Twins (DTs) first appeared in [130] and nowadays emerged as components in many types of projects, as they may help reduce the expenditures caused by unexpected disruptions [26]: a DT platform may show how given options would work in specific contexts of use. Human Digital Twins (HDT) is a subset of DT defined as capable of modeling the inner and outer self of a person and of assisting in everyday tasks [389]. While this is a fascinating perspective, there is little understanding of how HDTs could be socially accepted in virtual environments. On this line, further exploration of the role of HDT and how those are perceived by others within an x-Commerce environment is required. The same phenomenon was highlighted also for GCMs immersive content, for which no robust and standardized scales for their human evaluation were introduced.

In this chapter, we narrowed down this vast research space to three use cases that could partially answer all the aforementioned research questions. First, we designed and implemented a VR-based shopping experience for fashion retail, assessing the benefits of integrating an off-the-shelf VA: Amazon

Alexa [1]. This experience was designed to provide detailed insights regarding the technical solution along with a thorough assessment, given the possible use of VR-Voice Assistant (VR-VA) integrated solutions in x-commerce. We hence sought clues that could help answer an updated version of question #4: "Would an integrated VR-VA environment make fashion retail x-commerce viable?". To this aim, we approached the problem from two different perspectives, to understand: (a) the usability of the proposed VR-VA environment, and (b) its feasibility for a fashion retail x-commerce application. Such questions have been answered, by interviewing a special group of users, fashion experts, who fully tested the proposed x-commerce experience [251].

Based on the obtained results, particularly related to immersiveness and unnatural interaction with the VA, we considered HDTs carrying out a preliminary study to analyze their impact and social acceptance for future applications in virtual fashion shops. To this date, we took as a use-case a brick-and-mortar clothes shop where a clerk can receive requests from HDTs appearing on a screen. We hence created an avatar 3D-scanning a real person using the Artec Eva 3D scanner [13] and evaluated the impact of having this avatar ask a specific question to a human as a real customer visiting a fashion shop. Resorting to a group of users, we performed a user study to verify how such paradigms may be received in an x-commerce setting, as defined in [251].

Finally, we investigated the possible acceptance and integration of GCM immersive outcomes into x-Commerce environments [181], analyzing one of the key aspects behind it: human aesthetic and emotional evaluation, which demonstrated to be crucial for the acceptance of GCM virtual items in e-commerce environments [348, 126, 361]. An evaluation framework that implements such an aspect would only be possible with humans as the pivot in its flow, defining an urgency for novel forms of human-in-the-loop (HITL) in GCM training and evaluation pipelines [270, 184, 391, 348, 250, 251, 361, 439]. Extending such perspective to digital content that goes beyond 2D

[1]The code is available at: `https://github.com/Cancercookie/AVR-Thesis-LRPVR`

multimedia data, such as 3D models/scenes, and 360° panorama images and videos, it becomes clear that conventional display and interaction hardware may not be the most effective means for human evaluation. This was highlighted by previous research studies, where XR was proposed as a possible solution [223, 391, 30]. Considering this, we here introduce a preliminary framework that took into consideration all the mentioned aspects, taking as a specific use case 360° panorama musical videos experienced through a VR device [223, 391, 30], where the aesthetical and emotional factors were scored through the validated Aesthemos scale [330]. This particular item was selected as a pivotal example considering that is: (a) one of the most emotional and aesthetics activators [265, 82]; (b) it is one of the few that could be modulated in different degrees of realities; (c) it has a strong connection with fashion experiences like catwalks, nowadays approaching their maturity to show and sell automatic designed fashion garments [153, 254, 417].

Summing up, the main contributions of this chapter amount to:

- Introducing a VR environment integrating an off-the-shelf voice assistant technology (i.e., Alexa), thought for personal scenarios, in a fashion x-commerce to improve user shopping experience and action completion, while wearing a full-immersive head-mounted display. We then performed a mixed quantitative and qualitative analysis to measure the effects of such vocal assistance considering a group of 30 non-tech-savvy fashion domain experts. In particular, we made a comparison not only in terms of well-known usability and technology acceptance scales but also in terms of the perceived gain provided by the intelligent vocal assistant. We so compared this experience with a voice assistant-free fashion x-commerce environment. The obtained results provided the first empirical proof of the usefulness of intelligent vocal assistants in fashion immersive environments;

- Built upon the previous results, analyzing the criticalities that emerged with the Vocal assistant, like the unnatural voice and visual aspects, we

analyzed whether Human Digital Twins paradigms could be a solution. In particular, we verified if HDT could be positively applied to empower the naturalness of vocal assistance and at the same time, define self-shopping assistants. To this date, we considered a particular use case in the fashion world: a brick-and-mortar shop as a use case. In particular, we set an experiment where the HDT (built upon human 3D scanning, lips animation, and vocal synthesis) acts as a consumer who interacts with a real human store clerk asking to buy several fashion garments. In such a peculiar use case, we provided one of the first evidence on which factors should be considered to improve the realism and immersiveness of human avatars in metaverse-everyday experiences. At the same time, we set the stage for the acceptance of autonomous human digital twins that act as ourselves in the metaverse, for tasks like shopping. The experimental results highlighted the fundamental role of the voice and the positive acceptance of the aesthetic of the considered 3D-scanned human;

- Considering Aesthetics and Emotions, pivotal factors in the Creative Industry, we investigated the possible acceptance and integration of immersive (generated) content into immersive environments, like x-Commerce. To this date, we designed a human-in-the-loop framework that puts humans at its center, for content evaluation. This framework includes the usage of a well-known Aesthetical and Emotional evaluation scale, the Aesthemos. However, this evaluation framework could be less effective in case conventional display and interaction hardware were adopted. We so applied such a framework to verify whether XR immersive visualizations could be a solution to overcome the aesthetic perception limits of classical displays. To answer such a question, we considered a pivotal immersive multimedia experience for aesthetical and emotional perception: a musical concert, provided as a 360° video. To have a term of comparison, we execute an intensive user study considering different populations, that lived such an experience in the real

world, with full-immersive virtual reality, and classical displays. The statistical analysis highlighted how the experience lived through the HCT-Vive headset is the most aesthetically and emotionally similar to the live one. This is one of the first empirical evidence regarding the greater efficacy of VR in evaluating aesthetics and emotional components, rather than a classical 2D display.

In the following, we present the most relevant works related to the area of research so far described.

3.3 Related Works

3.3.1 An X-commerce VR Environment Integrated - With An Intelligent Vocal Assistant

The use of XR technologies has inspired many different initiatives and possible application scenarios in the fashion field, in both industrial [387, 97, 74, 101, 175] and academic [278, 427, 344, 78, 398, 277, 303] contexts. In the XR-AI-speech interaction domain, recent academic researchers have analyzed how verbal commands could be positively embedded in immersive experiences to facilitate the approach to new devices, making the interaction more natural [54, 287, 116, 11, 414, 449].

To the best of our knowledge, only two works have so far created and analyzed a VR shopping environment controlled by speech inputs [355, 354]. In such works, the authors described the design and implementation of a mobile VR shopping where product search was based on speech inputs (such inputs were processed at the word level using Google's speech recognition web service). In the evaluation study, the authors investigated the task of searching for a product in a VR web store with different combinations of hands-free input (pointing versus speech) and output (desktop versus VR) types. The combination of speech inputs and VR outputs resulted in working best in terms of user performance and preference.

Now, it is noteworthy that a new thrust to novel XR-based fashion initiatives may come from the growing popularity of verbal interaction services. Google Assistant (Android), Cortana (Windows), and Siri (Apple) provide useful and easy-to-use AI-based VA on smartphones (e.g., make a call, convert vocal messages to text) [98]. It is also worth noticing that, these services will soon be revolutionized by the integration of the more expressive power of Large pre-trained and Large Language Models [140].

3.3.2 Who Will Trust Human Digital Twins in Extended Reality?

While analyzing whether vocal assistants could be adopted in virtual fashion scenarios to guide users in their actions (Section 3.4) some criticalities emerged. Among them, the unnatural interactions users had with the considered VA and their visual aspects were bolded as key aspects [102, 427]. This particular aspect was worth additional investigation and provided us a starting line to examine which factors should be considered to improve the realism and immersiveness of intelligent virtual assistants but also human avatars in metaverse-everyday experiences. This line of work also set the stage to a broader analysis of factors enabling new digital avatar roles, that go beyond virtual assistants: a new digital entity system certification, allowing our avatar to autonomously act in our place within Metaverse experiences [423].

In such a scenario, the digitization of ourselves will integrate our behavior and thoughts, linked to our digital identity for data safeness and actions responsibility [423]. Through Human Digital Paradigms (HDT) we could envision a future where our digital self's interactions are not made directly by us, but by an intelligent agent that mimics our behavior [389, 263, 351]. HDTs are composed of mathematical models, all supported by a data lake of human personal data, integrated with IoT infrastructure, wearables, human visual modeling, explainable AI, and data visualization techniques. However, as stated by [423], one of the most important aspects of accomplishing our

self-avatar domination amounts to emotion projection, which can be analyzed as a multi-modal AI generation task, which however lacks validation settings [429].

In such a vast context research space, we here attempt to put the first brick in verifying the HDT acceptance factors considering a particular consumer task in the world of fashion: automatic shopping [65]. Considering the capability of HDTs to resemble the inner and outer self of a person, we could envision a Metaverse where our self HDT goes doing shopping for us, as it was our personalized fashion assistant [389, 351]. We analyzed which kind of factor could improve the acceptance, from both aesthetic and emotional perspective, of such an HDT-based fashion assistant. In particular, we analyzed the factor that was most criticized for the realism and the acceptance of the AI vocal assistant in our first defined digital environment (Section 3.4): the voice.

Following the brick-and-mortar shop analogy, the HDT covers the role of a consumer who interacts with a real human but instead covers the role of the store clerk. The former asks for information and buying methods for several fashion garments. Despite such a use-case being very specific to the world of fashion, it could be used as a pivotal example to understand which are the factors and behaviors that could lead to the definition and acceptance of HDT interactions. Moreover, the outcomes of such an analysis will be fundamental to increase the realism of AI-empowered VAs.

To the best of our knowledge, none of the previous related work has adopted HDT paradigms from such a perspective.

3.3.3 Evaluating Generated Immersive Content with eXtended Reality

In the context of digital e-commerce platforms, GCMs are nowadays applied for several purposes, among which their deployment in immersive retail experiences [324, 84]. Generative AI can, for example, enable novel forms of personalization, allowing sellers to instantly tailor their offerings to meet the

preferences of each user [417]. GCM based on GANs and Diffusion models were employed to develop fashion e-commerce assistive technology, for example, to generate garments from dressed people [441], and stylizing 3D meshes by text [430].

However, the current literature of GCMs evaluation lacks standardized scales and metrics for constructs beyond traditional quantitative metrics [270, 184, 391], like emotions and aesthetics, which connect high-level semantics to low-level computable visual features, useful to develop reliable systems for assessing digital multimedia, particularly in creative fields like art, and music. Those factors demonstrated to be key in virtual commerce experiences [184, 348, 250, 126, 251, 361, 269]. For example, [348] correlates the evaluations of functional value, emotional value and willingness to pay, highlighting a strong correlation while considering GAN generated items. To this date, we analyzed existing scales in the literature to capture those factors, and among many, the Aesthetic Emotions Scale (Aesthemos) emerged as one of the most adopted for visual media stimuli [231, 330, 335], since it was designed to evaluate the emotional responses to perceived aesthetic appeal across various domains.

Extending our perspective to encompass digital content that goes beyond traditional 2D media such as immersive content, conventional display hardware was highlighted as not effective visualization means for human evaluation [223, 391, 30]. A first solution to such a problem could consist of employing XR paradigms and devices to find better and more effective ways to present such visual/auditory stimuli to humans [30]. On this line, Spacedesign is one of the first frameworks introducing a mixed virtual environment to evaluate the aesthetic of 3D models [119], focusing on free-form curves and surfaces. The authors reported that the introduced system has been tested by experienced industrial designers who appreciated the 3D visualization and navigation, real-time editing, and intuitive interaction. In this scenario, designers and stylists play a significant role in the development process, enabling them to effectively steer their vision from start to

finish, resulting in the final product. Authors of [182] proposed using VR paradigms to emphasize aesthetic and emotional abilities in students for 3D design. Through a statistical assessment process, the authors reported that VR could have a positive impact on creative thinking while interacting with such multimedia data. On a similar line, some works have shown that VR can be effectively used while evaluating the aesthetic or the emotional effect of 3D scenes or 360° videos [391, 335]. For example, [69] studied the impact of VR in awe emotions. They considered 360° immersive videos, examining 42 participants who watched immersive and 2D videos displaying awe or neutral content, rating each their level of awe and sense of presence after the experiment. Results indicated that immersive videos significantly enhanced the self-reported intensity of awe as well as the sense of presence.

Concerning the current literature, we considered both aesthetic and emotional evaluation of immersive items by adopting XR paradigms, to overcome the limits of 2D displays.

3.4 An X-commerce VR Environment Integrated With An Intelligent Vocal Assistant

In this Section, we report the implementation of two different virtual experiences, to compare the utility of adopting an off-the-shelf voice assistant to foster interaction with a digital sales assistant [251]. We then provide a thorough description of the experimental settings we adopted along with the obtained results. To the best of our knowledge, the VR system proposed in this work is the first to integrate the use of an off-the-shelf voice assistant to foster interaction with a digital sales assistant. Concerning the state-of-art literature, in summary, it offers the following elements of distinction: (a) the use of a home run HMD device (i.e., HTC Vive headset), (b) the support of voice commands with Amazon Alexa, (c) the evaluation by a group of

non-tech-savvy fashion domain experts.

3.4.1 Experience Design

To quantify the users' satisfaction with x-commerce for fashion retail and the possible benefits that may derive from the integration of voice-based interaction into the VR environment, we built two different immersive experiences. Both may be considered prototypes of fashion x-commerce services and forerunners of a virtual shop to provide users with different experiences [244].

In the initial scenario, a user transitions to a small desert island, symbolizing solitude and seclusion, where they encounter a virtual dressing room. By manipulating hand controllers, users can choose and experiment with various clothing and accessories while exploring the surroundings. In the second scenario, users find themselves immersed in a traditional fashion shop setting, fostering dialogues with other patrons. Similar to the first scenario, users can try on fashion items, but in this case, they can engage with a sales assistant avatar using vocal commands. The sales assistant, in particular, can respond to user queries and translate spoken words into virtual actions. This mirrors the experience of seeking assistance or clarification from staff in a physical store. The integration of a voice assistant streamlines the interaction between the user and the VR environment, reinstating one of the most natural forms of human communication—voice. This feature also holds the potential to alleviate feelings of isolation and loneliness induced by the virtual realm.

To further elucidate the rationale behind our selections, we opted for the desert island scenario due to its stark contrast with the conventional VR shop environment. By choosing this setting, we aimed to underscore the distinctions between a remote and solitary setting versus a "typical" shop, embodied by the Virtual Store incorporating Alexa's vocal commands. In the latter, both the physical location and social interactions closely mimic real-world situations. This aspect holds significance for psychology and marketing research, as demonstrated by [193], who found that retail environments can

alleviate loneliness, attracting specific consumer demographics to spend more time and money. In essence, by having our participants undergo these two experiences sequentially, we sought to observe the impact of such variations on their responses. More detailed presentations of the two applications are reported in the following.

3.4.2 Fashion Island Application

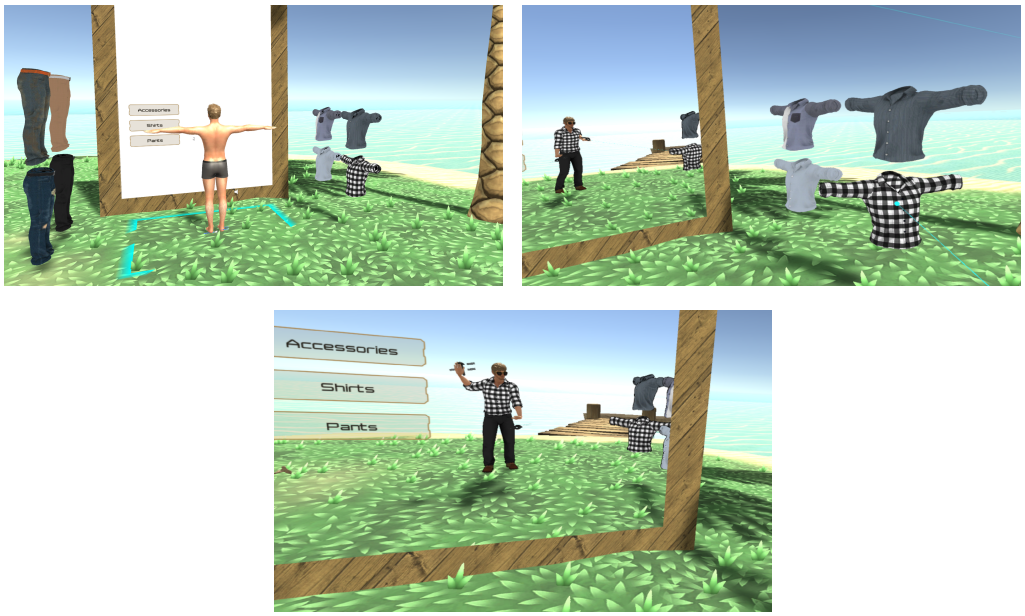


Figure 3.1: Some frames of the "Fashion Island" environment. At the top, is the third-person view of the user; at the center, the avatar explores the available shirts from the user's view; at the bottom, the avatar appraises the selected outfit in the mirror.

Our baseline application amounts to "Fashion Island", which may here be considered an introductory level VR-based fitting room [99]. As shown in Figure 3.1, a pre-modeled male avatar is placed on a desert island, barely dressed in front of a mirror. The game consists of dressing him up, by exploring and selecting the available shirts, pants, and accessories. With one hand controller, the user can teleport in the virtual environment, while the

other is used to select and wear clothing items. A simple graphical interface is placed in a fixed position on a large mirror, located at the center of the scene. The avatar can move his hands, arms, and head according to the HTC Vive tracking system and watch his movements in the mirror: this is implemented to improve the sense of immersion and presence. In essence, a user can somewhat experience a playful environment that, however, has been deprived of any social experience, as the user (avatar) is "abandoned" by him/herself on the desert island.

3.4.3 A Voice Assisted Fashion Store: Virtual Store

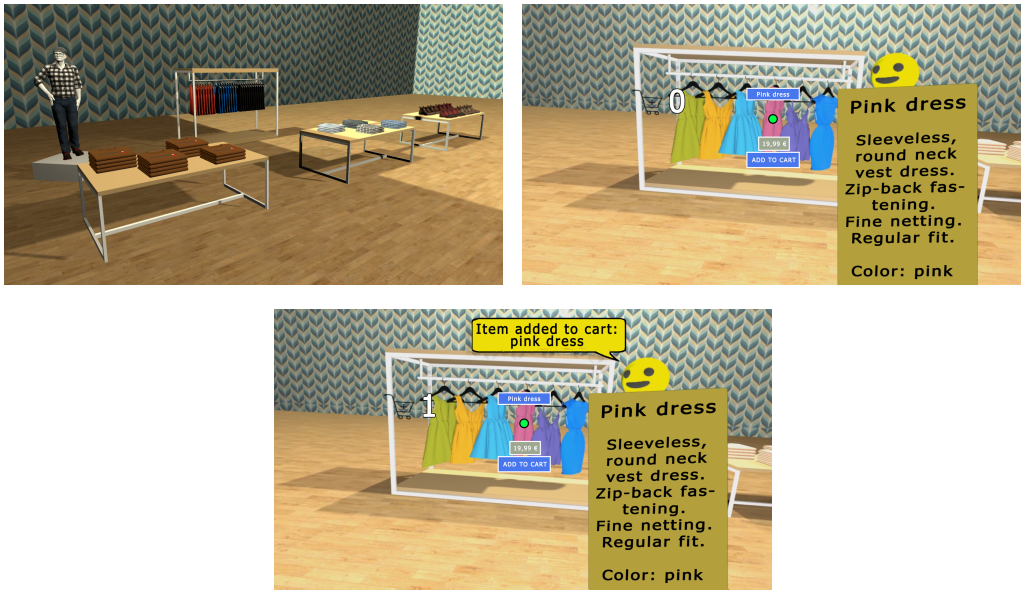


Figure 3.2: Some frames of the Virtual Store application. Top: the interior of the male sector. Center: pointing to a dress displays its price and the information that was previously asked of the shopping assistant. Bottom: the user add-to-cart action exploiting Alexa.

The second test application is called "Virtual Store", as it proposes a shopping experience in a fashion store where both male and female clothes are available for (virtual) shopping. The top image in Figure 3.2 shows the environment, designed with modern-style elements. A straightforward avatar,

depicted as a friendly and smiling emoticon, serves as the user's shopping assistant. This consistently appears within the user's field of vision, accompanied by the cart icon, forming an essential and unobtrusive interface. Leveraging integration with Amazon Alexa, the assistant adeptly interprets the customer's verbal commands by utilizing Alexa Skills for speech recognition and natural language understanding. Consequently, the assistant can guide the user in navigating the shopping platform and furnish information about selected items. This information is presented on a supporting panel (top-right image in Figure 3.2) and audibly conveyed by Alexa. Via vocal commands, the user can also add a product to the cart (bottom picture in Figure 3.2), control the cart itself, and finalize the purchase of the selected products. All of such functionalities may be activated also by hand controller commands; only navigation is exclusively managed via controllers.

3.4.4 VR-VA Architectural Framework

The virtual assistant incorporates the Amazon Echo Input speaker, recognized as one of the most cost-effective smart speakers available. Our decision to utilize this particular device stems from its widespread use, popularity, and ease of development within Amazon's supported framework [7]. Amazon offers device manufacturers free access to the Alexa Voice Service (AVS) [2], a cloud-based service providing APIs for exploiting Alexa's automatic speech recognition and natural language understanding, enabling voice recognition and vocal synthesis.

Our choice was to harness the Echo device and capitalize on the capabilities of Alexa Skills. These Skills, essentially specific apps for Alexa, make optimal use of the Amazon Web Services (AWS) [3] platform, serving as an infrastructure provider where services are accessible through REST API calls. Hosted and executed in the cloud directly within AWS, Skills adopts a server-

[2] https://developer.amazon.com/en-US/docs/alexa/alexa-voice-service/get-started-with-alexa-voice-service.html

[3] https://aws.amazon.com/?nc1=h_ls

less approach, utilizing abstract platforms with costs ranging according to usage. This approach ensures that all logic remains online, eliminating the need to manage computing infrastructure. As previously mentioned, AWS furnishes the complete stack of technologies necessary for the effectiveness of our Skills: (i) DynamoDB ensures data persistence within tables; (ii) Lambda executes the business logic in the cloud; and (iii) the NodeJS API Gateway facilitates the use of web sockets for establishing communication between endpoints.

In essence, Alexa natively generates a conversational model and detects the requests from the user's voice command, to execute the right branch of code. It is important to note that such an approach imposes some constraints, dictated by the design guidelines of Alexa's Skills, which may interfere with an application's flow. The first one is the impossibility of leaving Alexa on hold: if a user stays silent for over 30 seconds, the dialog session expires. In such cases, a new session can start by saying out loud one of Alexa's wake words ("Alexa" or "Computer"). Another restriction regards the case in which Alexa asks a question, and no answer is provided within 8 seconds: in this case, Alexa repeats the question only once, then the session expires. Furthermore, it is important to note that the Alexa account utilized in our setup has been configured with a male voice. This decision was influenced by our observations of recognition uncertainties experienced with a subset of female speakers, despite all experiments taking place in an open room with a noisy background. It is worth mentioning that this might not pose a significant issue since assistants are typically used by their respective account holders.

Finally, despite the limited customization opportunities set off by a proprietary solution, we preferred a proprietary voice assistant like Alexa over open-source ones, like Snips [347] or Mycroft [260], for the following reasons: even if the alternatives allow for greater freedom in designing the user experience, they still have some disadvantages, such as a lower adoption rate, a higher development time or even a lack of some components of the assistant

(e.g., the voice synthesis).

3.4.5 Participants

We recruited thirty-one subjects with at least one year of experience, with different roles, in the fashion field, including (a) researchers from the Fashion faculty at the University of Bologna, (b) Master students from the Design and Technology for Fashion Communication Master's program at the University of Bologna, (c) Bachelor students from the Fashion Cultures and Practices program and Master students from the Fashion Studies program, offered at the University of Bologna, and, (d) professionals (including product developers, photographers, and clerks) working in the fashion field. They tested the applications and answered a survey where they reported their opinions about their perceived comfort and XR usability. The considered population had an average age of 33 years and was composed of 25 female and 6 male students. Most of the participants declared themselves to be non-tech savvy, not having any expertise in 3D-model desktop applications, nor playing video games at all. None of them had tried the HTC-Vive device before, their opinions were hence well suited to evaluate the ease of use of our interfaces and experiences. The number of participants amounts to a trade-off between the necessity of acquiring sufficient feedback data from a specific population of participants (i.e., fashion experts) and the time spent for the evaluation phase. This number is higher than 10, which has repeatedly proven to be sufficient to discover over 80% of existing interface design problems [323, 173, 117].

3.4.6 Ethics & Apparatus

Written consent to participate in this experimental study was collected from each subject, in a controlled environment. The components adopted in the proposed system are:

(a) A head-mounted display, namely an HTC Vive headset. At the time of the experiments, this specific device slightly outperformed the Oculus

Rift in terms of perceived ease of use and perceived intuitiveness [378];

(b) A Dell Alienware Area 51;

(c) An Amazon Echo Input speaker powered by Alexa, the AI-based voice assistant developed by Amazon.

3.4.7 Assessment Model

Individuals' acceptance of technological innovation, such as x-commerce, may be analyzed through different theoretical approaches, already established in the literature, such as the notable Technology Acceptance Model (TAM) [88], which is based on two main key points, namely, perceived usefulness and perceived ease of use (PEEU). Perceived usefulness refers to the extent to which individuals believe that incorporating a specific technology will enhance their work performance, while perceived ease of use gauges the degree to which an individual anticipates that interacting with a particular technology will be user-friendly. As the concept of perceived ease of use encompasses both physical and mental efforts, our research endeavors to examine the influence of virtual interfaces on consumers' overall satisfaction. On such a line, we also designed an additional scale inspired by TAM, to evaluate the gain perceived by users while using the vocal assistant to control the system (Voice Gain).

In user-centered design and good practice in user interface design have already developed their international standards and general principles in well-established computing systems [32], the ease of use of XR devices is still a challenging issue [422, 255]. Above all, immersive VR devices may result in difficulties for non-expert users [99]. Nevertheless, more usable virtual content and scenes may positively affect consumers' perceptions and attitudes toward VR technologies. We have hence stated the following hypotheses:

H1. Fashion experts appreciate immersive VR interfaces;

H2. The VR headset and controllers, together with the voice command integration, provide a positive immersive experience.

Shifting our focus to the perceived usefulness of virtual experiences, we broadened our investigation to assess the viability of extended reality tools in the realm of fashion. As previously detailed in Section 3.3, x-commerce holds promise in bolstering conventional marketing approaches, be it through online platforms or physical retail outlets. Beyond facilitating user engagement, XR environments have the potential to enhance 3D product displays, addressing consumer skepticism arising from limited interactions with items and assistants. Therefore, we also considered the following hypotheses by distinguishing two applicability levels:

H3. XR represents an appealing channel for fashion communication;

H4. XR applications can be adopted to buy fashion products.

To answer such hypotheses, we designed two additional scales to evaluate named as Attitude Towards Using for Communication (ATUC), and Behavioural Intention (BI), formulated according to [309].

3.4.8 Assessment Survey

Once our participants tested both experiences, they were asked to complete a questionnaire. This has been designed to assess four constructs, namely, Perceived Ease and Enjoyment of Use, Voice Gain (VG), Attitude Towards Using for Communication (ATUC), and Behavioural Intention (BI), formulated according to [309].

3.4.8.1 Perceived Ease and Enjoyment of Use (PEEU) and Voice Gain (VG) constructs

The first set of questions aimed at investigating the PEEU (for both experiences) and the VG constructs (for the Virtual Store). For simplicity, a general overview of these constructs is reported in Table 3.1.

	Question	Evaluated on	Evaluation Method
PEEU	(I1) The interface is easy to use	Both VR experiences	5 point Likert scale
	(I2) Once I learned how to use the interface, it was simple to manipulate objects	Both VR experiences	5 point Likert scale
	(I3) I prefer the new interface to the mouse or keyboard	Both VR experiences	5 point Likert scale
	(I4) I enjoyed the overall experience	Both VR experiences	5 point Likert scale
VG	(A1) The introduction of voice commands makes the system easier to use	Virtual store	5 point Likert scale
	(A2) I would prefer to complete all tasks with voice commands and have no hand controllers at all	Virtual Store	5 point Likert scale

Table 3.1: Items and questions used in the survey to assess the Perceived Ease and Enjoyment of Use (PEEU) and Voice Gain (VG) constructs.

For what concerns the PEEU, a 5-point Likert scale [210] was used to quantify respondents' agreement to the following items, constructed based on those used by [309]:

I1. The interface is easy to use;

I2. Once I learned how to use the interface, it was simple to manipulate objects;

I3. I prefer the new interface to the mouse or keyboard;

I4. I enjoy the overall experience.

The I1 sentence immediately gets to the point; item I2 would result to be crucial in case I1 achieved a low score. On the contrary, in case I1 received a high agreement, question I3 would allow checking the users' willingness to adopt the proposed technologies, changing their habits. Through I4, we ask for a broad evaluation of each VR application. In this first Section, we also included some open questions to assess issues related to motion sickness.

Asking the users to evaluate the interfaces separately, we aimed at understanding whether the voice-command integration indeed facilitated their approach to a new device, hence a comparison of the average evaluations was considered into the VG construct. The VG construct contains the following specific sentences (evaluated through a 5 point Likert scale):

A1. The introduction of voice commands makes the system easier to use;

A2. I would prefer to complete all tasks with voice commands and have no hand controllers at all.

3.4.8.2 Attitude Towards Using for Communication (ATUC) and Behavioural Intention (BI) constructs

The second set of questions are meant to assess the ATUC and BI constructs. They exploit the participants' knowledge and studies to analyze the actual potential and feasibility of the whole XR-based fashion e-retailer. Such items have been inspired by previous studies and adapted to suit our case [309, 192]. These constructs were evaluated only on the Virtual store application. A general overview is reported in Table 3.2.

	Question	Evaluated on	Evaluation Method
ATUC	(F1) I would visit an XR retail store	Virtual Store	5 point Likert scale
	(Q1) Would you use an XR application to advertise your products?	Virtual Store	Yes/No question
	(Q2) Would you use an XR application to compose outfits?	Virtual Store	Yes/No question
	(Q3) Would you use an XR application to explore new fashion collections?	Virtual Store	Yes/No question
BI	(F2) I would buy products on an XR application	Virtual Store	5 point Likert scale
	(F3) I would buy more products on an XR application than on online stores	Virtual Store	5 point Likert scale
	(Q4) Would you use an XR application to buy online?	Virtual Store	Yes/No question
	(Q5) Would you use an XR application to buy a swimsuit?	Virtual Store	Yes/No question
	(Q6) Would you use an XR application to buy a sweater?	Virtual Store	Yes/No question

Table 3.2: Items and questions used in the survey to assess the Attitude Towards Using for Communication (ATUC) and Behavioural Intention (BI) constructs.

In particular, both constructs were investigated exploiting two groups of questions: the F and Q groups. The F group is formed by the following questions that were evaluated with a 5 point Likert scale:

F1. I would visit an XR retail store;

F2. I would buy products on an XR application;

F3. I would buy more products on an XR application than on online stores.

This group of items appears sufficient to answer and evaluate our constructs. Nevertheless, to avoid the neutral score problem related to the odd-Likert scales [76], we encouraged users to take a stronger stance by also submitting them the Q-question group, evaluated with a yes/no scale:

Q1. Would you use an XR application to advertise your products?

Q2. Would you use an XR application to compose outfits?

Q3. Would you use an XR application to explore new fashion collections?

Q4. Would you use an XR application to buy online?

Q5. Would you use an XR application to buy a swimsuit?

Q6. Would you use an XR application to buy a sweater?

As visible from Table 3.2, Q1-Q3 questions reinforce the opinion regarding the user perception on the XR retail store (F1) ensuring a good evaluation of the ATUC construct. In addition, Q1 puts the respondents in the shoes of a business owner who needs to advertise her/his products and reach as many customers as possible. All the remaining questions, listed in Table 3.2, adopt a consumer's perspective. Sentences F2-F3 and Q4-Q6 measure the Behavioural Intention (BI) from the consumers' point of view. Q4 is a control question to validate the F2 score regarding the purchase intention of a buyer in an XR application. We also included questions Q5 and Q6 concerning two practical scenarios, to understand whether XR can extend traditional e-commerce. Swimsuits and sweaters fit very differently. The former strictly depends on a client's body shape, whereas the latter is more easily predictable.

3.4.9 Results

The collected data has undergone a reliability check to test its internal consistency and validate our research. We computed the widely used Cron-

bach's alpha index, which corresponds to the Kuder-Richardson Formula 20 (KR-20) in the case of binary choice questions, such as our Q-items.

Construct	Items	α
PEEU	Is - Fashion Island	0.81
	Is - Virtual Store	0.77
VG	As	0.52
ATUC	Q1-Q3	0.81
BI	Q4-Q6	0.67

Table 3.3: Cronbach's α index for the considered constructs.

The results reported in Table 3.3 show that the PEEU and ATUC items may be deemed reliable (>= 0.70, as indicated by [383]). The BI items fell slightly below the 0.70 threshold value, whereas the VG ones cannot be accepted according to the utilized theoretical framework. The observed discrepancy might be attributed to questions that exhibited a degree of divisiveness. Specifically, the initial query (A1) inquired about the potential enhancement of system usability through the integration of a voice assistant with other interfaces. In contrast, the subsequent question (A2) sought to ascertain whether a participant would exclusively depend on voice, forgoing the utilization of other accessible interfaces.

3.4.9.1 Perceived Ease and Enjoyment of Use (PEEU) and Voice Gain (VG) Analysis

Figure 3.3 reports scores corresponding to the interface evaluations of the Fashion Island and the Virtual Store applications. We recall that the rule of the Likert scale voting system is the higher, the better, and any score above the average value of 3 indicates a satisfactory result.

The histogram reported in Figure 3.3 exhibits satisfactory results for both the Fashion Island and Virtual Store applications. However, the scores ob-

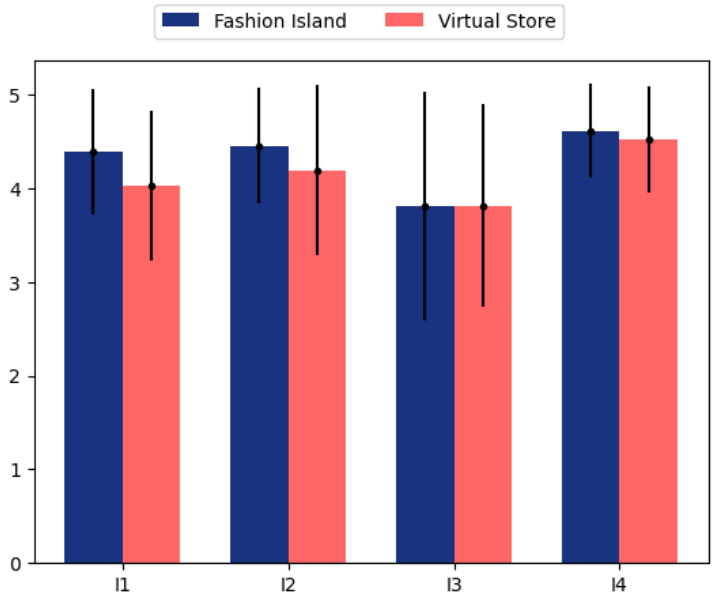


Figure 3.3: Histogram comparison of five-point Likert questionnaire results related to I-x items, which are relative to the Perceived Ease and Enjoyment of Use (PEEU). In blue and red we report the mean scores obtained by the Fashion Island application and the Virtual Store, respectively, along with their standard deviations.

tained for the I-x sentences show that the Virtual Store application registered an average value slightly lower than the one obtained for the Fashion Island (4.14 versus 4.32, with 0.3 and 0.35 the standard deviation values, respectively). In particular, Virtual Store scores slightly below Fashion Island ones for the I1 and I2 items: maybe voice commands increased the complexity of the interactions in such short experiments. We observe that the values obtained for the I3 and I4 items are roughly the same for the two applications: more complex interactions are, all in all, appreciated, and a similar enjoyment is reached. These observations are confirmed by the A1 item score reported in Table 3.4, which corroborates that the introduction of vocal commands and the presence of a virtual assistant made the VR store experience easier to use. Nevertheless, the users exhibited an interest in preserving the

use of hand controllers (item A2).

