European Commission
Horizon 2020
European Union funding
for Research & Innovation

H2020-ICT-46-2020
Robotics in Application Areas and Coordination & Support

Flexible Assembly Manufacturing with Human-Robot Collaboration and Digital Twin Models

FELICE

D3.1: State of the art report[†]

Abstract: This report reviews the state of the art in technologies and tools pertaining to topics investigated in the *FELICE* project. Its aim is to assess the wide range of related technologies and tools available today, in order to set the baseline for the further development of the *FELICE* eco-system. Different aspects related to subjects such as perception, cognition, human-robot collaboration, cognitive ergonomics, safety, orchestration and analytics are considered, the challenges of each are analyzed, whereas state of the art approaches are critically reviewed with respect to their applicability in the context of *FELICE*. More specifically, the following topics are discussed in dedicated sections: Scene and object perception; human behavior monitoring in assembly task execution; robotic hardware; adaptive workstations; human robot communication; cognitive ergonomics for human-robot dyads; safe robot operation; robot programming; synchronization of the human-robot dyad in taskable pipelines; prescriptive analytics in production system diagnosis, monitoring, and control; AI-driven digital twins and digital operators; orchestration of adaptive assembly lines; computing infrastructures; data privacy, vulnerability management, and security assurance; modular technologies and toolkits for agile production. Collectively, these topics lay the foundation for the innovations to be developed in the project.

[†]The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017151.

Contractual Date of Delivery:	30/06/2021
Actual Date of Delivery:	31/07/2021
Security Class:	Public (PU)
Editor:	UNISA
Contributors:	UNISA ICCS FORTH ACC PRO FHOOE TUD CRF IFADO FRAUNHOFER AEGIS CALTEK
Quality Assurance:	Manolis Lourakis (FORTH) Julian Eber (FRAUNHOFER)
Deliverable Status:	Final

The FELICE Consortium

Participant No.	Participant Organisation Name	Part. Short Name	Country
1 (Coordinator)	Institute of Communication & Computer Systems	*ICCS*	Greece
2	Centro Ricerche FIAT S.C.p.a.	*CRF*	Italy
3	FH OÖ Forschungs & Entwicklungs GmbH	*FHOOE*	Austria
4	AEGIS IT RESEARCH GmbH	*AEGIS*	Germany
5	Leibnitz Research Centre for Working Environment and Human Factors	*IFADO*	Germany
6	Foundation for Research and Technology – Hellas	*FORTH*	Greece
7	CAL-TEK S.r.l.	*CALTEK*	Italy
8	Technical University Darmstadt	*TUD*	Germany
9	University of Salerno	*UNISA*	Italy
10	Fraunhofer-Gesellschaft zur Foerderung der Angewandten Forschung e.v.	*FRAUNHOFER*	Germany
11	ACCREA Engineering	*ACC*	Poland
12	PROFACTOR GmbH	*PRO*	Austria
13	Eunomia Ltd	*EUN*	Ireland

Document Revisions

Ver.	Date	Editor	Overview
1.0	30/07/2021	UNISA, FORTH, FRAUNHOFER, ICCS	Final version approved by the reviewers
0.6	28/07/2021	FORTH, FRAUNHOFER	Second round of reviews
0.5	26/07/2021	UNISA, TUD, ACC, PRO, AEGIS, FRAUNHOFER	Second draft of the document addressing the comments of the reviewers
0.4	07/07/2021	FORTH, FRAUNHOFER	First round of reviews
0.3	28/06/2021	ICCS, CRF, FHOOE, FORTH, AEGIS, IFADO, CALTEK, TUD, FRAUNHOFER, ACC, PRO	Second draft partner contributions
0.2	26/05/2021	ICCS, CRF, FHOOE, FORTH, AEGIS, IFADO, CALTEK, TUD, FRAUNHOFER, ACC, PRO, UNISA	First draft partner contributions
0.1	01/02/2021	UNISA	First draft of the document

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADAPT Asset-Decision-Action-Property-RelaTionship

AED Attention-based Encoder-Decoder

AGV Automated Guided Vehicle

AI Artificial Intelligence

ALBP Assembly Line Balancing Problem

API Application Programming Interface

ASR Automatic Speech Recognition

AWS Adaptive Workstation

BA Bundle Adjustment

CAD Computer-Aided Design

CNN Convolutional Neural Network

COBOT Collaborative Robot

CPS Cyber-Physical System

CTA Cognitive Task Analysis

DIH Digital Innovation Hub

DL Deep Learning

DNN Deep Neural Network

DP Daisy Planner

DT Digital Twin

E2E End-to-End

EID Ecological Interface Design

GDPR General Data Protection Regulation

GNN Graph Neural Network

GRU Gated Recurrent Unit

HOI Human-Object Interaction

HRC Human-Robot Collaboration

HRI Human-Robot Interaction

HTA Hierarchical Task Analysis

IEC International Electrotechnical Commission

IIoT Industrial Internet of Things

IMU Inertial Measurement Unit

IoT Internet of Things

ISO International Organization for Standardization

KF Kalman Filter

LSTM Long Short-Term Memory

MES Manufacturing Execution System

ML Machine Learning

MLP Multi-Layer Perceptron

NDDL New Domain Definition Language

NLP Natural Language Processing

NLU Natural Language Understanding

OPC UA Open Platform Communications United Architecture

PA Prescriptive Analytics

PDDL Planning Domain Definition Language

REBA Rapid Entire Body Assessment

RFID Radio-Frequency IDentification

RL Reinforcement Learning

RNN Recurrent Neural Network

ROI Region Of Interest

ROS Robot Operating System

RULA Rapid Upper Limb Assessment

SDK Software Development Kit

SLAM Simultaneous Localization and Mapping

SMEs Small and Medium-sized Enterprises

SSL Sound and Source Localization

TFF Task Frame Formalism

ToF Time of Flight

TRL Technology Readiness Level

UCD User Centered Design

UWB Ultra-Wide Band

WMSD Work-related MusculoSkeletal Disorder

XML Extensible Markup Language

Executive Summary

FELICE aspires to develop a collaborative human-robot assembly line which will capitalise on the cognitive autonomy, endurance, repeatability and accuracy of robots in order to maximise the flexibility and productivity of assembly processes. By combining the skills of humans and robots, improved manufacturing performance and work ergonomics will be achieved, while a safe, mentally and socially satisfying environment for human-robot collaboration will be established. In this document, we report the state of the art tools and technologies and point out candidates that will be adopted in the different domains addressed by the *FELICE* project in order to design and realize the above mentioned system.

The document is organized in 18 different sections. Following an introduction in the first section, in the second one we summarize the architecture of *FELICE* and also remind the reader of the five pillars we have identified for the project. Recollecting the pillars and their main topics is important since this has been the driver for the definition of the sections we have included in this document. Indeed, each of sections 3 to 17 is devoted to discussing the state of the art tools and technologies for a specific identified topic. Finally, some conclusions are drawn in the last section. The rationale behind the chosen structure of the document is two-fold: from one side to enhance the readability so that the entire consortium can use it as a common reference; from the other side, to assess a wide range of technologies and tools in order to baseline the further development of the *FELICE* eco-system.

1 Introduction

1.1 Purpose of the document

The document describes the state of the art in technologies and tools regarding the domains investigated in the *FELICE* project. The activities conducted by the Consortium partners for compiling this deliverable can be thus considered as a preliminary yet sizeable activity for any R&D workpackage, since state of the art analysis and related discussion, both from a scientific and technological point of view, represent the basis and the driver for any subsequent R&D work.

More specifically, the following topics will be discussed in separate sections: Scene and object perception (Section 3); Human behavior monitoring in assembly task execution (Section 4); Robotic hardware (Section 5); Adaptive workstation (Section 6); Human robot communication (Section 7); Cognitive ergonomics for enhanced human-robot dyads (Section 8); Safe robot operation (Section 9); Robot programming (Section 10); Synchronization of the human-robot dyad in taskable pipelines (Section 11); Prescriptive analytics in production system diagnosis, monitoring, and control (Section 12); AI-driven digital twins and digital operators (Section 13); Orchestration of adaptive assembly line (Section 14); Computing infrastructure (Section 15); Data privacy, vulnerability management, and security assurance (Section 16); Modular technologies and toolkits for agile production (Section 17).

Despite that the topics covered in this document span a wide spectrum, there are some common questions that will be answered for each of the topics addressed: What are the challenges of a specific topic? Is there already a solution available in the literature (or in the market) for solving that specific problem? Or, at least a starting point to be considered as a baseline solution to the problem at hand?

1.2 Intended readership

Deliverable D3.1 is a public document (PU) and therefore is intended for online dissemination.

1.3 Relation to other FELICE deliverables

The analysis reported in this deliverable is related to all the R&D workpackages of *FELICE*, since it will assist in the identification of the tools, technologies, concepts that the project will build upon. Therefore, in terms of project development, the activities culminating to this deliverable took place in months 1-6 of the project's lifetime.

2 Overview

2.1 General concept and document rationale

The aim of the *FELICE* project is to develop a collaborative human-robot assembly line which will capitalise on the cognitive autonomy, endurance, repeatability and accuracy of robots in order to maximise the flexibility and productivity of assembly processes. By combining the skills of humans and robots, improved manufacturing performance and work ergonomics will be achieved, while a safe, mentally and socially satisfying environment for human-robot collaboration can be established. FELICE strives to go beyond traditional industrial automation systems in which robots are pre-programmed and re-programmed to carry out specific repetitive tasks with little variation but with high degrees of accuracy and precision.

In order to realize a collaborative assembly system in this project, an overview of its architecture is presented in Figure 1. It is important to point out that the definition and the description of the system's architecture is beyond the scope of this document. Nevertheless, it is important to report this information here for the sake of completion and clarity, since the different modules required in the system are discussed in terms of state of the art, tools and technologies in the remainder of this document.

As illustrated in the figure, the *FELICE* system is conceived as a two-layered architecture, where the local layer includes system components for perceiving the environment and facilitating human robot collaboration. The global layer comprises components for digital twin modelling, assembly orchestration, optimization, and prescriptive analytics. AI and machine learning algorithms are omnipresent throughout the system. Details about the architecture of the system can be found in deliverable *D2.1 - Robot architecture and system specifications*.

FELICE's architecture is supported by the following pillars:

- Pillar I: Smart process monitoring via integration of heterogeneous sensors and devices in industrial environments
- Pillar II: Collaborative robots with advanced cognitive capabilities, mobility and adaptability for joint task execution, addressing safety and fluency
- Pillar III: AI system for real-time orchestration and control of adaptive assembly lines
- Pillar IV: Distributed architecture computing paradigm and re-usable toolkits

The aforementioned pillars are detailed in the next section (i.e., 2.2), along with the specific topics related to each. The pillars have also driven the rationale behind the structure of this deliverable. An overview of the association among pillars and investigated topics, together with pointers to the Section where each topic is discussed, is provided in Figure 2.