Item	Mean	Std
A1	4.19	0.79
A2	3.26	1.26

Table 3.4: Mean and standard deviation (std) of the five-point Likert questionnaire results for the A-x items. A-x items are those related to the Voice Gain (VG).

To further analyze such results, we focused on those participants who assigned lower values to the Virtual Store I-x items than to the Fashion Island ones. In particular, we set $k = 2$ as the difference between the two 5-point Likert scales (e.g., a subject assigned 4 to a Fashion Island I-x item and 2 to the corresponding item of the Virtual Store). For these participants, we carried out a qualitative analysis based on the "thinking aloud" method, to catch the pros and cons of the Virtual Store experience [203]. To this aim, we here report the answers they gave to the following question: *"Why was the Virtual Store experience not as easy to use as the Fashion Island one?"*. The answers can be summarized with the following three arguments:

(a) Participants would have appreciated having conversations with Alexa, pausing after each sentence. However, Alexa was designed to continuously process sentences within a short time frame. So, a few seconds of hesitation made Alexa sessions expire. For example, participant # 11 stated that: *"It is annoying to call Alexa every 10 seconds!"*;

(b) Subjects are expected to be able to execute a greater variety of vocal commands for the same action (e.g., use words like "buy", "buy this", or "I want to buy that"). Indeed, subject # 17 expressed his/her will as follows: *"I think that just having two or three commands to act is too restrictive."*;

(c) Sometimes participants preferred hand controls because they found them more efficient. Among all the interviewees, the number #26 demonstrated the following concern: *"I think that the exclusive use of Alexa could slow down some operations"*.

These aspects might have impacted the perceived ease of use of the Virtual Store negatively. However, even participants with more critical perspectives acknowledged the utility of vocal commands during their initial experience with the application, as evidenced by the high rating for item A1. They also proposed the potential benefits of using vocal assistants for specific demographics, such as individuals with disabilities or certain medical conditions. Additionally, they emphasized the significant role of vocal commands when multitasking, such as simultaneously shopping for a dress while engaged in other activities.

3.4.9.2 Attitude Towards Using for Communication (ATUC) And Behavioural Intention (BI) Constructs Analysis

Now, we consider the responses to the five-point Likert sentences F1-F3, reported in Table 3.5, regarding both ATUC and BI constructs.

Item	Mean	Std
F1	4.74	0.73
F2	4.23	1.15
F3	3.59	1.05

Table 3.5: Mean and standard deviation (std) of the five-point Likert questionnaire results for the F-x items. F-x items are those related to the Attitude Towards Using for Communication (ATUC) and Behavioural Intention (BI).

The F1 and F2 scores reflect the results presented so far: the users positively accepted the idea of purchasing a fashion item using an XR application. However, they would slightly prefer XR systems over classical online stores

when buying multiple fashion products (F3). On the other hand, by asking the participants to better clarify their attitude towards XR-based shopping through closed questions Q1-Q6, we were able to draw the chart shown in Figure 3.4.

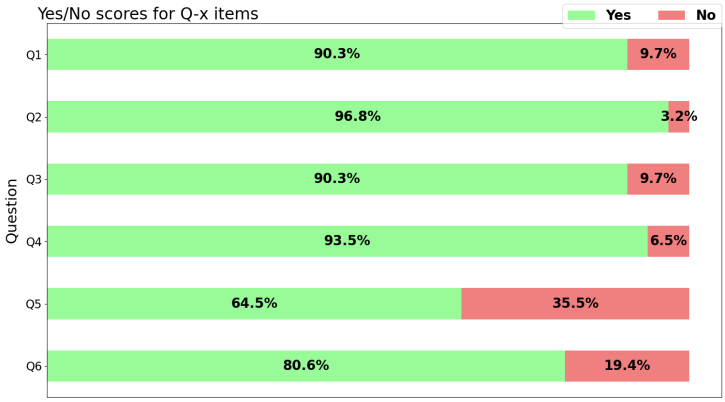


Figure 3.4: Yes/no answer percentages for the Q-x items. Q-x items are those related to the Attitude Towards Using for Communication (ATUC) and Behavioural Intention (BI).

Once forced to make a clear decision, the fashion experts confirmed the potential of XR technologies as tools for an expansion of marketing operations. Putting themselves in the shoes of a fashion manager, they would use XR as an innovative tool for their advertising campaign (90.3%) and conceive new outfits (96.8%). As customers, instead, they would use XR tools to explore new fashion collections (90.03%) and buy them (93.5%). When queried about the type of purchase they would conduct in XR, a noteworthy consideration emerged: their confidence levels varied depending on the nature of the products. They expressed a higher level of confidence in using XR for purchasing items like sweaters compared to swimsuits (80.06% vs. 64.5%, respectively): The latter might not instill as much trust due to potential concerns. As a final note, only one user declined all the options, likely due to reporting a slight headache and eye fatigue.

3.4.10 Discussion

To set the stage for a discussion around the contribution presented in this chapter, it is worth noticing that a very recent scientific survey has reviewed a body of seventy-two pieces of research that analyzed the use of VR in shopping [427]. Among these seventy-two it is worth mentioning that the two works that have been extended with our present contribution are included [250, 99], which were also the only ones to analyze the role of voice assistance in VR shopping [250].

Building upon the analysis performed in [427], the results presented in Section 3.4.9 let us highlight how the present contribution advances the field regarding the use of VR technologies in the fashion domain. We start observing that the PEEU (i.e., Perceived Ease and Enjoyment of Use) and the VG (i.e., Voice Gain) evaluations confirm both hypotheses H1 (i.e., fashion experts appreciate immersive VR interface) and H2 (i.e., the VR headset and controllers, together with the voice command integration, provide a positive immersive experience): in essence, even if the proposed implementation with Amazon Alexa as a voice assistant has imposed some design limitations, the presence of a voice assistant may improve the perceived immersion. It is worth noticing that such results were partially anticipated by [355], where the following two hypotheses were studied by implementing a WebVR shopping environment, empowered by a vocal search service, on a mobile VR platform: (H2 in [355]) "VR is preferred by the user in terms of user experience and usability" and (H4 in [355]) "VR with speech input outperforms the others in all aspects". Studying these hypotheses the authors found that: (a) the usage of voice inputs is preferred instead of head-pointing devices both in terms of usability and user experience, (b) the combination of VR and voice inputs outperforms all the tested combinations in terms of usability and user experience, and, (c) VR is not preferred over the classical keyboard/mouse inputs in terms of usability.

The findings provided in Section 3.4.9 reinforce and extend the reach of [355] in different directions, all contextualized in the fashion field. We did

this by constructing VR experiences related to fashion and performing an experimental campaign that included fashion domain experts. In terms of technology, our proposal exceeds the inclusion of voice inputs in VR scenarios by adding a voice assistant (i.e., Amazon Alexa), thus probably anticipating the near future: smart avatars may appear to assist users in their shopping activities, adding a social component to VR. Comparing now our results to finding (a) extracted from [355], our group of users appreciated an integrated approach that would include both a voice assistant and VR pointing inputs, whereas the voice was not preferred over pointing devices. Finding (b) is confirmed by our experiments. Finally, for finding (c), our results lead to an opposite conclusion, as we observe a preference for using VR interfaces instead of the classical keyboard/mouse one. This fact may be because a more advanced HMD device was employed in our work, thus equipped with easier-to-use controllers than a head-pointing mobile device. This aspect may lay the direction for future research in the field of Mobile VR.

Moving on to the usability and feasibility findings of XR, the questionnaire results for the ATUC and BI constructs confirm the H3 hypothesis: participants enjoyed the opportunity to explore virtual collections, compose new outfits, and buy products in an XR shop. Such findings confirm those reported by [278], where the effect of VR on fashion marketing was explored, and found to be positively related to pleasure and purchase intention We also observed that other researchers have already started to analyze the effects of XR on fashion marketing, using instead AR-based smartphone apps [303, 434, 304, 78, 43]. For example, [303] highlighted how consumer inspiration acts as a mediator between AR benefits and brand attitude changes. Our contribution may benefit such an existing line of work as it provides a new perspective (i.e., the role that a VA such as Alexa may have in an XR experience) for interaction with a complex digital system. In essence, it may help the user concentrate on what most inspires him/her, rather than on any interface-related details and technicalities (e.g., simply say "Alexa, I want to wear a purple shirt and blue jeans." rather than having to identify the

correct sequence of commands necessary to achieve the same goal). With hypothesis H4 (i.e., XR applications can be adopted to buy fashion products), we extend such lines of work. While the ATUC items confirm H4, we observe a controversial situation in the BI assessment: only two out of three BI yes/no questionnaire item results are strongly positive. We cannot confirm that x-commerce may overcome some of the well-known issues of fashion e-commerce (e.g., high return rate because of fitting problems). Indeed, the online selling of particular products (e.g., swimsuits, shoes, trousers, and dresses) remains a difficult task, as most customers would probably prefer the opportunity of a traditional shopping experience. Such challenges were not addressed in our prototype experiences, as realistic dress fitting was not possible.

3.5 Who will trust Human Digital Twins?

While analyzing whether vocal assistants could be adopted in virtual fashion scenarios to guide users in their actions (Section 3.4) some criticalities emerged, mostly regarding the unnaturality, from both visual and auditory perspectives, of the interactions with the VA [102, 427]. We here attempt to verify whether HDT could be applied to a particular use case in the fashion world: a brick-and-mortar shop as a use case, where the HDT is applied to a consumer, who interacts with a human store clerk asking to buy several fashion garments. With this use case, we here attempt to provide the first evidence on which factors should be considered to improve the realism and immersiveness of human avatars in metaverse-everyday experiences. At the same time, we here aimed at setting the stage for scenarios like autonomous and certified digital that could act as ourselves for different tasks in the metaverse, such as shopping [423].

We so leverage HDT paradigms, providing a preliminary study in a symmetric reality [448] setting: a virtual entity interacts with a non-virtual one. To this aim, we took as a use-case a brick-and-mortar fashion shop where a clerk can receive requests from HDTs appearing on a screen. We hence created an avatar 3D-scanning a real person [13] and tested the impact of having this avatar ask a specific question to a human as if the avatar were visiting a shop. Resorting to a group of users, we analyzed such impact, also verifying how such paradigm may be received in an x-commerce setting, as defined in [251]).

3.5.1 Method

We recruited a group of participants to analyze how receiving communication from an HDT may be accepted. Participants watched three videos: one portraying a real person (customer) and two of the HDTs of that person (the two HDTs differ in their voices, one natural and the other faked), each of which lasted 11 seconds. Each participant responded to a survey first

providing preliminary impressions of the HDTs and then walking through a process where s/he is asked to identify oneself as a clerk and answer an additional number of questions. Figure 3.5 shows the real and scanned image of the person who played the role of the customer in our experimental setting.

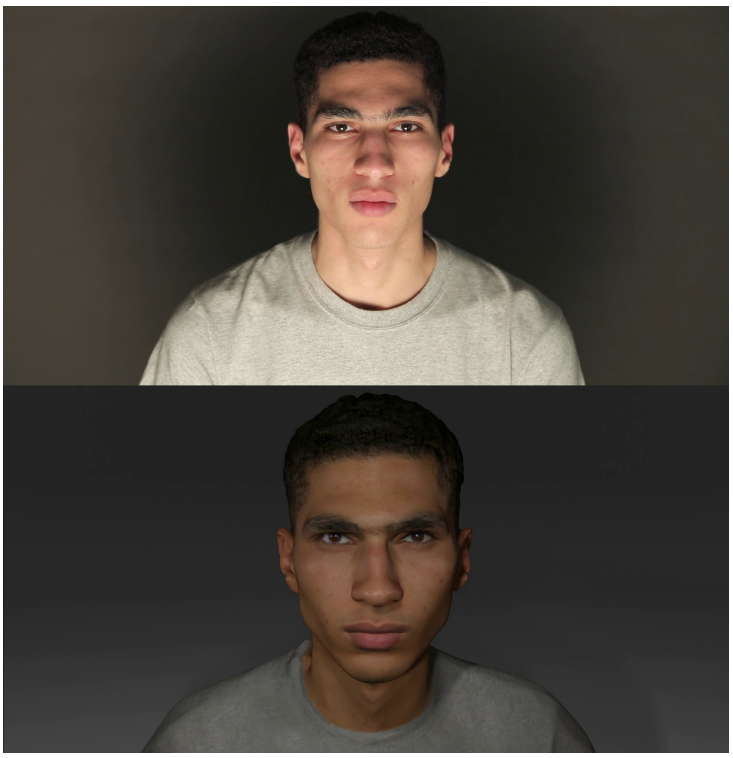


Figure 3.5: Real customer (top) and corresponding 3D model (bottom).

3.5.2 Material

After creating the 3D scan of the customer and producing the three videos, the experiment has been administered via the Google Survey platform[4], that also provides the aforementioned videos, as representative of a situation where a clerk interacts with a customer on a monitor and a trade-off between simplicity and recruiting participants during the COVID19 pandemic. The first video was taken by the independent artist Federica Sasso

[4] Link to the survey (Italian)

in a professional setting, fixing the camera in front of the professional model Ahmed Mesbah. His 3D model was then generated using Artec3D Eva scanner [13] and processed/animated with Blender [257]. Then, this model passed through a mouth rigging, followed by the application of a Lip-sync algorithm [5] to animate the avatar lips to pronounce the same statements formulated by the real model. Then, a frontal camera was set in front of its face to capture the frames, generating so the two HDT-analyzed videos.

3.5.3 Participants & Procedure

We recruited 31 participants for our experiment, 12 males and 19 females, 21 to 58 years old. The majority fell under the 35-year-old threshold (M = 27.77, SD = 8.93). They were either students, clerks, or professionals with little or no knowledge of digital avatars and 3D models. 15 declared to have no previous experience with digital avatars, while 10 out of 16 occasionally.

After completing consent and demographics questionnaires, participants answered preliminary questions about their occupation (student, worker, fashion shop clerk) and their relationships with digital avatars (specifying if they had at least one contact, the frequency, and their degree of comfort). Afterward, the experiment took place. Each participant was involved in a within-subject session requesting them to identify themselves as fashion clerks and respond to specific customer requests. The first request sounded as follows: *"Good morning. I need a shirt to wear with casual gray wool trousers. I usually wear size M, I am available to send you the 3D model of my body if you'd like"*.

Such a request is delivered using three different configurations: (a) natural person using natural voice (NN), artificial 3D model using natural voice (AN), and artificial 3D model using artificial, robotic, distorted voice (AA). After watching each of the three videos, a subject is asked to respond to the open question, (*"Which product would you propose to the customer?"*), in no more than one minute, to facilitate their deep processing state and, at the

[5] https://github.com/DanielSWolf/rhubarb-lip-sync

same time, focusing on the request instead of just on the customer's visual aspect. Then, a participant responded to different questions concerning the experience, adopting a 7-point Likert scale:

Q1 *"Understanding the customer's request was"*(complex vs easy);

Q2 *"How natural/spontaneous did you consider the interaction with the customer?"* (little vs lot);

Q3 *"During the sale, I felt comfortable"*.

Finally, each participant answered a few questions concerning the intention to work as a fashion shop assistant in an extended reality setting (measured with a 7-point Likert scale):

I1 *"Following this experience, would you work in a virtual store?"*;

I2 *"If you worked in a virtual store, would you like to interact with HDTs of your customers?"*.

3.5.4 Results and Discussions

The mean and standard deviation of all the Q-questions values are reported in Figure 3.6. On one hand, we observe that participants generally preferred watching the video in the NN configuration. On the other, the worst possible configuration is one that adopts the digital avatar along with an artificial and distorted voice (AA). The AN configuration exhibits lower mean values than the NN one, but this difference does not appear significant. However, we have also to point out that the mean overall results for question Q2 never exceed the value of 5.

These preliminary results suggest that to create a trustworthy and natural interaction between humans and HDTs in a fashion shop, the role of the voice remains fundamental. However, some concerns generally emerge regarding the interaction with an HDT. It is also worth highlighting that answers from items I1 and I2 provide a non-positive attitude on working in an

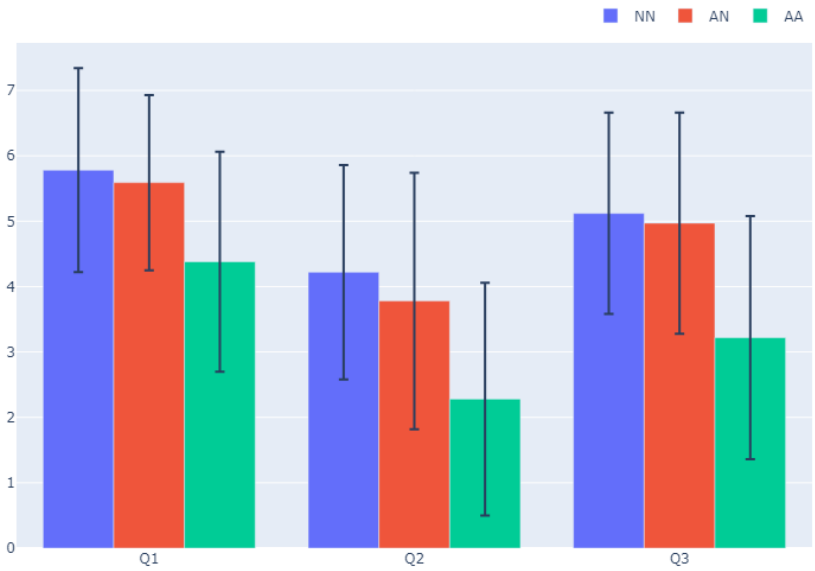


Figure 3.6: Mean and standard deviations of participants' answers to Q-questions.

extended reality setting as clerks (M=3.69, SD=1.71) while a non-negative one in communicating with customers' HDTs (M=4.5, SD=1.68). This fact underscores a bias that could have also influenced Q-questions. Finally, the majority of participants saw the merit of receiving the 3D model of the customer to better satisfy his request.

3.6 Evaluating Generated Immersive Content With eXtended Reality

Considering the number of DL-based GCMs, such as Text-To-Image (e.g., Imagen, Stable Diffusion, DALL·E), 2D-to-3D (e.g., NeRFs) and Text-to-Panorama (e.g., DiffCollage) to provide varied and high-quality multimedia content, including images, videos, and 3D assets it is possible to envision entire immersive XR environments created with automatic or nearly auto-

matic approaches [313, 320, 127, 60, 90, 198, 220, 111, 163, 143, 61, 316, 412, 316, 433]. 3D environments generated from basic descriptions, digital art paintings, and fashion 3D garments are just examples of virtual elements that could be automatically generated for the benefit of the entertainment and creative industry in virtual experiences. It is also worth noticing that, in the context of digital e-commerce platforms, all stakeholders are nowadays interested in integrating GCM for several purposes, with a particular focus on integrating them in novel and immersive commerce experiences [181, 324, 84]. For example, Generative AI can enable unprecedented forms of personalization (i.e., hyper-personalization), allowing digital sellers to instantly tailor their offerings to meet the preferences of each consumer [417]. On such a line, GANs and Diffusion models were employed in e-commerce environments, to improve image retrieval systems [15], and generate consumer profiles that could buy a certain product [196], garments from dressed people [441], fashion poses starting from a unique picture [188], entire fashion garments by sketches and prompts [79, 377], while also stylizing 3D meshes by text [430].

Given these capabilities, it is worth considering how humans (i.e., final consumers) perceive them. The goodness of such models is usually evaluated quantitatively with vastly adopted metrics, such as Learned Perceptual Image Patch Similarity (LPIPS), Fréchet Inception Distance (FID), Kullback–Leibler Divergence (KL), Minimum Matching Distance (MMD), Chamfer (pseudo)-distance (CD) and Minimum Matching Distance (MMD) [341, 160, 445, 425], Human evaluation takes place by adopting the Mean Opinion Score (MOS) or the Turing Test [191, 313, 90, 320, 198, 220, 111, 163, 143, 316]. However, those human evaluation method are not able to catch complex factors like the aesthetics and emotional ones, that demonstrated to be key in virtual commerce experiences, and also in controlling generative AI content experiences [184, 348, 250, 126, 251, 361, 269]. In fact, as reported in [184, 126], emotions and aesthetics bear high-level semantics that could be bonded to low-level computable visual features that could be put to good use to the positive control of the design and automatic generation of multime-

dia content, introducing novel forms of HITL. Modeling such a relationship would be particularly effective in creative fields like fashion, art, music, and literature, where human artists could collaborate with GCMs to maintain artistic vision and control to induce certain moods or emotions [270, 184, 391, 126]. When extending such a perspective going beyond traditional 2D images, videos, and audio, such as 3D models/scenes, and 360° panorama images and videos, conventional display hardware was highlighted to be less effective for human evaluation [223, 391, 30]. A first solution to such a problem could consist of employing XR paradigms to define more effective ways to present such visual/auditory stimuli to humans and ease their evaluation.

Considering all such aspects, we designed an evaluation XR-based framework (depicted in Figure 3.7), that could be deployed to measure immersive content aesthetic and emotional factors.

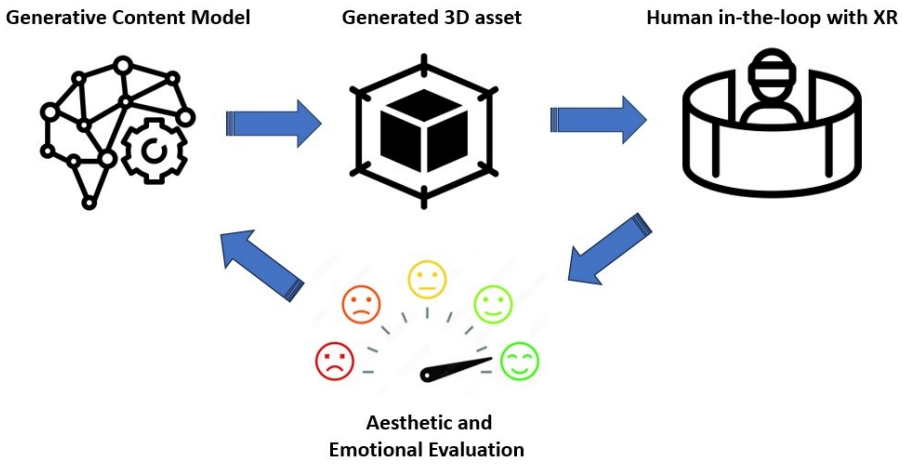


Figure 3.7: Schema of the proposed Aesthetic and Emotional Evaluation framework for 3D generated content with XR.

As described in Figure 3.7, a GCM produces some 3D immersive context that is experienced by a human which provides (synchronously or asynchronously), her/his emotional and aesthetic perception of that content (or part of it), providing useful feedback not only to steer the generation but also generating data for the next training iteration. It is worth noticing how such

a framework could define novel forms of HITL in the context of Generative AI [426].

Considering however its complexity, we here isolated and analyzed the right-most section of it, providing first insights about the possible adoption of XR paradigms to evaluate the aesthetical and emotional components of immersive items, comparing them to 2D displays [335] We contextualized such a framework by adopting the Aesthemos [330]) scale to evaluate the aesthetic and emotional experience of a pivotal 360° video of a musical concert [335]. This particular item was selected as a pivotal example considering that is: (a) one of the most emotional and aesthetics activators [265, 82]; (b) it is one of the few that could be modulated in different degrees of realities; (c) it has a strong connection with fashion experiences like catwalks, nowadays approaching their maturity to show and sell automatic designed fashion garments [153, 254, 417]. To the best of our knowledge, this is the first contribution that scratches an XR-based HITL framework to evaluate immersive (generated) content.

3.6.1 Applying The Framework: Aesthemos Scale Evaluated On a VR 360° Panorama Tango Concert Video

In evaluating digital multimedia content with VR, different scales could be adopted to assess the Aesthetic and Emotional dimensions related to it, even if the literature still lacks evidence on how consumers respond to VR experiences in the realm of aesthetic perception, for example, in the music domain [335]. In such direction, the Aesthetic Emotions Scale [330] emerged as the best candidate, considering that it provides structured 21 subscales covering prototypical aesthetic emotions, epistemic emotions, and emotions indicative of amusement. Adopting such a scale, VR could be exploited as a methodological boost to empirical aesthetics: virtual environments provide an excellent compromise between ecological validity and experimental

control, providing us control of the degree of immersiveness, allowing so to study its effect [401]. This means that we could examine the effect of the same visual and auditory stimuli, from different degrees of XR: from physical reality to full-immersive VR, providing us the perfect playground for inferential analysis [335]. However, also an item that could be deployed in any degree of reality should be considered to fully exploit such analysis flexibility. This motivated us to consider as the target immersive item a 360° video of a real-musical concert: this is one of the few items whose aesthetical and emotional effects could be measured in both physical reality and virtual one [224].

Thus, we addressed the aesthetic emotions evoked by this item in four conditions: (a) a live concert (LC, in presence); (b) the visualization of the corresponding 360° video experience in a 2D display (MV, the classic viewing of a concert on YouTube) and two different degrees of immersiveness, google cardboard (CVR) and a full-immersive HTC-Vive (HVR). The latter is depicted in Figure 3.8.

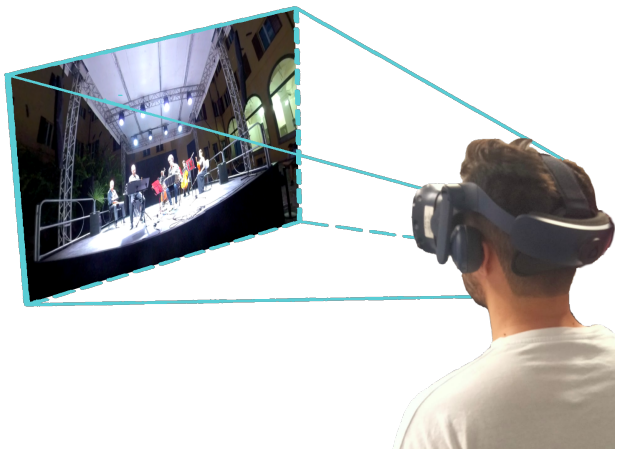


Figure 3.8: Exemplar visualization of a user experiencing the 360° virtual video concert with the HTC-Vive (HVR).

The CVR and the HVR allow respectively for a basic and easily accessible experience vs. a less affordable but more immersive experience. Both the

devices permitted a three-dimensional vision: by moving the head (3-DoF), participants could have a 360° view, therefore an overall vision of the concert venue, including musicians and audience, together with their possible reactions to a virtuosity or a false note played by the performers. These four conditions (LC, MV, CVR, HVR) efficacy, were evaluated through the administration of the Aesthemos questionnaire to a very well-studied kind of population: young students, who are not interested in attending a certain kind of cultural experience, namely a Tango music concert in a theatre, but that is prone to use Virtual Experience technologies [124, 131] and a strong comparing baseline, passionate adult people, which is usually the target audience for this kind of cultural activities [238]. In practice, we used LC participants' survey scores as a benchmark to compare the magnitude of interest of passionate adult people concerning the youngsters who lived instead of Virtual Experiences (MV, CVR, HVR). As an additional contribution, this framework could be adopted with any other kind of aesthetic experience (e.g., fashion catwalks, and tourist tours).

3.6.2 Experimental Session

The experimental condition live concert (LC) consisted of participation in the whole concert. From the 360° video recording of the concert, 9 significant minutes were extracted, including the introduction of the artists and the musical performance. The reason behind the 9-minute choice has its roots in both the literature and a questionnaire administered to participants online [335]. The young population is used to enjoying music videos on screen: several studies testify that the average time of such experiences is usually less than 9 minutes [264, 215, 432]. We also surveyed our young subjects about the topic, with F-X items that could be found in Table 3.6. The results of the F-X questionnaire show that 96% of the subjects had seen a video of a concert (item F1), and for these subjects, the responses to item F2 resulted in an average estimate of 19.6 minutes. Examining also the categorical item F3, the trend of the distribution is on 10 minutes (Figure 3.9). This is in line

with the results obtained in literature [264, 215]. Furthermore, the scores for item F4 suggested uncertainty about watching a video concert longer than 10 min, with a mean of 3.2 (S.D. ±1.30). Finally, the scores for item F5 revealed little interest in watching the video for a time longer than 9 minutes (68% of the sample answered "No"). Therefore, we considered the 9-minute video of the concert to be a fair trade-off for the youngsters to stay focused and, at the same time, to provide them with a music video slightly longer than the ones they are used to.

Question	**Question Type**	**Possible Answers**
F1. Have you ever seen a video of a concert?	Yes/No Question	{Yes, No }
F2. In your estimation, how many minutes would you spend watching the video of a concert?	Open Question	Open Question
F3. What are the minimum minutes that would make you consider a video to be long?	7-Point ordinal scale	Above {1,3,5,10,20,30,40} minute(s)
F4. Following the previous question, would you be predisposed to watch a concert video that exceeds 10 min?	5-Points Likert Scale	{1,2,3,4,5}
F5. After the experiment, would you have continued watching the video beyond the proposed 9 min?	Yes/No Question	{Yes, No }

Table 3.6: F-x items related to Musical video media habits.

In the experimental session, we tested 70 participants, 10 for the live concert condition and 20 different ones for the MV condition, CVR, and HVR for a total of 70 participants [335]. In the experimental session, the LC condition was executed at the Teatro Comunale "Pavarotti-Freni" in Modena, during the concert "Amarcord d'un Tango". The event took place outdoors, in the theatre courtyard. At the end of the concert, we asked volunteers to fill in a hard copy of both questionnaires. For MV, CVR, and HVR conditions, participants were tested at the University of Bologna.

The Aesthemos questionnaire was furnished in the form of forty-two 5-

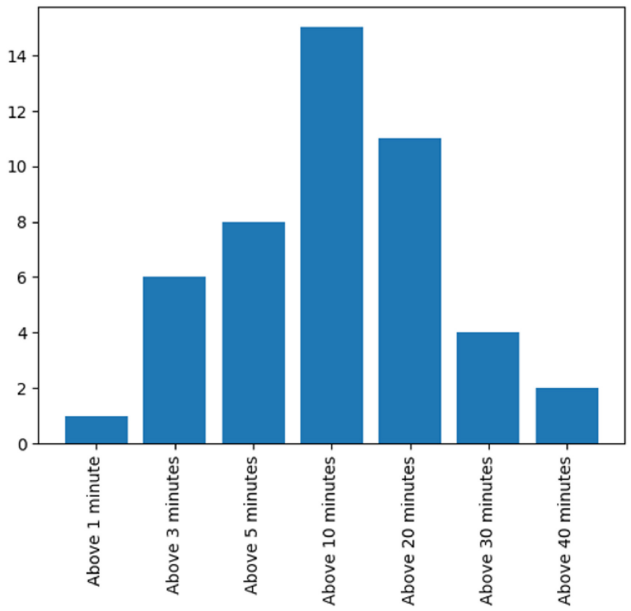


Figure 3.9: Histogram of answers to item F3

point Likert scale scores referring to twenty-one emotion subscale [330]: *Feeling of beauty; Fascination; Being moved; Awe; Enchantment; Nostalgia; Joy; Humor; Vitality; Energy; Relaxation; Surprise; Interest; Intellectual challenge; Insight; Feeling of ugliness; Boredom; Confusion; Anger; Uneasiness; Sadness* as described in [330, 335]. Additionally, we also defined a custom scale inspired by the Igroup Presence Questionnaire (IPQ) [334] to evaluate and compare the immersiveness provided by different devices, which is strongly related to people focus and evaluation capabilities (i.e., administered only to in-lab participants). [123, 170] (these scales are further detailed in Section 3.6.4).

3.6.3 Materials and Apparatus

For the Google Cardboard and HTC-Vive, the 360° video was recorded using the Insta360 Pro2 and post-processed with the Insta360 STITCHER that produced a 2K video [6]. We recorded the concert placing the camera

[6] https://www.insta360.com/it/product/insta360-pro2

between the stage and the audience, to capture as much detail as possible, without losing focus on the main content, i.e., the musicians on the stage. Two VR experiences were developed to present such video in both the CVR and HVR conditions.

We adopted Unity as the Game Engine to develop those VR experiences [7]. In particular, the CVR application was developed using the GoogleVR SDK, while the HVR one used the SteamVR SDK. Both applications allow to rendering and reproducing of 360° video on the target devices. The CVR and HVR settings (3D), different from the MV (2D), allowed participants to move their heads and explore the space around them while enjoying the show. This is possible considering that both CVR and HVR have human head rotation 3-degree of freedom devices (i.e., pitch, roll, and yaw). In any of the virtual conditions, no interaction was possible. The developed CVR application was executed on a Samsung S22 equipped with a Qualcomm SM8450 Snapdragon 8 Gen, 8 GB of RAM, and 1080×2340 pixels. The HVR one was instead executed on an HTC-VIVE Pro 1440×1600 connected to an Alienware Area 51 model. These devices were chosen based on their high performance to achieve the smoothest video reproduction along with the highest possible resolution.

3.6.4 Results and Discussions

In this Section, we reported the obtained results by subjecting our participants to those different experiences and then evaluating their aesthetic and emotional impact through the Aesthemos scale [330]. Moreover, since the immersiveness correlates with both of these factors, we also evaluated an additional short version of the IPQ scale [334].

Aesthemos The collected data has undergone a reliability check to test for internal consistency and validate the research and we further analyzed those constructs that exhibited a Cronbach's alpha index ≥ 0.70 [383]. Those are

[7] https://unity.com/

reported in Table 3.7, which also shows all the examined constructs included in the Aesthemos Scale [330].

Aesthemos Sub-groups	Items	Cronbach Alpha
1 Feeling of beauty - liking	E1-E6	**0.87**
2 Fascination	E7-E34	**0.81**
3 Being moved	E14-E36	0.57
4 Awe	E31-E40	-0.13
5 Enchantment	E8-E18	**0.81**
6 Nostalgia	E26-E28	0.59
7 Joy	E3-E39	**0.79**
8 Humor	E22-E42	**0.92**
9 Vitality	E9-E32	**0.76**
10 Energy	E16-E41	0.59
11 Relaxation	E4-E20	**0.86**
12 Surprise	E11-E29	0.21
13 Interest	E5-E38	**0.80**
14 Intellectual challenge	E2-E10	0.01
15 Insight	E13-E21	**0.73**
16 Feeling of ugliness	E12-E35	0.65
17 Boredom	E19-E33	**0.76**
18 Confusion	E24-E37	0.54
19 Anger	E17-E25	**0.87**
20 Uneasiness	E27-E30	0.68
21 Sadness	E15-E23	**0.78**

Table 3.7: Cronbach's alpha index for all considered sub-groups of the three questionnaires (please note that E1-E3 indicates two question items and not a range)). Twelve out of twenty-one constructs of the Aesthemos (in bold) passed the internal consistency test (in bold).

For the analysis of the accepted constructs, we first provided descriptive statistics about the distribution obtained for the valid question items plotting them via histograms, where mean scores and standard deviation are listed. Participants' scores for the Aesthemos, aggregated for each subgroup, are reported in Figure 3.10 (first six subgroups) and Figure 3.11 (last six subgroups).

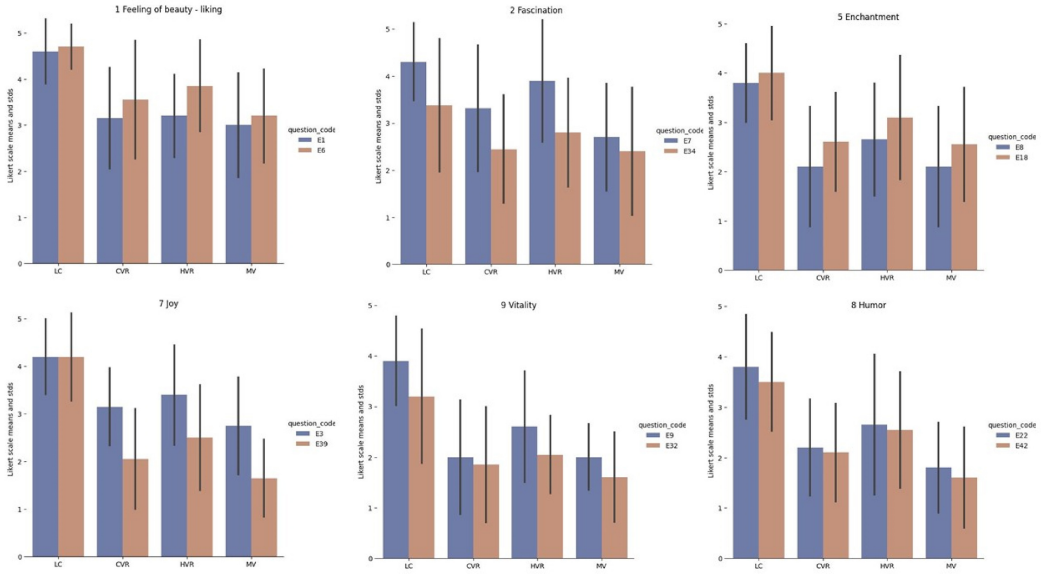


Figure 3.10: Histograms of first six sub-groups of the Emotional Scale: 1 Feeling of beauty – liking; 2 Fascination; 5 Enchantment; 7 Joy; 8 Humor; 9 Vitality. Error bars represent the standard deviation.