Each section is organized so as to provide a brief but comprehensive analysis of the state of the art, both from a scientific and a technological point of view. The length of each section is a few pages, except cases where more aspects need to be discussed under the same topic (i.e., categories of algorithms, tools, frameworks, etc., like in Sections 3,

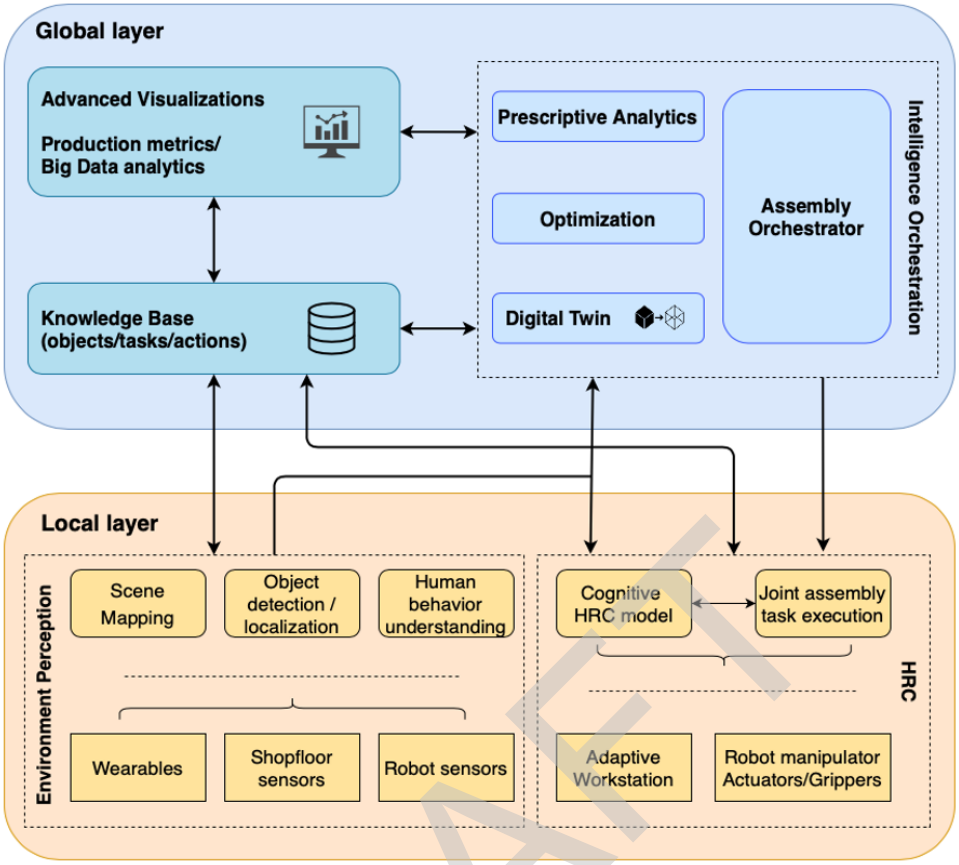


Figure 1: System architecture for *FELICE* project.

7), where the corresponding text is longer. In more detail, for every topic analysed, the following subsections are included:

- Overview: brief introduction of the topic discussed in the section;
- Relationship with *FELICE* project: Role of the tools/frameworks/algorithms described in the section, with reference to the whole *FELICE* project. In this manner, the role of the corresponding topic with reference to other topics is also clarified.
- State of the art: short description of the state of the art algorithms, tools and frameworks available for dealing with the specific problem at hand; the title of this section will be the specific name of the addressed topic; for the sake of clarity, it is also important to mention that this part is not intended to serve as an exhaustive review of the entire literature on the specific topic, but rather as a presentation of the most important and relevant trends/aspects. The state of the art sections are organised in two further subsections:
  - Baseline technologies and tools: since within the project it would be possible to exploit (whenever possible) existing algorithms from the literature, this subsection reports tools, libraries, frameworks already available for tackling the specific task.
  - Discussion: this subsection contains a critical analysis of the state of the art presented. The question that will be addressed in this section will be: *What*

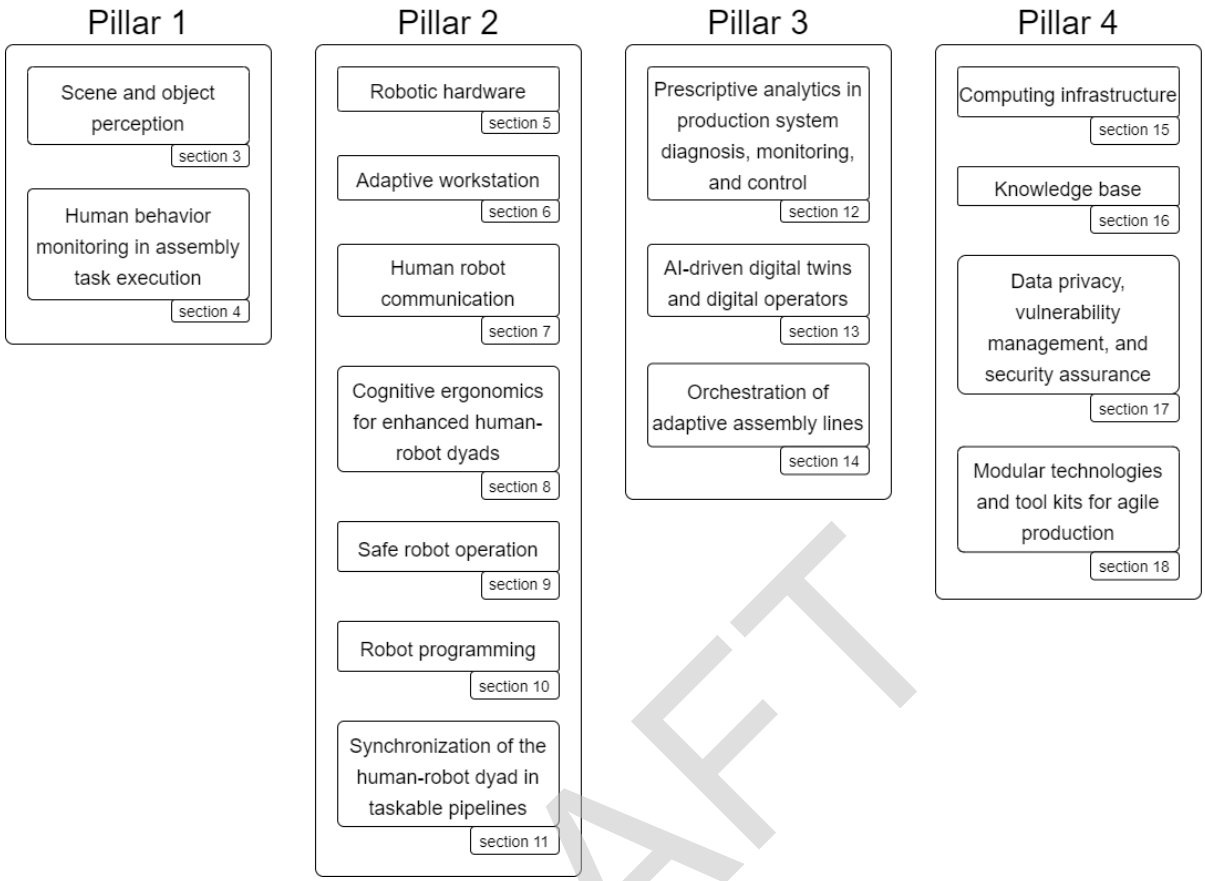


Figure 2: Pillars, topics and their corresponding sections.

is the best-suited algorithm(s) / library(ies) / tool(s) / framework(s) that can be used for solving that specific problem?

2.2 Pillars overview

Pillar I pertains to the incorporation of heterogeneous data acquired from different types of sensors, including visual, location, wearables, and actuators for acting upon the environment. Robot sensors/actuators are also considered part of a distributed IoT infrastructure that drives the extraction of low-level knowledge pertaining to the dynamic environment (see Section 3) and the human workers from different sensor streams (refer to Section 4). Information is shared with both the global and local layer whereas sensors can be configured, as instructed by the global layer.

Pillar II focuses on the fact that *FELICE* robots feature cognitive capabilities in order to deal with the uncertainty inherent in the human-occupied industrial environment by perceiving nearby events, planning and anticipating the outcome of their actions and the actions of other workers, and learning from their interaction with workers. Robotic hardware is the focus of Section 5, while the adaptive workstation is discussed in Section 6. Human robot communication by voice and gesture analysis is then reviewed in Section 7. Also, models from industrial and cognitive psychology will be exploited to assess the cognitive ergonomics of human-robot collaboration (detailed in Section

8). Mobility will extend the robots range of operation, endowing it with the ability to navigate to different workstations to support workers in a time-shared manner (refer to Section 11). Adaptability will be supported by task-level programming (see Section 10), achieving the operation of the robot amid changes in the environment (i.e., moving parts, uncertain locations, partial occlusions, failures), its ability to handle a wide variety of similar tasks and the rapid redeployment to new tasks. These abilities will enable robots and human workers to exist side-by-side, with robots adapting to the variability of tasks at each particular workstation, considering synchronization and safety in collaborative task accomplishment (more details in Section 9). Safety is one of the most important concerns when considering human-robot collaboration in industrial applications. Standardization activities, including risk assessment and control system validation will identify conditions and criteria for adherence to safety related standards (ISO 10218-1 & 2, ISO/TS 15066). This last aspect will be discussed in Section 8.

Pillar III refers to the capability of harnessing the combination of digital twins and AI to adaptively optimize on-line the entire production workflow and manufacturing execution system, improve resilience in the assembly line and mitigate the effects of error and faults. Digital twins (refer to Section 13) will support the modelling of the assembly line tasks and workflows at a fine-grained level considering the production resources, including the human workers, the robot and the tools based on "digital replicas" of machines/equipment and operators. AI methods, predictive and prescriptive analytics will orchestrate the collaboration between humans and robots, allocate tasks in the human-robot dyad towards enhancing performance through increasing the systems adaptive capacity and at the same time reducing physical and mental stress of the human operator. Prescriptive analytics in production system diagnosis, monitoring, and control will be the focus of Section 12, while state of the art, tools and technologies of the orchestration of the adaptive assembly lines will be discussed in Section 14.

Finally, as for Pillar IV, we have to take into account that *FELICE* project adopts a distributed architecture paradigm exploiting edge resources at the local workstation layer and the robotic platform, for sensor data collection and extraction of low-level cues and knowledge based on human behavioral, physiological and context related parameters as well as cloud resources. The computing infrastructure will be discussed in Section 15. Mechanisms for security and privacy-preservation across all data streams and modules and throughout the architecture will be adopted, supporting GDPR compliance (more details in Section 16). Furthermore, special emphasis is put on automatic vulnerability management, measurement and policy compliance evaluation of robotic systems.

FELICE will also actively contribute to European ecosystems and digital innovation hubs supporting the penetration of advanced digital solutions and robotic technologies in industrial production processes. The consortium will consider making the tools developed in the project open and freely available and will encourage end-users and robotics solution developers to use them in multiple application domains beyond the proposed work (refer to Section 17).

For the sake of completeness, it is also important to mention one last pillar of the *FELICE* project, namely *Pillar V: Technology validation in real industrial environments*. Indeed, during the project the consortium will systematically evaluate the approaches

proposed in Pillars I-IV, through validation and experimentation in different assembly line environments, namely a) a small-scale prototype environment (FHOOE) to evaluate the manufacturing execution system (MES) early in the project lifetime and elaborate the optimization algorithms incrementally; b) a large-scale demonstration and evaluation environment, set at the premises of the one of the largest car manufacturers in Europe (CRF). However, this aspect will not be discussed in this deliverable.

3 Scene and object perception

3.1 Overview

This section is concerned with technologies to be employed by the *FELICE* mobile robot in order for it to perceive its environment and the objects of interest that are present within it. Towards this end, two different basic competencies are discussed, namely localization and mapping as well as object detection and pose estimation.

3.2 Relationship with FELICE project

Scene perception will support the acquisition and fusion of information from various sensors distributed in the physical space of the shop floor in order to enable multi-scale perception for mapping environment changes and monitoring resources (human, robot, workpieces) in real time. At a macro level, the monitoring of the context parameters will enable the derivation of updated representations of the surroundings, including scene geometry, moving/static objects in the scene and occluded areas. Information will be exploited by the assembly orchestrator as well as by the robot providing a third-person perspective in path planning and navigation, resolving potential ambiguities and uncertainties. At a micro level, compliance to local constraints required by specific tasks will be enabled focusing on the detection and localization of textureless, reflective objects and object parts. These will mainly serve collision avoidance, grasping and quality check tasks during task execution.

3.3 Simultaneous localization and mapping

In order for the robot to navigate the dynamic and partially known environment of an assembly shop floor, it has to be aware of its position with respect to its surroundings. In more technical terms, the robot must use its on-board exteroceptive sensors to construct and constantly update a representation of its environment (i.e., a map), while simultaneously keeping track of its position and orientation within that environment (i.e., the robot localization state). This is the computational problem of *simultaneous localization and mapping*, abbreviated as SLAM. SLAM is a challenging problem whose difficulty stems from the fact that it represents a cause and effect dilemma: In order for the robot to localize itself a map is needed and for building a map, a pose estimate, that is localization, is necessary.