For the constructs shown in Figure 3.10 (Feeling of beauty – liking; Fascination; Enchantment; Joy; Humor; Vitality) there is a general agreement on the superiority of the Live condition over all others (MV, CVR, HVR). However, HVR is preferred to CVR and MV for the constructs of Fascination, Enchantment, and Joy.

For the constructs shown in Figure 3.11 (Relaxation; Interest; Insight; Boredom; Anger; Sadness), the advantage of the LC condition over all the others is no longer evident. In particular, the plot shows that Interest is similar for LC and HVR, and greater than for CVR and MV; Boredom is more pronounced for the MV condition than for the other three scenarios, especially when compared to LC. For the constructs of Anger and Sadness, the values are relatively consistent and similar for the different experimental conditions.

To validate such results, we subjected our data to statistical analyses to verify any significant differences among the four conditions (i.e., LC, MV,

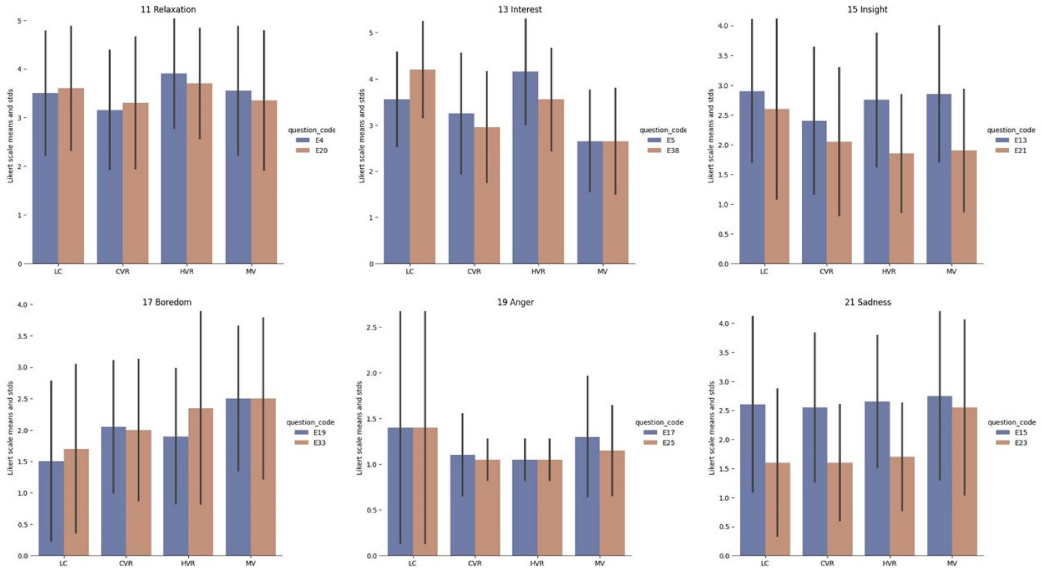


Figure 3.11: Histograms of the second six sub-groups of the Emotional Scale: 11 Relaxation; 13 Interest; 15 Insight; 17 Boredom; 19 Anger; 21 Sadness. Error bars represent the standard deviation.

CVR, and HVR). In particular, we performed an Adjusted Wald-Confidence Interval test [2], which allows us to check if the difference between two proportions is significant and how large the difference is. We selected this specific statistical test as we had four conditions and a low number of samples for each of them (i.e., ≤ 30) [328]. We considered a confidence level of 95%, thus the two-sided z critical value for the test corresponded to 1.96. To adapt our data for the specific test, we binarize the Likert-scale answers with a threshold: Likert scale answers with a score ≥ 4 were converted to 1, the lower to 0. Considering that the Adjusted-Wald Confidence Interval test compares two proportions, we proceeded to compare, two by two, all the scores in the different conditions for each subgroup of questions of the questionnaires. Thus, as an example, for the construct "X", composed of questions Ey – Ez, we run a comparison test between each possible pairing of our four conditions (LC, CVR, HCV, MV). Then we accumulated, for each question in each construct (subgroup), just the pairings that provided positive Wald Adjusted

Difference and a lower confidence interval greater than zero: the considered condition (e.g., LC) does have a higher score than the others (e.g., one among MC, CVR, HVR) for the question under examination (e.g., Ez). All significant comparisons for the Aesthemos are shown in Table 3.8, organized by construct and question-item, and reporting the Wald interval values (i.e., inferior bound, difference, and max bound).

Construct	Quest.	Comparisons	Wald Intervals
1 Feeling of beauty	E1	LC>CVR	0.18 — 0.47 — 0.76
		LC>HVR	0.18 — 0.47 — 0.76
		LC>MR	0.23 — 0.52 — 0.80
	E6	LC>CVR	0.16 — 0.42 — 0.68
		LC>HVR	0.03 — 0.28 — 0.54
		LC>MV	0.21 — 0.47 — 0.72
2 Fascination	E31	LC>CVR	0.02 — 0.34 — 0.66
		LC>MV	0.22 — 0.53 — 0.83
		HVR>CVR	0.04 — 0.32 — 0.60
		HVR>MV	0.25 — 0.50 — 0.76
5 Enchantment	E8	LC>CVR	0.08 — 0.40 — 0.73
		LC>HVR	0.08 — 0.40 — 0.73
		LC>MV	0.13 — 0.45 — 0.76
	E18	LC>CVR	0.03 — 0.36 — 0.69
		LC>MV	0.03 — 0.36 — 0.69
7 Joy	E3	LC>CVR	0.17 — 0.48 — 0.79
		LC>MV	0.17 — 0.48 — 0.79
	E39	LC>CVR	0.23 — 0.53 — 0.84
		LC>HVR	0.12 — 0.44 — 0.76
		LC>MV	0.29 — 0.58 — 0.87

Construct	Quest.	Comparisons	Wald Intervals
8 Humor	E22	LC>CVR	0.13 — 0.45 — 0.76
		LC>MV	0.19 — 0.49 — 0.80
		HVR>MV	0.04 — 0.27 — 0.51
	E42	LC>CVR	0.13 — 0.45 — 0.76
		LC>MV	0.13 — 0.45 — 0.76
9 Vitality	E9	LC>CVR	0.33 — 0.62 — 0.90
		LC>HVR	0.22 — 0.53 — 0.83
		LC>MV	0.45 — 0.71 — 0.97
	E32	LC>HVR	0.02 — 0.33 — 0.63
		LC>CVR	0.02 — 0.33 — 0.63
13 Interest	E5	CVR>MV	0.06 — 0.32 — 0.58
		HVR>LC	0.03 — 0.36 — 0.69
		HVR>CVR	0.00 — 0.27 — 0.55
		HVR>MV	0.36 — 0.59 — 0.83
	E38	LC>CVR	0.07 — 0.39 — 0.71
		LC>MV	0.22 — 0.53 — 0.83
		HVR>MV	0.05 — 0.32 — 0.59
21 Sadness	E23	MV>CVR	0.08 — 0.32 — 0.56
		MV>HVR	0.08 — 0.32 — 0.56

Table 3.8: All significant comparisons for the Aesthemos, organized by construct and question, reporting Wald values as (Wald inferior bound, Wald difference, Wald max bound). Conditions of advantage for HVR are bold.

Live Concert (LC) was found to be the most effective condition in activating aesthetic emotions since this condition resulted to be superior for 5 of the 7 constructs (acceptable Wald-difference) in at least one comparison with all the other three conditions (all of the above mentioned except *Interest* and *Sadness*). More specifically, LC was found to be significantly more activating (i.e., more enchanting, joyous, and amusing) when compared to MV and CVR, but not when compared to the HRV. Thus, even if the live condition remains the best way to enjoy a music concert, the difference with the same experience lived through the HCT-Vive headset is not meaningful.

This suggests that the HVR is the "artificial experience" that can offer the spectator an experience more like the "real-live one".

The only construct for which LC significantly differed from all the other conditions (thus also from HVR) is the Feeling of beauty: the LC was the most liked experience. Notably, looking at the construct *Interest*, the HVR condition resulted significantly more interesting than MV (underlined in Table 2). To the best of our knowledge, this is the first empirical evidence of a major interest in musical aesthetic experiences (i.e., experiences for which the acoustic component, and not the visual or social ones, should be predominant) when experienced in VR (HVR) than on a computer screen (e.g., on YouTube or Twitch). Moreover, for one of the items of the construct *Interest* (namely the E5: "It made me curious"), the HVR condition was found to be superior not only to MV but also to CVR and even to LC. Overall, our results are consistent in supporting a clear advantage for the Live Concert; nevertheless, they also emphasize that the experience with the HTC-Vive device is, immediately following, the most powerful in evoking aesthetic emotions, thus better than those generated with a device such as a computer (MV) or a google cardboard (CVR). This advantage is further confirmed if we look also at the scores for item E31, "I found it sublime", of the Fascination construct (for the complementary item, E34, no acceptable wald differences were found) and at the scores for the item E23, "It made me sad", of the *Sadness* construct (for the complementary item, E15, no acceptable wald differences were found). From this more comprehensive outlook, HVR emerges as significantly "more sublime" than MV and CVR, and MV appears as an experience significantly "sadder" than both HVR and CVR.

Thus, even if the LC remains the best way to enjoy a music concert, the difference with the same experience lived through the HCT-Vive headset is not statistically significant for certain aesthetic and emotional constructs. This suggests that the fully immersive virtual reality is the "artificial experience" that can offer the spectator an experience more like the "real-live one" concerning classical means such as 2D displays or mobile VR. To the

best of our knowledge, this is one of the first empirical evidence of a major interest in musical aesthetic experiences when experienced in VR than on a computer screen. This is also the first contribution that highlights how VR paradigms could be better than classical means to evaluate immersive (generated) multimedia data.

Short IPQ Scale For evaluating the immersiveness, we proposed a shorter version of the original IPQ scale [334], with three statements instead of the original fourteen, which however covers all the examined constructs. All the question items were furnished as a 5-Point Likert Scale and are here listed: (IP1) I felt present in the virtual environment; (IP2) I was not aware of the real environment around me; (IP3) I was completely captured by the virtual world. The mean scores and associated error bars are reported in Figure 3.12. The graph shows a clear tendency of immersiveness for the HVR concerning the other devices [157], even considering the variability.

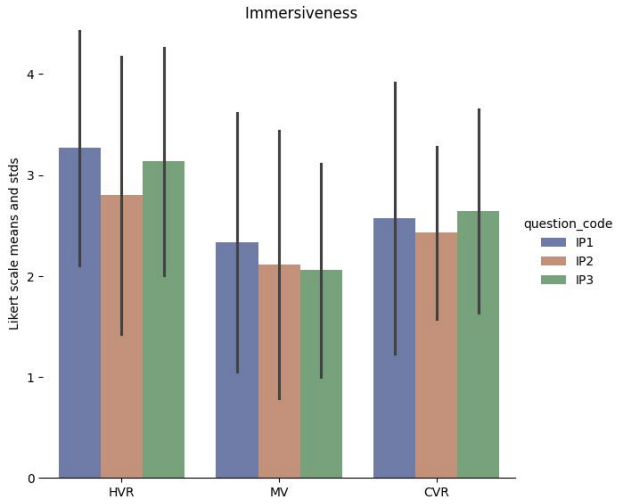


Figure 3.12: Results for the IPQ-inspired immersiveness scale.

3.7 Conclusions

Fashion, and in general the creative industry, is characterized by a search for technological innovation and digital presence, by a strong attachment to physical contact with products, and by a strong relationship with brands through their flagship stores. Nevertheless, despite initial skepticism, global trends exhibit an increasing acceptance of fashion e-commerce and x-commerce by a wide class of consumers.

We here considered the particular case of VR, which as well as the many other technologies involved in the provision of XRs, may also be exploited to further enhance customers' virtual experiences and in general strengthen the impact of brand retail strategies. In this chapter, we started analyzing one of the aspects that have restrained the diffusion of VR applications in retail, i.e. the difficulty of VR interfaces for non-expert users. As a viable solution, we identified the embedding of a Voice Assistant into fashion VR applications as a potential improvement of the users' perceived ease of use. We thus designed an experiment based on two VR immersive experiences simulating the processes typically involved in fashion stores. Only one, however, allowed its users to interact vocally with a virtual shopping assistant. Our findings demonstrate a high interest in the exploitation of voice-based interactions leveraging the use of a popular Voice Assistant such as Amazon Alexa. The Technology Acceptance Model has driven our work, linking the simplicity of the VR interface to the subsequent availability to adopt such a technology in a fashion retail setting. The group of users who tested the application confirmed the feasibility of x-commerce for fashion retail purposes, ranging from advertising to shopping platform providers (RQ-2, RQ-3).

These results, along with the theoretical concepts reported in [427], provided us with ideas regarding how x-commerce may be further empowered by: (a) exploiting the use of voice assistants, fully exploring their social dimension and their use in mobile settings, (b) using realistic and personalized avatars, and, (c) exploiting algorithms capable of providing a realistic fitting of clothes. Such thoughts provide us with a roadmap for future investigations:

this may be further pursued by increasing the number of users involved in the analysis and developing the role of the voice assistant in all of its aspects (e.g., appearance and interaction capabilities).

Following both (a) and (b), we presented a preliminary experiment in which we assessed the possible favor of HDTs as customers of a brick-and-mortar store. Our results suggest the adoption of such technologies could be positive (RQ-3). Yet, many aspects require further investigation to support any possible social interactions between humans and HDTs. In future works, we aim to improve the experimental design by examining different aspects (e.g., virtual gestures) contextualizing it in more structured x-commerce settings.

Then, we also tried to anticipate possible managerial implications deriving from the adoption of XR environments, for the benefit of the fashion domain [159, 428]. Unlike others, we involved fashion experts with little or no technical skills to assess our VR experiences and integrated the use of another technology that will likely occupy an important place on the stage. One of the results from our user study and also the ones obtained by related work [159, 162, 254], highlights how the low diffusion of such technologies and the lack of completely automatic methods aiming at supporting managerial directions will prevent, in the end, their global diffusion. Nevertheless, with an increase in the penetration rate of VR and novel performant DL models, such as Large Language or Diffusion models [143, 24], the present analysis may provide a valuable asset in the analysis of future trends.

Considering this, it is also worth analyzing whether managerial figures could steer and control such models to match a particular marketing or strategic campaign. In other words, verify the possibility for such professionals to automatically generate fashion items that could be sold in immersive environments [181]. This also requires an analysis from a consumer perspective: whether the generated content will have enough aesthetic and emotional appeal to convince consumers to buy it. In such a direction, we analyzed another aspect fundamental to enable this kind of x-Commerce sce-

nario: whether XR paradigms could be used to judge and enjoy such kind of immersive content through aesthetic and emotional constructs. To this date, we provided the first empirical result by considering a pivotal use case: a 360° musical concert video enjoyed in VR. Despite the obtained results being considered preliminary, those provided the first evidence that XR paradigms could be the best means to trigger emotional user response and, at the same time, to judge its quality, paving the path for new HITL paradigms (RQ-3). In future works, we intend to explore such an approach with generative models involving 3D models/scenes to measure how the aesthetic and emotional perception changes concerning classical fruition devices and paradigms [34, 158]. Despite the technological limitations of the analyzed approaches, x-commerce appears as a potential mainstream channel for fashion retail, among others, as soon as the readiness level and the costs of XR devices appeal to the mass market [250, 251].

In conclusion, this chapter introduced two different frameworks that exploit both AI and XR to support fashion x-commerce research. In particular how intelligent assistance and HDTs could be adopted to improve user interactions and the completion of fashion tasks (RQ-2, RQ-3). We also considered a novel human evaluation framework for immersive items' aesthetics and emotional factors, deployable in x-Commerce settings (RQ-3).

Chapter 4

eXtended Reality systems empowered with Artificial Intelligence to support humans in industrial use-cases

This chapter aims to analyze the role of AI & XR in the rapid development of Industry 4.0 with a user-centric approach, focusing on two main paradigms to empower smart workers: Digital Twins (DTs) and eXtended Reality (XR) immersive analytics [202, 58, 252].

DTs are computer models that simulate or mirror the life of a physical entity, which may be an object, a process, or a human [26], which are continuously connected to their physical counterparts, applied in fields ranging from industry to healthcare [385, 107]. DTs analyze data coming from the real world (e.g., IoT) with intelligent and adaptive methods (i.e., AI or statistical analysis) to provide insights about the current/future state of the DT and its physical counterpart [26, 40]. Despite DTs being often used as "invisible digital layers" for decision-making, a fundamental step, to analyze and control their behavior, amounts to smart visualization interfaces [180]. To this date, XR immersive analytics could be exploited to visualize information in

a physical space (AR/MR) or fully immersive settings (VR) [28, 195]. While many works on such a line focused on VR, AR/MR holds the main potential for industrial application, to empower workers with spatial information anchored to real-world objects, and so their DT [413, 217, 199]. However, to the best of our knowledge, we found a lack of (i) systems that integrate these technologies seamlessly into industrial products/processes, without increasing the operator's cognitive load; (ii) studies concerning the possible adoption of collaborative Human-in-The-Loop (HITL) role of operators, going beyond visualizations [217, 16, 370].

To this date, an interesting research question amounts to ergonomic interaction with DT-related information: how can XR interactions support humans in visualizing and manipulating the flow of information going back and forth between the physical and the virtual world? (RQ-2, RQ-4) [385, 358, 40, 195]. Following such a research question, in this chapter, we introduce a flexible HITL framework to allow users to access and manipulate flows of information in DTs, in any degree of reality by leveraging XR and AI paradigms. Considering its flexibility, we then provide a partial application of such a framework in three different contexts: industrial production, cultural heritage, and wine retail.

The rest of the chapter is organized as follows. Section 4.2 sets the stage for the theoretical background of this chapter while Section 4.3 refers to the related industrial use cases and adopted technologies. Section 4.4 provides the analysis, design, and contextualization of the collaborative human-intelligence DT-injected framework. Then, Section 4.5 details the domain knowledge and inspired implementation of a novel AR system for wine recognition based on CV and textual analysis. Finally, Section 4.6 concludes this chapter, providing new directions for future works.

I here declare that the content of this chapter is entirely based on my work and contains nothing that is the outcome of work done in collaboration with others and that all the contributions reported here have been partially published in [365, 358, 370, 10].

4.1 Research Questions

Considering the different aspects analyzed in the previous Sections, this chapter aims to answer the following research questions:

RQ-2 - Can Artificial Intelligence improve user interactions and task completion efficacy in eXtended Reality systems?

RQ-4 - Can industrial workers take advantage of eXtended Reality and Artificial Intelligence in their everyday activities?

4.2 Introduction

XR and AI technologies are playing a pivotal role in advancing and supporting Industry 4.0 across various contexts, for example, visualizing and creating new knowledge in real-time, decision-making, in rapidly changing environments [202, 58, 252]. In such a scenario, virtual representations of complex data and environments allow for faster and easier adoption of newer practices, which lead to higher value creation [202, 58, 252]. Such an integration diminishes the divide between the physical and digital realms, enabling direct interaction across remote physical locations [202, 58, 252].

This sets the stage for a new kind of operator characterized as an intelligent and skilled user, capable of engaging in collaborative and proactive efforts along machines [202, 58, 252]. This advanced human–machine interaction, coupled with adaptive automation, aims to establish systems where faster real-time decision-making and the generation of new knowledge are facilitated [314].

In such a context, DTs, XR, and AI represent pivotal technologies: DT links physical objects with their digital counterparts; XR enhances the user experience in terms of digital content visualization, interaction, and remote and collaborative operation; AI is the core of an intelligent software layer that allows to analyze and correlate data from various sources, providing knowledge to be visualized and manipulated [58, 252, 162]. Most of the

contributions employing these technologies in industrial settings hypothesize that the user is in a defined physical location. However, early forms of remote industrial operations are blooming [58, 252, 162, 195]. In both physical and remote contexts, an interesting research question amount to ergonomic factors between Humans and DT-related information [385, 358, 40, 195]: how could we support humans in visualizing and manipulating the flow of DT information going back and forth between the physical and the virtual world? In fact, despite DTs being used to act as invisible digital layers to model real-world entities and make simulations, a fundamental step, to analyze and control their behavior, amounts to smart visualization interfaces [180].

To this date, XR immersive analytics could be exploited, since allows users to visualize relevant immersive and spatially configurable information in a physical space (AR/MR) or in fully immersive settings (VR) [28, 195]. While many works focused on VR, AR/MR holds the main potential for industrial application, to empower workers with spatial information anchored to real-world objects, showing information related to it (i.e., situated visualizations) [413, 199]. It is worth noticing that, to provide a situated visualization, a user must begin her/his extended experience in the physical world but then, s/he can switch among any immersive visualization modalities (defined here as XR situated visualizations) [28, 195]. To achieve this ergonomically, an XR headset should possess sensors to furnish services like simultaneous localization and mapping, object recognition, and marker-based registration [323, 452].

In the context of DT, a careful combination of XR-situated visualizations and AI paradigms could be adopted to empower users with immersive digital holograms, allowing them to both visualize and modify information related to a DT in any degree of reality. This will also provide new forms of HITL related to DTs: the creation of a digital layer based on the actions and the information provided by users on the DT for the benefit of other humans (i.e., collaborative intelligence [109]). Using this research setting as our pillar, we here take three different use cases to apply XR-situated visualizations, AI

models, and DTs to improve operator roles in industrial workflows.

Firstly, we introduce a novel framework, H-CLINT-DT, to inject XR paradigms and Human Collaborative Intelligence (CLINT) in DTs through XR paradigms. Such a framework is reality-degree agnostic, indicating that if a user accesses the information related to DT through, AR, MR, and VR, she/he would be able to visualize and interact with it, through an adaptive interface. In all such realities, the user will always be able to manipulate and add knowledge related to a DT, which could have positive effects on improving decision-making and contextual information heritage [114, 77, 40, 195]. This framework could be applied in a wide range of scenarios and was here contextualized in both academic and industrial use cases: (i) memory preservation through annotations for family album photographs and (ii) a productive setting, designing an instance of H-CLINT-DT for a medium-sized electrical engineering firm (RQ-2, RQ-4). In both cases, we built custom AR and VR interfaces, to apply the HCLINT-DT framework, validating them through a user study.

The flexibility of such a framework allows its partial application: one could modularly implement a visualization and annotation tool in one or more degrees of reality (e.g., AR, MR), but not all of them. To demonstrate this, (iii) we considered a wine industry use case, where we introduced a novel AR/MR approach to support operators in wine recognition and cataloging. We leverage DL-based Optical Character Recognition (OCR) and mobile AR to define Augmented Wine Recognition (AWR), which overcomes the limits of marker or marker-less CV approaches (RQ-2, RQ-4), relying on the text within the label instead of the label itself to recognize a wine (which represent the variable digitally twinned information). Such an approach requires careful integration of methods, algorithms, and technologies to produce an efficient system (as detailed in Section 4.5). In particular, we developed a novel textual search algorithm to recognize a wine typology by the text retrieved on its bottle label. Then, we fully match the HCLINT-DT framework, by implementing an annotation system on top of this recognition, which was

further applied to empower the detection itself.

It is worth noticing that, both the discussed systems include a Multimedia Information Retrieval (MIR) module to retrieve information, visualized in immersive XR interfaces (defined as X-MIR) [202, 239, 373].

To summarize, this chapter introduced three different application systems that define novel interfaces and processes to support users in MIR through XR paradigms, while investigating the role of DT to empower such systems (RQ-2, RQ-4). Summarizing, the contributions of this chapter amount to [365, 358, 10, 370]:

- A theoretical discussion on how Human Collaborative Intelligence and eXtended Reality visualizations and annotations could be jointly injected into Digital Twins, highlighting the importance of Human-in-The-Loop approaches in Digital Twins infrastructures. The resulting design of the H-CLINT-DT framework exploits Human Collaborative Intelligence to improve knowledge sharing and heritage regarding Digital Twins, through eXtended Reality annotations;

- The contextualization of H-CLINT-DT in cultural heritage, considering the Digital Twin of a Family album pictures in which information can be accessed and integrated from both physical and virtual worlds. The developed system required the design and implementation of both Augmented and Virtual Reality interfaces, further validated through a user study;

- The contextualization of H-CLINT-DT, in a real industrial scenario, implementing a Mixed Reality system based on the Hololens 2 to optimize the time spent by workers in accessing and annotating information of Digital Twins of electrical circuits. In particular, we adopted a Lean Manufacturing perspective, concentrating on increasing at the same time worker's *Empowerment*, *Communication*, and *Training*;

- The implementation of the Augmented Wine Recognition (AWR) system, an AR app, based on deep learning OCRs, to recognize wine

typologies by the text lying on the their bottle labels. This system exploits a novel hierarchical textual search guided by wine domain knowledge. The AWR assessment employs a textual database of 2,426 wines belonging to the Italian Emilia-Romagna region (provided by Image-Line S.r.l.). AWR exhibited a performance of 91% in terms of recognition accuracy and an average of 2.37 seconds in terms of inference time, showing that this system may be acceptable from a user perspective. We also compared our solution to a naive one, built ignoring wine domain knowledge, showing that AWR can drop the wine type recognition time by two orders of magnitude. An HCLINT-DT adaptation for AWR introduced a novel annotation system to store images of wine bottles in a database that was further used to develop a Deep Learning-based image retrieval service for AWR, that could be used with the existing recognition service in a hybrid setting.

In the following, we present the most relevant works related to the area of research so far described.

4.3 Related Works

4.3.1 H-CLINT-DT: Inject Human Collaborative Intelligence In Digital Twins Through eXtended Reality

DTs are increasingly adopted in research and industry [385, 301, 338, 107, 256], aiming at replicating, twinning, or mirroring some physical entity. Interestingly, in this context, a central role is progressively being earned by XR paradigms, thanks to which it is possible to manipulate DTs, directly influencing the physical world and vice versa [301, 338, 256]. A guide in such space can be found in [413], where the authors focus on reviewing AR/MR remote collaboration contributions to physical tasks. Among the future research issues they sketch for this domain, they mention multi-modal interac-

tion, hybrid interfaces, and industrial AR-based remote collaboration, also citing DTs. In the following, we will focus on the research that falls closest to ours, using XR paradigms to maintain and sustain the growth and sharing of collaborative intelligence.

In [125], the authors proposed VirCA, namely the Virtual Collaboration Arena, to implement a complex vision by adopting a shareable and fully customizable 3D virtual workspace as a central idea. They aim to enable users, not necessarily co-located, to collaboratively create ideas and then design and implement them in a shared virtual space. The authors of [9], instead, focused on providing a surgical telementoring tool where annotations are superimposed directly onto the surgical field using an AR-simulated transparent display. With their system, annotations stick to the surgical field as the trainee display moves, and the surgical field deforms or becomes occluded. Finally, in [253] an on-spot technician-manufacturer remote maintenance tool is proposed. The system can record the malfunctions reported by end-users and provide, using an AR display, the maintenance actions suggested by an expert.

Concerning these contributions, we here concentrated on how humans could help others, adopting a Human Collaborative Intelligence (HCLINT) approach that assigns a central role to DTs. An HCLINT could help humans in supporting other humans in their activities. As reported by [67], HCLINT *involves an extensive collaboration of different team members to solve problems while giving a non-stop real-time learning opportunity.* Following this line of thought, we also resorted to a well-known knowledge transfer strategy commonly adopted by humans: asynchronous, persistent annotations. However, we moved a further step in this process: in our contribution, we support providing and sharing human annotations made of text, voice, or videos aligning both the physical and the virtual worlds utilizing XR (AR + VR in our case) paradigms. This amounts to a unique way to make humans learn from each other, fostering efficiency. Indeed, providing seamless exploitation of annotations provided via AR and VR paradigms can empower

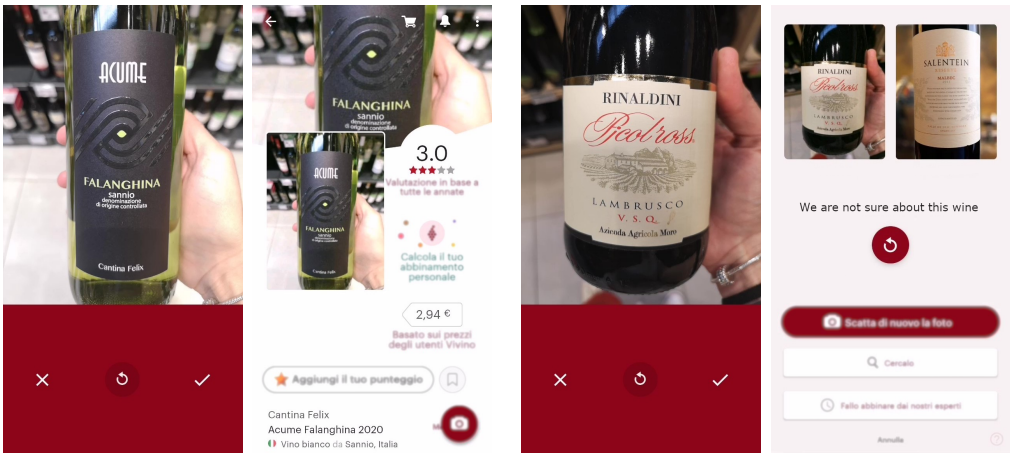
those who interact with the physical object, and receive feedback from those who exploit its twin and vice versa. Hence, when compared to the works cited in this Section, ours is the only one that aims at designing an HCLINT module to empower DTs through annotation while maintaining contact with both the physical and virtual space using XR paradigms. In addition, we here pursued a user-centric approach by providing an assessment adopting two different strategies, one using Technology Acceptance Model (TAM) constructs and a second conducting a short-term observational study within a manufacturing plant [438, 103]. In Section 4.4, we propose supporting such a collaborative environment through XR paradigms.

4.3.2 AWR: An OCR-based AR System to Recognize Wine Typologies From Bottle Labels Text

Food and beverage product identification is often performed with bar codes or QR codes. Many AR applications exploit image detection and recognition paradigms, which may also be based on codes or the recognition of a product as it appears [349, 279, 376, 322]. Identifying and exploiting visual cues in food product pictures is an approach that appeared in various research contributions [146, 167, 212, 450]. In this scenario, some works, from both industrial and academic contexts, focused on wine label recognition and its application in the AR realm [388, 404, 128, 424, 49, 186, 5, 206].

Regarding commercial solutions, *WineEngine* is an online wine label recognition service [388] which exploits a combination of OCRs and image-retrieval-based approaches using the wine bottle front label. This approach requires adding reference label images to the considered database and does not provide an AR interface. Another interesting system to recognize wine bottles is *Living Wine Labels* [214]. It also uses image retrieval to recognize the front label of a particular wine bottle and subsequently present customized AR animations. Mostly used for story-telling purposes, it supports eleven brands and requires a database of images for each of the different front-label bottles. Finally, Vivino is the most downloaded app with a com-

munity comprising 20 million users around the globe [404, 290] and provides features such as wine exploration, evaluations, and a wine bottle front label recognizing service. Vivino does not provide an AR interface and implements an image retrieval approach based on the Vuforia Cloud Recognition service that compares incoming front-label scans uploaded by a user to the ones stored in a custom database, to discover the closest match [290]. Given the high number of downloads and its large community, and considering that the approach it adopts is entirely based on image retrieval, we performed a simple experiment with Vivino to assess its performance in an everyday life scenario. We visited a local supermarket and tested its performance with 60 bottles of wine: 47 were correctly recognized (78% accuracy) taking an average time of 2.05 seconds (with a standard deviation of 0.65). Examples of wine labels correctly/wrong detected are reported in Figure 4.1.



Correctly recognized. Not recognized.

Figure 4.1: Examples of wine recognition with Vivino.

Considering now academic contributions, mostly have followed image retrieval-based approaches [186, 128, 261, 424, 5, 206, 49]. In [128], the authors implemented a front-label recognition method computing SURF key points and label descriptors and comparing such descriptors to pre-computed ones in a label database to search for a match. Similar approaches can be found in [186, 424]. In [206] the authors proposed a CNN-SIFT framework

for wine label retrieval, where a trained CNN model recognizes the wine producer to narrow the search range, while a SIFT descriptor empowered with RANSAC and TF-IDF mechanisms matches the final sub-brand. In [5], the authors presented an AR system running on a Microsoft HoloLens, making use of the Vuforia SDK to recognize markers attached to wine bottles and to display information concerning those bottles [408]. It is also possible to find other approaches in literature that concentrate on recognition sub-problems. In [261], for example, the authors concentrated on a preliminary step, a region of interest extraction method (GrabCut algorithm) for front labels, that may serve subsequent ones such as image analysis, recognition, and retrieval. All of the aforementioned academic contributions rely on image-retrieval-based approaches, and so present the main limit of requiring an extensive image database, which may be very difficult if not impossible considering old, out-of-production, or new wine types (i.e., long-tail samples). Differently, [49] implemented an OCR-based solution to read serial numbers from wine labels to provide counterfeit prevention and brand protection. However, this would be required to have access to all the correspondences between serial numbers and related bottle wine types.

References	IR	AR	AReT	OCR	TDO	LoTE
[388]	✓	✗	✓	✓	✗	✗
[214]	✓	✓	✗	✗	✗	✗
[404]	✓	✗	✗	✗	✗	✗
[128, 186, 261, 424, 206]	✓	✗	✗	✗	✗	✗
[5]	✓	✓	✓	✗	✗	✗
[49]	✗	✗	✓	✓	✓	✗
AWR	✗	✓	✓	✓	✓	✓

Table 4.1: Comparison between the characteristics of the different wine recognition systems and AWR.

Differently from the presented related works, AWR entirely relies on text, as it solely employs an OCR and a custom database to recognize a wine type

from the text reported on the back label of a bottle. Table 4.1 compares the characteristics of our solution against existing ones, where IR stands for Image Retrieval, AR for Augmented Reality, AReT for Almost Real-Time, OCR indicates the usage of an OCR, TDO for Textual Database only, and LoTE for Long-Tail Extensible.

4.4 H-CLINT-DT: Inject Human Collaborative Intelligence in Digital Twins Through eXtended Reality

The versatility of DTs has led their applications in diverse fields, from industry to healthcare [385, 107]. In [384], authors explored a novel concept of DT shop-floor, discussing four key components, including physical shop-floor, virtual shop-floor, shop-floor service system, and shop-floor DT data. They aimed to find a convergence of the manufacturing physical and the virtual worlds to realize smart interconnections, interactions, control, and management. XR paradigms may add depth to such discussion with the advent of the Metaverse ecosystem and an increased offer of applications [259]. In particular, recent works have already explored the effectiveness of the combination of DTs and VR models in several different application domains [256, 301, 338]. AR techniques have also been put to good use to visualize DTs, assisting users in their everyday activities [333, 451, 291]. It is worth noticing that most of such works focused on the DTs of humans and objects and their manipulation through VR paradigms, directly influencing the physical world and vice versa. However, a more limited number of works have focused on the role of Human-Machine Interactions (HMIs) in such models.

HMI focuses on finding natural ways to communicate, cooperate, and interact between humans and machines. In this domain, it is worth citing [185] who introduced a DT-empowered multi-modal UI framework to adapt assistance systems to different environmental conditions and human workers. In

such a proposal, the DT of a human served the purpose of modeling in a systematic and fine-grained way specific human abilities, peculiarities, and preferences. In [222], instead, the authors explored the adoption of DTs in the product lifecycle of an HMI, including its design, manufacturing, and service. Nevertheless, despite the progress made in automation, sensing, and automated learning, today it is widely accepted that *"machines cannot fully replace the unique perception and communication skills of humans"* [139]. None of such works, to the best of our knowledge, has considered leveraging the potential of XR and DTs technologies to deliver into HMIs a primary component of any system, the collaborative intelligence created and molded by humans while involved in given activities. The focus of this work is hence to consider the dynamics of workers interacting with physical objects and virtual models from a Collaborative Intelligence (CLINT) perspective. CLINT is not novel to the scientific community, and many different definitions have been provided depending on the specific perspective [149, 109, 421, 139]. To define our proposal, we started from the definition of CLINT introduced by [67]: *it involves an extensive collaboration of different team members to solve problems. It can provide more information for designing better solutions than any single member could while giving a non-stop real-time learning opportunity. Moreover, such collaboration has the potential of integrating diverse contributions (different members provide different information/knowledge, skill, and experience to a problem resolution) into a platform to produce a creative solution for successfully solving a problem.*

Based on those considerations, we focus on the problem of embedding a Human Collaborative Intelligence (HCLINT) paradigm within a DT using the most ergonomic way to interact with digital information: XR paradigms-[195]. To this date, we present a CLINT-empowered DT-based framework for HMIs where the DT of a process can integrate human annotations (e.g., textual and vocal annotations), called HCLINT-DT. The role of this framework is not only to support the exploration of information but also to allow the creation of an all-in-one-place resource to preserve human knowledge.