The vast amount of relevant research during the last three decades has led to significant progress and SLAM systems that perform well in certain domains. Nowadays, SLAM is a fundamental building block for a wide range of technologies that are in need of mapping unknown environments and range from wearable computing and virtual or augmented reality to self-driving vehicles. In particular, SLAM is one of the key competences towards the realization of truly autonomous mobile robots since it provides a means of tracking their location and identifying key landmark locations. Availability of a map is essential to perform path planning, support visualization and limit the localisation drift. The latter is achieved by a process known as *loop closing*, which refers to the assertion that a robot has returned to a previously visited location, hence revisits

previously seen parts of its environment. SLAM without loop closing is often referred to as *odometry*. Another variation is *visual-inertial SLAM* that fuses images and inertial measurement unit (IMU) data to provide high-precision odometry for SLAM.

SLAM is a broad topic and relevant solutions come in different variants, depending on the algorithmic approach, the type of sensors (e.g., sonar, laser, ToF, monocular or stereo cameras and RGBD) and the data representation used, as well as the particularities of an application. Therefore, a representative overview of the topic is well beyond the scope of this document and the reader is referred to more extended reviews available in the surveys found in [83, 203, 80, 407, 53]. In the following, we provide a brief general overview with slightly more emphasis on *visual SLAM* (vSLAM) [407]. The latter has attracted strong interest in recent years owing to the advantages of camera sensors (i.e., low cost, power consumption, mass, form factor, etc).

SLAM was introduced in the mid '80s [420] and in earliest approaches, was formalized in a probabilistic fashion that aims to estimate the model parameters (map and robot state) that maximize the probability of obtaining the actual measurements. A SLAM system involves two main components, namely the front and back ends. The front-end transforms sensor data so that they can be more suited to estimation, while the back-end performs inference on the data provided by the front-end [83]. In the case of visual SLAM, the front-end extracts the pixel locations of discriminant points in the environment and associates them with specific landmarks, e.g. 3D points; this task is known as data association. Estimating the pose of a moving camera and the 3D geometry of a scene has been studied by the computer vision community under the name of *structure from motion* (SfM). Despite their similarities, SfM and vSLAM approached the problem from different perspectives: SfM emphasized accuracy of reconstruction at the cost of batch operation whereas vSLAM sought incremental camera motion estimation with online performance. Starting in the early 2000s, the research on SfM and vSLAM has converged and the generic character of the localisation and reconstruction problems has been understood better [123].

A possible formulation of SLAM is the full one, where the entire robot trajectory and the map are estimated given all the control inputs and all sensor measurements. Although this approach can lead to highly accurate results, it suffers from the drawback that the problem grows unboundedly with the number of the considered variables. Thus, it cannot be solved in real time, which in turn limits its practical applicability. Another approach is that of online SLAM, where the problem is solved incrementally by updating the estimated model using recently acquired sensor information only. Estimation techniques can be classified as either filter or optimization based.

Filter-based methods originate from Bayesian filtering and sequentially fuse image measurements by updating probability distributions over the map and pose parameters. They operate in two steps: First, a prediction of the model is made and then the current sensor measurements are used to correct the previously predicted state. A Kalman Filter (KF) relies on the premise that the pose model is a linear one. The KF couples a prediction stage that uses the previous values to predict the current state, followed by an update phase that combines the predicted state with current sensor data. As the linearity assumption does not hold in practice, derivatives of the KF are used to handle nonlinear systems. The extended KF (EKF) performs linearization around the current estimate via first-order Taylor expansion. The EKF update time depends quadratically on the size of the state vector, hence EKF cannot support the growing map size of large-

scale SLAM. To deal with this, new empty sub-maps (linked via a higher-level nap) are introduced when the map size becomes large [322]. Still, the EKF approach has high computational demands. The unscented KF (UKF) performs better than the EKF for highly non-linear systems and does not require the calculation of the Jacobians, however it is also computationally costly. An alternative approach to KF filtering is provided by particle filters (PF), which use particles (i.e., samples) to represent the posterior distribution of a stochastic process given noisy or partial observations. PFs sample the state with a set of particles and perform displacement predictions and updates for each particle. During the update, particles are weighted according to their likelihood and then the most likely ones are retained while the rest are eliminated and replaced by new particles which are generated near the high-weight ones. PFs avoid making any linearity or Gaussianity assumptions and can generate samples from a required distribution without any assumptions about the state-space model or the state distributions. FastSLAM is the best-known algorithm based on particle filtering [316].

Optimization-based methods also comprise two parts. The first finds a correspondence between new observations and the map, thus providing constraints. Then, the model is updated to accommodate the constraints provided by the new observations. Optimization can be carried out with Bundle Adjustment (BA) [277] and its variants or graph SLAM [241, 140]. Bundle adjustment simultaneously refines the sensor pose and scene geometry so as to optimize a criterion involving the reprojection error of all points in all images. BA performs batch optimisation on a large number of variables and yields accurate results at the cost of a high computational cost. To limit this cost, BA is performed on selected images such as those in a sliding window or on spatially distributed keyframes [237]. A comparison of filter-based and optimization techniques for monocular SLAM concluded that a large number of keypoints is more beneficial to the accuracy of SLAM compared to having a large number of frames [431]. Furthermore, since the computational cost of BA grows less with the number of keypoints compared to EKF approaches, the former outperforms the latter. In the author's words, keyframe BA is preferable "since it gives the most accuracy per unit of computing time. Graph SLAM employs a graphical representation in which nodes represent poses and connecting edges correspond to spatial constraints between them. Constraints are linearized to obtain a sparse matrix corresponding to a sparse graph map. A reduction process removes the redundant map variables and optimization strives to find a configuration of nodes that minimize the error induced by the constraints [251].

Another taxonomy divides SLAM algorithms into direct and indirect [141]. Indirect methods preprocess raw sensor measurements and generate an intermediate representation (e.g. feature correspondences [451]) that is next interpreted in a model to estimate geometry and camera motion [123, 237, 323]. Direct methods bypass the preprocessing step and use the sensor values directly in a probabilistic model [142, 141]. Yet another classification discriminates between sparse and dense methods. The former employ and reconstruct a selected set of independent image points (e.g. corner keypoints) [237, 323], whereas dense methods attempt to use and reconstruct all pixels in a 2D image [142]. Intermediate, semi-dense approaches refrain from reconstructing the complete surface, but still aim at using and reconstructing a largely connected subset.

3.3.1 Baseline technologies and tools

A plethora of SLAM algorithms have been proposed in recent years and many of them have publicly available implementations. Extensive lists can be found online, e.g. [487]. Other online resources provide links to public datasets annotated with ground truth [106]. In the interest of saving space, we just mention LSD-SLAM [142], ORB-SLAM [323, 86], SVO [150], and Kimera [386] as the most prominent and up-to-date open source pipelines for visual SLAM that implement combined functionality for tracking, localization, mapping and loop closure.

ORB-SLAM is a feature-based approach which employs the ORB feature detector [387] and produces maps in the form of sparse point clouds. LSD-SLAM is a direct approach that uses the image intensities to optimize geometry, hence generates a semi-dense map of the environment. SVO uses a hybrid approach to estimate camera motion and uses both pixel intensities and features. Most of the aforementioned pipelines support different types of cameras in addition to ordinary pinhole ones. In certain cases, additional abilities provided such as IMU support (SVO, ORB-SLAM) and semantic mesh classification of the reconstruction [486, 163] bear significance for the goals of *FELICE*. In other cases, the underlying approach to tracking such as the so-called direct SLAM (LSD-SLAM) can potentially prove to be highly resilient in indoor environments wherein surfaces may be smooth and generally feature-less. It is therefore prudent to assess the suitability of each of the aforementioned tools individually and/or in groups with regards to the challenges discussed below.

3.3.2 Discussion

This section discusses some challenges affecting the performance of SLAM systems and discusses them in relation with the particularities of the *FELICE* deployment environments. Additionally, it identifies certain design choices and their applicability/impact on *FELICE* environments.

Robust data association: A critical factor for the performance of any SLAM system relates to the robustness of data association (i.e. matching). This refers to the the extend to which the system correctly associates image landmarks across images in the sequence [153]. Apart from observations originating from different parts of the environment but being similar in their sensed appearance (a phenomenon known as *perceptual aliasing*), matching can also fail due to sensor noise or environmental changes such as illumination.

Typically, the employed landmarks are either image points with distinctive local RGB patterns (feature-based SLAM), or entire image regions along edges (direct SLAM). It has been observed that representatives of the latter methodology (e.g. LSD-SLAM) are usually more effective in man-made indoor environments wherein the scene may contain texture-less surfaces, such as tiles, desktops, walls, etc. On the other hand, feature-based methods (e.g., ORB-SLAM) can be more advantageous when rich-textured objects are present, or when prior knowledge of distinctive landmarks in the environment is employed to enhance global localization. It may therefore desirable, in the context of the current project, to pursue a hybrid approach (e.g., SVO) which will allow both agile tracking and mapping as well as landmark identification for global localization in the work-space.

Moving objects: SLAM algorithms typically assume that the environment is mostly static. As a result, different motion patterns in dynamic environments often pose challenges for visual SLAM algorithms [398]. Typical remedies aim at filtering out the motion of dynamic objects as statistically inconsistent with the dominant motion of the scene (e.g., ORB-SLAM, LSD-SLAM). However, such solutions suffer from bias associated with the spatial extent of true camera motion evidence between images, which is particularly affected by occlusions and partial overlap between multiple moving objects.

In the specific large-scale use case pursued by *FELICE* at CRF, certain objects and parts of the environment move independently. Human workers performing their tasks are an obvious such example. Additionally, independent motion is exhibited by an automated guided vehicle (AGV) transporting dollies supporting assembled components, a conveyor belt on the shop floor and dollies carried by it. Furthermore, in the context of *FELICE*, holistic scene understanding could be beneficial and therefore the motion of dynamic objects should be further analyzed as opposed to simply being labeled as outlying. Towards this end, Kimera [386] presents additional potential for more comprehensive analysis of dynamic objects.

Drift over long distances: Another common issue with SLAM algorithms is the accumulation of small errors in the pose estimate over time, which adversely affects localization. Typically, this accumulated error is compensated for by using foreknowledge of the locations of landmarks in the work-space, or by means of associating features detected in the latest image with features already stored in tha spatial memory of the current SLAM session, a process also known as *loop closure*. Detecting when known areas are revisited is also known as place recognition [280]. Feature-based loop closing is implemented in ORB-SLAM [160], LSD-SLAM [170] and Kimera [292] and the performance of each variant should be assessed in the context of *FELICE*'s work-space environment and in view of the necessary enhancements for accommodating identification of previously known landmarks in different representations (e.g. as features, shapes, geometric spatial arrangements of markers, etc.). A complementary solution for countering drift is to incorporate non-visual indoor localization sensors, for example based on Ultra Wideband (UWB) radio technology [499].

We next move on to present certain preliminary design choices for the adaptation/development of *FELICE*'s SLAM pipeline:

- **Prior map constructed offline**: Considering that the collaborative robot will repeatedly navigate the same environment, robustness and performance improvements could be gained by relying on a prior map constructed offline. In this respect, it is relevant to investigate potential representations for work-space structure in a way that will accommodate fast look-up and robust matching with the spatial memory acquired during a running SLAM session. Such representations may comprise CAD models, feature-based dictionaries, geometric arrangements of features, or combinations thereof.

- **Passive vs. depth camera**: Although there is a general consensus that depth maps can largely improve a visual SLAM estimate, the adoption of active depth cameras however raises the issue of potential trade-offs in terms of execution time and association with the stereo disparity estimates from the RGB images. The limitations in the effective range and resolution of depth images suggest that depth can be used as an additional measurement to enhance the disparity map

obtained from plain RGB stereo images. A more detailed discussion regarding the choice of camera sensors can be found in Section 3.5.

- **Sparse vs. dense SLAM**: Leveraging stereo vision and depth sensing can accommodate dense reconstructions, which in turn lead to a finer-grained scene understanding. Clearly, storing dense maps can be advantageous for tasks pertinent to semantic understanding of the environment. However, storage and access time of a dense map may impose significant computational burden to the system. A potential solution is to employ semi-dense representations which focus on edges combined with triangulated visual features-only SLAM. In this way, it is possible to achieve the best of both worlds, i.e., a finer-detail map containing visual features that promote the loop-closing abilities of SLAM. Furthermore, fully dense maps may be fragmented in a way only small portions of them will be loaded into memory, thereby allowing access to high-detail 3D information when necessary.