The flexibility of such a structure depends on how human annotations are gathered, retained, and accessed. To simplify such operations we resorted to XR paradigms, defining VR and AR interfaces to annotate a DT, in both the virtual and physical space (e.g., annotating objects and actions performed in the virtual or physical world). Such a framework supports asynchronous cross-reality access to such information [332]. In other words, those who work on the physical object may benefit from feedback and advice received from those who have worked with their digital counterpart in the virtual realm and vice versa. An additional feature consists of visualizing a particular step of the process history, using vocal or visual markers that act as "revive" triggers.

An assessment of the HCLINT-DT framework could be performed considering any scenario involving physical objects/processes and their DTs, where annotations are cross-provided, stored, and accessed. Such a type of scenario is very general, as it spans from industrial to commercial and household contexts. To demonstrate such flexibility, we here provide two kinds of contextualization, implementation, and assessment for this framework.

Firstly, we considered a specific use case: memory preservation through annotations for family album photographs. Previous works have experimented with a HoloLens 2-based interface and Deep Learning (DL) paradigms for easy annotation and cataloging of pictures to revive the phenomena of exploring families' past [360, 363]. Such works amount to the first step of the HCLINT-DT framework: the annotation one. We expanded such works by developing a DT formed by the album's cyber-physical counterpart, along with the annotations made by the subject and the subject itself, all in the virtual realm. It is interesting to note that this use case shares traits common to other settings.

For example, in manufacturing, workers often learn how to carry out an unknown process either by studying or by accessing the experiences shared by others [133]. The opportunity of anchoring the annotations to a physical object or its twin can foster the latter. This is not the only positive aspect

though. Factors that influence employee involvement are, according to [230]: empowerment (sharing power with employees and increasing their level of autonomy), training (developing workforce shared abilities and a better understanding of the processes in which they participate), communication, and remuneration. To assess whether a framework like HCLINT-DT could contribute to any of such factors in an industrial productive setting, we also performed an on-site observational analysis of a small and medium-sized enterprise in the business of building electrical systems. The analysis resulted in the implementation of a professional tool used in the company with which we collaborated, named AnnHoloTator.

4.4.1 Framework Design

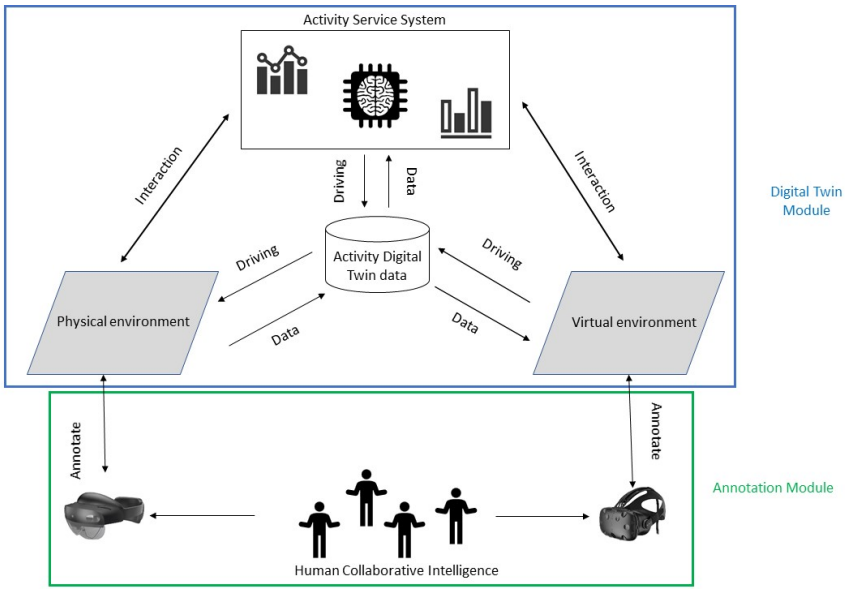


Figure 4.2: Main components of the HCLINT-DT framework.

In this Section, we describe the framework that aims at exploiting Human Collaborative Intelligence (HCLINT) to empower the five-dimensional Digital Twin (DT) model introduced in [384], as depicted in Figure 4.2. For the sake of clarity, in Table 4.2 are reported the principal acronyms, along with the corresponding full-name, used in Sections 4.4.1 and 4.4.2.

Acronym	Full-name	Acronym	Full-name
ADTD	Activity Digital Twin Data	**HCLINT**	Human Collaborative Intelligence
ASSYST	Activity Service System	**AM**	Annotation Module
PE	Physical Environment	**XR**	eXtended Reality
VE	Virtual Environment	**AR**	Augmented Reality
DT	Digital Twin	**VR**	Virtual Reality

Table 4.2: Table of acronyms along with their full textual description.

4.4.1.1 Digital Twin Model

The five-dimensional DT model adopted in this work builds upon the framework presented in [384], comprising the following components: (a) a Physical Environment (PE) that includes a series of entities, such as humans, machines, and documents; (b) a Virtual Environment (VE) consisting of models built in multiple dimensions, including geometry, physics, behavior, and rules, evolving according to the PE, and the (c) Activity Service System (ASSYST). ASSYST amounts to an integrated service module, which encapsulates the functions of data analytics, models, algorithms, etc., into sub-services, and combines them to form composite services for specific demands from the PE and the VE. Finally, (d) the Activity Digital Twin Data (ADTD) includes the PE, VE, and ASSYST data, their aggregations, and the existing modeling methods, optimizing and empowering both the PE and VE. The data in ADTD communicates in real-time with all other modules to eliminate possible islands of information to provide a comprehensive, synchronized, and consistent vision.

4.4.1.2 Annotation Module

The Annotation Module (AM) emerges as a companion to the five-dimensional DT module. The AM acts as a plug-and-play module that exploits HCLINT-DT together with XR paradigms. In practice, each user that interacts with the PE or the VE can read or produce annotations. Such annotations can originate from both AR and VR. In the first case, a user wearing an

AR head-mounted display (e.g., the Hololens) can: (a) individuate those real-world objects that also possess a cyber-physical counterpart in the VE, and (b) provide annotations, with the production of multimedia content (e.g., textual, visual and vocal data). The ADTD and the ASSYST modules process and replicate such annotations, making them available to the corresponding elements in the VE. The VR setting offers instead the possibility to directly provide annotations in the VE, exploiting a VR head-mounted display. The updates shared in VR are processed by the ADTD and the ASSYST modules to export such annotations to AR. The proposed AM is agnostic concerning the specific type of DT (object, human, or process).

To visually depict the operation mechanism of the AM, and how it interacts with the reference DT model, we introduce Figure 4.3. As illustrated, the AM works in three stages: before, during, and after inserting an annotation. In this Figure, the blue, purple, yellow, and green blocks represent PE, VE, ASSYST, and AR/VR interfaces, respectively. Their operations and interactions are supported by the ADTD.

The system identifies objects in the user's view through a snapshot captured from either the PE or the VE, depending on the chosen AR or VR interface. An object recognition service, powered by the ASSYST module, utilizes DL-based detectors for AR or examines metadata for VR to recognize and highlight objects. For AR, a photo is taken from the camera while for VR it includes the set of objects that lie in the viewing frustum. An object recognition service, included in the ASSYST module is invoked, and, in the case of AR, employs one or more DL-based object detectors to identify known objects, along with their spatial coordinates, while for VR, the recognition consists of examining meta-data associated with the set of considered objects. Subsequently retrieving related annotations and displaying an interactive menu upon user interaction.

Now, a given user could decide to provide a new annotation or visualize one of the existing ones (the annotations could be many). In the first case, the user exploits virtual keyboards or recorders based on the type of annotation

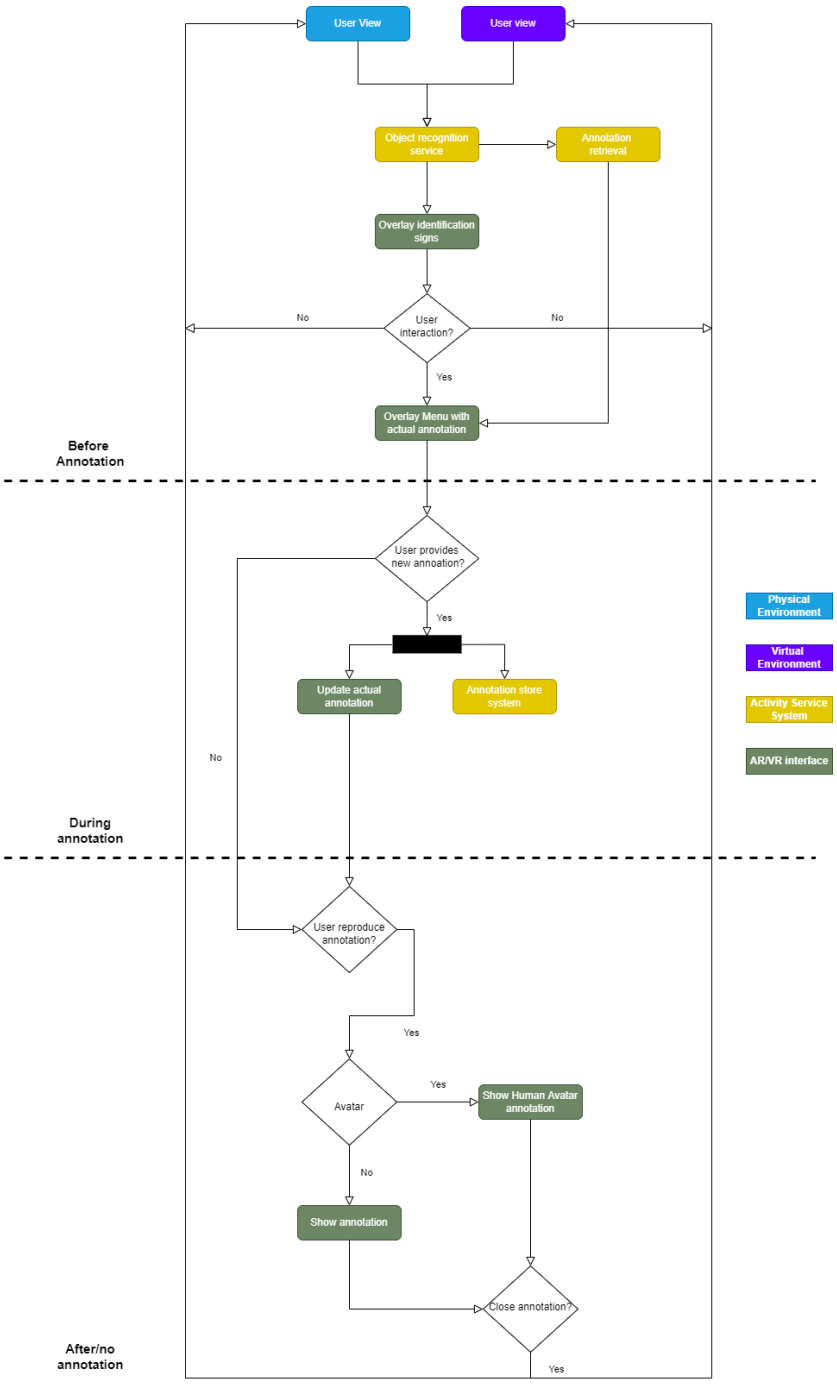


Figure 4.3: HCLINT-DT workflow: interactions between AM and the DT model.

(textual vs. vocal). In the second, the ASSYST module stores the new annotation and the identity of its author. Finally, further annotations could be either provided or reproduced.

When the user decides to reproduce an actual annotation, it may simply visualize a textual item or pick a more natural way of acquiring the information inside. To this aim, a user can listen to the contents of a given annotation directly from the avatar of the person who shared it. After this step, the process loops from the start.

4.4.2 Experimenting with HCLINT-DT: A Use Case For Family Photo Albums

In this Section, we provide a practical use case for the HCLINT-DT module: the DT of a family photo album and annotation process. Family photo albums provide an unrepeatable chance to revive old memories about social events, affections, relatives, friends, special events, etc. [393]. Throughout the 20th century, people printed photos and collected them in family albums. Despite the spread of digital photography and social media, people still look back and discover their families' pasts, often sharing think-aloud thoughts and memories that typically survive for the time of the conversations where they were exchanged [326, 360, 363]. Digital technologies may provide viable ways of retaining memories, annotating, and reviving such elements in an easy and meaningful way. Applying the HCLINT-DT framework it is possible to experiment with its ability to store and provide relevant annotations. Often, when a photo is placed in an album, a few annotations are written on its back. Those who will browse those pictures afterward will be able to discover what that picture portrayed thanks to those annotations. This amounts to a typical example of HCLINT, where the production and consumption of information occur collaboratively. We designed the DT of the family photo album, applying the HCLINT-DT module and implementing a two-sided AR and VR application for annotation sharing.

4.4.2.1 AR Interface

The AR interface resorts to models developed in previous research [360, 363], where a system was developed for the digitization and cataloging of collections of family album photographs exploiting the HoloLens 2 [394] as a wearable device, and DL models to catalog family album photos. In particular, we fine-tuned a well-known object detector, YOLOv5 [392], to identify the pictures within a given user's view. Such pictures are then classified according to socio-historical labels, using the IMAGO models provided in [367]. These models provide the prediction of the date and the socio-historical context of an analog family album photo, respectively. We decided to extend the AR system introduced in such work, taking YOLOv5 as a picture detector and adding a module to identify which pictures are already in the ADTD [221]. In this way, it is possible to recognize the pictures that are included in the database of the DT of the photo album and retrieve the associated annotations. Hence, all models run as part of the ASSYST module defined in Section 4.4.1. The AR interface comprises three modules: picture detection and matching, annotations retrieval, and interaction menu. Firstly, the AR interface captures the user's view to individuate any family album photos using object detection and classification framework introduced in [360], to predict the bounding box coordinates of the different pictures in the scene and crop them accordingly. However, additional work is required to implement the feature that allows users to match the detected pictures to ones that were already scanned and annotated. To this date, once an image is identified, its crop is passed to a feature descriptor algorithm, namely SIFT (Scale Invariant Feature Transform) [218]. SIFT is used for detecting and describing local features in images, designed to be robust to various transformations such as scaling, and rotation. It identifies key points or interest points in an image and extracts distinctive descriptors around those points. Those are used to define a database of feature descriptions and then used to match incoming images. The workflow appears in Figure 4.4. In any case, once the recognition and retrieval steps are completed, the user can revive

the annotations made by previous users on the considered photos (if present) or add new ones.

Figure 4.4: Example of SIFT execution, ranking, and subsequent database search for annotation retrieval.

The AR menu was thought to be as simple as possible to let the user focus on pictures, without being distracted. Such a menu is reported in Figure 4.5.

Figure 4.5: AR interface main Menu.

Recalling that, in the previous step, the bounding boxes of pictures were obtained, it is possible to create an invisible interaction area, amounting to the rectangles enclosing the photos. Exploiting this method, when the user touches one of the pictures, the menu appears. At this point, the user could decide to write a new annotation or reproduce an existing one. In Figure 4.6

the procedure to create a new annotation is reported, considering the textual type.

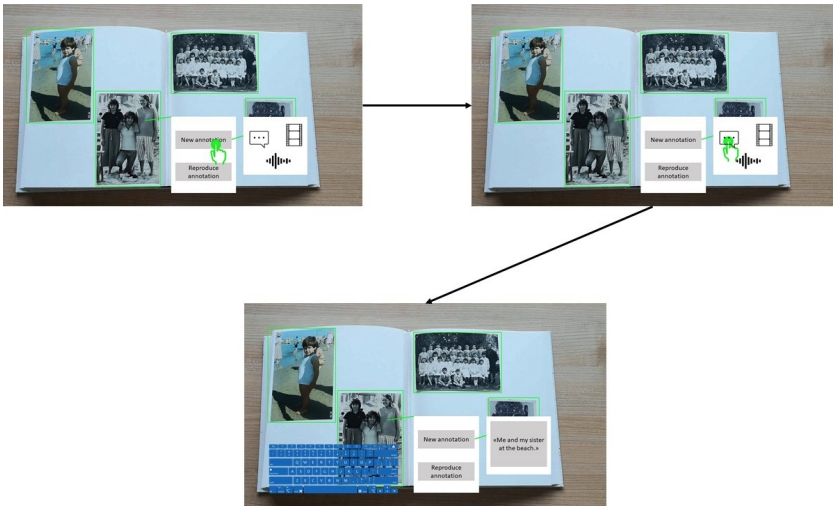


Figure 4.6: AR interface Menu: annotate a picture.

In practice, the user interacts with the menu by touching the *New Annotation* button and then decides if s/he wants to leave a textual, vocal, or video annotation. For each type of annotation, a different menu appears: considering the textual one, a virtual keyboard, used as an input device, will materialize next to the menu; for the vocal and video ones, the user could register a vocal note about a particular picture, associating its 3D avatar. The latter may be implemented by adopting Deep Learning models like the one provided in [183] and tools like [174] (included in the ASSYST).

All new annotations are immediately sent to the ADTD module to support their access to the VE while having the chance to reproduce the new or old annotations at the end of the process. In this case, as reported in Figure 4.7, s/he may select the *Reproduce Annotation*, and then the kind of annotations s/he wants to reproduce, finally picking one from the list.

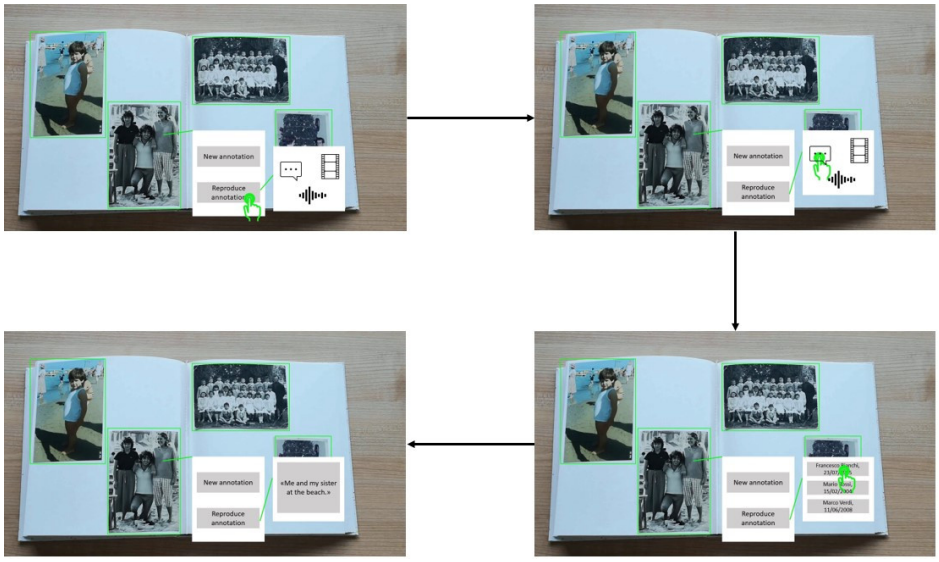


Figure 4.7: AR interface main Menu: user reproduces textual annotations of a picture.

4.4.2.2 VR Interface

The VR application carries out the same actions as the AR one with different paradigms, being in a fully immersive environment. All the following environments, objects, and interactions were developed using the Unity Game Engine [1] while using the Google VR SDK, to deploy it on mobile phones, possessed by the majority of the population [2]. All the user interactions were implemented through ray-casting, allowing the selection of items and actions by a simple point-and-click mechanism.

The environment that hosts the DT is minimal (visually reported in Figure 4.8). An empty room with a table, a chair, and a family album DT. The visualized pictures could be changed with a shift-like command as if the user were browsing a family album.

The main activity that the user could carry out is browsing the family album to revive memories behind the picture. The user could so click on one

[1] https://unity.com/

[2] https://github.com/googlevr/gvr-unity-sdk

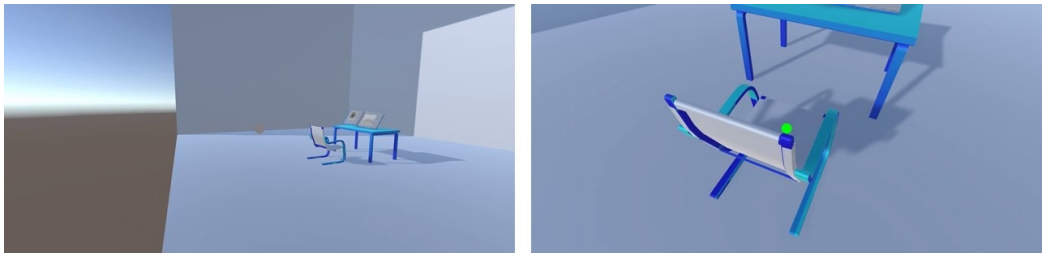


Figure 4.8: Initial VR user view and user seat.

of them, to let a menu appear. As for the AR interface, the menu provides the possibility to create a new annotation or reproduce already existing ones about that particular photo (as graphically reported in Figure 4.9).

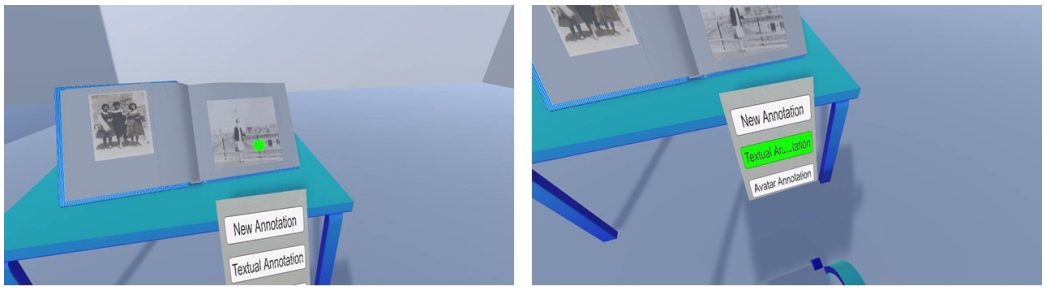


Figure 4.9: VR family album and Menu.

Figure 4.10: Textual annotation of a particular picture, reported in its original language.

As previously stated, the annotation could be textual or vocal/visual. Starting from the first, once the *Textual Annotation* button is selected, the

menu shows all the textual annotations made by other users. In case there is a unique annotation, this is directly reported on top of the picture, as depicted in Figure 4.10.

A user could also visualize a 3D Avatar annotation, which corresponds to a textual/audio one but reproduced with the 3D avatar of the user who created it (see Figure 4.11).

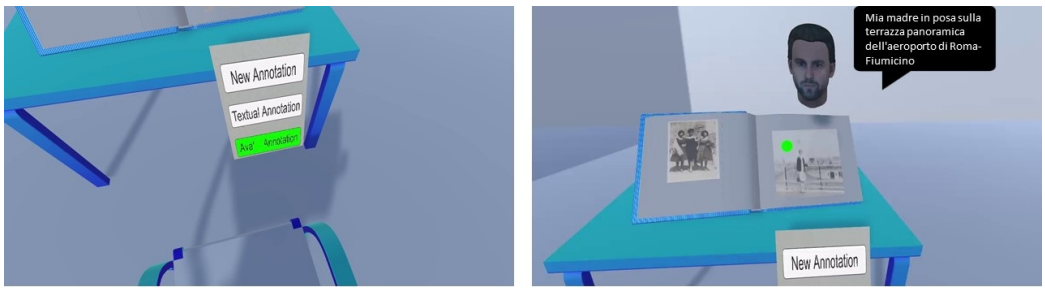


Figure 4.11: Reproducing Avatar annotation in VR.

Beyond visualization, the user could provide new annotations: as for the AR interface, s/he could provide a textual and a vocal/visual annotation following the same steps, but in VR. For example, considering the vocal annotation process, the user should select the *New Annotation* button and then the *Vocal mode*; a Sub Menu will appear as reported in Figure 4.12.

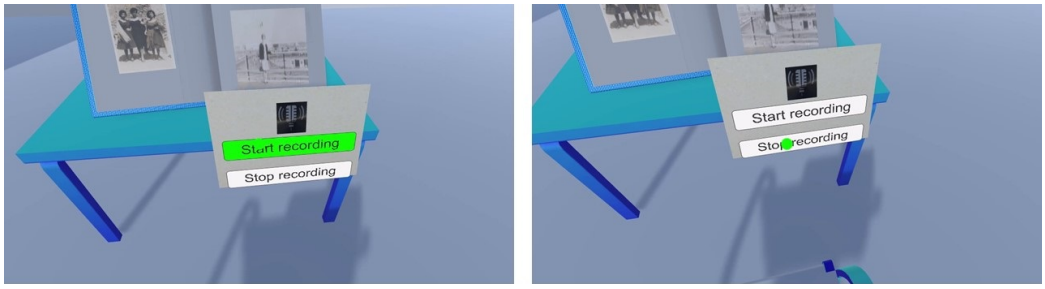


Figure 4.12: Vocal Sub_Menu and recording mechanism.

At this point, s/he could click *Start recording* while contextually talking. Once finished, the *Stop recording* button provides the possibility to save the vocal note. All the produced annotations are then sent to the ADTD

module, which will relay them to the ASSYST module and the PE (making them available to the AR interface).

4.4.3 Assessment Model and Results

We here report the results obtained with the assessments of the AR interface as a tool employed to insert annotations. The outcomes of such analysis were collected after the user experienced the interaction with the physical counterpart of the considered DT (i.e., a family photo album), through the app. The most important and so worth analysis use case for our framework is the one involving a user making and enjoying annotations on physical objects. Others should be able to easily access that information, even if not in the physical place, from the VE. The evaluation was performed with a survey in which the participants were first asked to watch a video of how the AR interface worked. The survey was administered to a group of 30 participants online during the coronavirus-19 pandemic. The group included a gender-balanced number of participants: 15 males and 15 females. The average age of the participants corresponds to 28.17 (S.D. $\pm$3.12), where the youngest and oldest were 22 and 37 years old, respectively. The number of participants has been chosen as a trade-off between the necessity of acquiring sufficient feedback data from a population and the time spent for the evaluation phase [323, 117].

4.4.3.1 Assessment Model

The assessment model was designed to evaluate: (a) the ease and intention of use of such an interface and (b) its usefulness/adoption. To design such a survey, we took inspiration from the TAM [88] used to measure user intentions in terms of their attitudes, subjective norms, perceived ease of use, and usefulness. From the TAM, we derived the following questions:

I1. I found the new interface easy to understand (5-point Likert scale);

I2. I would prefer watching an Augmented Family Photo Album instead of a normal one (5-point Likert scale);

I3. I appreciated the automatic identification of the pictures (5-point Likert scale);

I4. Would you use this AR interface to share your annotations? (Yes/No question);

I5. I enjoyed the overall experience (5-point Likert scale).

The I1 sentence aims at evaluating the easiness of use of the AR interface design; item I2 lets us understand whether the users prefer to live an augmented experience or a classical one. Item I3 aims to measure the usefulness of one of the main features of the AR interface: the automatic identification of pictures utilizing YOLOv5 and SIFT. Then, I4 serves the purpose of understanding how much our users want to use this AR interface to share their memories, and so the annotations of their pictures. Finally, through I5, we ask for a broad evaluation. As reported, four out of five items are measured with a five-point Likert scale, except for item I4. This item was formulated as a Yes/No question because we wanted to emphasize the direct intention of the users to use the annotation system.

4.4.3.2 Results

Item	I1	I2	I3	I4	I5
Mean	4.37	3.70	4.37	0.77	4.30
Std	1.00	1.06	0.93	0.43	0.88

Table 4.3: Mean and Standard deviation for all the considered items.

In Table 4.3 are reported the means along with the standard deviation obtained by surveying our subjects on the proposed items. The internal

consistency of the questionnaire was verified by adopting Cronbach's alpha index, which corresponds to the Kuder-Richardson Formula 20 (KR-20) in the case of binary choice questions, such as our I4 item. The returned Cronbach's alpha index corresponds to 0.751, which can be considered reliable ($\geq$ 70, as indicated by [383]). This is a valuable outcome, considering that some items could be divisive (e.g., I2 vs. I3). After checking the internal consistency and validity of our questionnaire, we report in Table 4.3 all the means and standard deviations obtained by surveying our subjects on the proposed items, along with the item response histograms in Figures 4.13 and 4.14.

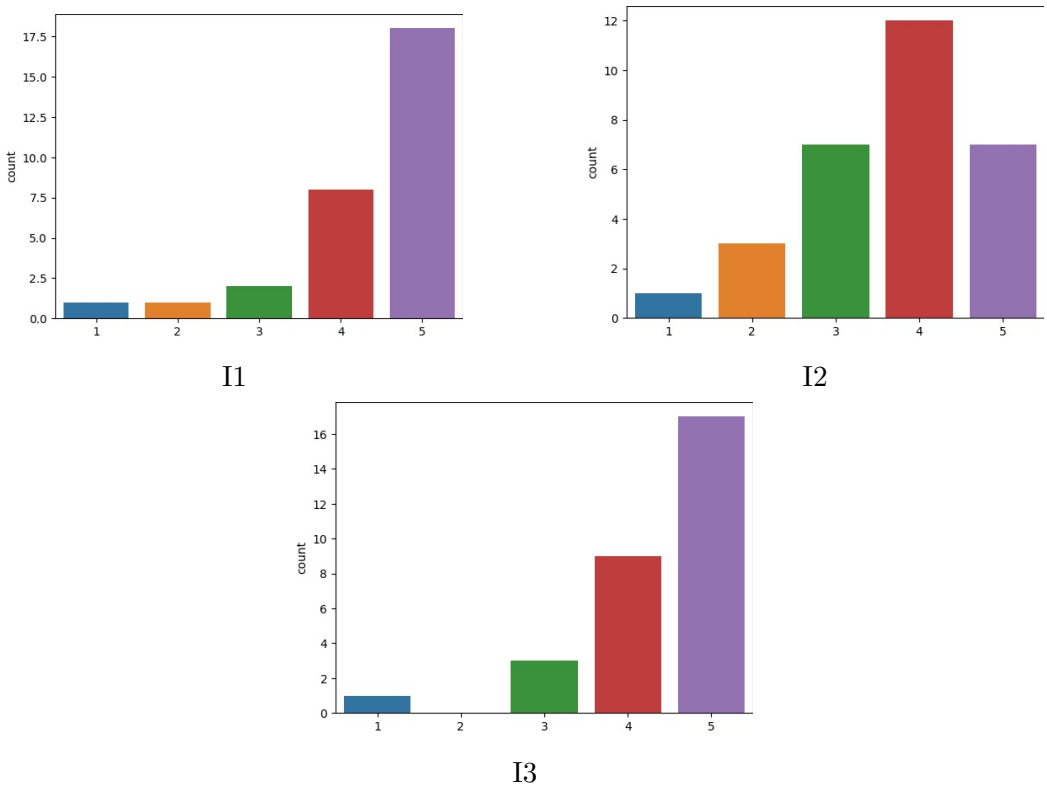
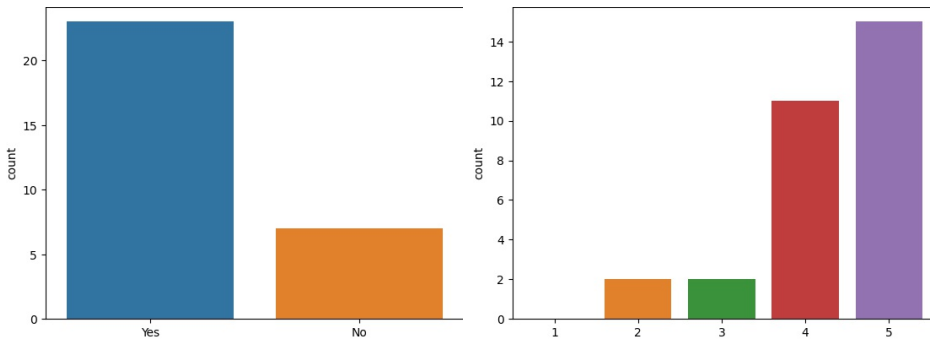


Figure 4.13: Histograms for answers in 5-point Likert scale items I1, I2 and I3.

Table 4.3 and Figure 4.13 highlight a general agreement about the ease of use of the interface design (I1), and a slightly positive outcome when com-



I4 I5

Figure 4.14: Histograms for answers Yes/No and 5-point Likert scale items I4 and I5.

pared to the use of modern technologies in the given application scenario (I3). However, this last outcome contrasts with answers for item I2, where we can only appreciate a partial agreement in the preference of watching a family photo album through the lens of modern technology. This is probably because some of the respondents continue to prefer reviving their old memories physically without filters. For what concerns instead of the usage of the proposed AR interface for picture annotations (I4), Table 4.3 and Figure 4.14 exhibit an agreement level that supports the initial aim of our application. Finally, scores from I5 highlight that all the subjects appreciated the AR interface.

To further confirm our results, and test the statistical significance of the obtained answers, we performed a one-sample t-test [132] over all the five-point Likert scale items. The one-sample t-test compares the mean to a hypothesized mean value as long as the samples follow a normal distribution. Since the sample size is thirty, it approximates the standard normal distribution according to the central limit theorem [197]. However, the classical two-tailed one-sample t-test does not highlight the direction of the difference between the sample mean and the hypothesized mean value. For this reason, we performed a one-tailed one-sample t-test in which the null hypothesis H_0 assumes that the mean (μ) is lower or equal to the fixed mean value, while

the alternate H_1 that (μ) is higher. The compared value was set as three considering the five-point Likert scale items since any value greater than three demonstrates a partial agreement in a five-point Likert scale. We set the parameter $\alpha = 0.05$ as the significative threshold for the p-value. The results of the performed test on all the considered five-point Likert scale items are reported in Table 4.4. From Table 4.4 it is evident that, for all the considered items, the null hypothesis H_0 can be rejected, considering both the significative threshold for the p-values, but also the critical t-values [374], thus confirming the analysis carried out so far.

Item	I1	I2	I3	I5
Value	7.4898	3.6329	8.0676	8.1199
P-value	1.48e-08	5.3e-4	3.4e-09	3e-09

Table 4.4: One-tailed one-sample t-test performed over all the considered 5-point Likert scale items. For each of them, the null hypothesis can be rejected.

For what concerns, instead, of the unique Yes/No question item (I4), we performed a Binomial Test returning the probability for the assumption that the observed frequencies are equal to the expected frequencies [409]. Also, in this case, we adopted a one-tailed perspective fixing the expected probability of the positive outcome to 0.6 (positive agreement). The null hypothesis H_0 assumes so that the probability $P(I4 = 1) \leq 0.6$, while the alternate hypothesis H_1 assumes that $P(I4 = 1) > 0.6$. We set the parameter $\alpha = 0.05$ as the significative threshold for the p-value. The returned proportion estimate value (0.77), and the p-value = 0.044, demonstrate the statistical significance of our test and so the positive outcome of Item I4.

It is worth noticing that all the questions provided to our subjects were very general: this choice was taken to evaluate an abstract scenario, that can be applied in any context in which a person is watching an object and wants to take/retrieve annotations on it. This can be seen as a simplified form of

any watch-and-annotate scenario, where the user can share but also retrieve annotations from other remote users, taking full advantage of HCLINT-DT.

4.4.4 AnnHoloTator: An HCLINT-DT Application Based On Industrial Observational Analysis

The HCLINT-DT may be generally adapted to different contexts. To explore this possibility, we carried out a short-term observational study at Elettrotecnica Imolese S.U.R.L. [112], whose mission is to produce industrial electrical systems, starting from their design to the final installation and testing. In particular, we observed and interviewed professional workers in their everyday activities to understand how HCLINT-DT could improve their processes from a lean manufacturing perspective [148, 230]. We concentrated on how to improve *Empowerment*, *Communication*, and *Training* factors using HCLINT-DT where the main worker collaboration activities take place. Professional workers engage in collaborative activities during the assembly, installation, and testing of electrical switchboards. Workers use portable computers to access and visualize documents describing electrical schemes and circuits during the assembly process. They make annotations on these documents (only one per electrical switchboard), noting missing materials or errors, and mark components they correctly assemble. These annotations are shared through cloud technologies with engineers who correct electrical schemes. This iterative process continues until the assembly phase is complete. The entire workflow is illustrated in Figure 4.15.

After the installation, the testing step takes place when the professional testers verify, following strict protocols, that the switchboards are correctly installed, analyzing each of the components and their interconnections. For each component, information is annotated on the same shared document with positive or negative checks. Regardless of the outcome (success or failure), the annotated documents are shared with the assemblers.