- **Incorporation of non-visual data**: The possibility of integrating measurements originating from non-visual sensors in order to simplify or speed-up SLAM tasks will be considered. Such data, for instance, might come in the form of IMU or UWB measurements for use in relative pose estimation and loop closure, respectively.

3.4 Object detection and pose estimation

Object detection and pose estimation deal with the problems of identifying and estimating the location of the objects present in a scene and are usually dealt with by computer vision methods. There are different approaches that rely either on known or unknown objects. Object detection usually regards unknown objects and does not concern the estimation of an object location. It usually estimates only a generic region where the object of interest lies, usually a bounding box, and serves better cases of unknown surrounding scenes. Object detection can be either 2D or 3D, providing respectively 2D or 3D bounding boxes in the image. On the other hand, in 6D object pose estimation, also met in the literature as localization with known objects, the 3D target objects are known. This allows for the accurate estimation of their 3D location and 3D orientation in the coordinate system of the camera. Both object detection and 6D pose estimation are exploited in robotic applications. However, 6D pose estimation can serve better the accuracy requirements of industrial applications, where the target objects are known, facilitating the successful, safe, accurate and real time human-robot interaction, e.g. to support robot grasping.

There is a great variety of 6D pose estimation methods, which are extensively reviewed in [110, 134, 231, 185, 394, 377, 484]. These methods are commonly categorised into those using 2D or 3D information derived by different sensors [193, 418, 185], single or multiple views [252], classic or deep learning methods [193, 134, 185] and correspondence, template or voting based approaches [134].

Regarding the information used, 2D color images are widely employed for 6D pose estimation. Moreover, depth information can also be exploited as an additional cue, either in the form of 3D point cloud or polygon mesh of the known objects. This information can be derived by either active or passive sensors, which in the case of close range applications rely on Kinect-like [211, 197, 440, 248, 179], structured light [390],

etc or stereo and monocular camera [328, 63, 495, 179] sensors. However, active depth sensors cannot cope with reflective, shiny, texture-less, very dark (highly absorbing) and transparent objects, which are very common in industrial applications, because they do not reflect properly the active illumination. For that reason, passive sensors are more widely used for this purpose. Nowadays though, the use of depth information as an additional imaging modality is not always necessary. Instead, the combination of single or multiple RGB images (2D case) and Deep Learning methods for predicting 2D-3D correspondences [376, 358, 349, 155, 344, 500, 269] or applying regression [488, 293] or classification [228, 434] on detected bounding boxes, has overcome the restrictions and performance achievements of depth-based methods. In this document, a short overview of the different methods either using or not depth information and also using classic or deep-learning based methods will be presented, according to the categorisation of [134] into correspondence, template and voting-based ones.

Correspondence-based methods define correspondences between the known 3D object model and the captured RGB images or the 3D information from depth images. In the case of RGB images, usually 2D image keypoints are detected, described and matched to the corresponding points on the known 3D geometry of the known object. Then a Perspective-n-Point (PnP) [260, 444] algorithm is used for the final 6D pose estimation. However, such methods use 2D feature-based keypoint detectors and descriptors (e.g., SIFT [279], SURF [59], etc.) that fail to handle objects with texture-less and homogeneous surfaces. Nowadays though, deep learning has overcome this restriction and detects characteristic keypoints on the images by predicting the 2D projection of 3D points, edges, corners of the bounding boxes, etc in the 3D space [442, 376]. Some methods use representative 3D control points of the known object to predict the pose of the different parts of the object [118] or take advantage of other intermediate representations of the object like edges and symmetries [422]. On the other side, other methods make predictions of the 3D positions of the 2D image points of the object [500], applying pixel-wise regression for texture-less [344] or symmetric objects with multiple instances [193]. Some of the methods of this category could be beneficial for the scenarios of the *FELICE* project, since they are able to handle several challenging objects and conditions, while providing real-time performance [344, 193]. One of the most well performing, with respect to accuracy and speed, methods for 6D object pose estimation in this direction is the recently developed CosyPose algorithm [252], which matches the input image to a rendered one [268]. This is a method that is applicable to both RGB and RGBD images with comparable results. It offers the possibility of estimating the object pose on a single image and also to further optimize the estimate on a global level using multiple images (object-level correspondence), especially in the case of multiple symmetric overlapping objects co-existing in the scene. Regarding the correspondence-based methods for the 3D case [503, 179], the correspondences are defined between the known and the created from the input depth images point clouds. Such methods though, fail to handle texture-less and reflective objects that are very common in the industrial environments, as in the case of *FELICE* applications.

Template-based methods compute the 6D pose of objects by applying template matching of an input images against a library of templates. Those templates are known images, classified according to the known 6D pose of the depicted object, but require a lot of training data which are not always easily available. However, template-based methods that do not require depth information can successfully cope with texture-less

objects [197, 191]. However, to overcome the problem of obtaining the training data, it is very common to use rendered images that are created using a 3D model of the known object. For the methods that require depth information (3D case), the whole 3D object is considered as a template and the problem yields in the calculation of the transformation between the created and the known 3D information of the object, often achieved with ICP [66]. Template-based methods are highly efficient in the presence of noise, great variety in object poses and texture-less objects. However, they cannot handle occluded objects since they rely on global features and they fail to meet the real-time performance requirements of the industrial assembly lines, especially in cases of large object databases. As it can be concluded, it seems that template-based methods will not be an appropriate solution for *FELICE*, where timely performance is very crucial for the preservation of the natural speed of the assembly line.

Finally, in voting-based methods, every pixel or 3D point votes for a correspondence-candidate keypoint (2D [358] or 3D [184]) or directly for a 6D object pose [440, 133, 195, 463, 240]. However, the most well performing voting-based methods are inappropriate for reflective objects because they use RGBD images and require depth information, making them probably less appropriate for *FELICE* applications.

3.4.1 Baseline technologies and tools

There exists a wide variety of technologies and tools for 6D object pose estimation, that even provide publicly available source code implementations [455] and annotated datasets [196]. However, this section will analyse shortly only some of those methods, that are the most recent and well performing regarding run time and accuracy. These methods are CosyPose [252] and EPOS [193], which calculate simultaneously the pose of all the present objects using a single shared DNN model. CosyPose [252] presents a single-RGB 6D pose estimation method, that handles symmetric and occluded objects, using a render-and-compare DNN method inspired by DeepIM [268] (using EfficientNet-B3 [437] instead of FlowNet [130]). Initially, it detects all known objects in the image and then for each object it assumes a canonical pose for which it creates a rendered image. Comparing this rendered image to the input, it estimates a coarse pose, which is then refined using a similar iterative refiner DNN network. The pose hypothesis can be further optimized by matching pose hypotheses across the different views and applying global scene refinement. EPOS [193] is a single-RGB, correspondence-based method that defines object-surface fragments to handle both global and partial symmetries. It predicts pixel-fragment correspondences, using an encoder-decoder network. Finally, EPOS overcomes the many-to-many correspondences issue and recovers all the object instances using a PnP algorithm [260] within a RANSAC framework [115].

Table 1 summarizes information from the most recent and well performing methods, as compared and presented in the context of the 2020 BOP challenge [196]. It presents the datatype and the average recall (AR) and run times of those methods for all the datasets examined in BOP challenge 2020 [196]. Table 1 also presents separately the accuracy results for some of the most interesting and challenging datasets. For example, for datasets T-LESS [194] and LM-O [77] which contain texture-less, symmetric, occluded objects and ITODD [132] which has co-existing, identical, overlapping metallic objects. From the table it can be concluded that the accuracy seems to depend highly on

the examined dataset, with all methods providing relatively good performance on some datasets, like T-LESS [194] and LM-O [77], but unacceptable on ITODD [132]. This is most probably because ITODD describes a very hard case of bin-picking of identical, shiny, texture-less and flat objects, under imperfect illumination conditions. However, it also seems that the accuracy tendency is more or less consistent throughout the different datasets, with CosyPose [252] outperforming all RGBD and RGB methods. However, testing CosyPose using RGBD data ends up being very slow with more than 13s processing time per image, compared to less than 0.5s for the RGB case. Other methods, also tested on both RGB or RGBD data, showed that with the use of Deep Neural Networks the additional depth information is no longer a critical accuracy factor. On the contrary, the use of photo-realistic physically-based rendered data has proved very crucial for the 6D pose accuracy. Regarding processing time, the methods exceeding the 1min per image, especially that of Drost [133], are considered inappropriate for real time applications. On the other hand, the methods that achieve 0.5-2s run time per image, like CosyPose [252], EPOS [193] and Leaping from 2D to 6D [274], can be considered potential candidates for real time applications like *FELICE*.

Table 1: Performance results of the most recent and best performing methods of the BOP challenge 2020 [196].

Method	Data type	Accuracy (AR)				Run time per image (s)
		Average on all BOP data	T-LESS [194]	LM-O [77]	ITODD [132]	
CosyPose [252]	RGBD (rendered+real)	69.8	70.1	71.4	31.3	13.74
Koenig-Hybrid [240]	RGBD (synthetic+real)	63.9	65.5	63.1	48.3	0.63
CosyPose [252]	RGB (rendered+real)	63.7	72.8	63.3	21.6	0.45
Pix2Pose [344]	RGBD (rendered+real)	59.1	51.2	58.8	35.1	4.84
CosyPose [252]	RGB (rendered)	57.0	64.0	63.3	21.6	0.47
Vidal-Sensors18 [462]	D	56.9	53.8	58.2	43.5	3.22
CDPN [269]	RGBD (rendered+real)	56.8	46.4	63.0	18.6	1.46
Drost [133]	RGBD	55.0	50.0	51.5	57.0	87.57
CDPN [269]	RGBD (rendered)	53.4	43.5	63.0	18.6	1.49
CDPN [269]	RGB (rendered+real)	52.9	47.8	62.4	10.2	0.94
Drost [133]	D	50.0	40.4	46.9	46.2	80.06
Drost [133]	D	48.7	44.4	52.7	31.6	7.70
CDPN [269]	RGB (rendered+real)	47.9	49.0	56.9	6.7	0.48
CDPN [269]	RGB (rendered)	47.2	40.7	62.4	10.2	0.98
Leaping from 2D to 6D [274]	RGB (rendered+real)	47.1	40.3	52.5	7.7	0.42
EPOS [193]	RGB (rendered)	45.7	46.7	54.7	18.6	1.87

3.4.2 Discussion

6D pose estimation requires a highly accurate and precise determination of the location and orientation of known objects in real time. This might include a lot of challenges, especially in cases of objects with special geometry and texture characteristics and changing scene conditions, like those below:

- **Object properties and occlusions**: initially developed feature-based methods using hand-crafted descriptors fail to detect and localise objects with weak texture, reflective surfaces and symmetries. Correct detection can also be prevented by challenging scene conditions, like overlapping and occluded objects, multiple instances of the same object, objects in different locations (boxes, dollies, trolleys, etc) with different backgrounds and changing illumination conditions. Additionally, acquisition parameters like unknown camera positions, camera noise and low

image resolution or large object-camera distance with respect to the object size, can also lead to failure. However, nowadays, Deep Learning methods cope successfully with these challenges without necessarily requiring depth information, outperforming classic methods using or not additional depth information. For example, Cosypose [252], EPOS [193] and Pix2Pose [344] (see Table 1) are some of the most well performing recent deep learning methods that handle such challenging cases.

- **Time and memory cost**: A successful human-robot interaction has to be real time, so that it offers a natural user experience and unobstructed assembly flow. To achieve this, 6D pose estimation would better be completed in less than 1s per image and run on on-board machines, which however might have limited memory and processing capabilities. This is a core challenge for time and memory consumption, that has to be repeated several times for different tasks and objects and also requires the connection to a database with known 3D object information.

- **Displaced or missing objects**: For reasons of time efficiency, the robot will try to detect an object not anywhere in the assembly line but within a predefined area / table / trolley / dolly, that this objects is expected to be found. It could happen though that an object has incorrectly been placed in a spot different than the expected one or it is totally missing. In such a case, the robot will have to provide the information of the missing object.