Considering this context of use, the HCLINT-DT framework, and the three factors that most influence lean manufacturing, we designed a lean

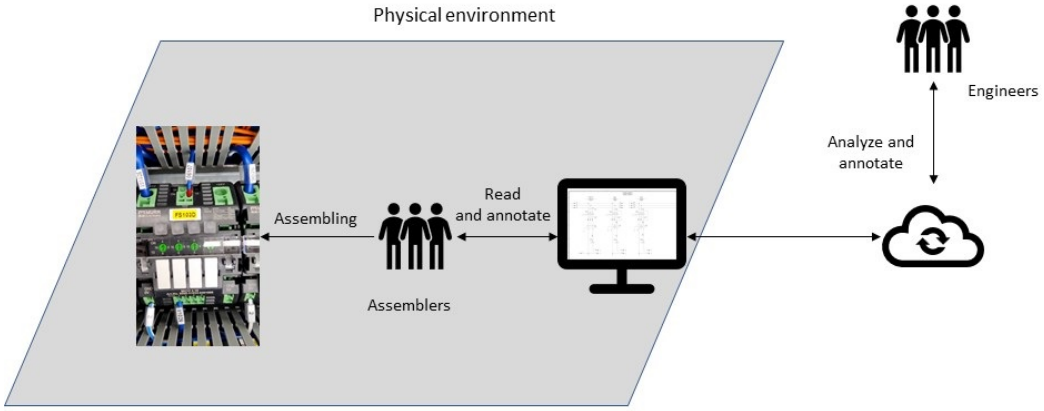


Figure 4.15: In the assembling phase, assemblers read and annotate shared documents of electrical switchboards and engineers eventually analyze and correct errors or suggestions made by the assemblers.

manufacturing optimization process, taking inspiration from the model provided by [148]. Firstly, we identified the weaknesses of the existing process. We found out that the main one amounts to the superfluous time spent by the assemblers while: (a) physically moving from the portable PC to the physical switchboards and (b) searching for the correct page in the document to read the instruction or annotate a particular components and its connections. We then sketched the implementation of the HCLINT-DT to erase these time-wasters by simultaneously improving *Empowerment*, *Training*, and *Communication* factors.

In practice, we started defining the DT of an electrical switchboard by considering all its components and interconnections. As for the considered family album photos use cases (described in Section 4.4.2.1), an AR interface should recognize each component in the real world, providing the possibility to make or read annotations by different users. In this case, however, the read/write annotation process would be mediated by digital documents (e.g., PDFs). This means that, when a component is recognized, the document pages containing relevant information will be visualized in the user's view and manipulated by employing AR paradigms.

In the virtual realm, instead, an exact reproduction of the switchboards along with all the annotations posed by different figures is defined. This approach should improve on the existing process. Firstly, the AR interface would run on an AR device like the Hololens, so the assemblers could keep going with their work without moving from the electrical board to the personal computer. Secondly, object recognition would prevent wasting time searching for the instructions/annotations regarding a particular component by immediately visualizing all the info in the shared document. The adaptation of the HCLINT-DT framework for this use case is visually reported in Figure 4.16.

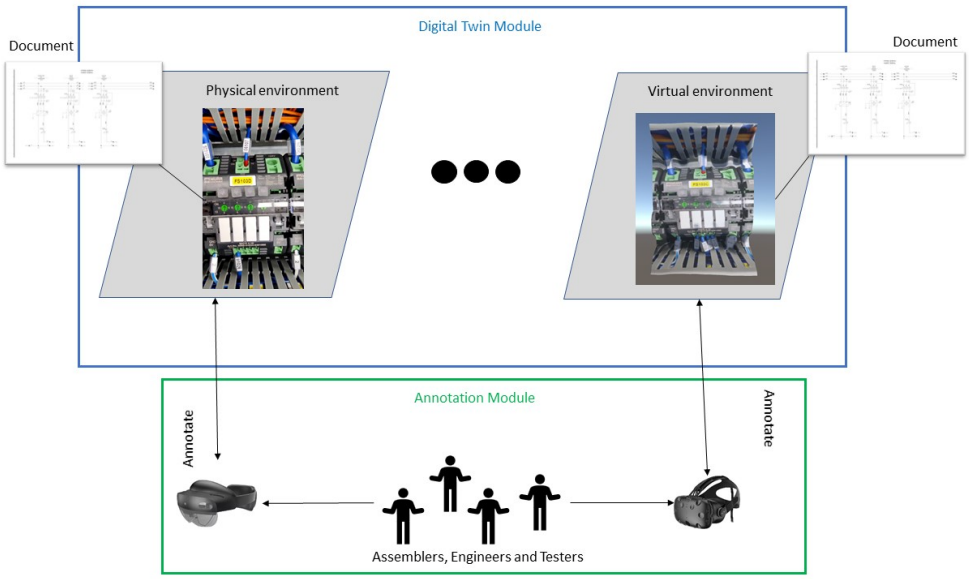


Figure 4.16: HCLINT-DT framework adapted to the Elettrotecnica Imolese use case: physical components of electrical switchboards are recognized giving the chance to read or write annotations that are mirrored in its digital twin in the virtual space, which also provides the same possibilities.

In addition, thanks to the adoption of the HCLINT-DT framework, the workers' *Empowerment* is provided with the use of state-of-the-art tools and XR devices (e.g., HoloLens 2 and HTC-Vive), as confirmed by the assemblers, engineers, and testers which were interviewed during our analysis. The

Communication improvement is given by the natural working mechanism of HCLINT-DT shared annotations. Finally, *Training* is a natural consequence of the same HCLINT-DT shared annotation mechanism since workers could analyze previously annotated documents to learn from others' experiences.

All of these design factors were put to good use to implement AnnHoloTator, which was designed for AR/MR but can be easily ported to fully immersive settings [370] [3]. AnnHoloTator is an AR/MR collaborative platform designed and developed to provide a way for workers to browse, visualize, and annotate useful information from digital documents providing a high level of customization. The system provides two situated visualization mechanisms based on OCR [4] retrieval (HCLINT-DT compatible) and remote textual search. In practice, while wearing the Hololens 2, a user could retrieve information related to a particular item in the physical space (Figure 4.17a) and search for documents that contain that information. S/he could do the same while using a virtual keyboard. In any case, the user visualizes the resulting document as depicted in Figure 4.17b.

(a) Deep learning-based OCR applied to physical electrical board codes.

(b) AnnHoloTator: Key-Search Mode linked to the OCR retrieved code.

Figure 4.17: AnnHolotator views.

One of the key features of the platform is the ability for users to arrange elements in the virtual space following a situated visualization approach. This includes the ability to position elements anywhere in the space and

[3] A demo of the AnnHoloTator system is available here

[4] https://github.com/JaidedAI/EasyOCR

adjust their size to fit the needs of the specific use case. This level of customization allows users to create a virtual workspace that is tailored to their specific needs and preferences. For example, a user may choose to arrange elements such as wiring diagrams, instructions, and tools in a way that is most convenient for their workflow.

After positioning the retrieved documents, the user could also put annotations directly on a page of the document's hologram. Various kinds of annotations could be inserted within a document: *Drawing mode*, *Check mode*, *Uncheck mode*, *Text mode*. Each of the annotation modes shares the same interaction logic: using the Hand-Tracking provided by the Mixed Reality Toolkit of the Hololens 2, to identify when and where a collision between the user's finger and the virtual document happens to know the absolute position of the novel annotation. Any of the annotations could be saved, undo, and redo using the buttons in the top-right section of the document's hologram.

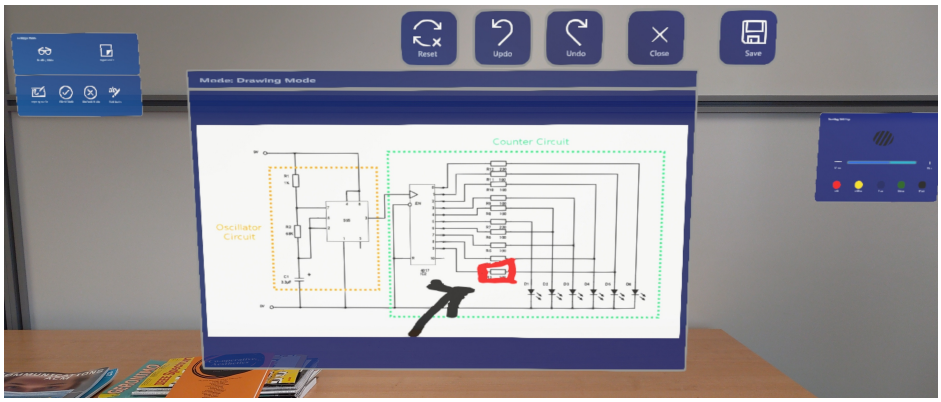
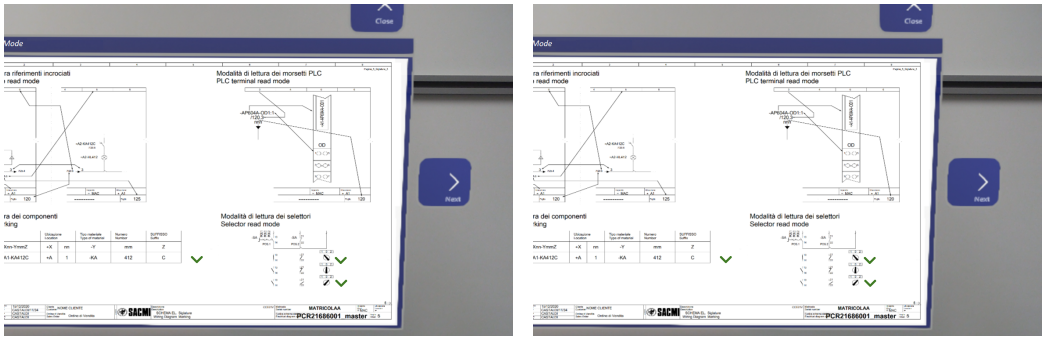


Figure 4.18: AnnHolotator: Drawing Mode Example

Drawing Mode By selecting "Drawing mode" from the menu a user can take free handwritten notes on the current page of the document by using his/her fingers. An example is shown in Figure 4.18. This type of annotation provides a high level of flexibility for the user to add notes, underline text, highlight sections, and even draw diagrams or other explanatory elements that enhance the information in the document. This could be put in combi-

nation with all the other modes to create even more meaningful annotations.

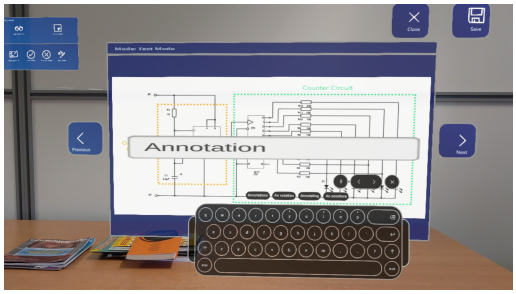
Check/Uncheck Mode The Check Mode (Uncheck mode) feature allows users to add affirmative (negative) annotations within documents by simply tapping with their finger where they want to insert the annotation. In particular, the application will spawn a 3D model for the related symbol overlaying it in the desired position (Figure 4.19). These annotations can serve several purposes, such as marking completed (uncompleted) tasks.



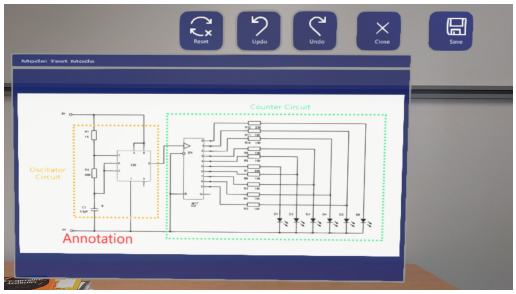
(a) AnnHolotator: Check mode. (b) AnnHolotator: Uncheck mode.

Figure 4.19: AnnHolotator check/uncheck views.

Text Mode Finally, the text mode allows users to insert textual annotations. With the same finger-tapping mechanism, a user chooses the location where s/he wants to put the textual annotation and the system spawns a virtual keyboard (as reported in Figure 4.20a). When the user presses the enter button, a textual graphical field will be overlayed in the virtual document. After the insertion, the user can move the text note to a different location within the document by finger-tapping (Figure 4.20b).



(a) AnnHolotator: Keyboard spawned in Text Mode.



(b) AnnHolotator: Textual annotation performed.

Figure 4.20: AnnHolotator check/uncheck views.

4.5 AWR: an OCR-based AR System to Recognize Wine Typologies from Bottle Labels Text

Situated visualizations (SV) and reality-based information retrieval systems aim at superimposing context-based digital information to real-world entities, such as food, people, buildings, photos [46, 44, 232]. Considering the advancements and affordability of AR, SV becomes viable in several domains, providing information regarding physical objects, and chaperoning a user through a specific process (e.g., learning, making a choice). In the food and beverage sector, for example, many initiatives have worked on scanning packages to visualize augmented information related to its contents (e.g., nutritional information, reviews) [273, 308, 376, 437, 404, 407, 62]. Such applications are built on top of marker or marker-less computer vision approaches for object detection, recognition, and tracking [349], which often require high computational power and the construction of large-scale datasets of reference images. Such approaches, however, may hardly scale with long tail products [161] as, for example, wines [375, 4]. In fact, in many cases, it may be difficult to acquire wine labels beforehand making them impossible to be recognized. For example, a winery may have changed its labels or stopped

the production of specific wines or there could be a shortage of pictures of old wine labels. It is very challenging to create a complete database covering the present and the past of wine labels.

Considering this scenario, Optical Character Recognition (OCR)-based techniques [45, 419] may represent a solution. In fact, given a wine bottle label, identifying and recognizing it by exploiting the text within, could allow us to overcome all the aforementioned limits. Nevertheless, this approach is practically challenging because of the label visual features, such as complex backgrounds, text of different fonts, and distortions due to the curved surface of bottles, which could mislead the OCR predictions (to minimize such events, we resorted to a DL-based OCR). Even with perfectly recognized text, an automatic process should be able to focus on only those words that can be used as cues to identify a wine. This step, alone, could require an amount of time not compatible with the implementation of a situated visualization. Therefore, such an approach requires careful integration of methods, algorithms, and technologies to produce an effective system in terms of detection efficacy and efficiency.

To this aim, we designed and implemented Augmented Wine Recognition (AWR), an AR system to automatically identify a wine type by recognizing and analyzing the text within its corresponding bottle back label. In particular, we tailored AWR on the Italian wine domain knowledge, considering that Italy is a wine top producer and follows European regulations regarding label organization and design [63, 4]. In particular, we leveraged the regulations regarding labels' textual information structure, placement, and content, to model a custom tree-data structure defining a hierarchy of textual features that discriminates different types of wine. Then, we defined a custom hierarchical search algorithm that explores such structure by matching and branching the tree itself, using as criteria the distance between a sought feature and the words found on a label. The process ends when the best wine-type candidates are found.

The AWR system [365][5], has been assessed employing a textual database of 2,426 wines belonging to the Italian Emilia-Romagna region (provided by ImageLine S.r.l.). AWR exhibited a performance of 91% in terms of recognition accuracy and an average of 2.37 seconds in terms of inference time, showing that this system may be acceptable from a user perspective. We also compared our solution to a naive one, built ignoring wine domain knowledge, showing that AWR can drop by two orders of magnitude the wine type recognition time. It is worth noticing that the AWR system could support humans in recognizing wine bottles but also experts in adding new items having at their disposal a base textual knowledge.

Finally, we also applied HCLINT-DT in the AWR system, by introducing an annotation system to store images of wine bottles in a database that was further used to develop a Deep Learning-based image retrieval service for AWR, that could be used with the existing recognition service in a hybrid setting.

4.5.1 Wine Domain Knowledge

A wine bottle usually includes two labels, a front and a back one. The front one is often devoted to brand communication, whereas the back one reports all the information characterizing a given wine, displayed according to its home-country regulations [63]. It is also possible to find wines bearing a single label: such a label will include all the required information in a compressed form. From now on, we will use the term "label" to indicate the ones that contain information to discriminate wine types, as required by Italian regulations [4]. According to Italian regulations, specific information (e.g., wine appellation, winery) must appear on a wine bottle in the same field of view (i.e., a consumer should not have to turn a bottle to read them all). Italian labels report different information, some mandatory and some not [55, 240, 403, 118]. A list of the most important ones is provided in the following (i.e., the attribute that was also used in our custom hierarchical

[5]An online demo visualization of the system, is accessible here

search algorithm):

Name: the wine name, typically found at the label top-center; **Type:** wine, varietal wine, appellation wine. In the case of appellation wine, this is related to the geographical area of production; **Appellation:** Appellation wines are classified as Protected Geographical Indication (PGI) and Protected Designation of Origin (PDO). Italian PDO wines fall into DOC or DOCG, now part of DOP. PGI wines are categorized as IGT, now included in IGP. DOP (European) signifies products reliant on a specific geographical environment for all production phases, while IGP designates products where at least one phase must occur in specific areas. Italian acronyms translated include DOC, DOCG, DOP, IGT, and IGP. It's mandatory to provide this information for appellation wines, allowing flexibility in choosing either the appellation (e.g., DOC, DOCG, IGT) or the corresponding European category (e.g., DOP, IGP). **Appellation Value:** In addition, the appellation value is the "proper name" of the class and it is unique for each wine type (e.g., Pignoletto and Romagna are the concrete appellation Values of wines categorized as DOC). The label should report the appellation field near the appellation value; **Effervescence:** still, sparkling, spumante. This information could appear with synonyms also: stationary, moved, etc. If nothing is specified the wine is assumed to be still; **Sweetness:** The terms change according to the wine effervescence. For still and sparkling wines *Secco*, *Semisecco*, *Abboccato*, *Amabile* and *Dolce* are used. For spumante, however, there are many more possible terms, including *Brut nature*, *Extra Brut*, *Brut*, *Extra dry*, *Sec*, *Demi-sec*, *Doux*. Such information is mandatory only for spumante wines. In addition, if the sugar content of the products justifies the use of two terms, the choice is up to the manufacturer; **Color:** red, white, rosé. It could also appear with synonyms: red/black, etc.

The information introduced to this point is valuable to identify wine types: *wine name*, *appellation*, *appellation value*, *effervescence*, *sweetness*, and *sweetness*, which are also the wine descriptors adopted in our system to discriminate among wine types. It is worth noticing that those features are

usually placed in the label top area using a font that is larger than the rest of the text [55, 240, 403].

4.5.2 AWR System Architecture

The proposed AWR system, visually depicted in Figure 4.21, includes two main components: (a) a client AR interface running on a mobile device, used to take pictures of the wine label and present AR content after wine type identification, and (b) a server that executes an algorithmic pipeline. The latter employs an OCR at two different stages to retrieve the text within the image sent by the mobile device and filter the relevant words retrieved by the OCR. These words are fed to a custom hierarchical search algorithm that skims a hierarchical textual database (textual DB) according to them, providing the best wine-type candidates. In the following, we will discuss each of the main components of such a system: AR interface, OCR module and retrieved text processing, Wine database, and Textual Search Algorithm.

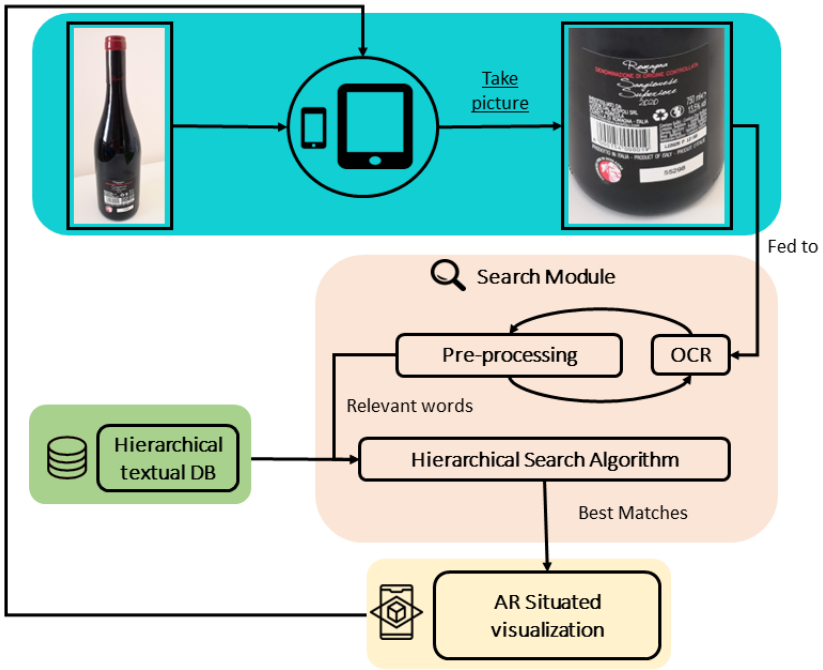


Figure 4.21: AWR system.

4.5.3 AR Interface

The client side of AWR has been implemented adopting an AR approach for Android-based smartphones with the Unity Game Engine and the Vuforia SDK[6] (Figure 4.22, which shows four different views of the AWR interface).

(a) (b) (c) (d)

Figure 4.22: AR interface: (a)-(b) Wrong and Correct results, (c) Correct confirmation, (d) After the confirmed identification, the scan stops.

Once activated, the AR interface starts to continuously scan what is framed by the device camera, collecting more and more frames. When a certain number of frames are collected, the system picks the less blurry one, by taking the frame which presents the smallest variance of the Laplacian, as reported in [23]. This frame is then used by the system to verify whether the camera is pointing at a known label. During this recognition process, a spinning loading icon appears on the screen's bottom left corner. Once a label is recognized, the interface shows the wine name, appellation, region, and region image (if available) related to the first query result. The interface also lists other possible candidates on its right panel (the here-introduced algorithm is a ranking one). If the right answer is present in this list, a user

[6] https://developer.vuforia.com/

can select it. In this case, the interface opens a dialogue box asking the user to save the selected result and, if confirmed, only the label related to that result is shown until the "Close" button is pressed. In case the back-end recognition service is not able to match the targeted wine because the related entry is not included in the textual DB, an alert is displayed. In addition, at any moment, the user can stop the interface scan through a toggle on the screen's bottom right corner.

4.5.4 Back-end Components

The system back-end (developed in Python 3.6 [7]) components include an algorithmic pipeline employing an OCR (detailed in Section 4.5.4.1) to retrieve the text within the image sent by the mobile device and a custom hierarchical search algorithm that skims a hierarchical textual DB based on the previously extracted words, providing the best candidate wine types. In particular, the OCR is involved multiple times in the text extraction step.

At first, during the cropping stage, the OCR is used as an object detector, to reduce the visual search area that encloses the most relevant words, taking advantage of Italian wine label regulations (Section 4.5.1): the most discriminative information appears with larger font sizes than any other text. Given this axiom, it is possible to define a variable bounding box that encloses all relevant text pieces. After this first stage, the OCR predicts the words within the bounding box. The retrieved text is then used to individuate within the hierarchical textual DB the path identifying the sought wine type. This is simplified with the use of a tree characterized by mutually exclusive paths. For example, if a wine is "dolce", it cannot be characterized by any other sweetness value, therefore the sub-trees not connected to "dolce" will be pruned. In addition, the tree size depends on the presence of specific values. For example, if "DOC" appears on a label, the same label should also report other mandatory information (e.g., harvest year).

Leveraging on such type of information, the search space for a given wine

[7] https://www.python.org/downloads/release/python-360/

type may shrink and its search running time may dramatically drop when compared to a naive approach (details in Section 4.5.5).

In summary, the algorithm that searches for the wine type in the hierarchical textual DB (detailed in Section 4.5.4.3) first individuates the possible relevant words using an OCR-based cropping and decoding approach. The algorithm then iterates starting at the first level of the hierarchical textual DB, executing a linear search and computing the best match according to a pre-defined textual distance (e.g., Levenshtein, Hamming, Cosine). The linear search compares the text with the term(s) stored at every given level of the hierarchy. In addition, each layer of the hierarchy is composed of a finite set of possible terms, and the algorithm picks the most probable one, which depends on the textual distance computed between the retrieved words and the words contained at the current tree level. Then, the algorithm proceeds to analyze the successive level of the hierarchy only after individuating the best match and after pruning the other branches.

4.5.4.1 OCR Module

An OCR algorithm is employed to find the possible area where relevant words lie and recognize them, which accuracy should be as high as possible since those retrieved words will be used by the textual search algorithm (Section 4.5.4.2) to identify the best match wine type. To this date, many OCRs are available in literature [343], and we picked an off-the-shelf DL-based OCR, namely EasyOCR [8]. The EasyOCR detection component employs the CRAFT algorithm [20], while the recognition model amounts to a Convolutional Recurrent Neural Network (CRNN) [340], trained with the same pipeline reported in [19]. Finally, the sentence decoding step utilizes Connectionist Temporal Classification (CTC) [144]. EasyOCR appeared particularly suited to our use case as its model is trained on images from heterogeneous environments (i.e., not only scanned documents) and was identified EasyOCR as the best OCR for images of true-to-life scenes by [346]. Easy-

[8] https://github.com/JaidedAI/EasyOCR

OCR retrieves as output the detected text and the relative bounding boxes encapsulating it, defined as coordinates in the original image space. It is worth noticing that, in all of our experiments, we executed EasyOCR DL models, by selecting those weights optimized to recognize Italian and English words from a custom vocabulary defined by its authors (release version 1.2.2). Qualitative examples of executing EasyOCR on two Italian wine bottles from the Emilia-Romagna area are reported in Figure 4.23.

Figure 4.23: Examples of EasyOCR retrieved words, without considering the confidence factor: all of the retrieved words are visualized.

This first experiment showed the positive and negative aspects of the adopted EasyOCR baseline: many relevant keywords were recognized, even with different kinds of fonts and backgrounds. However, some text was not detected or correctly transcribed. This can be due to several reasons: the text does not lie on a planar surface, color contrast varies between background and text, poor light conditions, or the words on the label weren't included in the ones used to train the EasyOCR decoder (e.g., IGP). In conclusion, EasyOCR may return values that are mistaken and do not hence support a direct querying mechanism. It may be possible to overcome such limits by

exploiting a fine-tuning approach to optimize the EasyOCR model for the wine domain. However, a dataset should be defined from scratch, labeling wine bottle label pictures with the coordinates of the relevant text and its corresponding characters, resulting in a costly procedure in terms of time and workforce. So, a different path was taken: we designed an algorithm that corrects wrong text predictions by leveraging pre-defined wine domain dictionary terms. To clarify the entire algorithmic pipeline, we named the process of predicting the words with the OCR module as **OCR words inference**.

4.5.4.2 Wine Database

The information (features) that could uniquely identify a type of wine corresponds to the wine name, appellation, appellation value, effervescence, sweetness, and sweetness (check Section 4.5.1). These features are not independent: a certain value of feature $i \in [1, n]$ defines a subset of all the other possible values for the other features $j \in [1, n], j \neq i$ where n is the total number of features. For this reason, wines are grouped firstly by one feature, like the appellation, and then sub-grouped based on each of the other ones, like the effervescence, and this process continues until one has grouped all the possible features, defining a hierarchical tree data structure. For example, both the *Lambrusco di Sorbara rosato* and the *Reggiano Lambrusco rosato* wine typologies have common appellation value (DOC), color (rosé), and effervescent (sparkling wine). In Figure 4.24, an example shows how to transform features from a table-like format to a hierarchical tree structure. This kind of transformation is used to convert our table-like DB composed of 2,426 wine types into a hierarchical one.

In particular, the hierarchical textual DB follows a classical Non-Binary Tree structure: each layer of the tree is composed of k nodes, one per each value of a particular feature (e.g., DOC, DOCG, DOP, IGT, IGP for the appellation). Recognizing a wine type means visiting the last level of the hierarchy, returning all the possible hits (i.e., more than one leaf). The considered database follows a specific nested key-value data model in which the key is the

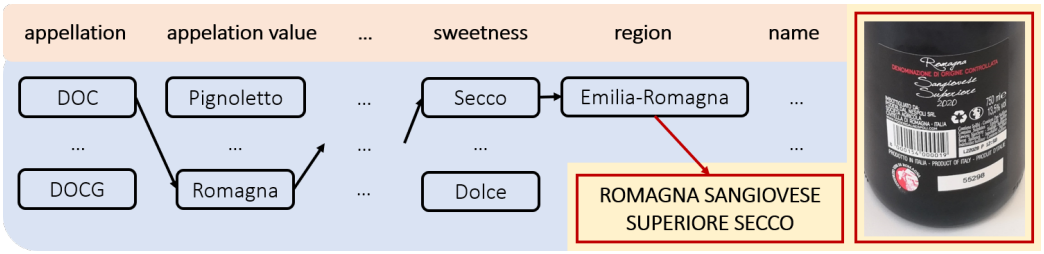


Figure 4.24: An example of features conversion from a table-like to a hierarchical-tree-structure.

value of a feature, and the respective value is the subset of all the wines sharing that particular value. In this way, a single wine tree traversal depends on the number of hierarchy levels (i.e., the features), given a priori by the values of the visited features, and the computational cost in a one-level key-value database is constant. For this reason, efficient NoSQL-like databases appear well-suited for this scenario. The schema of the hierarchical-tree database follows hence the feature order.

The 2,426 textual descriptions of different wines from the Emilia-Romagna region are hence converted in a hierarchy based on the chosen and sorted characterizing features: *appellation*, *appellation value*, *sweetness*, *color*, *effervescence*, *wine name*. This order amounts to a trade-off between (a) pruning as many leaves as possible at the higher node in the tree hierarchy, depending on how many values a feature may take, and, (b) the complexity of searching for a given feature, taking into account the average number of words composing a feature by its possible values. In other words, for (a) we considered the number of nodes that may be severed when the value of a feature is known: the more nodes, the higher the position of the feature in the hierarchy. At the same time, we considered the complexity of searching for a match for a given feature with (b). For example, considering (a), in our dataset, there are three possible values for the appellation and thirty for the appellation value. The number of wine types that would remain when choosing an appellation value is, hence, lower than the number that would remain to choose the ap-

pellation, therefore this indicates that the former should be assigned a higher place in the tree hierarchy. When considering (b), instead, the average number of words used in the appellation values requires a higher matching time than the corresponding appellation (lower number of words), thus indicating the appellation should be placed at a higher level of the tree hierarchy.

We, therefore, see that (a) and (b) push the organization of the hierarchical tree in different directions. This contrast has been solved using a single factor to evaluate (a) and (b). In fact, we defined for a given feature f_i a cost function $c_{f_i}(c_{v_{f_i}}, c_{s_{f_i}})$ that synthesizes the contributions of these two components. In particular, the cost function is computed as follows, $c_{f_i} = (w_v \times c_{v_{f_i}}) + (w_s \times c_{s_{f_i}})$, where, $c_{v_{f_i}}$ amounts to the average number of leaves (wine types) remaining in the tree once given feature f_i is chosen, $c_{s_{f_i}}$ to the average number of words used in f_i by the number of values f_i can take, and, w_v and w_s to two constants whose role is to balance the contribution of the two cost components (in our work, these have been set to 0.4 and 0.6, respectively). With the use of $c_{f_i}(c_{v_{f_i}}, c_{s_{f_i}})$, features are ordered within the hierarchical tree with the rationale that a higher hierarchy value is assigned in correspondence of a lower cost.

4.5.4.3 Textual Search Algorithm

We here describe the full algorithmic pipeline used in the AWR system to recognize a wine type from a picture of the back label of a wine bottle. In particular, we here describe how the domain-specific wine hierarchical textual DB, EasyOCR, OCR, and a novel text pre-processing and a hierarchical search are orchestrated together to return the best possible matches.

As mentioned before, the relevant information is usually placed in the top area of the label with a font larger than other text [403]. For the words that compose that text retrieved by EasyOCR, the area is usually larger than any other retrieved one. Assuming this, a pre-processing step composed by the **OCR words area detection**, that automatically detects and crops the areas that enclose the words of interest, is implemented after having executed

the first **OCR words inference**. Not performing this pre-processing step, would require accounting for all of the words identified on a label, requiring a higher cost for both time and space complexity. More in detail, only the words enclosed in the bounding boxes larger than the sample median are in the end selected, and the bounding box that includes all of such words becomes the one used to crop the label. With these conditions, it is possible to detach the word detection phase from the textual inference phase and reduce the number of processed words. At the same time, the words that could be misinterpreted by the OCR due to their size and location are discarded (e.g., small area words placed near the boundaries). An example of the execution of such pre-processing is reported in Figure 4.25

Original frame Area of interest Cropped area

Figure 4.25: Example of cropping the area of interest.

The **OCR words inference** is then again executed, returning all the words found in the cropped image. The pipeline hence only considers a subset of the words appearing on the label, reducing the computational cost. To further reduce the set, duplicates and stopwords are also removed.

The obtained set of words is then processed in the last step of the entire pipeline: the *Hierarchical search algorithm* whose Python-like pseudo-code is detailed in Algorithm 1. The algorithm searches for the best wine type by exploring each level of the hierarchical tree database and searching for a match

Algorithm 1 Hierarchical search algorithm

```
1: procedure HIERARCHICAL_SEARCH(
   bottle_retrieved_words, h_db, bottle_features, threshold, distance_method)
2:     for f in bottle_features do
3:         dict_match ← {}
4:         for v in f.possible_values() do
5:             matched_terms ← Algorithm 2(
                                bottle_retrieved_words, v,
                                threshold, distance_method)
6:             dict_match[v] ← matched_terms
7:         end for
8:         correct_value ← highest_score(dict_match)
9:         bottle_retrieved_words, h_db ← branch_db(
           h_db, bottle_retrieved_words, correct_value)
10:    end for
11:    return h_db
12: end procedure
```

within the set of words retrieved by the OCR stopping at the node exhibiting the text value with the minimum textual distance (check Algorithm 1) [201]. In brief, this algorithm takes as parameters the words returned by the OCR, a copy of the hierarchical textual DB, the sorted list of features, and a distance threshold percentage, which will be used in the Algorithm 2 (line 1). Subsequently, it cycles on the relevant features to initialize the features dictionary (lines 2 and 3). The second cycle (line 4), instead, iterates on the possible values of the considered feature (e.g., DOC, DOCG, DOP, IGT, IGP for the appellation).

Then, the *Linear search (post-OCR) correction algorithm* finds any existing matches for the given feature term (line 5), and all matches are added to the dictionary (line 6). Now, the algorithm selects the feature value that received the highest number of matches. For example, between "Denominazione origine controllata" and "Denominazione origine controllata e garantita", the latter would be chosen in the case that all the terms are matched with some word in the OCR retrieved words. In case of a tie between two features, both would be selected. Then, the dictionary contains all the matches found for the given feature values (line 8). Once a particular feature value(s) is picked, the hierarchical DB is skimmed, branching the specific sub-tree(s)

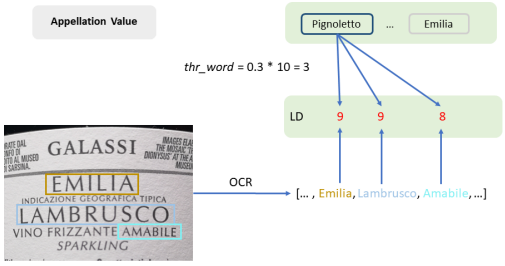
involving that value or those values. After applying the skimming step for all the wine features, the remaining elements of the hierarchical DB contain only one or more elements that possibly include the correct wine (line 9), and the hierarchical DB is returned (line 11). Given this, the *Hierarchical search algorithm* computational cost depends on the number of considered features f (i.e., the height of the tree), and on the cost of the *Linear search correction algorithm* which implements two different sub-tasks: detection and correction. The detection task identifies incorrect tokens, and the correction task tries to correct the errors found by the previous one. The algorithm adopts an isolated-word approach, relying on specific lexicons, or word unigram language models, and a distance for selecting candidates of OCR errors. In this case, the Levenshtein distance metric has been utilized [201]. The corresponding Python-like pseudo-code appears in Algorithm 2.