- **6D pose accuracy**: 6D pose estimation accuracy has been evaluated according to several evaluation metrics in the literature [196]: (i) Average Distance for distinguishable (ADD or ADD(-S)), symmetric distinguishable (ADD-S) and indistinguishable (ADI) objects, which calculates the distance between the vertices of the predicted and ground truth objects in the 3D space. (ii) Visible Surface Discrepancy (VSD), which also calculates error in the 3D space, but uses distance maps of the input and rendered image to define the object location. It is invariant to symmetries and it encounters only the visible object parts. (iii) Maximum Symmetry-Aware Surface Distance (MSSD) takes into account the maximum instead of the average distance (as in ADD), making it independent of the object geometry. (iv) Maximum Symmetry-Aware Projection Distance (MSPD) is similar to MSSD but it does not depend on the Z misalignment and for that reason provides 2D instead of 3D space error. All these metrics evaluate a different aspect of pose estimation. For that reason, combinations of such metrics are commonly used to define more representative pose estimation accuracy scores [196], like the recall, the average recall for varying thresholds and the average of the average recalls for different metrics.

FELICE will adopt a 6D pose estimation procedure as follows:

- A database of known 3D object models will be used. These models will be derived from passive sensors and will be as light (sparse) and representative as possible, to avoid unnecessary memory consumption and computational cost.

- Deep Learning methods able to handle cases of objects with special characteristics of geometry and texture will be used. RGB information from the input images will

be used as a base, but additional depth information, derived from passive sensors, could also be exploited.

- The implementation will comply with the memory, computational and time restrictions of the moving, assisting robot in the CRF assembly line.

3.5 Camera sensors for scene and object perception

The earliest visual methods for both SLAM and object detection and localization employed plain RGB cameras and were based on local and low-level features. However, to extract accurate depth information, they usually required multiple images with large baselines, increasing complexity and costs. Additionally, passive RGB cameras have difficulties in reliably handling objects and scenes with weak texture and occlusions. To overcome these issues and facilitate 3D perception and reconstruction [109], active 3D cameras [42] were examined.

Structured light and time of flight (ToF) are nowadays the two prominent 3D camera technologies [219]. A structured light camera operates by projecting an active pattern and then analyzing its deformations on the scene surfaces to recover depth. Kinect I, introduced in 2010, is the most popular structured light camera. A ToF camera measures the time that light has been in transit to estimate distance. The second-generation Kinect II is a ToF camera [400]. Applications of structured light cameras in robotics and vision are described in [390, 211, 197, 440, 248, 179], whereas of ToF in [42].

ToF cameras are robust to occlusions, shadows and sharp depth edges because they determine depth from a single view [219]. However, they provide low spatial resolution, fail to handle motion because it requires multiple shots and might also require manual focus adjustment [42, 219]. Structured-light sensors, on the other hand, are more appropriate for controlled environments due to their low frame rate but are more prone to giving rise to depth map holes due to obstruction of the line-of-sight [219]. Moreover, all active RGBD cameras face difficulties in dealing with materials that do not properly reflect the camera-emitted active illumination, like for example reflective, shiny, dark and transparent ones. Such objects though are very commonly found in industrial applications and indoor environments like in *FELICE*, where the illumination conditions are not optimal. Space limitations and accessibility are also critical aspects in such applications, with active sensors having a limited range of operation [400], like 0.7m–5.0m and 0.5m–4.5m for Kinect I and ToF Kinect II, and 0.2m–10m for the Intel RealSense D435. On the contrary, the operating range and field of view of passive RGBD cameras depends on the lenses that they carry, so it can be flexibly adjusted to the needs of a particular application, by exchanging lenses. Due to their low-cost, small physical size, simplicity and wide applicability, RGBD cameras have been widely used for several years, especially for indoor applications with textured and deformable objects. In recent years though, both RGB and RGBD cameras are being combined with machine and deep learning methods, overcoming many of the restrictions of the past and enabling the perception of complex scenes and challenging objects.

4 Human behavior monitoring in assembly task execution

4.1 Overview

Human-behavior monitoring in *FELICE* it is of interest a) to differentiate humans from the environment, being able to track and follow their motion and detecting their actions to realize an "anticipatory control" by robots and b) to detect behaviour patterns and support more precise inferences about the subjects state. To this end, an overview of research in human pose estimation, action recognition is provided. Furthermore, relevant research in vision-based human activity monitoring and posture-based ergonomic risk assessment is summarised.

4.2 Relationship with FELICE project

Acknowledging the importance of human-centric environments and the human resource in a hybrid HRC assembly scenario, behaviour monitoring in *FELICE* aims to (i) facilitate a better understanding of the activities/actions, health, and risks faced by a worker by detecting behaviour patterns and supporting more precise inferences about the worker's situation and environment, (ii) to enhance the synergy between robots and humans during assembly task execution. For the former, the aim is to automatically detect human actions and identify abnormalities in task cycle execution which may relate to abnormal body postures or assess spatio-temporal variations in and between assembly actions performed by the worker and further associate them with indicators for ergonomics analysis to support decisions for directing the robot to a specific worker or for configuring the workstation components. Information will aim at supporting a per-worker profiling on postural deviations and physical stress aggregating information from a series of work task cycles. *FELICE* relies primarily on visual sensors deployed across the assembly line and secondarily on non-visual sensors, such as smartwatches, which might provide additional cues (i.e. acceleration, number of steps and heart rate measurements) to resolve spatial ambiguities in human body segmentation and temporal ambiguities in action recognition, which are challenging due to the particularities of industrial environments.

4.3 Human pose estimation

The human pose estimation task aims to recover the posture of the human body from sensor inputs. Vision based approaches exploit camera inputs to provide an estimate. Human pose estimation is a very important research field related and applied to action/activity recognition [282, 262], action detection [261], human tracking [208]. The surveys of [367, 314, 37, 165, 506, 256] reviewed the early work of human motion analysis in many aspects (e.g., detection and tracking, pose estimation, recognition) and described the relation between human pose estimation and other related tasks.

More recent surveys mainly focused on subdomains, such as RGB-D-based action recognition [101, 473], 3D human pose estimation [200, 399], model-based human pose estimation [200, 360], body parts-based human pose [275], and monocular- based human pose estimation [174]. The most recent survey by Chen (2020), summarizes the

deep learning-based human pose estimation methods, which were mainly published from 2014 onwards. The methods can be categorized into a) generative methods (model-based) and discriminative methods (model-free), based on whether they use designed human body models or not; b) top-down and bottom-up methods according to the starting point of the prediction: high-level abstraction or low-level pixel evidence; c) regression- and detection-based methods, where the former directly map the input image to the coordinates of body joints or the parameters of human body models and the later treat the body parts as detection targets based on two widely used representations: image patches and heatmaps of joint locations; d) One-stage vs. Multi-stage deep learning methods, with the former aiming to map the input image to human poses by employing end-to-end networks, and the later predicting human pose in multiple stages, accompanied by intermediate supervision.

Works address both directions of 2D and 3D single person pose estimation in addition to multi-person pose estimation. The later is out of the scope of this overview as the human behavior analysis considers a single person conducting the assembly task at each workstation. 2D human pose estimation calculates the locations of human joints from monocular images or videos, whereas 3D human pose estimation predicts locations of body joints in 3D space from images or other input sources. Methods using CNNs can be categorised in regression and detection-based, the former attempting to learn a mapping from image to kinematic body joint coordinates by an end-to-end framework and produce joint coordinates, whereas the later predict approximate locations of body parts or joints, and usually are supervised by a sequence of rectangular windows or heatmaps. Each category bears advantages and disadvantages, though heatmap learning results in better robustness. 3D human pose estimation is more challenging since it needs to predict the depth information of body joints, plus, the training data for 3D human pose estimation are not easy to obtain. For this category a kinematic model is widely used (e.g. [306]. Most existing datasets are obtained under constrained environments with limited generalizability.

The state of the art has been advanced significantly with the introduction of both depth cameras and deep learning techniques. Despite the fairly accurate performance of state of the art algorithms in controlled or semi controlled settings, coping with complex, realistic scenarios exposes the limits of these algorithms, particularly their effectiveness in handling occlusions. Efficient networks and adequate training data are the most important requirements for deep learning-based approaches. Furthermore, to support the processing on low-capacity devices, the network parameters need to be reduced. Regarding the choice of appropriate camera sensors, the discussion of Section 3.5 on passive RGB and active RGB-D systems is also relevant here, thus the reader is referred to it.

4.4 Action recognition and understanding in videos

Human motions extend from the simplest movement of a limb to complex joint movement of a group of limbs and body. Although the concept of “action” is intuitive, the difficulty in providing a single definition to what an “action” constitutes and the differentiation from an “activity” is evident from a number of different works providing examples to these definitions [314, 367, 474]. Understanding human actions in visual

data is tied to advances in complementary research areas including human dynamics, domain adaptation and semantic segmentation and extends over a broad range of application areas from video surveillance to human-computer and human-robot interaction, quality-of-life improvement for elderly care, and sports analytics. A large body of research exists in the domain, and continues to expand. Relevant survey papers on related methodologies follow a different taxonomy though with the latest advances in deep learning, the most recent ones provide a separate category for deep learning based techniques, discussing various architectures and training methods as in [190].

Recently, a significant amount of research has been dedicated to visual understanding of human actions and human-object interactions (HOI) using deep neural network models [255, 266, 144]. Most of these approaches focus on capturing and modelling coarse geometric and appearance representations of the whole human body using coarse spatial regions-of-interest (ROIs) [169] and classify human actions and HOI in short video clips. More elaborate methods rely on fine representations of the temporal and spatial structure of acting entities, i.e. 2D/3D skeletal or mesh body models [89, 373], 2D hand or object masks [56], 3D poses of hand(s)-object(s) [441, 162], 3D mesh hand models [412]. Spatiotemporal relationships of human-object(s) are modelled using attention mechanisms [284], Graph Neural Network-based (GNN) methods [266] or their combination [186, 103], Convolutional Neural Network-based (CNN) methods [232], Recurrent Network-based (RNN) methods [507] or the recently proposed, versatile and powerful Transformer model and its variants [505, 168, 457].

While successful in learning and recognizing HOI, most of these methods treat HOI as non-composite, monolithic activities classifying them into single-layered classes. Moreover, they do not generalize well over the number of actions they are able to model and learn to discriminate. Another body of research considers HOI as fine-grained, composite activities involving multifaceted spatio-temporal human-object(s) relations [300] by (i) integrating high level semantic information [54], (ii) formal logic rules and knowledge-based graphs [490], (iii) graph-based methods [341]. A novel HOI representation [214] relies on spatio-temporal graphs to encode human-object relationships across time. Other recent approaches shift their attention to the challenging tasks of visual reasoning [56, 176] to discover causal relationships in space and time between interacting entities.

Motivated by the emerging applications of Human-Robot Interaction and Collaboration [476, 117, 113, 233, 234] and fine-grained Human Behavior Monitoring, researchers in computer vision and robotics [156] have recently joined their efforts to tackle the special and challenging problem of fine-grained recognition of compositional activities in assembly videos. In this context, the fine-grained recognition problem refers to joint temporal segmentation (action detection) [222] and classification [505] of a sequence of assembly actions that comprise a complex and possibly long assembly activity. A series of methods have been proposed that are able to model both the temporal and spatial structure of assembly procedures in a fine-grained manner in realistic scenarios [144, 215, 216, 385]. Several features can also be integrated to advance the functionality of methods and the recognition performance regarding the semantics of manipulation actions [492, 69], the 3D human hand motion and grasp types [501], generic object-level information (3D/2D shape, pose and motion) [433] but also the object contact points [216] and affordances [436, 117, 201]. Most of these methods showcase their performance using videos of furniture construction tasks [216], cooking

activities [492], toy block building tasks [215] and simple human-robot collaborative assembly tasks [476]. Jones et al. [216] have recently proposed a novel generative approach on fine-grained activity recognition for assembly videos by employing the notion of kinematic state as a graphical model along with observation features that take advantage of a spatial assemblys special structure. A probabilistic, segmental Conditional Random Field is trained to infer the temporal boundaries and categories of actions in a video demonstrating an assembly process and enforce temporal consistency in the models output. Previous methods have also employed hierarchical graphical models to encode fine-grained manipulation activities in assembly or cooking videos using action grammars [433] and probabilistic context-free grammars [466]. In essence, a parse tree is built whose terminal and non-terminal nodes represent primitive objects and composite actions of the overall activity.