Algorithm 2 Linear search correction algorithm

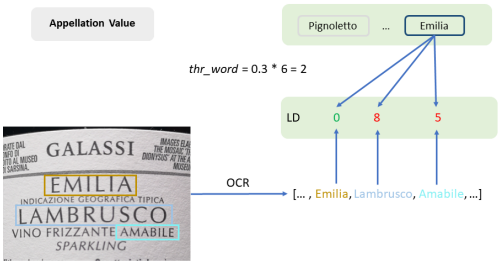
```
1: procedure LINEAR_SEARCH_CORRECTION(
   ocr_retrieved_words, feature_words, threshold, distance_method)
2:     matched_words ← []
3:     for w in feature_words do
4:         for w_ocr in ocr_retrieved_words do
5:             distance ← distance_method(w_ocr, w)
6:             thr_word ← threshold × len(w)
7:             if distance ≤ thr_word then
8:                 matched_words.append(w)
9:                 break
10:            end if
11:        end for
12:    end for
13:    return matched_words
14: end procedure
```

The algorithm works as a two-level nested loop that iterates over the words retrieved by the OCR (first level, line 3), and the values of the word that define a feature (second level, line 4). The algorithm computes the distance adopting the `distance_method` per each couple (`w`, `w_ocr`) (line 5), and if the distance is less or equal to a certain threshold, `w_ocr` is considered to be `w` (lines 6, 7 and 8). When this happens, the algorithm interrupts the

inner loop (line 8) and continues to search for the next relevant word inside the label text (line 3). To note that, here, the threshold represents a value of accepted distance: a lower distance indicates that it is more likely that `w_ocr` matches `w`. More in detail, `thr_word` is calculated based on the length of the word to match (`w`) as a simple percentage of the considered `threshold` (line 6). A `threshold` of 0 indicates that the two words must be the same (i.e., no difference), while, if set to 1, indicate that the `w_ocr` would match `w` if their distance is less or equal than p, where p is the number of character of `w`. Figure 4.26 visually reports a sample *the Linear Search Algorithm* applied to two words of the **Appellation Value** feature with a `threshold` of 0.3 (i.e. 30%): on the left side, the OCR retrieved words are compared with a non-matching value while on the right side, a matching word was detected.



Wrong correspondence produces high Levesthein Distances.



A perfect correspondence was found between the OCR-detected words and the searched ones.

Figure 4.26: Example of Linear Search correction algorithm iterating over OCR retrieved words and node values for the **Appellation Value** feature (threshold is set to be equal to the 30% of the label word). In this figure, LD stands for Levesthein Distance.

The computational cost of Algorithm 2 depends on the two lists of size n and m, respectively (i.e., number of `ocr_retrieved_words` and `feature_words`), which are constants as these may be both upper bounded by some fixed value (named as α). Considering now the cost of the full *Hierarchical Search Algorithm* (Algorithm 1), the height of the traversed hierarchical tree

database corresponds to f by construction, while an additional cost is provided to the maximum number of children to match at each level, named z (upper-bound). Considering that the cost to search for a match on a single node corresponds to α, the Algorithm 1 total cost corresponds to $f \times z \times \alpha$. It should be noted that the values at stake are all constants that do not exceed a few tens of units, simply meaning that an asymptotic analysis does not apply.

Using this *Search algorithm* it is possible to recognize terms belonging to the wine domain also when distorted by the OCR. In addition, the *Search algorithm* exploits, in general, the wine domain knowledge (Section 4.5.1): some features (e.g., effervescence) possess a default value, implying that if none of the non-default values are detected, the default one is taken. Again, failures may be due to mistaken feature identification (i.e., no relevant word appears in the OCR list or the words are in part wrongly predicted), and/or relevant information is not reported on the label. Figure 4.27 visually depicts an example of the Hierarchical Search Algorithm matching all the considered features.

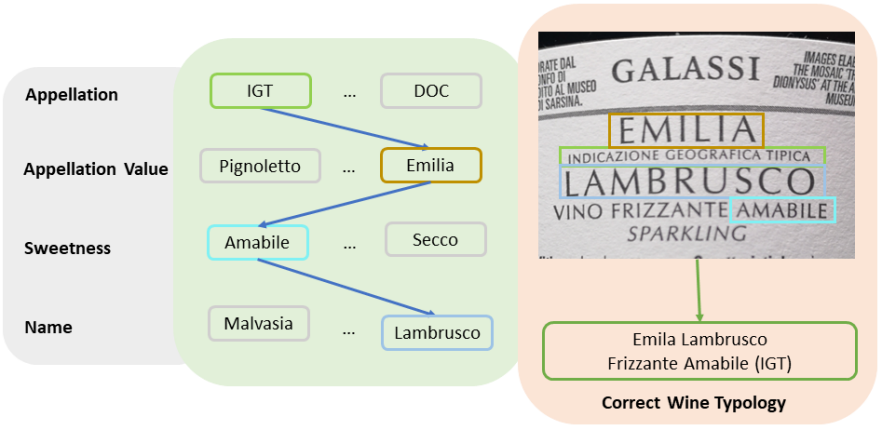


Figure 4.27: Example of a correctly matched hierarchical search algorithm traversal considering *appellation*, *appellation value*, *sweetness*, and *wine name*.

4.5.5 Experiments and Results

We here report the results obtained by applying the *Hierarchical Search Algorithm* (Algorithm 1, Section 4.5.2) to 45 different wine bottles coming from the Emilia-Romagna region in Italy, and considering a hierarchical textual DB containing 2,426 entries. The evaluation has been performed in a multi-frame setting, collecting and using multiple frames per wine.

More in detail, we collected 45 videos, one per different bottle, lasting an average of 8 seconds, rotating the camera around the bottle label. Then, we selected from each video the less blurred frame per second, adopting the variance of image Laplacian, as reported in [23]. The use of multiple frames is motivated principally by the expectation that employing videos may be possible to reduce OCR detection errors caused by environmental problems (e.g., light conditions), and the fact that taking a video does not require the user to take a picture of what s/he is seeing each time. Intending to provide a complete picture of the AWR system performance, we considered and reported the results both in terms of *efficacy* and *efficiency*. Per each considered video frame, the algorithmic pipeline has been applied to the words retrieved by EasyOCR. The result is a set composed of all the wines that expose an equal difference between the original name length and the number of matched words. To compute **efficacy**, wine bottles are considered to be correctly recognized if their names are included in the set of retrieved ones. Now, before presenting the final results, it is necessary to highlight that by tuning the hyperparameters it was possible to individuate the threshold values that best identify a word as recognized (Algorithm 2).

Recalling that a word retrieved by EasyOCR is considered to match one value for each different feature if their Levenshtein distance falls below a given threshold, the selection of the threshold value has been performed by assessing the pipeline for all the values between 0 and 1 with a step of 0.1. In particular, two distinct thresholds have been set for words and acronyms (e.g., Denominazione Origine Controllata vs DOC), 0.3 and 0.1, respectively, capable of returning the best results: with these values, 41/45 bottles were

correctly recognized (91% of accuracy). This accuracy was computed considering that the algorithm could provide more than one wine as output, retrieving on average 1.41 bottles per analyzed video (min 1, max 15). It is worth noticing that the errors could depend on the EasyOCR performance. To verify this, we defined 45 different textual files (one per bottle) that contain each word reported in the wine bottle label, to simulate the performance of a perfect OCR, transcribing all the words appearing on labels, respecting their order. With such data and a threshold set to 0, as a perfect OCR should match the exact words, the algorithm reached a 100% accuracy. Always considering that the algorithm could provide more than one wine as an output, the linear search algorithm applied to the perfect OCR setting retrieved on average 1.01 bottles per examined video (min 1, max 8). All such results are summarized and reported in Table 4.5.

Performance Metric	**AWR** (words = 0.3; acronyms=0.1)	**AWR-POCR** (words = 0.0; acronyms=0.0)
Guessed bottles (correct/total)	41/45	45/45
Retrieved bottles (min, mean, max)	1, 1.41, 15	1, 1.01, 8

Table 4.5: Results obtained with specified confidence values include AWR performance using selected OCR thresholds for words and acronyms (0.3 and 0.1), while AWR-POCR evaluates performance with a simulated perfect OCR, accepting only perfect matches.

We then measured the system **efficiency**, i.e., the time taken to return results. The algorithm is composed of four steps (Section 4.5.2): OCR words area detection, cropping by detected areas (image cropping), OCR words inference, and the *Hierarchical search algorithm*. The captured times regarding EasyOCR do not include the deep learning model initialization times, as not relevant. Time measurements come from executing a single trial of the complete algorithmic pipeline on the selected video frames for all the bottles. Figure 4.28 depicts, the min, mean, and max times, provided in seconds, for each algorithm component, averaged over ten trials.

We then compared the efficiency of the hierarchical tree search and the

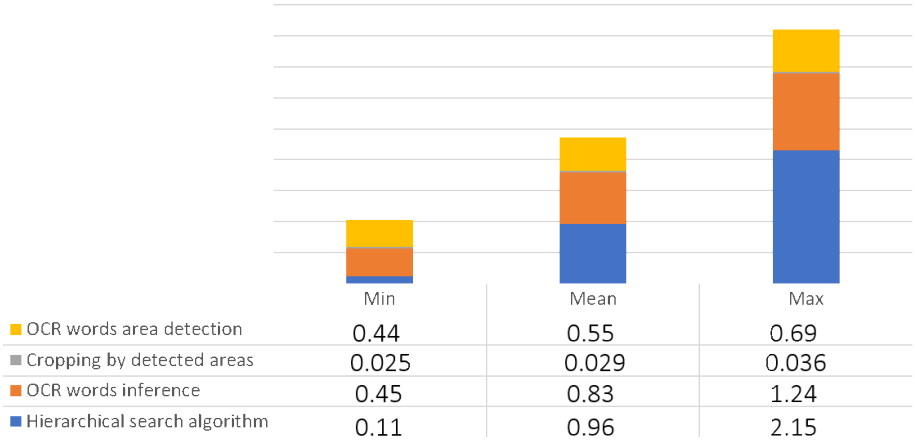


Figure 4.28: Total time in seconds over different Hierarchical Mode algorithm steps.

classical linear one. We defined a flat database containing all the relevant values of the different distinguishing features, and we applied Algorithm 1 at each step of the data structure. So, the final output includes the wine(s) with the highest number of matching features.

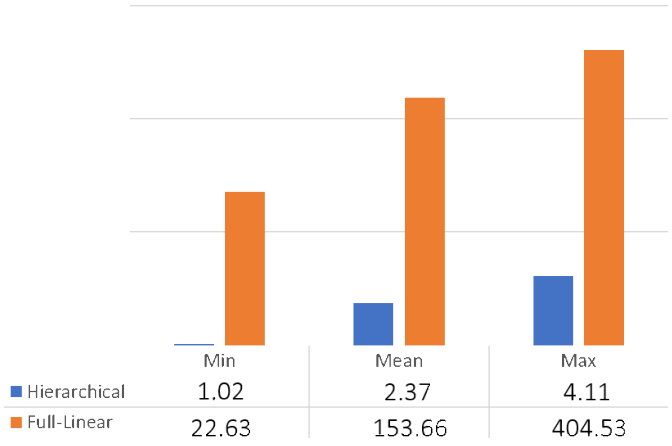


Figure 4.29: Hierarchical vs Full-Linear Total time in seconds (plotted in log scale).

We report in Figure 4.29 the computed min, mean, and max time values obtained over the same wine bottle set used for the previous evaluations, again over ten trials. It is possible to note that a naive approach, like the

Full-Linear one, is roughly a hundred times slower than the *Hierarchical* proposed approach. This fact is justified since in the *Full-Linear* mode the tree is never pruned, so the algorithm compared all the retrieved words with all the words describing the $2,426$ wine types.

4.5.6 Annotation and Image Retrieval Systems

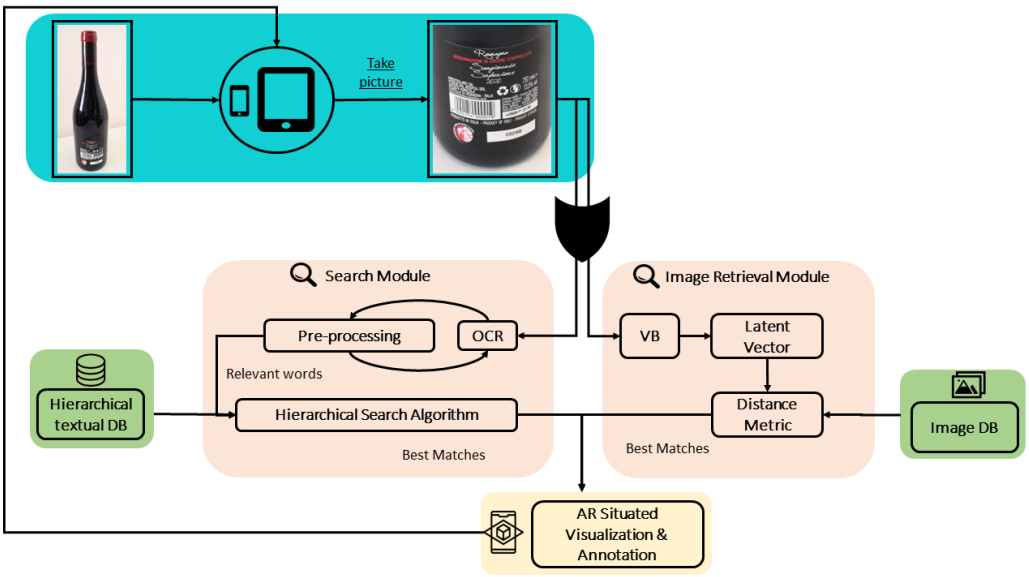


Figure 4.30: AWR architecture, extended with an Image Retrieval Service.

The AWR system was developed upon the system architecture described in Section 4.5.2, in which a flexible structure allows (a) straightforwardly substitute or joint exploit different recognition methods; (b) extend the user interface mobile application with novel UI components. Exploiting such flexibility, we introduced novel features to match the annotation system introduced in our HCLINT-DT framework, while also empowering the wine search mechanism (see Figure 4.30). In particular, we here extend the AWR architecture and implementation with two additional features:

- Define an annotation system to add novel wine bottles taking pictures around the world to generate a new image database;

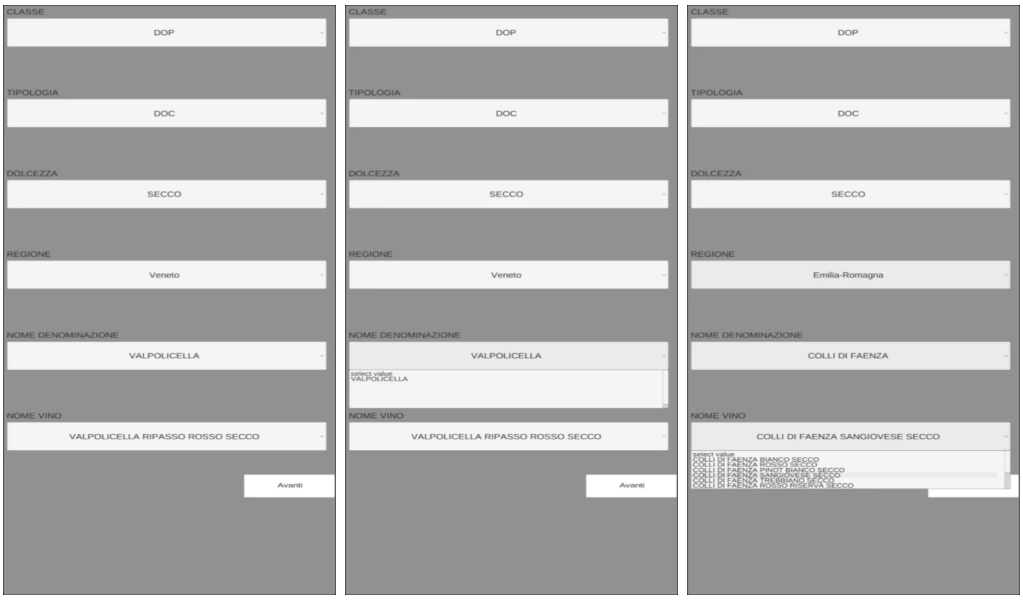
- Use those data to train and define a DL-based image retrieval service, that could be used with the existing recognition service in a hybrid setting.

4.5.6.1 Annotation System

As mentioned in Section 4.4 and also stated in relevant works [227, 77], annotations amount to a fundamental source of information to be attached to real-world objects, processes, or people, to provide context and interpretation also to whom never interacted with them, particularly exploiting HCLINT paradigms. However, our AWR application lacked a simple way to provide annotations on the recognized bottle, that can later be retrieved and visualized in both physical and virtual worlds. Considering our particular use case, such an annotation system could be specialized to accomplish additional responsibilities [357]: speed up the cataloging process of new bottles and provide image annotations to improve wine recognition through image retrieval paradigms.

As detailed in 4.5, our system requires as a precondition a hierarchical textual database describing a tree of features and as a postcondition returns all the best-ranked wine typologies concerning the extracted words from the wine label image. Along with the wine typologies, one could attach the tree path that provided that particular match, furnishing so all the information needed to pre-fill a form, built based on the defined hierarchical tree. This process would serve the purpose of defining annotation for a wine bottle that is already registered in the database, but it could also be used to add novel records to it. This is because a bottle not present in the database could share a lot of properties with the ones included in it. To clarify, we report here a concrete example. The wine "Colli Bolognesi Barbera Abboccato", correctly registered in the database, has common features concerning the non-registered "Colli Bolognesi Bologna Bianco Secco", like the Denomination and the Appellation, but differs in their sweetness (Abboccato vs Secco) and name (Barbera vs Bologna). Suppose that a user now attempts to recognize

the second: the match would fail, recognizing the former instead. However, the user could insert this new bottle typology in the DB, and while doing that, it could avoid the burden of filling the form: a pre-filled one could be generated based on the data returned by the ranking algorithm. When a user changes one of the proposed fields, all the child features will change, according to the hierarchical database structure, to guide the user in filling the form quickly.



(a) (b) (c)

Figure 4.31: AR Annotation interface: (a) After picking a wine from the returned rank, a form is prefilled with its information (b)-(c) A user changes the attribute of the Denomination and Name features.

Such a workflow was implemented in both the front-end and the back-end of AWR, and we report the UI interfaces generated from this aim in Figure 4.31. Such an interface, could be easily extended, to provide multi-media annotations, such as images.

4.5.6.2 Image Annotations and Retrieval System

We extended the detailed form Annotation System, including also images. This aims at satisfying the need to define a novel image database linked to each record in the textual one, that could be then used to train DL models to provide an image retrieval service. We first defined a new UI interface (depicted in Figure 4.32) to let the user capture n pictures of the considered wine.

(a) (b)

Figure 4.32: AR Image Annotation interface: (a) After taking a photo, the UI spawns 4 vertexes, representing the bounding box s/he wants to enclose. (b) The user modifies the bounding box to enclose the wine bottle label.

Per each picture, the user can define and manipulate 4 points, detailing the vertexes of a convex plane, enclosing the label of the wine, through drag-and-dropping. This easy-to-use interface provides two advantages: (a) directly saving the wine bottle label instead of the entire picture, excluding irrelevant and misleading visual details; (b) furnishing more data to improve the image retrieval system (i.e., it leverages both the picture of the entire bottle and its user-made crop). The storing mechanism was implemented in

the back-end by storing each picture, cataloged in a folder structure generated based on the user cataloging information.

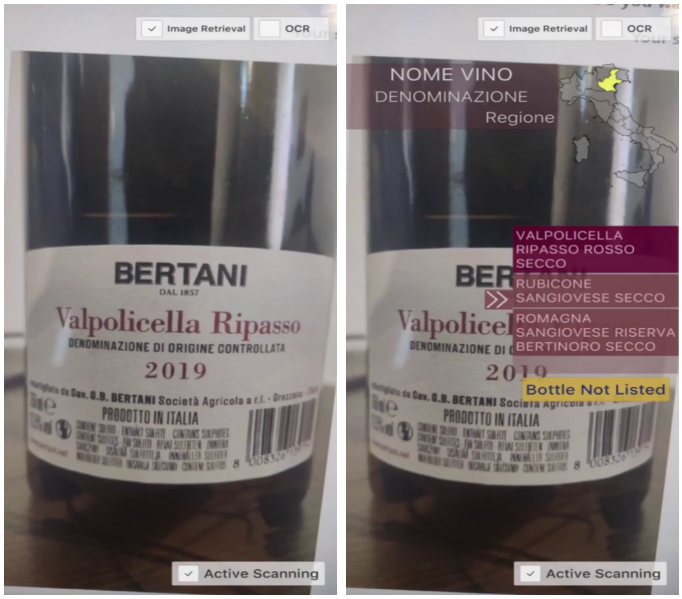
At this point, we could leverage those collected images to define an image retrieval service [205]. To this date, we resorted to a well-known DL Vision Backbone: Vision Transformer (ViT) [100, 151]. ViT applies a pure transformer directly to sequences of image patches to classify the full image, achieving state-of-the-art performance in image classification, including object detection, and representation learning [151, 295]. In particular, [295] trained ViT in a multi-modal self-supervised setting, to learn rich features that emerged from the complementarity of different modalities, through the InfoNCE loss. Those learned representations were used out-of-the-box for different tasks, such as image generation and retrieval [295, 327]. We so adopted the same approach, using the extracted features from the ViT version that exposes the lower number of parameters, for a faster inference (16 patches of 224×224 model [9]). In particular, for each wine picture available, we proceeded by first pre-processing it to match the input domain of ViT (i.e., size, transformations, and data range) and then executing a forward pass, extracting its calculated class token, which became the latent representation vector $v \in R^{384}$ for the considered image. Concatenating all the extracted representations, we defined our latent vector database $v_b \in R^{N \times 384}$, with N representing the cardinality of the collected images, which can be used for visually querying a new image. As also represented in Figure 4.30, once a new image is taken from the mobile app, ViT is used to calculate the embedding $q \in R^{384}$ corresponding to it and then calculate its Euclidean distance concerning each $v_i \in v_b$, taking the nearest $k = 3$ vectors, which defines the best matches.

$$\operatorname*{top\,k}_{i=1}^{k} \min \|q - v_i\|$$

Figure 4.33, reports the mobile app views of a correctly matched bottle

[9] https://github.com/huggingface/pytorch-image-models/blob/main/timm/models/vision_transformer.py

with the image retrieval service. Despite this recognition component being used as an alternative to the textual-based one, we plan to implement a hybrid sequential approach: the textual algorithm would prune the search space to reduce the wine candidates, and then the image retrieval query would be executed on this subset to get the final results. This approach could overcome both the limitations of textual-only and image-only approaches (i.e., OCR fails vs Similar visual labels).



(a) (b)

Figure 4.33: AR Image Retrieval interface: (a) The user scans a wine bottle which is then (b) correctly retrieved by the image retrieval system.

4.6 Discussion and Conclusions

In this chapter, we introduced the first version of a DT-injected Human Collaborative Intelligence framework, that leverages XR and AI paradigms to empower Human-DT interactions (HCLINT-DT). We contextualized the H-CLINT-DT framework with a cultural heritage and industrial use case, demonstrating its flexibility. Its main functionalities regard the fruition and

manipulation of human DT annotations from anywhere, at any time, and with any degree of reality. To validate such an approach, we assessed a use case involving family photo albums through a user study. The results showed a general agreement on the ease of use of the proposed interface. We also explored the adaptability of the proposed approach considering a use case drawn from an industrial electrical engineering context. We implemented an AR/MR interface to manipulate information related to the DT of electrical boards, highlighting the adaptability of HCLINT-DT (RQ-2, RQ-4). We also validated this system in a real industrial context, along with workers using it. Further investigations could involve (a) improving annotation systems retrieval and the possible impact of natural languages processing technologies, such as speech-to-text and text parsing (to provide vocal commands), and word embeddings (to ease the query of a particular annotation); (b) analyzing how annotations may automatically influence the physical space exploiting AI; (c) examine the possible injection of crowd intelligence which allows interweaving crowd and machine capabilities to address challenging computational problems [204].

Exploiting again the flexibility of HCLINT-DT we applied it to the wine retail domain. To this date, we defined AWR, an AR wine recognition app that could be used to retrieve wine bottles from the text lying on their labels, This approach overcomes limitations posed by classical image-retrieval approaches, such as scalability and long-tail distributions (RQ-2) [365, 10]. We showed that by adopting a textual-based approach, it is possible to overcome such limitations while maintaining reasonable performances. We also implemented an annotation system that was used to define a novel image database. This was then adopted to develop an image retrieval system that could be used in a hybrid way with the introduced textual search. However, additional work is needed to generalize our system in a more varied context. It is possible to envision a hybrid approach using both textual and image retrieval, to first prune the hierarchical textual DB and then execute an image query on the remaining wine label images [205]. Secondly, we could

exploit OCRs to detect additional information within labels (e.g., bottle capacity and alcohol content), to present additional information in AR, such as caloric intake, and maximum recommended dose. Thirdly, the performance of the adopted DL-based OCR could be improved by (i) introducing image pre-processing techniques such as warping and rectification [137, 440, 120, 194]; fine-tuning it with a dataset composed of wine-label pictures, tagged with bounding boxes and corresponding textual annotations. These would ameliorate the erroneous word prediction highlighted in the results reported in Table 4.5. Then, a more general and applicable detection pipeline that exceeds the Italian wines domain could be implemented (e.g., fine-tuning with different countries' bottles).

Hence, there are many possible future research directions involving the interaction between Humans and DTs modulated by both XR and AI. It is believed that this research direction could play a pivotal role in advancing and supporting Industry 4.0 across various contexts to support humans, defining novel and improved ways of working, decision-making, and practices, increasing also value creation [202, 314, 58, 252, 162].

Chapter 5

Conclusions

The rapid evolution of XR and AI algorithmic pipelines, along with the higher availability of dedicated hardware, will reshape our future interactions with digital information, impacting all sectors of our society. This includes but is not limited to, creative fields, industry, economics, and cultural heritage [84, 62, 324, 162]. AI will provide an intelligent layer to make faster and better decisions, defining automatic methods to assist humans in everyday tasks. In contrast, XR will provide interfaces to empower human-machine interactions, creating adaptive methods wrapped around this intelligent layer.

In this thesis, we examined so how to adopt "Extended Artificial Intelligence" paradigms [420, 162], offering high-level reflections based on carefully selected case studies from three main research areas: Cultural Heritage, Creative Industries, and Industrial workflows. This concluding chapter aims at synthesizing its diverse contributions, starting from the research background and derived questions outlined in Chapter 1.

In Chapter 2, we investigated RQ-1 (*Can Artificial Intelligence and eXtended Reality be combined to respond to unanswered cultural heritage questions?*) and RQ-2 (*Can Artificial Intelligence improve user interactions and task completion efficacy in eXtended Reality systems?*) by proposing a multimedia and DL-based tool application to assist socio-historians in cataloging family album photos and support socio-historical research. We first intro-

duced a novel dataset called IMAGO, where each picture was labeled by shooting date and socio-historical context. Then, we developed DL classifiers to build the mentioned tool, resorting to both Convolutional Neural Networks and Vision Transformers, demonstrating state-of-the-art performances. We also assessed the performance of such models against a human expert, highlighting how those models were similar or more accurate. We then used this tool to analyze their possible contributions in socio-historical investigations, like mixed quantitative-qualitative studies and cross-cultural influences. Finally, we showed how this DL tool could be synergized with AR paradigms, defining a system to automatically scan and catalog physical family album pictures, exploiting the Hololens 2. We assessed this application with a user study, which points towards its positive adoption.

Then, Chapter 3 contributed to RQ-2 and RQ-3 (*Which kind of eXtended Reality and Artificial Intelligence technologies could be applied in the creative industry?*) by exploring how AI and XR could support fashion x-commerce. Firstly, we introduced an intelligent Voice Assistance (i.e., Alexa) in a virtual shop environment, to improve its usability. We thus designed an experiment based on two VR immersive experiences simulating the processes typically involved in fashion stores, demonstrating a high interest and a positive adoption attitude toward voice-based interactions. Then, hypothesizing that voice covers a fundamental role in the perceived realism of assistants, we leveraged Human Digital Twins paradigms, by modeling an avatar of a real human, that could communicate with real fashion shop clerks. The avatar made shopping requests using fake or natural voices and fake or natural aspects. This experimental setting was then applied in a user study, highlighting how natural voice is a pivotal aspect in accepting Human Digital Twins. Finally, we introduced an XR-based framework to provide a better means for humans to judge the aesthetic and emotional constructs for immersive (generated) content, considering as a use-case 360° musical concert videos (i.e., pivotal for aesthetics and emotional triggers). Those factors are fundamental to enabling automatic fashion x-commerce scenarios. To validate such

a framework, we carried out a thorough user study, demonstrating that XR paradigms could be more effective, concerning classical displays, in evaluating such constructs.

Finally, Chapter 4, outlined our contributions for RQ-2, and RQ-4 (*Can industrial workers take advantage of eXtended Reality and Artificial Intelligence in their everyday activities?*). We first introduced a DT-injected Human Collaborative Intelligence framework (H-CLINT-DT), to empower Human-DT interactions at any level of the XR spectrum. To assess the feasibility and flexibility of such a framework, we contextualized it with both a cultural heritage and industrial use case. The main functionalities of the developed human-DT interfaces involve the fruition and manipulation of human annotations from anywhere, at any time, and with any degree of reality. To this date, we developed AR, MR, and VR interfaces, designed to be used in the considered settings. The results coming from their validation showed a general agreement on HCLINT-DT ease of use and adoption gain. Considering the highlighted flexibility, we also applied it in the wine retail domain, where we defined AWR, an AR wine recognition app that could retrieve annotations from the text on wine bottle back labels. In such a system, we showed that by adopting a textual-based approach, it is possible to overcome limitations provided by classical Computer Vision approaches, while maintaining reasonable performances. We also showed how AWR could be extended with an annotation system that could define a new wine image database. The latter was then adopted to implement an Image Retrieval system, to improve wine typology recognition with a hybrid textual-visual approach.

As a final reflection, it is worth highlighting that, despite being confined to the three main research areas of this thesis (Cultural Heritage, Creative Industries, and Industrial applications) the methodological findings and contributions made in each of them can be re-adapted to other disciplines and use cases.

Considering the methodology developed in the context of Cultural Her-

itage for family album pictures, the analysis of cultural relationships could be re-adapted to study past, and present, cross-cultural influences for a large and varied set of visual material [165, 444]. The same is exposed for the developed DL-driven AR catalog application, which could be re-adapted for almost any physical visual archive [166].

Considering instead the contributions made in Creative Industries, the Vocal Assistance findings could be re-applied in many other guidance-driven contexts, ranging from industrial production to education [216, 189]. The fashion customer HDT considerations have instead relevant ethical implications concerning digital identity in the Metaverse, a concept recently examined by public and private authorities [235]. Also, the XR aesthetic and emotion-driven evaluation framework could impact several cultural heritage and creative use cases, including fashion, art, and museum experiences [12, 96].

Finally, focusing on the advancements and analysis made for industrial use cases, we feel that the introduced HCLINT-DT framework already demonstrated a good level of flexibility, being applied in two different use cases without any kind of modification. This exhibited flexibility level allows its application for other objectives like the preservation of cultural heritage material, education, and creative design process optimization [110, 211, 389].

Looking ahead, numerous unexplored applications could leverage Extended Artificial Intelligence in the examined areas, with many potential investigations. Among those, one of the possible paths of exploration regards the definition of AI and data-driven adaptive XR experiences to support humans in contexts with variable environments and objectives. Such adaptive mechanisms could, in particular, be applied to human data generated in both implicit and explicit ways [249]. For such a research space, AI paradigms represent an intelligent layer to infer which kind of adaptation must be performed and how.

On top of the previous considerations, the HCLINT-DT framework could provide a unique opportunity to define a foundational human-knowledge base

to teach AI how to make such decisions, directly from the experience of past and current users (i.e., crowd intelligence AI) [204]. However, such a framework must be generalized to be applicable in any context of use, involving objects, people, and processes. Such extensions could have relevant impacts on several fields of study, like medicine and engineering [272]. To this date, future works will involve the study of novel XR devices and AI paradigms under this lens. In particular, the examination of efficient and few-shot AI paradigms to project any object from the physical world in the digital realm (e.g., few-shot 3D generation [1]).

Finally, investigations regarding the alignment between the multi-modal AI and XR multi-modal interactions will be carried out [22, 299, 162]. Following the same rationale of [162], we could find novel complementarity usages of XR and AI, but in the multi-modal arena. In particular, we will investigate how to project humans into multi-modal spaces of analysis, generated by AI methods regarding a certain context of use, object, or process, while exploiting XR manipulations to have an immersive and adaptive space to multi-modally interact with them [22, 299, 209].

To conclude, the contribution of this thesis, along with planned future investigations, amounts to digital tiles guiding us toward a pervasive Digital World: the Metaverse.

References

[1] Milad Abdollahzadeh et al. "A survey on generative modeling with limited data, few shots, and zero shot". In: *arXiv preprint arXiv:2307.14397* (2023).

[2] Alan Agresti and Brian Caffo. "Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures". In: *The American Statistician* 54.4 (2000), pp. 280–288.

[3] Mariano Alcañiz, Marco Sacco, and Jolanda G Tromp. *Roadmapping Extended Reality: Fundamentals and Applications*. John Wiley & Sons, 2022.

[4] Julian M Alston and Davide Gaeta. "Reflections on the Political Economy of European Wine Appellations". In: *Italian Economic Journal* 7.2 (2021), pp. 219–258.

[5] Jesús Omar Álvarez Márquez and Jürgen Ziegler. "Improving the Shopping Experience with an Augmented Reality-Enhanced Shelf". In: *Mensch und Computer 2017-Workshopband* (2017).

[6] Amazon. *Amazon SageMaker Ground Truth*. https://aws.amazon.com/it/sagemaker/groundtruth/. 2021.

[7] *Amazon Alexa*. https://developer.amazon.com/en/alexa. Accessed: 2020-02-20. 2020.

[8] Eleftherios Anastasovitis and Manos Roumeliotis. "Creative Industries and Immersive Technologies for Training, Understanding and Communication in Cultural Heritage". In: *Euro-Mediterranean Conference*. Springer. 2020, pp. 450–461.

[9] Daniel Andersen et al. "Virtual annotations of the surgical field through an augmented reality transparent display". In: *The Visual Computer* 32.11 (2016), pp. 1481–1498.

[10] Alessia Angeli et al. "Making paper labels smart for augmented wine recognition". In: *The Visual Computer* (2023), pp. 1–13.

[11] Tiago Araújo et al. "Aspects of Voice Interaction on a Mobile Augmented Reality Application". In: *International Conference on Virtual, Augmented and Mixed Reality*. Springer. 2016, pp. 199–210.

[12] Alice Arnold, Susan Martin Meggs, and Annette G Greer. "Empathy and aesthetic experience in the art museum". In: *International Journal of Education Through Art* 10.3 (2014), pp. 331–347.

[13] Artec3D. *Artec Eva - Fast 3D scanner for professionals*. https://www.artec3d.com/portable-3d-scanners/artec-eva. 2021.

[14] Luca Asunis et al. "HOCTOPUS: An Open-Source Cross-Reality tool to Augment Live-Streaming Remote Classes". In: *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 2023, pp. 29–34. DOI: 10.1109/ISMAR-Adjunct60411.2023.00014.

[15] Betül Ay et al. "A visual similarity recommendation system using generative adversarial networks". In: *2019 international conference on deep learning and machine learning in emerging applications (Deep-ML)*. IEEE. 2019, pp. 44–48.

[16] Cheick T Ba et al. "Web3 Social Platforms: Modeling, Mining and Evolution". In: *CEUR WORKSHOP PROCEEDINGS*. Vol. 3340. CEUR-WS. 2022, pp. 168–179.

[17] Cheick Tidiane Ba, Matteo Zignani, and Sabrina Gaito. “Cooperative behavior in blockchain-based complementary currency networks through time: The Sarafu case study”. In: *Future Generation Computer Systems* (2023).

[18] Sujin Bae et al. “The influence of mixed reality on satisfaction and brand loyalty in cultural heritage attractions: A brand equity perspective”. In: *Sustainability* 12.7 (2020), p. 2956.

[19] Jeonghun Baek et al. “What is wrong with scene text recognition model comparisons? dataset and model analysis”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4715–4723.

[20] Youngmin Baek et al. “Character region awareness for text detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9365–9374.

[21] Haythem Bahri, David Krčmař’ik, and Jan Koč’i. “Accurate object detection system on hololens using yolo algorithm”. In: *2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)*. IEEE. 2019, pp. 219–224.

[22] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. “Multimodal machine learning: A survey and taxonomy”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.2 (2018), pp. 423–443.

[23] Raghav Bansal, Gaurav Raj, and Tanupriya Choudhury. “Blur image detection using Laplacian operator and Open-CV”. In: *2016 International Conference System Modeling Advancement in Research Trends (SMART)*. 2016, pp. 63–67. DOI: 10.1109/SYSMART.2016.7894491.

[24] Omer Bar-Tal et al. “Multidiffusion: Fusing diffusion paths for controlled image generation”. In: (2023).

[25] S Barba et al. “An application for cultural heritage in Erasmus Placement. Surveys and 3D cataloguing archaeological finds in Merida (Spain)”. In: (2011).

[26] Barbara Rita Barricelli, Elena Casiraghi, and Daniela Fogli. “A survey on digital twin: definitions, characteristics, applications, and design implications”. In: *IEEE access* 7 (2019), pp. 167653–167671.

[27] F Basura et al. “Color features for dating historical color images”. In: *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2014, pp. 2589–2593.