4.5 Posture-based ergonomic risk assessment

Automatic, vision-based postural assessment is an important task in the broad area of activity recognition and understanding with numerous solutions and important applications from video retrieval, robotics to surveillance, human performance evaluation/analysis and automatic ergonomic risk assessment in the workplace [454, 464]. In particular, the emerging application of automatic ergonomic risk assessment refers to the evaluation of potential risks for work-related musculoskeletal disorders (WMSDs) by observing the human body configuration and motion during work activities. The overarching goal of ergonomic risk assessment is the prevention of WMSDs towards the improvement of occupational safety and health of workers in real working environments [471, 414]. To achieve this objective, an efficient solution is required to visually identify prolonged suboptimal (also noted as abnormal) working postures, motion patterns, material handling, forces exerted to the human body and possibly combine those findings with information acquired by physiological or other type of measurements of the workers body during strenuous, labor-intensive work activities that often attribute numerous repetitive tasks during a work shift. Based on these observations, the assessment of ergonomic risks for physical strain and overloading of body joints is considered a significant indicator for preventing potential muscular injuries in the workplace and WMSDs.

An ergonomic risk index is an assessment tool that involves an observational and graphical protocol or a checklist that associates the intensity, frequency and duration of suboptimal working postures or other observations with the level of physical strain and risk for WMSDs as a single valued score or set of scores. Some ergonomic risk assessment methods that are commonly used in the industry are the Rapid Entire Body Assessment (REBA) [303], the Rapid Upper Limb Assessment (RULA) method [302], the European Assembly Worksheet [402] (EAWS), the OCRA checklist [334], the MURI risk analysis [483] and others. Conducting an ergonomic risk assessment is a foundational element of the ergonomic process that drives the analysis, design and optimization of manufacturing or construction activities towards effective and ergonomically safe workflows [113]

Ergonomic risk assessment methods are categorized into three main groups based on the type of measurements acquired to evaluate the working postures and risk factors

for WMSDs [35, 264]. Those refer to self report assessment based on questionnaires, rating scales, checklists and interviews of workers that are considered as subjective observations, direct methods that rely on analysis of physiological measurement acquired by wearable sensors on the worker's body and screening or observation methods that consist of manual or computer-aided analysis of visual data acquired by cameras during work activities [249, 424]. Analysis of visual information can be performed offline and manually or in a semi-automatic manner by experts or can be automatically computed in real-time or offline using vision-based methods for monitoring human behavior and postural assessment of physical strain that provide indicators to increased ergonomic risk of WMSDs in the workplace.

Based on recent advances in vision-based human pose estimation and tracking in 3D, even from a single RGB image, as well as skeleton-based action recognition [190, 506, 256], various effective solutions have been proposed for human performance analysis that associate postural assessment and ergonomic risk assessment. We summarize previous work related to those important tasks below.

The work proposed in [35] focuses on the task of human posture analysis using a deep learning approach to assess the risk of WMSDs during manufacturing activities. The proposed method relies on motion capture data to drive synthetic human models for representing workers motion, while a data generation pipeline is also presented to synthesize a dataset of depth frames that features simulations of manual tasks performed by different workers. A deep residual convolutional neural network model is trained using the synthetic data to predict body joint angles of manufacturing workers from a single depth image and predict the Rapid Upper Limb Assessment (RULA) metric [302] for the investigation of work-related upper limb disorders. Parsa et al. [347] introduced a novel approach for action segmentation and subsequently for predicting the ergonomic risk of object manipulation actions according to the REBA ergonomic risk index using spatial and temporal visual features from RGB-D frames. A Temporal Convolutional Network (TCN) is trained to semantically segment action into a hierarchy of actions, which are either ergonomically safe, require monitoring, or need immediate attention. Given the input skeleton and the recognized activity for a video the REBA score is averaged over all subjects offline, while one score is reported for each activity class. A new dataset was also introduced comprising twenty individuals picking up and placing objects of varying weights to and from cabinet and table locations at various heights and relevant annotations according to the REBA ergonomic model [303].

Apart from the visual data modalities based on the appearance and depth of the human body and the observed scene, the estimation of the human body skeleton and the extraction of relevant features in the form of 2D or 3D coordinates or angles of body joints or parts/limbs, provide significant information that can efficiently encode the spatial body configuration, the coarse as well as the fine human motion during activities. Recently proposed methods leverage the use of skeletal features for the risk assessment of WMSDs showing improvement over the methods that merely use appearance and/or depth features. To this end, Shafti et al. [409] focus on a real-time human-robot interaction scenario during welding actions driven by ergonomics following the RULA posture monitoring method. The proposed method automatically extracts the 3D skeletal information of the upper body using RGB-D frames to continuously analyze the users posture and understand the safe range of arm motions during welding actions in order to further calculate appropriate robot responses. In the work of Li et

al. [263], the 3D skeletal pose of the worker in videos of construction activities is estimated per frame by combining the 2D poses computed from two different viewpoints using deep neural networks. The method introduced by Yan et al. [491] aims to recognize awkward postures of construction workers using view-invariant features from 2D skeleton motion data captured by a single ordinary RGB camera in the context of construction hazard prevention, Mehrizi et al. [305] have also proposed a deep learning approach for markerless 3D pose estimation optimized in the context of object lifting tasks by using RGB images from two different viewpoints. Kim et al. [235] introduced a novel framework for adaptable workstations in manufacturing, aiming to improve worker ergonomics and productivity based on a reconfigurable human-robot collaboration setting. To achieve this goal, the tasks of overloading assessment of the workers body joints and user intention recognition are applied in real time to monitor and adjust the reduction of ergonomic risks of working with power tools using a stereo camera. A cobot is then programmed to preemptively act helping the worker to perform the intended manipulation task in configurations where the effect of external loads on body joints is at a minimum. The method introduced by Plantard et al. [365] aims to evaluate potential WMSDs using 3D skeletal pose estimation information computed from a single RGB-D camera to evaluate the RULA ergonomic assessment in real workstations of a car manufacturing factory.

Recently, Parsa et al. [346] proposed a novel variant of Graph Convolutional Networks, noted as Spatio-Temporal Pyramid Graph Convolutional Network (ST-PGN, to segment and recognize action classes in videos using 3D skeleton information from video sequences and also predict the REBA ergonomic risk score. The proposed model combines Pyramidal GCNs (PGNs) and Long Short-Term Memory Units (LSTMs) to learn an hierarchical representation of spatial features corresponding to different areas of the human skeletal body model comprising different body joints. Each level of this hierarchical representation is then used as input to an individual LSTM unit to learn the temporal aspect of the input sequence for this body area at different spatially semantic layers. Then, estimation of the REBA score is performed online as a second processing step, given the skeletal features and the computed activity labels for each recognized activity towards the assessment of ergonomic risk for musculoskeletal disorders and occupational safety. An extension of the latter work [345] employs a multitask learning paradigm to simultaneously detect actions in videos and predict the REBA ergonomic risk score for each frame in videos demonstrating object manipulation tasks (lifting, moving boxes etc.). Action detection is powered by an Encoder-Decoder Temporal Convolutional Network to semantically segment long videos into distinct activity classes, whereas online regression of the ergonomic risk per frame relies on GCN and LSTM models to embed the spatiotemporal relationships between human joints in the features.

Finally, another recently proposed method by Konstantinidis et al. [242] introduces a novel multi-stream deep network that acquires 3D skeletal data sequences extracted from videos to compute the REBA score regardless of the activities performed in a video. Each stream is responsible for predicting a partial score that corresponds to a predefined set of body parts prior to their aggregation for the computation of the total REBA score.

4.6 Datasets

With the advancement of motion capture systems and crowd sourcing the available data have expanded in terms of quantity and acquisition context, i.e. outside the lab environment. Papers [45, 174] provide a summary of early datasets though these are mainly characterised by limited number of images and activities serving specific applications. Among the state of the art benchmarks is the Max Planck Institute for Informatics (MPII) Human Pose Dataset [45] which includes rich annotations, the joint-annotated Human Motion Database (J-HMDB) [213]. Other recent state-of-the-art datasets for 3D action recognition is the large-scale action recognition NTU RGB+D dataset (120 action classes and 114.480 samples in total)[273], the Kinetics dataset [92], the recently proposed BABEL large scale dataset with language labels describing the actions being performed in mocap sequences and frame-level annotations for fine-grained action analysis [371] and the FineGym dataset providing a insightful hierarchical representation of gymnastic activities for fine-grained action understanding and performance evaluation in sports [413].

Among the few existing datasets related to action recognition or posture-based ergonomic risk assessment in assembly videos, the UW-IOM dataset [347] features a limited number of object manipulation actions involving awkward poses and repetitions and provides frame-level annotations for scores according to the REBA ergonomic risk index, while the existing TUM Kitchen dataset [443] was also annotated with respect to the REBA scores in the same work. The IKEA furniture-assembly demonstration dataset [476] provides multifaceted annotation data for a realistic scenario of chair assembly actions in video. The most relevant dataset to action recognition and ergonomic risk assessment that was recently introduced regards a set of action sequences demonstrating door assembly scenarios captured in a real industrial workplace [340]. It provides rich annotation data towards assembly action detection, classification and ergonomic risk assessment based on the MURI risk analysis method [483], as provided by experts in ergonomics for manufacturing activities.

4.7 Discussion

Open challenges in visual understanding of human motion refer, among others, to the intra-class variability, the complex and modular structure of human actions, visual polysemy of actions, hand-object occlusions, multiple scales (near, far) and viewpoints (first-, third-person) of observations, multi-functionality of objects, etc. Innovative ideas to tackle these challenges will be sought towards the integration of rich, multi-level semantic information on the spatio-temporal relationships of human, objects and actions based on domain expert knowledge, rich vision-based extracted information regarding the pose and motion of objects and of the acting human body using 3D models for pose estimation and tracking across time and an efficient representation of the temporal structure of the assembly activities.

Moreover, an open research challenge to consider regards the use of powerful deep neural network models and effective attention mechanisms to learn the temporal multi-level semantic structure of fine-grained HOI in the context of assembly actions in manufacturing. The analysis of the current state of art methodologies suggest to search

for efficient solutions to the tasks of action recognition and postural assessment for ergonomic risk analysis towards deep learning approaches based on elaborate Graph Neural Networks, enriched with attention mechanisms to efficiently represent and learn the temporal structure of complex assembly activities, model the human body configuration and the spatio-temporal relationships of human and objects in videos. Efficient marker-less estimation and tracking of the full articulated 3D human body pose and of the 2D or 3D poses of objects in the scene is a fundamental step to consider towards these goals.

Within *FELICE*, we will consider using RGB and depth-based visual information acquired by a multi-camera system in order to enable the acquisition of rich visual data for the interacting entities in a video and to overcome occlusions and will possibly occur in the real settings of the cluttered and complex industrial environment. Another important aspect to consider regards the estimation of smooth and physically-principled 3D body configurations and motions by learning the correlation between the human body dynamics and a set of biomechanically inspired constraints. Such constraints can be integrated to an appropriate 3D human body model and in the training of a neural network model to enable effective learning of anatomically and physically plausible body poses and motions, thus facilitate effective, real-time postural analysis and ergonomic assessment for a user during work activities. The rich information extracted for the poses and trajectories of the human body and interacting objects throughout the video will be combined with prior, multi-level semantic information regarding the temporal structure of the activities to train an attention-based graph neural network model for learning to segment and classify assembly actions during composite activities in long videos. A Deep Multi-Task Learning (DMTL) approach will also be considered in order to jointly learn and bring together the challenging tasks of activity recognition and ergonomic assessment. Integration of non-visual sensory data, such as accelerometer data based on worker's hand motion and of heart-rate data, acquired by wearable devices, will also be considered towards multi-modal human activity recognition and ergonomic assessment.

5 Robotic hardware

5.1 Overview

This section gives an overview of the commercially available manipulators that could be potentially employed in scenarios similar to those pursued within the *FELICE* project. The basic characteristics of these products and the technologies used within will be discussed in the context of their applicability to the *FELICE* scenario. Furthermore, a brief description of the robot under development in the context of *FELICE* is given. More details can be found in "D2.1: Robot, architecture and system specifications I".