[28] Andrea Batch et al. “Evaluating View Management for Situated Visualization in Web-based Handheld AR”. In: *Computer Graphics Forum*. Vol. 42. 3. Wiley Online Library. 2023, pp. 349–360.

[29] Eric PS Baumer et al. “Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?” In: *Journal of the Association for Information Science and Technology* 68.6 (2017), pp. 1397–1410.

[30] Alice Bellazzi et al. “Virtual reality for assessing visual quality and lighting perception: A systematic review”. In: *Building and Environment* 209 (2022), p. 108674.

[31] K Bentein. “Minor complementation patterns in Post-classical Greek (I–VI AD): A socio-historical analysis of a corpus of documentary papyri”. In: *Symbolae Osloenses* 89.1 (2015), pp. 104–147.

[32] NIGEL BEVAN. “International standards for HCI and usability”. In: *International Journal of Human-Computer Studies* 55.4 (2001), pp. 533–552. ISSN: 1071-5819. DOI: https://doi.org/10.1006/ijhc.2001.0483. URL: http://www.sciencedirect.com/science/article/pii/S1071581901904835.

[33] Pronaya Bhattacharya et al. “Towards future internet: The metaverse perspective for diverse industrial applications”. In: *Mathematics* 11.4 (2023), p. 941.

[34] Simone Bianco et al. “Predicting image aesthetics with deep learning”. In: *Advanced Concepts for Intelligent Vision Systems: 17th International Conference, ACIVS 2016, Lecce, Italy, October 24-27, 2016, Proceedings 17*. Springer. 2016, pp. 117–125.

[35] C M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[36] Nick Black. “Why we need qualitative research.” In: *Journal of epidemiology and community health* 48.5 (1994), p. 425.

[37] Hennie R Boeije. “Analysis in qualitative research”. In: *Analysis in qualitative research* (2009), pp. 1–240.

[38] Luciano Bononi et al. “Digital twin collaborative platforms: Applications to humans-in-the-loop crafting of urban areas”. In: *IEEE Consumer Electronics Magazine* (2022).

[39] E Borcoci, D Negru, and C Timmerer. “A novel architecture for multimedia distribution based on content-aware networking”. In: *2010 Third International Conference on Communication Theory, Reliability, and Quality of Service*. IEEE. 2010, pp. 162–168.

[40] Monica Bordegoni and Francesco Ferrise. “Exploring the intersection of metaverse, digital twins, and artificial intelligence in training and maintenance”. In: *Journal of Computing and Information Science in Engineering* 23.6 (2023), p. 060806.

[41] L Bosi and H Reiter. “Historical Methodologies”. In: *Methodological practices in social movement research* (2014), pp. 117–43.

[42] P Bourdieu. “On the family as a realized category”. In: *Theory, culture & society* 13.3 (1996), pp. 19–26.

[43] Malaika Brengman, Kim Willems, and Helena Van Kerrebroeck. “Can’t touch this: the impact of augmented reality versus touch and non-touch interfaces on perceived ownership”. In: *Virtual Reality* 23.3 (2019), pp. 269–280.

[44] Nathalie Bressa et al. “What’s the Situation with Situated Visualization? A Survey and Perspectives on Situatedness”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.1 (2021), pp. 107–117.

[45] Thomas M Breuel. “High performance text recognition using a hybrid convolutional-lstm implementation”. In: *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*. Vol. 1. IEEE. 2017, pp. 11–16.

[46] Wolfgang Büschel, Annett Mitschick, and Raimund Dachselt. “Here and now: Reality-based information retrieval: Perspective paper”. In: *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 2018, pp. 171–180.

[47] M Á Cabrera. *Postsocial history: An introduction*. Lexington Books, 2004.

[48] M A Cabrera. “On language, culture, and social action”. In: *History and Theory* 40.4 (2001), pp. 82–100.

[49] Stevan Čakić et al. “The Use of Tesseract OCR Number Recognition for Food Tracking and Tracing”. In: *2020 24th International Conference on Information Technology (IT)*. IEEE. 2020, pp. 1–4.

[50] D Calanca. “Percorsi di storia della famiglia”. In: *Rivista di storia e storiografia* 5.5 (Nov. 2004), pp. 203–210.

[51] D Calanca. “Album di famiglia. Autorappresentazioni tra pubblico e privato (1870-1950).” In: *Storia e Futuro - N° 8-9* (2005).

[52] D Calanca. “Fotografie amatoriali e fotografie professionali nell’Italia del boom economico”. In: *Storia e Futuro - N° 12* (2006).

[53] D Calanca. “Italians posing between public and private. Theories and practices of Social Heritage”. In: *Almatourism-Journal of Tourism, Culture and Territorial Development* 2.3 (2011), pp. 1–9.

[54] Michael James Callaghan et al. “Voice Driven Virtual Assistant Tutor in Virtual Reality for Electronic Engineering Remote Laboratories”. In: *Smart Industry & Smart Education*. Ed. by Michael E. Auer and Reinhard Langmann. Cham: Springer International Publishing, 2019, pp. 570–580. ISBN: 978-3-319-95678-7.

[55] Camera di Commercio Molise. *Guida etichettature vino*. https://www.molise.camcom.gov.it/sites/default/files/guida_etichettatura_vino.pdf. 2016.

[56] Vincent J. Cannato. *How America became Italian*. t.ly/fUKb. Washington Post, 2022.

[57] Roberta Capello, Silvia Cerisola, and Giovanni Perucca. “Cultural heritage, creativity, and local development: A scientific research program”. In: *Regeneration of the built environment from a circular economy perspective* (2020), pp. 11–19.

[58] Leonor Adriana Cardenas-Robledo et al. “Extended reality applications in industry 4.0.-A systematic literature review”. In: *Telematics and Informatics* (2022), p. 101863.

[59] Julie Carmigniani and Borko Furht. “Augmented reality: an overview”. In: *Handbook of augmented reality* (2011), pp. 3–46.

[60] Pasquale Cascarano et al. “On the First-Order Optimization Methods in Deep Image Prior”. In: *Journal of Verification, Validation and Uncertainty Quantification* 7.4 (2022), p. 041002.

[61] Pasquale Cascarano et al. “Constrained and unconstrained deep image prior optimization models with automatic regularization”. In: *Computational Optimization and Applications* 84.1 (2023), pp. 125–149.

[62] Seong-Soo CHA. “Metaverse and the evolution of food and retail industry”. In: *The Korean Journal of Food & Health Convergence* 8.2 (2022), pp. 1–6.

[63] Steve Charters, Larry Lockshin, and Tim Unwin. “Consumer responses to wine bottle back labels”. In: *Journal of Wine Research* 10.3 (1999), pp. 183–195.

[64] Jinjun Chen and Honggang Wang. “Guest editorial: Big data infrastructure i”. In: *IEEE Transactions on Big Data* 4.2 (2018), pp. 148–149.

[65] Juan Chen et al. “Paying attention in metaverse: an experiment on spatial attention allocation in extended reality shopping”. In: *Information Technology & People* 36.8 (2023), pp. 255–283.

[66] X Chen et al. “Learning and Fusing Multiple User Interest Representations for Micro-Video and Movie Recommendations”. In: *IEEE Transactions on Multimedia* (2020).

[67] Yuanfang Chen et al. “Industrial internet of things-based collaborative sensing intelligence: framework and research challenges”. In: *Sensors* 16.2 (2016), p. 215.

[68] Wen-Huang Cheng et al. “Fashion meets computer vision: A survey”. In: *ACM Computing Surveys (CSUR)* 54.4 (2021), pp. 1–41.

[69] Alice Chirico et al. “Effectiveness of immersive videos in inducing awe: an experimental study”. In: *Scientific reports* 7.1 (2017), p. 1218.

[70] Looi Theam Choy. “The strengths and weaknesses of research methodology: Comparison and complimentary between qualitative and quantitative approaches”. In: *IOSR Journal of Humanities and Social Science* 19.4 (2014), pp. 99–104.

[71] E S Clemens and M D Hughes. “Recovering past protest: Historical research on social movements”. In: *Methods of social movement research* 16 (2002), pp. 201–230.

[72] E Coburn et al. “The Cataloging Cultural Objects experience: Codifying practice for the cultural heritage community”. In: *IFLA journal* 36.1 (2010), pp. 16–29.

[73] William Jay Conover. *Practical nonparametric statistics*. Vol. 350. USA: john wiley & sons, 1999.

[74] *CoverGirl offers augmented reality makeup trials in Times Square*. https://www.springwise.com/covergirl-opens-high-tech-flagship-store-in-new-york-city. Accessed: 2019-12-12. 2019.

[75] L Criscenti, G D'autilia, and G De Luna. *L'Italia del Novecento: Le fotografie e la storia*. Giulio Einaudi editore, 2005.

[76] James T Croasmun and Lee Ostrom. "Using likert-type scales in the social sciences." In: *Journal of adult education* 40.1 (2011), pp. 19–22.

[77] Valeria Croce et al. "Semantic annotations on heritage models: 2D/3D approaches and future research challenges". In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 43.B2 (2020), pp. 829–836.

[78] Edmanuel Cruz et al. "An augmented reality application for improving shopping experience in large retail stores". In: *Virtual Reality* 23.3 (2019), pp. 281–291.

[79] Yi Rui Cui et al. "FashionGAN: display your fashion design using conditional generative adversarial nets". In: *Computer Graphics Forum*. Vol. 37. 7. Wiley Online Library. 2018, pp. 109–119.

[80] E Culurciello. *Neural Network Architectures*. https://towardsdatascience.com/neural-network-architectures-156e5bad51ba. 2021.

[81] Stuart Cunningham. "From cultural to creative industries: theory, industry and policy implications". In: *Media International Australia* 102.1 (2002), pp. 54–65.

[82] Anna Czepiel et al. "Aesthetic and physiological effects of naturalistic multimodal music listening". In: *Cognition* 239 (2023), p. 105537.

[83] RALPH D'AGOSTINO and Egon S Pearson. "Tests for departure from normality. Empirical results for the distributions of b 2 and b". In: *Biometrika* 60.3 (1973), pp. 613–622.

[84] Muhammet Damar. "Metaverse shape of your life for future: A bibliometric snapshot". In: *Journal of Metaverse* 1.1 (2021), pp. 1–8.

[85] Robertas Damaševičius. "From E-commerce to V-commerce: Understanding the Impact of Virtual Reality and Metaverse on Economic Activities". In: *Journal of Information Economics* 1.3 (2023), pp. 55–79.

[86] Sebastian Damrich et al. "From t-SNE to UMAP with contrastive learning". In: *The Eleventh International Conference on Learning Representations*. 2022.

[87] Shaveta Dargan et al. "Augmented Reality: A Comprehensive Review". In: *Archives of Computational Methods in Engineering* 30.2 (2023), pp. 1057–1080.

[88] Fred D Davis. "Perceived usefulness, perceived ease of use, and user acceptance of information technology". In: *MIS quarterly* (1989), pp. 319–340.

[89] J Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[90] Nianchen Deng et al. "Fov-nerf: Foveated neural radiance fields for virtual reality". In: *IEEE Transactions on Visualization and Computer Graphics* 28.11 (2022), pp. 3854–3864.

[91] J Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[92] Andrew Dewdney. "More than black and white: The extended and shared family album". In: *Family Snaps: The Meaning of Domestic Photograph, London: Virago* (1991).

[93] Giuseppe Di Maria, Lorenzo Stacchio, and Gustavo Marfia. "Unity-VRlines: Towards a Modular eXtended Reality Unity Flight Simulator". In: *International Conference on Entertainment Computing*. Springer. 2023, pp. 241–250.

[94] R DIAgostino. "An omnibus test of normality for moderate and large sample sizes". In: *Biometrika* 58.34 (1971), pp. 1–348.

[95] *Digital Market Outlook*. https://www.statista.com/outlook/244/100/fashion/worldwide. 2019.

[96] Roberto Diodato. "Virtual reality and aesthetic experience". In: *Philosophies* 7.2 (2022), p. 29.

[97] *Dior Eyes: Virtual Reality Headset*. https://luxuryretail.co.uk/dior-eyes-virtual-reality-headset/. Accessed: 2019-04-19. 2019.

[98] *Do you currently ever use a voice-operated personal assistant?* 2020. URL: https://www.statista.com/statistics/1171363/share-of-voice-assistant-users-in-the-us-by-device/ (visited on 12/04/2021).

[99] Lorenzo Donatiello et al. "Exploiting Immersive Virtual Reality for Fashion Gamification". In: *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE. 2018, pp. 17–21.

[100] A Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[101] *Double the fun - The world's first online-to-offline virtual fitting solution*. http://www.fxmirror.net/en/main. Accessed: 2018-07-22. 2018.

[102] Mihai Duguleană et al. "A virtual assistant for natural interactions in museums". In: *Sustainability* 12.17 (2020), p. 6958.

[103] Andreas Dünser and Eva Hornecker. "An observational study of children interacting with an augmented story book". In: *International Conference on Technologies for E-Learning and Digital Entertainment*. Springer. 2007, pp. 305–315.

[104] Daniele Duranti, Davide Spallazzo, and Daniela Petrelli. "Smart Objects and Replicas: A Survey of Tangible and Embodied Interactions in Museums and Cultural Heritage Sites". In: *ACM Journal on Computing and Cultural Heritage* (2023).

[105] Yogesh K Dwivedi et al. "Setting the future of digital and social media marketing research: Perspectives and research propositions". In: *International Journal of Information Management* 59 (2021), p. 102168.

[106] Yogesh K Dwivedi et al. "Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy". In: *International Journal of Information Management* 66 (2022), p. 102542.

[107] Haya Elayan, Moayad Aloqaily, and Mohsen Guizani. "Digital Twin for Intelligent Context-Aware IoT Healthcare Systems". In: *IEEE Internet of Things Journal* (2021).

[108] K J Enns and M J Martin. "Gendering Agricultural Education: A Study of Historical Pictures of Women in the Agricultural Education Magazine." In: *Journal of Agricultural Education* 56.3 (2015), pp. 69–89.

[109] Susan L Epstein. "Wanted: collaborative intelligence". In: *Artificial Intelligence* 221 (2015), pp. 36–45.

[110] Itxaro Errandonea, Sergio Beltrán, and Saioa Arrizabalaga. "Digital Twin for maintenance: A literature review". In: *Computers in Industry* 123 (2020), p. 103316.

[111] Patrick Esser et al. "Structure and content-guided video synthesis with diffusion models". In: *arXiv preprint arXiv:2302.03011* (2023).

[112] ETI. *ELETTROTECNICA IMOLESE*. https://www.eti.it/index. 2022.

[113] *Extended Reality Convergence*. https://www.qualcomm.com/news/onq/2017/05/31/extended-reality-convergence. 2019.

[114] Stephen Fai et al. "Building information modelling and heritage documentation". In: *Proceedings of the 23rd International Symposium, International Scientific Committee for Documentation of Cultural Heritage (CIPA), Prague, Czech Republic*. 2011, pp. 12–16.

[115] Ana Paula Faria and Joana Cunha. "Extended reality (XR) in the digital fashion landscape". In: *International Conference on Fashion communication: between tradition and future digital developments*. Springer. 2023, pp. 49–56.

[116] Valéria Farinazzo Martins et al. "Usability and Functionality Assessment of an Oculus Rift in Immersive and Interactive Systems Using Voice Commands". In: *Virtual, Augmented and Mixed Reality*. Ed. by Stephanie Lackey and Randall Shumaker. Cham: Springer International Publishing, 2016, pp. 222–232. ISBN: 978-3-319-39907-2.

[117] Laura Faulkner. "Beyond the five-user assumption: Benefits of increased sample sizes in usability testing". In: *Behavior Research Methods, Instruments, & Computers* 35 (2003), pp. 379–383.

[118] FEDERDOC. *I VINI ITALIANI A DENOMINAZIONE D'ORIGINE 2020*. https://www.federdoc.com/new/wp-content/uploads/2020/06/vini_italiani_denominazione_origine_2020.pdf. 2021.

[119] Michele Fiorentino et al. "Spacedesign: A mixed reality workspace for aesthetic industrial design". In: *Proceedings. International Symposium on Mixed and Augmented Reality*. IEEE. 2002, pp. 86–318.

[120] Patrick Follmann, Bertram Drost, and Tobias Böttger. "Acquire, Augment, Segment and Enjoy: Weakly Supervised Instance Segmentation of Supermarket Products". In: *Pattern Recognition*. Ed. by Thomas

Brox, Andrés Bruhn, and Mario Fritz. Cham: Springer International Publishing, 2019, pp. 363–376. ISBN: 978-3-030-12939-2.

[121] Alessandro E Foni, George Papagiannakis, and Nadia Magnenat-Thalmann. "A taxonomy of visualization strategies for cultural heritage applications". In: *Journal on Computing and Cultural Heritage (JOCCH)* 3.1 (2010), pp. 1–21.

[122] R Franzosi. "Narrative as data: linguistic and statistical tools for the quantitative study of historical events". In: *International review of social history* 43.S6 (1998), pp. 81–104.

[123] Joakim Grant Frederiksen et al. "Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: a randomized trial". In: *Surgical endoscopy* 34 (2020), pp. 1244–1252.

[124] Julián de la Fuente Prieto, Pilar Lacasa, and Rut Martıńez-Borda. "Approaching metaverses: Mixed reality interfaces in youth media platforms". In: *New Techno Humanities* 2.2 (2022), pp. 136–145.

[125] Péter Galambos et al. "Design, programming and orchestration of heterogeneous manufacturing systems through VR-powered remote collaboration". In: *Robotics and Computer-Integrated Manufacturing* 33 (2015), pp. 68–77.

[126] Theodoros Galanos, Antonios Liapis, and Georgios N Yannakakis. "Affectgan: Affect-based generative art driven by semantics". In: *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE. 2021, pp. 01–07.

[127] Kyle Gao et al. "Nerf: Neural radiance field in 3d vision, a comprehensive review". In: *arXiv preprint arXiv:2210.00379* (2022).

[128] Timnit Gebru, Oren Hazi, and Vickey Yeh. "Mobile Wine Label Recognition". In: (2022).

[129] R Stuart Geiger et al. “Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?” In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 325–336.

[130] David Gelernter. *Mirror worlds: Or the day software puts the universe in a shoebox... How it will happen and what it will mean*. Oxford University Press, 1993.

[131] Lili Geng, Yufei Li, and Yongji Xue. “Will the interest triggered by virtual reality (VR) turn into intention to travel (VR vs. Corporeal)? The moderating effects of customer segmentation”. In: *Sustainability* 14.12 (2022), p. 7010.

[132] Banda Gerald. “A brief review of independent, dependent and one sample t-test”. In: *International Journal of Applied Mathematics and Theoretical Physics* 4.2 (2018), pp. 50–54.

[133] Rod Gerber. “How do workers learn in their work?” In: *The learning organization* (1998).

[134] Vladimir Geroimenko. *Augmented Reality and Artificial Intelligence: The Fusion of Advanced Technologies*. Springer Nature, 2023.

[135] Osvaldo Gervasi, Damiano Perri, and Marco Simonetti. “Empowering knowledge with Virtual and Augmented Reality”. In: *IEEE Access* (2023).

[136] S Ginosar et al. “A century of portraits: A visual historical record of american high school yearbooks”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015, pp. 1–7.

[137] Chris A Glasbey and Kantilal Vardichand Mardia. “A review of image-warping methods”. In: *Journal of applied statistics* 25.2 (1998), pp. 155–171.

[138] Brent G Goff et al. “The influence of salesperson selling behaviors on customer satisfaction with products”. In: *Journal of retailing* 73.2 (1997), pp. 171–183.

[139] Ken Goldberg. “Robots and the return to collaborative intelligence”. In: *Nature Machine Intelligence* 1.1 (2019), pp. 2–4.

[140] Sashank Gondala et al. “Error-driven pruning of language models for virtual assistants”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 7413–7417.

[141] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[142] Google. *AI Platform Data Labeling Service*. https://cloud.google.com/ai-platform/data-labeling/docs. 2021.

[143] Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. “ChatGPT is not all you need. A State of the Art Review of large Generative AI models”. In: *arXiv preprint arXiv:2301.04655* (2023).

[144] Alex Graves et al. “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks”. In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 369–376.

[145] Amy Grech, Jörn Mehnen, and Andrew Wodehouse. “An extended AI-experience: Industry 5.0 in creative product innovation”. In: *Sensors* 23.6 (2023), p. 3009.

[146] Venugopal Gundimeda et al. “An automated computer vision system for extraction of retail food product metadata”. In: *First International Conference on Artificial Intelligence and Cognitive Computing*. Springer. 2019, pp. 199–216.

[147] Stephen Gundle and Marco Guani. “Lámericanizzazione del quotidiano. Televisione e consumismo nellÍtalia degli anni Cinquanta”. In: *Quaderni storici* (1986), pp. 561–594.

[148] Shaman Gupta and Sanjiv Kumar Jain. “A literature review of lean manufacturing”. In: *International Journal of Management Science and Engineering Management* 8.4 (2013), pp. 241–249.

[149] J Richard Hackman. *Collaborative intelligence: Using teams to solve hard problems*. Berrett-Koehler Publishers, 2011.

[150] K Han et al. “A survey on visual transformer”. In: *arXiv preprint arXiv:2012.12556* (2020).

[151] Kai Han et al. “A survey on vision transformer”. In: *IEEE transactions on pattern analysis and machine intelligence* 45.1 (2022), pp. 87–110.

[152] Ummu Hani et al. “Preserving cultural heritage through creative industry: A lesson from Saung Angklung Udjo”. In: *Procedia Economics and Finance* 4 (2012), pp. 193–200.

[153] Jacqueline-Nathalie Harba. “New approaches to customer experience: where disruptive technological innovation meets luxury fashion”. In: *Proceedings of the International Conference on Business Excellence*. Vol. 13. 1. 2019, pp. 740–758.

[154] Holger Harreis et al. “Generative AI: Unlocking the future of fashion”. In: *McKinsey & Company* (2023).

[155] Mahmoud Hassaballah and Khalid M Hosny. “Recent advances in computer vision”. In: *Studies in computational intelligence* 804 (2019), pp. 1–84.

[156] Anne-Cecilie Haugstvedt and John Krogstie. “Mobile augmented reality for cultural heritage: A technology acceptance study”. In: *2012 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE. 2012, pp. 247–255.

[157] Linjia He et al. “Am I in the theater? usability study of live performance based virtual reality”. In: *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*. 2018, pp. 1–11.

[158] Simon Hentschel, Konstantin Kobs, and Andreas Hotho. “CLIP knows image aesthetics”. In: *Frontiers in Artificial Intelligence* 5 (2022), p. 976235.

[159] Marc Herz and Philipp A Rauschnabel. “Understanding the diffusion of virtual reality glasses: The role of media, fashion and technology”. In: *Technological Forecasting and Social Change* 138 (2019), pp. 228–242.

[160] Martin Heusel et al. “Gans trained by a two time-scale update rule converge to a local nash equilibrium”. In: *Advances in neural information processing systems* 30 (2017).

[161] Oliver Hinz, Jochen Eckert, and Bernd Skiera. “Drivers of the long tail phenomenon: an empirical analysis”. In: *Journal of management information systems* 27.4 (2011), pp. 43–70.

[162] Teresa Hirzle et al. “When XR and AI Meet-A Scoping Review on Extended Reality and Artificial Intelligence”. In: *Proceedings of the 2023 CHI Conference on Human Factors in ComputingmSystems*. 2023, pp. 1–45.

[163] Jonathan Ho et al. “Imagen video: High definition video generation with diffusion models”. In: *arXiv preprint arXiv:2210.02303* (2022).

[164] *Hololens*. 2021. URL: https://docs.microsoft.com/en-us/hololens/ (visited on 04/13/2021).

[165] Wei-Lin Hsiao and Kristen Grauman. “From culture to clothing: Discovering the world events behind a century of fashion images”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1066–1075.

[166] Shun-Hsiang Hsu et al. “Defect inspection of indoor components in buildings using deep learning object detection and augmented reality”. In: *Earthquake Engineering and Engineering Vibration* 22.1 (2023), pp. 41–54.

[167] Bin Hu et al. “DiffNet: a learning to compare deep network for product recognition”. In: *IEEE Access* 8 (2020), pp. 19336–19344.

[168] Hong-zhi Hu et al. “Application and prospect of mixed reality technology in medical field”. In: *Current medical science* 39 (2019), pp. 1–6.

[169] G Huang et al. *Densely Connected Convolutional Networks*. 2018. arXiv: 1608.06993 [cs.CV].

[170] Wen Huang et al. “Motivation, engagement, and performance across multiple virtual reality sessions and levels of immersion”. In: *Journal of Computer Assisted Learning* 37.3 (2021), pp. 745–758.

[171] Matthias Husinsky et al. “Situated visualization of iiot data on the hololens 2”. In: *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE. 2022, pp. 472–476.

[172] Thien Huynh-The et al. “Artificial intelligence for the metaverse: A survey”. In: *Engineering Applications of Artificial Intelligence* 117 (2023), p. 105581.

[173] Wonil Hwang and Gavriel Salvendy. “Number of people required for usability evaluation: the 10±2 rule”. In: *Communications of the ACM* 53.5 (2010), pp. 130–133.

[174] itSeez3D Inc. *Avatar Maker - 3D avatar from a single selfie*. https://assetstore.unity.com/packages/tools/modeling/avatar-maker-free-3d-avatar-from-a-single-selfie-134782. 2022.

[175] *Increase your sales and reduce your returns thanks to accurate sizing*. https://fitle.com/en. Accessed: 2018-07-22. 2018.

[176] Medet Inkarbekov, Rosemary Monahan, and Barak A Pearlmutter. "Visualization of AI Systems in Virtual Reality: A Comprehensive Review". In: *arXiv preprint arXiv:2306.15545* (2023).

[177] *Introducing Varjo XR-3, the only true mixed reality headset.* https://varjo.com/products/xr-3/. 2021.

[178] Leila Ismail and Rajkumar Buyya. "Metaverse: A Vision, Architectural Elements, and Future Directions for Scalable and Realtime Virtual Worlds". In: *arXiv preprint arXiv:2308.10559* (2023).

[179] J Redmon. *YOLO: Real Time Object Detection.* https://github.com/pjreddie/darknet/wiki/YOLO:-Real-Time-Object-Detection. Online; accessed 3 August 2020. 2019.

[180] Stanislav Jeršov and Aleksei Tepljakov. "Digital twins in extended reality for control system applications". In: *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*. IEEE. 2020, pp. 274–279.

[181] Zhoumingju Jiang et al. "Data-driven generative design for mass customization: A case study". In: *Advanced Engineering Informatics* 54 (2022), p. 101786.

[182] Antonio Jimeno-Morenilla et al. "Using virtual reality for industrial design learning: a methodological proposal". In: *Behaviour & Information Technology* 35.11 (2016), pp. 897–906.

[183] Hai Jin et al. "Robust 3D face modeling and reconstruction from frontal and side images". In: *Computer Aided Geometric Design* 50 (2017), pp. 1–13.

[184] Dhiraj Joshi et al. "Aesthetics and emotions in images". In: *IEEE Signal Processing Magazine* 28.5 (2011), pp. 94–115.

[185] Klementina Josifovska, Enes Yigitbas, and Gregor Engels. “A digital twin-based multi-modal ui adaptation framework for assistance systems in industry 4.0”. In: *International Conference on Human-Computer Interaction*. Springer. 2019, pp. 398–409.

[186] Jeong-Mun Jung et al. “Wine Label Recognition System using Image Similarity”. In: *The Journal of the Korea Contents Association* 11.5 (2011), pp. 125–137.

[187] H Kaiming et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[188] Johanna Karras et al. “DreamPose: Fashion Video Synthesis with Stable Diffusion”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 22680–22690.

[189] Daniel Katz et al. “Utilization of a voice-based virtual reality advanced cardiac life support team leader refresher: prospective observational study”. In: *Journal of medical Internet research* 22.3 (2020), e17425.

[190] Jaspreet Kaur et al. “Consumer behavior in the metaverse”. In: *Journal of Consumer Behaviour* (2023).

[191] Melissa G Keith, Louis Tay, and Peter D Harms. “Systems perspective of Amazon Mechanical Turk for organizational research: Review and recommendations”. In: *Frontiers in psychology* 8 (2017), p. 1359.

[192] Jihyun Kim, Ann Marie Fiore, and Hyun-Hwa Lee. “Influences of online store perception, shopping enjoyment, and shopping involvement on consumer patronage behavior towards an online retailer”. In: *Journal of retailing and Consumer Services* 14.2 (2007), pp. 95–107.

[193] Youn-Kyung Kim, Jikyeong Kang, and Minsung Kim. “The relationships among family and social interaction, loneliness, mall shopping motivation, and mall spending of older consumers”. In: *Psychology & Marketing* 22.12 (2005), pp. 995–1015.

[194] Alexander Kirillov et al. “Segment anything”. In: *arXiv preprint arXiv:2304.02643* (2023).

[195] Elif Hilal Korkut and Elif Surer. “Visualization in virtual reality: a systematic review”. In: *Virtual Reality* (2023), pp. 1–34.

[196] Ashutosh Kumar, Arijit Biswas, and Subhajit Sanyal. “ecommercegan: A generative adversarial network for e-commerce”. In: *arXiv preprint arXiv:1801.03244* (2018).

[197] Sang Gyu Kwak and Jong Hae Kim. “Central limit theorem: the cornerstone of modern statistics”. In: *Korean journal of anesthesiology* 70.2 (2017), p. 144.

[198] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. “Diffusion models already have a semantic latent space”. In: *arXiv preprint arXiv:2210.10960* (2022).

[199] Benjamin Lee, Michael Sedlmair, and Dieter Schmalstieg. “Design Patterns for Situated Visualization in Augmented Reality”. In: *IEEE Transactions on Visualization and Computer Graphics* (2023).

[200] J Lemley, S Bazrafkan, and P Corcoran. “Deep Learning for Consumer Devices and Services: Pushing the limits for machine learning, artificial intelligence, and computer vision.” In: *IEEE Consumer Electronics Magazine* 6.2 (2017), pp. 48–56.

[201] Vladimir I Levenshtein et al. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet physics doklady*. Vol. 10. 8. Soviet Union. 1966, pp. 707–710.

[202] Michael S Lew et al. “Content-based multimedia information retrieval: State of the art and challenges”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 2.1 (2006), pp. 1–19.

[203] Clayton Lewis. *Using the" thinking-aloud" method in cognitive interface design*. IBM TJ Watson Research Center Yorktown Heights, NY, 1982.

[204] Wei Li et al. "Crowd intelligence in AI 2.0 era". In: *Frontiers of Information Technology & Electronic Engineering* 18 (2017), pp. 15–43.

[205] Xiaoqing Li and Jinwen Ma. "Distributed search and fusion for wine label image retrieval". In: *PeerJ Computer Science* 8 (2022), e1116.

[206] Xiaoqing Li, Jiansheng Yang, and Jinwen Ma. "CNN-SIFT consecutive searching and matching for wine label retrieval". In: *International Conference on Intelligent Computing*. Springer. 2019, pp. 250–261.

[207] Y Li et al. "Deep Metric Learning With Density Adaptivity". In: *IEEE Transactions on Multimedia* 22.5 (2019), pp. 1285–1297.

[208] Zewen Li et al. "A survey of convolutional neural networks: analysis, applications, and prospects". In: *IEEE transactions on neural networks and learning systems* (2021).

[209] Paul Pu Liang et al. "Multiviz: Towards visualizing and understanding multimodal models". In: *arXiv preprint arXiv:2207.00056* (2022).

[210] Rensis Likert. "A technique for the measurement of attitudes." In: *Archives of psychology* (1932).

[211] Kendrik Yan Hong Lim et al. "A digital twin-enhanced system for engineering product family design and optimization". In: *Journal of Manufacturing Systems* 57 (2020), pp. 82–93.

[212] Mingyuan Lin, Longhua Ma, and Binchao Yu. "An Efficient and Lightweight Detector for Wine Bottle Defects". In: *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE. 2020, pp. 957–962.

[213] M Lincoln et al. "CAMPI: Computer-Aided Metadata Generation for Photo archives Initiative". In: (2020).

[214] livingwinelabels. *livingwinelabels*. https://www.livingwinelabels.com/. 2021.

[215] Frank Loh et al. "Youtube dataset on mobile streaming for internet traffic modeling and streaming analysis". In: *Scientific Data* 9.1 (2022), p. 293.

[216] Francesco Longo, Letizia Nicoletti, and Antonio Padovano. "Smart operators in industry 4.0: A human-centered approach to enhance operators' capabilities and competencies within the new smart factory context". In: *Computers & industrial engineering* 113 (2017), pp. 144–159.

[217] Francesco Longo, Letizia Nicoletti, and Antonio Padovano. "New perspectives and results for Smart Operators in industry 4.0: A human-centered approach". In: *Computers & Industrial Engineering* 163 (2022), p. 107824.

[218] David G Lowe. "Object recognition from local scale-invariant features". In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee. 1999, pp. 1150–1157.

[219] Leanne Luce. *Artificial intelligence for fashion: How AI is revolutionizing the fashion industry*. Apress, 2018.

[220] Zhengxiong Luo et al. "VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 10209–10218.

[221] Jiayi Ma et al. "Image matching from handcrafted to deep features: A survey". In: *International Journal of Computer Vision* 129.1 (2021), pp. 23–79.

[222] Xin Ma et al. "Digital twin enhanced human-machine interaction in product lifecycle". In: *Procedia Cirp* 83 (2019), pp. 789–793.

[223] Ardeshir Mahdavi and Hesham Eissa. “Subjective evaluation of architectural lighting via computationally rendered images”. In: *Journal of the illuminating Engineering Society* 31.2 (2002), pp. 11–20.

[224] Alexis DJ Makin. “The gap between aesthetic science and aesthetic experience”. In: *Journal of Consciousness Studies* 24.1-2 (2017), pp. 184–213.

[225] Henry B Mann and Donald R Whitney. “On a test of whether one of two random variables is stochastically larger than the other”. In: *The annals of mathematical statistics* (1947), pp. 50–60.

[226] Steve Mann et al. “eXtended meta-uni-omni-Verse (XV): Introduction, Taxonomy, and State-of-the-Art”. In: *IEEE Consumer Electronics Magazine* (2023).

[227] Adeline Manuel, Philippe Véron, and Livio De Luca. “2D/3D semantic annotation of spatialized images for the documentation and analysis of cultural heritage”. In: *14th EUROGRAPHICS Workshop on Graphics and Cultural Heritage*. Eurographics. 2016.

[228] Dietrich Manzey, Juliane Reichenbach, and Linda Onnasch. “Human performance consequences of automated decision aids: The impact of degree of automation and system experience”. In: *Journal of Cognitive Engineering and Decision Making* 6.1 (2012), pp. 57–87.

[229] Gustavo Marfia and Lorenzo Stacchio. “Fashion in the Metaverse: Technologies, Applications, and Opportunities”. In: ().

[230] Juan A Marin-Garcia and Tomas Bonavia. “Relationship between employee involvement and lean manufacturing and its effect on performance in a rigid continuous process industry”. In: *International Journal of Production Research* 53.11 (2015), pp. 3260–3275.

[231] Slobodan Marković et al. “Aesthetic experience and the emotional content of paintings”. In: *Psihologija* 43.1 (2010), pp. 47–64.

[232] Nuno Cid Martins et al. “Augmented reality situated visualization in decision-making”. In: *Multimedia Tools and Applications* 81.11 (2022), pp. 14749–14772.

[233] Mercedes Marzo-Navarro, Marta Pedraja-Iglesias, and Ma Pilar Rivera-Torres. “The benefits of relationship marketing for the consumer and for the fashion retailers”. In: *Journal of Fashion Marketing and Management: An International Journal* 8.4 (2004), pp. 425–436.

[234] Marta Massi, Marilena Vecco, and Yi Lin. “Digital Transformation in the Cultural and Creative Industries”. In: *Digital Transformation in the Cultural and Creative Industries* (2020), pp. 1–9.

[235] McGill and Mark. “White Paper-The IEEE Global Initiative on Ethics of Extended Reality (XR) Report–Extended Reality (XR) and the Erosion of Anonymity and Privacy”. In: *Extended Reality (XR) and the Erosion of Anonymity and Privacy-White Paper* (2021), pp. 1–24.

[236] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[237] Leland McInnes et al. “UMAP: Uniform Manifold Approximation and Projection”. In: *The Journal of Open Source Software* 3.29 (2018), p. 861.

[238] Suzanne Meeks, Sarah Kelly Shryock, and Russell J Vandenbroucke. “Theatre involvement and well-being, age differences, and lessons from long-time subscribers”. In: *The Gerontologist* 58.2 (2018), pp. 278–289.

[239] D Menaga and S Revathi. “Deep learning: a recent computing platform for multimedia information retrieval”. In: *Deep learning techniques and optimization strategies in big data analytics*. IGI Global, 2020, pp. 124–141.

[240] Michele A. Fino. *Questione di Etichetta*. https://www.spazioprever.it/salabar/vino/pdf/Questione_di_etichetta.pdf. 2013.

[241] Joel Michell. *Measurement in psychology: A critical history of a methodological concept*. Vol. 53. Cambridge University Press, 1999.

[242] Paul Milgram and Fumio Kishino. "A taxonomy of mixed reality visual displays". In: *IEICE TRANSACTIONS on Information and Systems* 77.12 (1994), pp. 1321–1329.

[243] Georgios Minopoulos and Konstantinos E Psannis. "Opportunities and challenges of tangible XR applications for 5G networks and beyond". In: *IEEE Consumer Electronics Magazine* (2022).

[244] Silvia Mirri, Marco Roccetti, and Paola Salomoni. "Collaborative design of software applications: the role of users". In: *Human-centric Computing and Information Sciences* 8.1 (2018), pp. 1–20.

[245] G Mitman and K Wilder. *Documenting the world: film, photography, and the scientific record*. Univ. of Chicago Press, 2016.

[246] Emmanuel Mogaji, Yogesh K Dwivedi, and Ramakrishnan Raman. "Fashion marketing in the metaverse". In: *Journal of Global Fashion Marketing* (2023), pp. 1–16.

[247] Adrià Molina et al. "Date Estimation in the Wild of Scanned Historical Photos: An Image Retrieval Approach". In: *International Conference on Document Analysis and Recognition*. Springer. 2021, pp. 306–320.

[248] MoMA. *Vernacular photography*. https://www.moma.org/collection/terms/vernacular-photography. 2020.

[249] M Moranges et al. "Explicit and implicit measures of emotions: Data-science might help to account for data complexity and heterogeneity". In: *Food Quality and Preference* 92 (2021), p. 104181.

[250] Elena Morotti, Lorenzo Donatiello, and Gustavo Marfia. "Fostering fashion retail experiences through virtual reality and voice assistants". In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE. 2020, pp. 338–342.

[251] Elena Morotti et al. “Exploiting fashion x-commerce through the empowerment of voice in the fashion virtual reality arena: Integrating voice assistant and virtual reality technologies for fashion communication”. In: *Virtual Reality* (2022), pp. 1–14.

[252] Dimitris Mourtzis, John Angelopoulos, and Nikos Panopoulos. “Operator 5.0: A survey on enabling technologies and a framework for digital manufacturing based on extended reality”. In: *Journal of Machine Engineering* 22 (2022).

[253] Dimitris Mourtzis, Vasilios Zogopoulos, and E Vlachou. “Augmented reality application to support remote maintenance as a service in the robotics industry”. In: *Procedia Cirp* 63 (2017), pp. 46–51.

[254] Xiangyu Mu et al. “Fashion Intelligence in the Metaverse: Promise and Future Prospects”. In: (2023).

[255] Muhanna A Muhanna. “Virtual reality and the CAVE: Taxonomy, interaction challenges and research directions”. In: *Journal of King Saud University-Computer and Information Sciences* 27.3 (2015), pp. 344–361.

[256] Abhishek Mukhopadhyay et al. “Virtual-reality-based digital twin of office spaces with social distance measurement feature”. In: *Virtual Reality & Intelligent Hardware, XXXX, XX (XX)* (2021), pp. 1–21.

[257] Tony Mullen. *Mastering blender*. John Wiley & Sons, 2011.

[258] Eric Müller, Matthias Springstein, and Ralph Ewerth. ““When was this picture taken?”–Image date estimation in the wild”. In: *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings 39*. Springer. 2017, pp. 619–625.

[259] Luis Muñoz-Saavedra, Lourdes Mir’o-Amarante, and Manuel Dom’ınguez-Morales. “Augmented and virtual reality evolution and future tendency”. In: *Applied sciences* 10.1 (2020), p. 322.

[260] *Mycroft site*. 2019. URL: https://mycroft.ai/ (visited on 12/03/2019).

[261] In Seop Na, Yan Juan Chen, and Soo Hyung Kim. "Automatic Segmentation of Product Bottle Label Based on GrabCut Algorithm". In: *International Journal of Contents* 10.4 (2014), pp. 1–10.

[262] Hyeonseob Nam et al. "Reducing domain gap by reducing style bias". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 8690–8699.

[263] Tam N Nguyen. "Toward human digital twins for cybersecurity simulations on the metaverse: Ontological and network science approach". In: *JMIRx Med* 3.2 (2022), e33502.

[264] Kjetil Liestøl Nielsen. "Students' video viewing habits during a flipped classroom course in engineering mathematics". In: (2020).

[265] Sirke Nieminen et al. "The development of aesthetic responses to music and their underlying neural and psychological mechanisms". In: *Cortex* 47.9 (2011), pp. 1138–1146.

[266] Huansheng Ning et al. "A Survey on the Metaverse: The State-of-the-Art, Technologies, Applications, and Challenges". In: *IEEE Internet of Things Journal* (2023).

[267] Louis Nisiotis et al. "Interwoven Spaces with XR, AI, and Robots: Merging Realities in Space and Time". In: *Museums and Technologies of Presence*. Routledge, 2023, pp. 243–261.

[268] Yves S Nkulu-Ily. "Combining XR and AI for Integrating the Best Pedagogical Approach to Providing Feedback in Surgical Medical Distance Education". In: *International Conference on Intelligent Tutoring Systems*. Springer. 2023, pp. 452–466.

[269] NVIDIA. *Understanding Aesthetics in Deep Learning*. https://developer.nvidia.com/blog/understanding-aesthetics-deep-learning/. 2016.

[270] Heather L O'Brien and Elaine G Toms. "The development and evaluation of a survey to measure user engagement". In: *Journal of the American Society for Information Science and Technology* 61.1 (2010), pp. 50–69.

[271] Justin O'Connor. *The cultural and creative industries: a review of the literature*. Arts Council England, 2007.

[272] Samuel D Okegbile et al. "Human digital twin for personalized healthcare: Vision, architecture and future directions". In: *IEEE network* 37.2 (2022), pp. 262–269.

[273] Alessandro Orsini et al. "Augmented reality enhanced cooking with Microsoft HoloLens". In: *Rutgers, State University of New Jersey* (2017).

[274] Krishan Pal and Mayank Sharma. "Performance evaluation of non-linear techniques UMAP and t-SNE for data in higher dimensional topological space". In: *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. IEEE. 2020, pp. 1106–1110.

[275] F Palermo, J Hays, and A A Efros. "Dating historical color images". In: *European Conference on Computer Vision*. Springer. 2012, pp. 499–512.

[276] Julie Pallant. "SPSS survival manual 4th edition". In: *Everbest Printing* (2011).

[277] Savvas Papagiannidis et al. "To immerse or not? Experimenting with two virtual retail environments". In: *Information Technology & People* 30.1 (2017), pp. 163–188.

[278] Minjung Park, Hyunjoo Im, and Do Yuon Kim. "Feasibility and user experience of virtual reality fashion stores". In: *Fashion and Textiles* 5.1 (2018), pp. 1–17.

[279] Lara Penco et al. “Mobile augmented reality as an internationalization tool in the “Made In Italy” food and beverage industry”. In: *Journal of Management and Governance* 25.4 (2021), pp. 1179–1209.

[280] M R Peres. *The concise Focal encyclopedia of photography: from the first photo on paper to the digital revolution*. CRC Press, 2014.

[281] Aravin Prince Periyasamy and Saravanan Periyasami. “Rise of digital fashion and metaverse: influence on sustainability”. In: *Digital Economy and Sustainable Development* 1.1 (2023), pp. 1–26.

[282] Adolfo Perrusquiá and Wen Yu. “Human-in-the-loop control using euler angles”. In: *Journal of Intelligent & Robotic Systems* 97 (2020), pp. 271–285.

[283] T H Phan and K Yamamoto. *Resolving Class Imbalance in Object Detection with Weighted Cross Entropy Losses*. 2020. arXiv: 2006.01413 [cs.CV].

[284] Colin Phelan and Julie Wren. “Exploring reliability in academic assessment”. In: *UNI Office of Academic Assessment* (2006), pp. 92005–2006.

[285] Flavia Piancazzo. “Developments of Cultural Appropriation in Fashion: An In-Progress Research”. In: *International Conference on Fashion communication: between tradition and future digital developments*. Springer. 2023, pp. 136–143.

[286] Theodora Pistola et al. “Creating immersive experiences based on intangible cultural heritage”. In: *2021 IEEE International Conference on Intelligent Reality (ICIR)*. IEEE. 2021, pp. 17–24.

[287] D. Polap. “Voice Control in Mixed Reality”. In: *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*. Sept. 2018, pp. 497–500.

[288] Giorgia Profumo et al. “Metaverse and the fashion industry: A systematic literature review”. In: *Journal of Global Fashion Marketing* (2023), pp. 1–24.

[289] Jon Prosser. “The status of image-based research”. In: *Image-based research: A sourcebook for qualitative researchers* (1998), pp. 97–112.

[290] PTC. *Vivino and Vuforia’s Image Recognition Solution Make a Great Pairing*. https://www.ptc.com/en/case-studies/vivino. 2022.

[291] Chan Qiu et al. “Digital assembly technology based on augmented reality and digital twins: a review”. In: *Virtual Reality & Intelligent Hardware* 1.6 (2019), pp. 597–610.

[292] X Qiu et al. “Ensemble deep learning for regression and time series forecasting”. In: *2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL)*. 2014, pp. 1–6. DOI: 10.1109/CIEL.2014.7015739.

[293] Joaquin Quinonero-Candela et al. *Dataset shift in machine learning*. Cambridge: Mit Press, 2008.

[294] Filip Radenović, Giorgos Tolias, and Ondřej Chum. “Fine-tuning CNN image retrieval with no human annotation”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.7 (2018), pp. 1655–1668.

[295] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[296] Jason Radford and Kenneth Joseph. “Theory in, theory out: the uses of social theory in machine learning for social science”. In: *Frontiers in big Data* 3 (2020), p. 18.

[297] M Raghu et al. “Do Vision Transformers See Like Convolutional Neural Networks?” In: *Thirty-Fifth Conference on Neural Information Processing Systems*. 2021.

[298] B Rainer et al. “Statistically indifferent quality variation: An approach for reducing multimedia distribution cost for adaptive video streaming services”. In: *IEEE Transactions on Multimedia* 19.4 (2016), pp. 849–860.

[299] Ismo Rakkolainen et al. “Technologies for multimodal interaction in extended reality—a scoping review”. In: *Multimodal Technologies and Interaction* 5.12 (2021), p. 81.

[300] Monika Raková. “Using of intellectual property rights in a creative industry in a global dimension”. In: *SHS Web of Conferences*. Vol. 92. EDP Sciences. 2021, p. 03024.

[301] Adil Rasheed, Omer San, and Trond Kvamsdal. “Digital twin: Values, challenges and enablers from a modeling perspective”. In: *Ieee Access* 8 (2020), pp. 21980–22012.

[302] Jay Ratican, James Hutson, and Andrew Wright. “A proposed meta-reality immersive development pipeline: Generative ai models and extended reality (xr) content for the metaverse”. In: *Journal of Intelligent Learning Systems and Applications* 15 (2023).

[303] Philipp A Rauschnabel, Reto Felix, and Chris Hinsch. “Augmented reality marketing: How mobile AR-apps can improve brands through inspiration”. In: *Journal of Retailing and Consumer Services* 49 (2019), pp. 43–53.

[304] Philipp A Rauschnabel et al. “Fashion or technology? A fashnology perspective on the perception and adoption of augmented reality smart glasses”. In: *i-com* 15.2 (2016), pp. 179–194.

[305] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[306] Alan J Reid and Alan J Reid. “A brief history of the smartphone”. In: *The Smartphone Paradox: Our Ruinous Dependency in the Device Age* (2018), pp. 35–66.

[307] Dirk Reiners et al. “The Combination of Artificial Intelligence and Extended Reality: A Systematic Review”. In: *Frontiers in Virtual Reality* 2 (2021), p. 721933.

[308] Abderahman Rejeb, Karim Rejeb, and John G Keogh. “Enablers of augmented reality in the food supply chain: a systematic literature review”. In: *Journal of Foodservice Business Research* 24.4 (2021), pp. 415–444.

[309] Alexandra Rese et al. “How augmented reality apps are accepted by consumers: A comparative analysis using scales and opinions”. In: *Technological Forecasting and Social Change* 124 (2017), pp. 306–319.

[310] Taina Ribeiro de Oliveira et al. “Systematic Review of Virtual Reality Solutions Employing Artificial Intelligence Methods”. In: *Symposium on Virtual and Augmented Reality*. 2021, pp. 42–55.

[311] Marina Ricci et al. “Assessing the impact of immersive versus desktop virtual reality shopping experiences in the fashion industry metaverse”. In: *Connectivity and creativity in times of conflict*. Academia Press, 2023, p. 5. URL: https://library.oapen.org/handle/20.500.12657/85848.

[312] Marco Roccetti et al. “Potential and limitations of designing a deep learning model for discovering new archaeological sites: A case with the Mesopotamian floodplain”. In: *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*. 2020, pp. 216–221.

[313] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[314] David Romero et al. “Social factory architecture: Social networking services and production scenarios through the social internet of things, services and people for the social operator 4.0”. In: *Advances in Production Management Systems. The Path to Intelligent, Collaborative*

and Sustainable Manufacturing: IFIP WG 5.7 International Conference, APMS 2017, Hamburg, Germany, September 3-7, 2017, Proceedings, Part I. Springer. 2017, pp. 265–273.

[315] Daniela Rosner, Marco Roccetti, and Gustavo Marfia. "The digitization of cultural practices". In: *Communications of the ACM* 57.6 (2014), pp. 82–87.

[316] Ludan Ruan et al. "Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 10219–10228.

[317] Olga Russakovsky et al. *ImageNet Large Scale Visual Recognition Challenge*. 2015. arXiv: 1409.0575 [cs.CV].

[318] Stuart J Russell and Peter Norvig. *Artificial intelligence a modern approach*. London, 2010.

[319] J Tumblin S Paris P Kornprobst and F Durand. "A Gentle Introduction to Bilateral Filtering and Its Applications". In: *ACM SIGGRAPH 2007 Courses*. SIGGRAPH '07. San Diego, California: Association for Computing Machinery, 2007, 1–es. ISBN: 9781450318235. DOI: 10.1145/1281500.1281602. URL: https://doi.org/10.1145/1281500.1281602.

[320] Chitwan Saharia et al. "Photorealistic text-to-image diffusion models with deep language understanding". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 36479–36494.

[321] T Salem et al. "Analyzing human appearance as a cue for dating images". In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2016, pp. 1–8.

[322] Nareen OM Salim et al. "Study for food recognition system using deep learning". In: *Journal of Physics: Conference Series*. Vol. 1963. 1. IOP Publishing. 2021, p. 012014.

[323] Paola Salomoni et al. “Diegetic user interfaces for virtual environments with HMDs: a user experience study with oculus rift”. In: *Journal on Multimodal User Interfaces* 11 (2017), pp. 173–184.

[324] Hani Sami et al. “The Metaverse: Survey, Trends, Novel Pipeline Ecosystem & Future Directions”. In: *arXiv preprint arXiv:2304.09240* (2023).

[325] Ali Samini, Karljohan Lundin Palmerius, and Patric Ljung. “A Review of Current, Complete Augmented Reality Solutions”. In: *2021 International Conference on Cyberworlds (CW)*. IEEE. 2021, pp. 49–56.

[326] M Sandbye. “Looking at the family photo album: a resumed theoretical discussion of why and how”. In: *Journal of Aesthetics & Culture* 6.1 (2014), p. 25419.

[327] Aditya Sanghi et al. “Clip-forge: Towards zero-shot text-to-shape generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18603–18613.

[328] Jeff Sauro and James R Lewis. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016.

[329] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. “Do datasets have politics? Disciplinary values in computer vision dataset development”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–37.

[330] Ines Schindler et al. “Measuring aesthetic emotions: A review of the literature and a new assessment tool”. In: *PloS one* 12.6 (2017), e0178899.

[331] C Schreiber. “The construction of ‘female citizens’: a socio-historical analysis of girls’ education in Luxembourg”. In: *Educational Research* 56.2 (2014), pp. 137–154.

[332] Jan-Henrik Schröder et al. “Collaborating Across Realities: Analytical Lenses for Understanding Dyadic Collaboration in Transitional Interfaces”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–16.

[333] Greyce Schroeder et al. “Visualising the digital twin using web services and augmented reality”. In: *2016 IEEE 14th international conference on industrial informatics (INDIN)*. IEEE. 2016, pp. 522–527.

[334] Thomas W Schubert. “The sense of presence in virtual environments: A three-component scale measuring spatial presence, involvement, and realness.” In: *Z. für Medienpsychologie* 15.2 (2003), pp. 69–71.

[335] Claudia Scorolli et al. “Would you rather come to a tango concert in theater or in VR? Aesthetic emotions & social presence in musical experiences, either live, 2D or 3D”. In: *Computers in Human Behavior* 149 (2023), p. 107910.

[336] T J Sejnowski. *The deep learning revolution*. MIT press, 2018.

[337] R R Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL: http://dx.doi.org/10.1007/s11263-019-01228-7.

[338] Samad ME Sepasgozar. “Digital twin and Web-based virtual gaming technologies for online education: A case of construction management and engineering”. In: *Applied Sciences* 10.13 (2020), p. 4678.

[339] E Serafinelli. *Digital life on Instagram: New social communication of photography*. Emerald Group Publishing, 2018.

[340] Baoguang Shi, Xiang Bai, and Cong Yao. “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.11 (2016), pp. 2298–2304.

[341] Jonathon Shlens. “Notes on kullback-leibler divergence and likelihood”. In: *arXiv preprint arXiv:1404.2000* (2014).

[342] Manuel Silva and Luıś Teixeira. “eXtended Reality (XR) experiences in museums for cultural heritage: A systematic review”. In: *International Conference on Intelligent Technologies for Interactive Entertainment*. Springer. 2021, pp. 58–79.

[343] Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin. “A survey of OCR applications”. In: *International Journal of Machine Learning and Computing* 2.3 (2012), p. 314.

[344] Harry Singh, Chetna Singh, and Rana Majumdar. “Virtual reality as a marketing tool”. In: *Emerging Trends in Expert Applications and Security*. Springer, 2019, pp. 445–450.

[345] Khushboo Singh et al. “Evaluation Planning for Artificial Intelligence-based Industry 6.0 Metaverse Integration”. In: *Intelligent Human Systems Integration (IHSI 2023): Integrating People and Intelligent Systems* 69.69 (2023).

[346] Kirill Smelyakov et al. “Effectiveness of Modern Text Recognition Solutions and Tools for Common Data Sources”. In: *CEUR Workshop Proceedings*. 2021, pp. 154–165.

[347] *Snips*. 2019. URL: https://snips.ai/ (visited on 12/03/2019).

[348] Kwonsang Sohn et al. “Artificial intelligence in the fashion industry: consumer responses to generative adversarial network (GAN) technology”. In: *International Journal of Retail & Distribution Management* 49.1 (2020), pp. 61–80.

[349] Andreas Sonderegger et al. “Food talks: visual and interaction principles for representing environmental and nutritional food information in augmented reality”. In: *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE. 2019, pp. 98–103.

[350] Junrong Song et al. “From Expanded Cinema to Extended Reality: How AI Can Expand and Extend Cinematic Experiences”. In: *Proceedings of the 16th International Symposium on Visual Information Communication and Interaction*. 2023, pp. 1–5.

[351] Yu Song. “Human digital twin, the development and impact on design”. In: *Journal of Computing and Information Science in Engineering* 23.6 (2023), p. 060819.

[352] P Sorcinelli. “Imago. Laboratorio di ricerca storica e di documentazione iconografica sulla condizione giovanile nel XX secolo”. In: *Rivista di storia e storiografia* 5.5 (Nov. 2004), pp. 200–202.

[353] Luiś Fernando de Souza Cardoso, Flávia Cristina Martins Queiroz Mariano, and Ezequiel Roberto Zorzal. “A survey of industrial augmented reality”. In: *Computers & Industrial Engineering* 139 (2020), p. 106159.

[354] Marco Speicher. “Shopping in Virtual Reality”. In: *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE. 2018, pp. 1–2.

[355] Marco Speicher, Sebastian Cucerca, and Antonio Krüger. “VRShop: a mobile interactive virtual reality shopping environment combining the benefits of on-and offline shopping”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (2017), pp. 1–31.

[356] Maximilian Speicher, Brian D Hall, and Michael Nebeling. “What is mixed reality?” In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–15.

[357] Lorenzo Stacchio, Alessia Angeli, and Gustavo Marfia. “Empowering Locksmith Crafts via Mobile Augmented Reality”. In: *Proceedings of the Conference on Information Technology for Social Good*. 2021, pp. 305–308.

[358] Lorenzo Stacchio, Alessia Angeli, and Gustavo Marfia. “Empowering digital twins with extended reality collaborations”. In: *Virtual Reality & Intelligent Hardware* 4.6 (2022), pp. 487–505.

[359] Lorenzo Stacchio, Giuseppe Di Maria, and Gustavo Marfia. “Flying in XR: Bridging Desktop applications in eXtended Reality through Deep Learning”. In: *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE. 2024, to appear.

[360] Lorenzo Stacchio, Shirin Hajahmadi, and Gustavo Marfia. “Preserving family album photos with the hololens 2”. In: *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE. 2021, pp. 643–644.

[361] Lorenzo Stacchio, Claudia Scorolli, and Gustavo Marfia. “Evaluating Human Aesthetic and Emotional Aspects of 3D generated content through eXtended Reality”. In: (2023).

[362] Lorenzo Stacchio et al. “Imago: A family photo album dataset for a socio-historical analysis of the twentieth century”. In: *arXiv preprint arXiv:2012.01955* (2020).

[363] Lorenzo Stacchio et al. “Revive family photo albums through a collaborative environment exploiting the HoloLens 2”. In: *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE. 2021, pp. 378–383.

[364] Lorenzo Stacchio et al. “Applying deep learning approaches to mixed quantitative-qualitative analyses”. In: *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*. 2022, pp. 161–166.

[365] Lorenzo Stacchio et al. “Rethinking Augmented Wine Recognition”. In: *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE. 2022, pp. 560–565.

[366] Lorenzo Stacchio et al. “Searching for cultural relationships through deep learning models”. In: (2022).

[367] Lorenzo Stacchio et al. “Toward a Holistic Approach to the Socio-historical Analysis of Vernacular Photos”. In: *ACM Transactions on Multimedia Computing, Communications and Applications* 18.3s (2022), pp. 1–23.

[368] Lorenzo Stacchio et al. “Who will Trust my Digital Twin? Maybe a Clerk in a Brick and Mortar Fashion Shop”. In: *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE. 2022, pp. 814–815.

[369] Lorenzo Stacchio et al. “Analyzing cultural relationships visual cues through deep learning models in a cross-dataset setting”. In: *Neural Computing and Applications* (2023), pp. 1–16.

[370] Lorenzo Stacchio et al. “AnnHoloTator: A Mixed Reality Collaborative Platform for Manufacturing Work Instruction Interaction”. In: *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE. 2023, pp. 418–424.

[371] Lorenzo Stacchio et al. “M-AGEW: Empowering Outdoor Workouts with Data-Driven Augmented Reality Assistance”. In: *International Conference on Artificial Intelligence and Virtual Reality*. IEEE. 2024, to appear.

[372] Lorenzo Stacchio et al. “WiXaRd: Towards a holistic distributed platform for multi-party and cross-reality WebXR experiences”. In: *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE. 2024, to appear.

[373] Evgeny Stemasov et al. “ShapeFindAR: Exploring in-situ spatial search for physical artifact retrieval using mixed reality”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–12.

[374] Manfred Stommel and J. Dontje Katherine. *Statistics for Advanced Practice Nurses and Health Professionals - Appendix C*. Springer Publishing Company, 2014, p. 349. ISBN: 978-0-8261-9824-2. DOI: 10.1891/9780826198259. URL: https://connect.springerpub.com/content/book/978-0-8261-9825-9/back-matter/bmatter3.

[375] Susanne Stricker, Rolf AE Mueller, and Daniel A Sumner. "Marketing wine on the web". In: *Choices* 22.316-2016-6376 (2007), pp. 31–34.

[376] Georgios D Styliaras. "Augmented Reality in Food Promotion and Analysis: Review and Potentials". In: *Digital* 1.4 (2021), pp. 216–240.

[377] Zhengwentai Sun et al. "SGDiff: A Style Guided Diffusion Model for Fashion Synthesis". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 8433–8442.

[378] Mirko Suznjevic, Matija Mandurov, and Maja Matijasevic. "Performance and QoE assessment of HTC Vive and Oculus Rift for pick-and-place tasks in VR". In: *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE. 2017, pp. 1–3.

[379] Niladri Syam and Arun Sharma. "Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice". In: *Industrial marketing management* 69 (2018), pp. 135–146.

[380] C Szegedy et al. *Rethinking the Inception Architecture for Computer Vision*. 2015. arXiv: 1512.00567 [cs.CV].

[381] L Tilton T Arnold. "Distant viewing: analyzing large visual corpora". In: *Digital Scholarship in the Humanities* 34.Supplement_1 (2019), pp. i3–i16.

[382] T Nguyen. *Yolo face implementation*. https://github.com/sthanhng/yoloface. Online; accessed 3 August 2020. 2018.

[383] Keith S Taber. "The use of Cronbach's alpha when developing and reporting research instruments in science education". In: *Research in science education* 48.6 (2018), pp. 1273–1296.

[384] Fei Tao and Meng Zhang. "Digital twin shop-floor: a new shop-floor paradigm towards smart manufacturing". In: *Ieee Access* 5 (2017), pp. 20418–20427.

[385] Fei Tao et al. "Digital twin in industry: State-of-the-art". In: *IEEE Transactions on Industrial Informatics* 15.4 (2018), pp. 2405–2415.

[386] Abbas Tashakkori and John W Creswell. *The new era of mixed methods*. 2007.

[387] *The Topshop Virtual Reality Experience AW14*. https://www.youtube.com/watch?v=lUal_Lrhec0. Accessed: 2019-09-02. 2019.

[388] TinEye. *WineEngine is image recognition for the beverage industry*. https://services.tineye.com/WineEngine. 2021.

[389] Iwaki Toshima et al. "Challenges facing human digital twin computing and its future prospects". In: *NTT Technical Review* (2020). ISSN: 13483447.

[390] H Touvron et al. "Training data-efficient image transformers & distillation through attention". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10347–10357.

[391] Stefano Triberti et al. "Developing emotional design: Emotions as cognitive processes and their role in the design of interactive technologies". In: *Frontiers in psychology* 8 (2017), p. 1773.

[392] Ultralytics. *Yolo v5*. https://github.com/ultralytics/yolov5. Online; accessed 06 June 2021. 2021.

[393] UNESCO. *World heritage, humanity's gift to future*. https://whc.unesco.org/en/activities/487/. Online; accessed 06 June 2021. 2021.

[394] Dorin Ungureanu et al. “HoloLens 2 Research Mode as a Tool for Computer Vision Research”. In: *arXiv preprint arXiv:2008.11239* (2020).

[395] *US Virtual and Augmented Reality Users 2020*. https://www.emarketer.com/content/us-virtual-and-augmented-reality-users-2020. 2021.

[396] F Vaccaro et al. “Image Retrieval using Multi-scale CNN Features Pooling”. In: *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 2020, pp. 311–315.

[397] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[398] Helena Van Kerrebroeck, Malaika Brengman, and Kim Willems. “When brands come to life: experimental research on the vividness effect of Virtual Reality in transformational marketing communications”. In: *Virtual Reality* 21.4 (2017), pp. 177–191.

[399] A Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[400] Viswanath Venkatesh, Venkataraman Ramesh, and Anne P Massey. “Understanding usability in mobile commerce”. In: *Communications of the ACM* 46.12 (2003), pp. 53–56.

[401] Sara Ventura et al. “Virtual reality as a medium to elicit empathy: A meta-analysis”. In: *Cyberpsychology, Behavior, and Social Networking* 23.10 (2020), pp. 667–676.

[402] Irene Viola and Maria Torres Vega. “On the Impact of Interactive eXtended Reality: Challenges and Opportunities for Multimedia Research”. In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 9707–9708.

[403] Vittorio Portinari. *Elementi di Legislazione Vitivinicola: le norme per l’etichettatura e la tracciabilità dei vini*. t.ly/MrIly. 2016.

[404] Vivino. *Vivino*. https://www.vivino.com/. 2021.

[405] *Voice Assistant Use Reaches Critical Mass*. https://www.emarketer.com/content/voice-assistant-use-reaches-critical-mass. 2021.

[406] Athanasios Voulodimos et al. "Deep learning for computer vision: A brief review". In: *Computational intelligence and neuroscience* 2018 (2018).

[407] Michalis Vrigkas et al. "Augmented reality for wine industry: Past, Present, and Future". In: *SHS Web of Conferences*. Vol. 102. EDP Sciences. 2021, p. 04006.

[408] Vuforia. *Vuforia SDK*. https://developer.vuforia.com/downloads/SDK. 2022.

[409] Michaela M Wagner-Menghin. "Binomial test". In: *Encyclopedia of statistics in behavioral science* (2005).

[410] *Waking up to a new reality*. https://www.accenture.com/gb-en/insights/technology/responsible-immersive-technologies?c=acn_glb_g20responsibleftwitter_10950166&n=smc_0519. 2019.

[411] Ziyu Wan et al. "Bringing old photos back to life". In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 2747–2757.

[412] Anqi Wang et al. "Towards Computational Architecture of Liberty: A Comprehensive Survey on Deep Learning for Generating Virtual Architecture in the Metaverse". In: *arXiv preprint arXiv:2305.00510* (2023).

[413] Peng Wang et al. "AR/MR remote collaboration on physical tasks: a review". In: *Robotics and Computer-Integrated Manufacturing* 72 (2021), p. 102071.

[414] Ranran Wang et al. "Research on Voice Interaction for Augmented Reality assisted Maintenance". In: *in Industrial Maintenance and Reliability Manchester, UK 12-15 June, 2018* (2018), p. 160.

[415] X Wang et al. *ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks*. 2018. arXiv: 1809.00219 [cs.CV].

[416] Yannick Weiß et al. "What User Interface to Use for Virtual Reality? 2D, 3D or Speech–A User Study". In: *2018 International Conference on Cyberworlds (CW)*. IEEE. 2018, pp. 50–57.

[417] Michael Wessel et al. "Generative AI and its transformative value for digital platforms". In: *Journal of Management Information Systems* (2023).

[418] M Wevers and T Smits. "The visual digital turn: Using neural networks to study historical images". In: *Digital Scholarship in the Humanities* 35.1 (2020), pp. 194–207.

[419] Christoph Wick, Christian Reul, and Frank Puppe. "Calamari-a high-performance tensorflow-based deep learning package for optical character recognition". In: *arXiv preprint arXiv:1807.02004* (2018).

[420] Carolin Wienrich and Marc Erich Latoschik. "extended artificial intelligence: New prospects of human-ai interaction research". In: *Frontiers in Virtual Reality* 2 (2021), p. 686783.

[421] H James Wilson and Paul R Daugherty. "Collaborative intelligence: humans and AI are joining forces". In: *Harvard Business Review* 96.4 (2018), pp. 114–123.

[422] Maja Wrzesien et al. "Towards a Virtual Reality-and Augmented Reality-Mediated Therapeutic Process model: a theoretical revision of clinical issues and HCI issues". In: *Theoretical Issues in Ergonomics Science* 16.2 (2015), pp. 124–153.

[423] Hong Wu and Wenxiang Zhang. "Digital identity, privacy security, and their legal safeguards in the Metaverse". In: *Security and Safety* 2 (2023), p. 2023011.

[424] Mei-Yi Wu, Jia-Hong Lee, and Shu-Wei Kuo. "A hierarchical feature search method for wine label image recognition". In: *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE. 2015, pp. 568–572.

[425] Tong Wu et al. "Density-aware chamfer distance as a comprehensive metric for point cloud completion". In: *arXiv preprint arXiv:2111.12702* (2021).

[426] Xingjiao Wu et al. "A survey of human-in-the-loop for machine learning". In: *Future Generation Computer Systems* 135 (2022), pp. 364–381.

[427] Nannan Xi and Juho Hamari. "Shopping in virtual reality: A literature review and future agenda". In: *Journal of Business Research* 134 (2021), pp. 37–58.

[428] *XR Technology Survey: Key Stakeholders Optimistic About Mass Adoption*. https://arpost.co/2019/03/26/xr-technology-survey-stakeholders-optimistic-mass-adoption/. 2019.

[429] Chao Xu et al. "High-fidelity Generalized Emotional Talking Face Generation with Multi-modal Emotion Space Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 6609–6619.

[430] Haibo Yang et al. "3dstyle-diffusion: Pursuing fine-grained text-driven 3d stylization with 2d diffusion models". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 6860–6868.

[431] Riyan Yang et al. "The Human-Centric Metaverse: A Survey". In: *Companion Proceedings of the ACM Web Conference 2023*. 2023, pp. 1296–1306.

[432] Shiyu Yang et al. "The science of YouTube: What factors influence user engagement with online science videos?" In: *Plos one* 17.5 (2022), e0267697.

[433] Guy Yariv et al. *Diverse and Aligned Audio-to-Video Generation via Text-to-Video Model Adaptation*. 2023. arXiv: 2309.16429 [cs.LG].

[434] Mark Yi-Cheon Yim and Sun-Young Park. ""I am not satisfied with my body, so I like augmented reality (AR)": Consumer responses to AR-based product presentations". In: *Journal of Business Research* 100 (2019), pp. 581–589.

[435] W Yin et al. "Socialized mobile photography: Learning to photograph with social context via mobile devices". In: *IEEE Transactions on Multimedia* 16.1 (2013), pp. 184–200.

[436] Jason Yosinski et al. "How transferable are features in deep neural networks?" In: *Advances in neural information processing systems* 27 (2014).

[437] Yuka. *Yuka*. https://yuka.io/it/. 2021.

[438] Dwi Yuniarto, Esa Firmansyah, and Muhammad Helmiawan. "Technology Acceptance in Augmented Reality". In: vol. 3. June 2018. DOI: 10.15575/join.v3i1.158.

[439] I de Zarzà et al. "Emergent Cooperation and Strategy Adaptation in Multi-Agent Systems: An Extended Coevolutionary Theory with LLMs". In: *Electronics* 12.12 (2023), p. 2722.

[440] Fangneng Zhan and Shijian Lu. "Esir: End-to-end scene text recognition via iterative image rectification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2059–2068.

[441] Haijun Zhang et al. "CascadeGAN: A category-supervised cascading generative adversarial network for clothes translation from the human body to tiled images". In: *Neurocomputing* 382 (2020), pp. 148–161.

[442] K Zhang. *Image Restoration Toolbox*. https://github.com/cszn/KAIR. 2019.

[443] K Zhang, W Zuo, and L Zhang. *FFDNet: Toward a fast and flexible solution for CNN-based image denoising*. IEEE Transactions on Image Processing. 2018.

[444] Qing Zhang, David Elsweiler, and Christoph Trattner. "Understanding and predicting cross-cultural food preferences with online recipe images". In: *Information Processing & Management* 60.5 (2023), p. 103443.

[445] Richard Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[446] W Zhang et al. "Deep Learning–Based Multimedia Analytics: A Review". In: *ACM Trans. Multimedia Comput. Commun. Appl.* 15.1s (Jan. 2019). ISSN: 1551-6857. DOI: 10.1145/3279952. URL: https://doi.org/10.1145/3279952.

[447] Xujie Zhang et al. "Armani: Part-level garment-text alignment for unified cross-modal fashion design". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 4525–4535.

[448] Zhenliang Zhang. "Symmetrical Cognition Between Physical Humans and Virtual Agents". In: *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE. 2021, pp. 587–588.

[449] Pengyuan Zhou. "Unleasing chatgpt on the metaverse: Savior or destroyer?" In: *arXiv preprint arXiv:2303.13856* (2023).

[450] Lili Zhu et al. "Deep learning and machine vision for food processing: A survey". In: *Current Research in Food Science* 4 (2021), pp. 233–249.

[451] Zexuan Zhu, Chao Liu, and Xun Xu. "Visualisation of the digital twin data in manufacturing by using augmented reality". In: *Procedia Cirp* 81 (2019), pp. 898–903.

[452] Stefanie Zollmann, Christian Poglitsch, and Jonathan Ventura. "VISGIS: Dynamic situated visualization for geographic information systems". In: *2016 international conference on image and vision computing New Zealand (IVCNZ)*. IEEE. 2016, pp. 1–6.

[453] Zhengxia Zou et al. "Object detection in 20 years: A survey". In: *Proceedings of the IEEE* (2023).

[454] Mark Zuckerberg et al. "Meta Platforms". In: (2022).