5.2 Relationship with FELICE project

FELICE will follow an inter-disciplinary approach in studying the various aspects of Human-Robot and Robot-Environment Interaction in industrial environments, towards advancing the way that service robots co-work with humans, not only as computer-based servants solely capable of obeying commands, but as true companions, adapting and evolving along with the human. The proposed system's core will be an autonomously moving service robot with an on-board touch screen and a dexterous robotic hand and arm. Following the paradigm of similar research projects, the robot will build upon existing knowledge so as to include advanced versions of features already researched and developed.

5.3 State of the art

5.3.1 Commercially available mobile manipulators

FELICE project targets a specific scenario of a robotic assistant that will share the working space with human workers supporting them in their daily tasks. Implicitly, this scenario alone poses important requirement for the robotic hardware. It has to have the capabilities similar to those of a human worker in terms of locomotion, manipulation, perception and interaction with the environment and the other workers. As indicated in the project proposal, no off-the-shelf robots can fulfill all the requirements, and therefore only the most suitable candidates are presented here. The *FELICE* consortium has already identified the leading competitors in the market. A direct comparison of robotic platforms can be found in Figure 3 considering among others, physical characteristics, potential to fulfil the *FELICE* user requirements, user friendliness, and design features related to industrial environments.

		Tiago	Tiago ++	Movo	Fetch	Toyota HSR	Pepper	CareObot4	Kompai	Aeolus	RB-1
GENERAL	footprint axb (cmxcm) /Φ (cm)	Φ54	Φ54	50,8x81	50,8x55,9	Φ43	48x42,5	Φ72	42 x 49,5	N/A	Φ50
	weight (kg)	70	85	120	113,3	37	29,1	140	45	N/A	54
	max body height (cm)	145	145	158	149,1	135	121	148	118,6	N/A	102,9
	torso lift (cm)	35	35	48	39,5	34,5	N/A	N/A	N/A	N/A	40
ARM	ARM no x DoF no.	1 x 7	2 x 7	2 x 7	1 x 7	1 x 4	2 x 6	2 x 7	N/A	2 x 7	1 x 7
	single arm payload (kg)	3	3	2,4	6	1,2	N/A	5	N/A	N/A	2,1
	arm reach without end effector (cm)	87	87	98,4	94,05	60	N/A	90	N/A	N/A	98,4
	griper type (grasping force N x opening cm)	pararell 2 fingers / 5 fingers	pararell 2 fingers / 5 fingers	2/3 fingers adaptive	pararell 2 fingers (245Nx10cm)	2 jaw (N/Ax13cm)	5 fingers	1 finger	N/A	2 jaws	2/3 fingers
	sensors	optional 6 DoF F/T sensor	6 DoF F/T sensor	torque sonsors (every joint)	6 Dof IMU (gripper)	6 DoF F/T sensor	N/A	N/A	N/A	N/A	N/A
PLATFORM	drive type	differential	differential	holonomic (mecanum wheels)	differential	holonomic	N/A	N/A	differential	N/A	N/A
	speed (m/s)	1	1	2	1	0,2	1,39	1,1	N/A	N/A	1,5
	operational env.	indoor	indoor	indoor	indoor	indoor (5 deg inclines and 5 mm thresholds)	N/A	N/A	N/A	N/A	indoor
	sensors	laser + sonars + 6 Dof IMU	laser + sonars + 6 Dof IMU	2x 2D planar laser	2D laser + 6 Dof IMU	6 Dof IMU, Laser range	6 lasers + 2 sonars, gyrosensor	3 lasers	2 lasers, 3D cam and infrared	N/A	laser
HEAD	head DOF no	2	2	2	2	2	2	N/A	N/A	N/A	2
	display	optional	optional	no	no	yes	no	yes	no	no	no
	sensors	RGB-D camera	RGB-D camera	Kinect for Xbox One	3D RGB-D camera	stereo camera, 3d sensor, wide camera	N/A	N/A	360° 2d lidar	N/A	ORBBED camera, Asus Xtion RGBD
CONNECTIVITY	connectivity	Wi-fi/bluetooth 4.0	Wi-fi/bluetooth 4.0	WI-fi, ethernet	Wi-fi	N/A	Wi-fi, Ethernet, Bluetooth 4.0	Ethernet	N/A	N/A	Wi-fi, Ethernet, USB, HDMI
BATTERY	Voltage and Capacity	36V 20Ah/ 36V 40Ah	36V 20Ah/ 36V 40Ah	N/A	12V / N/A	N/A	N/A/30Ah	48V/N/A	N/A	N/A	24V / 30Ah
COMPUTING	parameters (CPU, RAM, SSD)	i5/i7, 8GB/16GB, 250GB/500GB	i5/i7, 8GB/16GB, 250GB/500GB	2x (NUC5i7RYH, 16GB, 128 GB)	i5, 16GB, 120GB	N/A	ATOM E3845, 4 GB, 32 GB eMMC	4-6 Intel NUC i5, 8GB, 256 GB	N/A	N/A	4 Generation Intel i7, 120 GB, 8 GB

Figure 3: An overview of commercially available mobile manipulators.

The manipulators listed above are at various levels of their maturity vs availability. at the moment, only the Tiago, Movo Fetch, Pepper and RB-1 are practically commercially available. However they do not fulfill all the consortium requirements regarding computing power, human-robot interaction, physical safety, connectivity etc. The possibilities of adaptation of the readily available products to the project needs are also very limited. The *FELICE* consortium sees the need for custom building of the experimental platform and the crucial robotic components.

5.3.2 The FELICE robot concept

FELICE aims to develop a robotic system that will introduce solutions beyond the industry state-of-the-art, to face the challenges identified in the targeted market. Practically speaking, ACCREA will elaborate on its existing RAMCIP robot[1] adapting it to the in-

[1] https://ramcip-project.eu/

dustrial scenario. A comprehensive list of such challenges, the solutions envisaged and the ambition involved, indicated by their technology readiness level (TRL), is provided below:

Modularity: *FELICE* will address increased personalization needs in the target market by developing a highly modular system. Each hardware component will be designed as a standalone unit accompanied by its low level control interface implemented in ROS [375]. ROS is a middleware capable of integrating asynchronous sensors and H/W components (TRL 9). Many of the existing robots, including RAMCIP robot (TRL 6) has been developed on ROS. The design of ultra-modular hardware components that can be seamlessly assembled and one can recognize the physical existence of the other through ROS, eases the entrance of the *FELICE* in the research market (e.g. a customer already owns a robot platform and needs only the manipulator and the gripper from *FELICE*) and allows the existence of flexible production lines to host various robot versions.

Safe human robot interaction: Another significant challenge for the contemporary service robots that target operation in real house environments with real users is safety. RAMCIP had a significant amount of safety features (TRL 6) including compliance manipulation, panic buttons, bumpers and human aware navigation (TRL 6). In *FELICE*, safety will be inherited to the robots from the design phase. The mobile platform along the existing features will also have compliance control coupled with human aware navigation to consider also human comfort. The manipulator will be integrated with f/t sensors in each joint to enhance compliance manipulation and back drivable attitude. Such hardware features integrated with active vision will increase the safety of manipulation while the human is at close distance to the robot, by dynamically updating its collision space.

Smart factory connectivity: It is anticipated that within the next years the amount of smart devices integrated in factory environments will increase. Thus, smart factory environments will be also the physical environments that a robot should live. *FELICE* will transfer this technology to the envisioned robot and will provide it with the capacity to connect and interact with diverse IoT devices in smart factories.

Specifically, the robot under development in *FELICE* consists of a differential mobile platform providing the locomotion and navigation functionality, a manipulator with exchangeable effectors including a dexterous gripper, a sensorised head with cameras, speakers and a face display, and an elevation mechanism regulating the height of the robot body. The robot is human-scaled, both in terms of its physical size and performance, and built with ultra-lightweight materials. All actuated parts of the robot are equipped with force-sensitive actuators recognising contacts with environment, thus ensuring safe interaction with humans. The attached power supply, communications and computing units constitute the robot a fully integrated and self-contained entity. In accordance to the requirements of anticipated use-cases, various dedicated tool and effectors have been selected, including a vacuum gripper. These effectors and a corresponding tool change mechanism will be custom developed and built in the course of the project. In addition to the above robot, FHOOE's PlugBot mobile platform will also be used for more constrained experimentation, (see Figure 4).

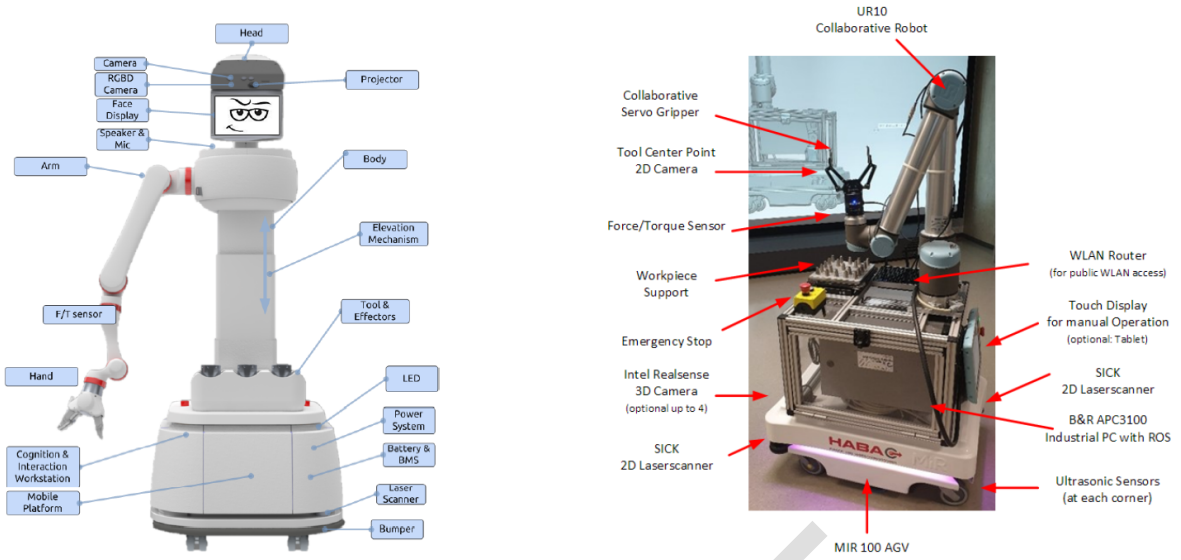


Figure 4: The robotic platform built by ACCREA (left) and FHOOE's PlugBot mobile robot platform (right).

6 Adaptive workstation

6.1 Overview

Adaptive workstations (AWS) allow individualized customization of the workspace to improve physical and cognitive ergonomics, productivity/efficiency, and quality. This section provides an overview of the most recent developments (both publications and projects) in adaptive workstation systems for assembly operations. Additionally, the demonstrators currently in use by *FELICE* partners will be described. The relevant elements or aspects of a workplace and workstation that can be adapted will be presented and the corresponding adaptation processes relevant to the *FELICE* use case (for instance, vehicle door assembly) will be discussed.

6.2 Relationship with FELICE project

The *FELICE* adaptive workstation is going to be developed with safety, physical, cognitive and environmental ergonomics as well as productivity in mind. The development process includes three phases, starting with the definition of requirements by an analysis of the state-of-the-art systems and technologies, followed by the development of prototypes, and ending with the evaluation of the design against these requirements. The design of the workstation will accommodate ergonomic posture and movement towards improved productivity/efficiency of the system during work processes by physically adapting to the user's body proportions based on a set of actuators. Furthermore, it will assist worker deliberation by providing visual or auditory information. Other types of support will include the improvement of environmental conditions, e.g. by adapting the local illumination intensity. The AWS will be based on CRF's current workstation for assembly operations of car doors (located at the Melfi training center) and TUD's adaptive workstation prototypes for assembly operations. The AWS will be able to adapt itself in real-time based on the input received from the AI-driven Manufacturing Execution System (MES). In particular, it will use its video and audio system to inform human workers about changes in the assembly process, detected ergonomic issues (for example prolonged bad posture), and adapt parameters such as the height or inclination of the workpiece (e.g. car door) in order to improve posture and cope with worker fatigue. An early concept of the workstation is depicted in Figure 5.

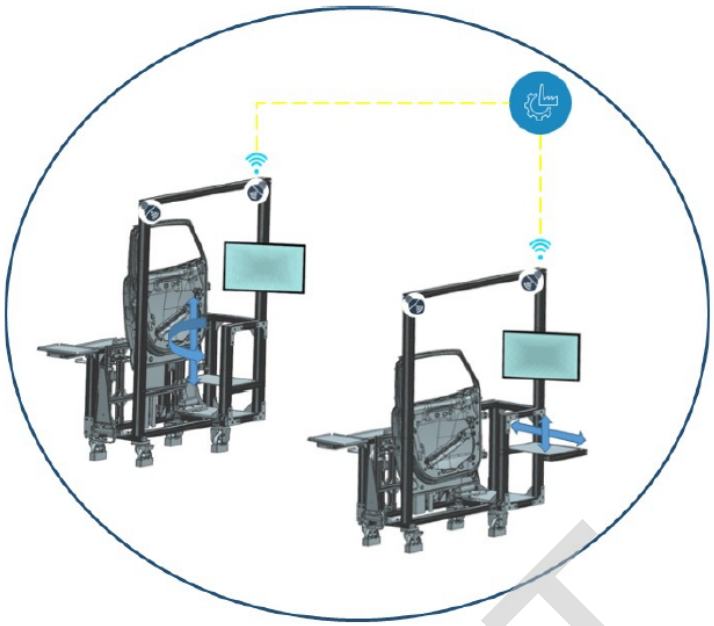


Figure 5: CRF's adaptive workstation.

6.3 State of the art

Adaptive workstations are already deployed on an industrial scale, e.g. at the FCA automotive plant in Cassino, Italy. The design of most assembly systems is generally based upon average worker body proportions. Through the use of big data, the analysis of weight, height, origin and other parameters of 3000 male and 3000 female workers of 13 automotive plants, led to the creation of models capable of representing the real features of the workers to study ergonomic workstations, to allocate personnel in an efficient way according to their anthropometric characteristics, or to better manage PPE (personal protective equipment) without waste, thus further reducing risks. Using this data, the workstation can raise or lower itself automatically based on information that is stored on the worker's badge, e.g. [408].

Adaptive or self-adaptive systems are also the topic of several research projects in academia. For example, there are two prototype demonstrators currently in use at TUD's Institute of Ergonomics and Human Factors (IAD) laboratories. The ergonomic adaptive workstation with automatic height and reach adjustment for precise assembly tasks was developed in 2016 as part of the Mittelstand 4.0 project (see Figure 6). It improves the worker's posture by automatically adjusting the height of the working desktop and the distance of the material storage into an optimal configuration based on the user's anthropometric data (i.e., body proportions) and the assembly task requirements. The adaptation process starts when the worker is in close proximity to the AWS by obtaining the worker's ID from an RFID tag. Personal information of the worker (e.g. height, age, qualification) can be retrieved from a database using the read ID. An adaptive lighting system adjusts the light intensity with respect to the personal preferences of the worker in regards to the particular task [391, 392].

The IAD human-robot collaboration workplace (see Figure 6) combines a height-adjustable working desktop with a static cobot that supports the operator by supplying

parts and delivering heavy task-related objects. The demonstrator provides different modes of operation to assess how various conditions during human-robot collaboration affect the user's interaction with the system in order to provide data for basic & applied lab studies [459].

Figure 6: Adaptive workstation prototypes at the TUD/IAD laboratories.

Other recent developments in the field of adaptive workstations in the EU include the "Adaptive Assembly Workstation" of the CO-ADAPT project[2], which is capable of adjusting the desktop height and the light intensity of the workplace. The workstation supports the worker via an integrated cobot that can be easily programmed with a tablet and is able to handle payloads of up to 10 kg [111].

Other technologies pertinent to adaptive workplaces have been developed by 2016 in the "MAN-MADE" project[3] which, amongst others, encompasses a 3D camera system to measure anthropometric data, sensors to assess the physical and cognitive capacity of a worker, tools to detect ergonomic risks at manual assembly stations, etc. [335].

Bortolini et.al. present a prototype of a "Self-Adaptive Smart Assembly System", able to reconfigure the position of a material storage area in front of the worker along two Cartesian axes. The materials needed to assembly variants of a product can e.g. be moved towards the worker to support him during the picking task. The aim of this system is to improve posture and to reduce movements and picking-time. The working desktop, where the parts are assembled, can be moved vertically as well. The system adapts based on the user's body proportions, the work cycle, and the product, but is fully customizable by the worker via a Graphic User Interface. Tracking of the worker and performance monitoring is achieved through a marker-less motion capture system comprised of at least four cameras. An increase of productivity by over 60% is reported during a selected assembly task while using this system [75].

6.4 Adaptivity of the workplace

The *FELICE* AWS will adapt to a user's individual needs. Ten aspects (dimensions) of the workplace can be designed in a user-adaptive manner as described in Table 2 using the taxonomy by Rupprecht and Schlund [389].

[2]https://cordis.europa.eu/project/id/826266
[3]https://cordis.europa.eu/project/id/609073

Table 2: Dimensions and options of the workstation user-adaptivity [389].

Dimensions	**Configuration options**
DIM 1 - Working desktop	- height - rotation/angle
DIM 2 - Arrangement of physical objects and provision of materials	- physical objects - materials - control elements
DIM 3 - Arrangement and design of digital elements/user interface	- software selection - UI design - implementation
DIM 4 - Lighting parameters	- illuminance - light intensity - lighting type - colour
DIM 5 - Climatic parameters	- temperature - humidity - particle concentration
DIM 6 - Acoustic parameters	- noise - music or acoustic information
DIM 7 - (Work-) Information Systems / Digital Assistance Systems	- information content - type/location - scope
DIM 8 - Physical assistance systems	- usage (YES/NO) - duration - strength - type of assistance
DIM 9 - Human-system interaction	- type - design - height of the default configuration - level of division of labour
DIM 10 - Work organisation	- work cycle length - working hours - work content - general labour division

6.4.1 Baseline technologies and tools

To achieve user adaptivity, a variety of technologies and tools can be employed. Table 3 highlights some configuration options for selected dimensions of user adaptivity [389]. Deploying these configuration options can have positive outcomes on both the physical and cognitive ergonomics of the workplace and/or workstation.

Table 3: Dimensions and options of the workstation user-adaptivity [389].

Dimensions	**Configuration options**
DIM 1 - Working desktop	- electric engines/drives - electro-pneumatic or electro-hydraulic cylinders and drives
DIM 2 - Arrangement of physical objects and provision of materials	- material carriers - flexible and collaborative robots - automated guided vehicle systems - automatic cable pulls/overhead conveyors
DIM 3 - Arrangement and design of digital elements /user interface	- dynamic projection systems/ Spatial Augmented Reality systems - individual user interfaces, designs on tablets/screens, etc. - individual elements when using data glasses/ AR systems, wearables - individual APPs for Smartphones, etc.
DIM 4 - Lighting parameters	- individual, digital lighting control systems and APPs - intelligent, smart LED panels, LED inserts - modern daylight supply systems - sensors for weather, day and night - adaptive blue light filters, grey light filters
DIM 6 - Acoustic parameters	- digital control of noise protection measures - self-driving noise protection systems on e.g. AGV - moving noise protection devices with robots - sound sensors at the workplace - automatic Noise Cancelling Headphones - digital control of acoustic information systems
DIM 7 - (Work-) Information Systems/ Digital Assistance Systems	- through individual information systems (Spatial AR, AR, wearables, screens) - individual information provision locations (directly at the workplace, central terminals, mobile with wearables, etc.) - including individual information content, granularity, depending on qualifications and experience)
DIM 8 - Physical assistance systems	- as decision assistance for use YES/NO/Optional - with individual strength of the assistance through automatic adjustment actuators - as automated just-in-time assistance system

6.4.1.1 Physical ergonomics

The field of physical ergonomics studies anatomical, anthropometric (body proportions and measurements), physiological (body functions) and biomechanical characteristics of physical activity with the aim to analyse and design work environments or products in relation to the body posture, materials handling, repetitive movements or action

forces to improve the safety and health of a worker [209]. As outlined in the first two dimensions in Table 3 above, this includes the design of the workplace and the arrangement of tools and materials, e.g. by adjusting the height of the desktop based on the anthropometry of the user to reduce stress on the body.

An adaptive workstation could fulfill such tasks automatically based on the user's body proportions and measurements using a set of actuators. Actuation is hereby defined as a process that converts energy to mechanical form based on a control input. An actuator can therefore be defined as a device that performs this conversion [71]. An actuator is characterised by the type of energy it receives and the mechanical movement it outputs. The energy receiving part of an actuator can e.g. draw energy from electrical, chemical or mechanical sources. These include pneumatic, hydraulic forms of energy and their combinations. An actuator most frequently outputs a linear or rotational mechanical movement, which can be used to e.g. change the height of a desktop, a workpiece, or the distance of the tools and materials needed by the worker. Combinations of linear and rotational actuators can be used to perform movements of higher complexity as well [297].

6.4.1.2 Cognitive ergonomics - User interface design

Complementary to physical ergonomics, applied cognitive ergonomics for system design consider the necessary mental processes related to task performance e.g. perception, memory, reasoning, or motor response and include the field of human-machine interaction [209]. The user interacts with the workstation using an interface, i.e. a device with multiple components (e.g. buttons, light indicators, displays) that enables communication between the human operator and the system [175]. Interfaces that are appropriately designed according to principles of cognitive ergonomics allow users to be aware of the respective automation modes and the current as well as future states of the system and to act accordingly [175]. In other words, interfaces provide meaningful support in terms of context-relevant guidance for understanding and action [64].

From the perspective of cognitive ergonomics, the User Centered Design (UCD) [245] and the Ecological Interface Design (EID) [64] paradigms are suitable for the design of visual information (e.g., screen characteristics and content – DIM3) and acoustic information (auditory content – DIM6). Both UCD and EID emphasize on the user at every stage of the design process but differ in their focus for interface design. The concept of UCD assumes that user information should be given considerable attention at every stage of the design process. It improves product performance in terms of how users can, want, or need to use a product, rather than forcing a user to change his or her behaviour to accommodate the product. UCD focuses on the interaction of single user with the system on the task level and on the end user's requirements, needs, and constraints for individual actions for the purpose of interface design [485]. However, the classical UCD approach does consider the overall contextual conditions of work for the display of information and for shaping the interaction process. The EID approach intends to close this gap for interface design. It has been specifically developed for human-computer interaction in complex real-time systems and has been applied in various contexts such as process control, aviation, network management, and tactical systems in command and control operations [460]. The focus of analysis of EID lays on the overall work domain and considers the relationships of the respective system components in the interface design process [65]. The approach offers the possibility to analyse complex socio-technical

systems in a user-oriented way and accommodate problem-solving and decision making behaviour though the appropriate display of interdependent system processes and constraints [64, 65]. Interface design is based on a so called abstraction hierarchy that represents a functional decomposition of the work domain (i.e. the overall work environment in question) in functional (goals and purposes, salient rules) as well as physical (tools, components) parts [485, 65]. EID primarily aims to make relationships of system components of the work domain and respective constraints explicitly accessible to the user and to ensure maintenance of control as well as guide appropriate action.

6.4.2 Discussion

Several of the adaptation processes described by Rupprecht and Schlund [389] can be implemented in the concept of the *FELICE* adaptive workstation. The physical adaptation processes (related to DIM 1) can be divided into user-specific adaptation processes and task-specific adaptation processes. During user-specific adaptation, the optimal configuration of the workpiece is derived from the user's anthropometric features. Additional task-specific adaptation processes adapt the height and inclination of the workpiece based on the specific challenges of assembly sub-tasks to further reduce physical strain. As each physical adaptation process takes time, a compromise between the ideal ergonomic configuration and the productivity must be reached. Ideally, physical adaptation processes will be triggered automatically without user intervention. However, IAD research has shown that providing the user with the option to readjust the adapted AWS parameters manually is important for the acceptance of the system [391]. User readjustment could potentially be implemented via the user interface or speech commands. The physical adaptation of the second dimension in the FELICE system, namely the adaptation of physical objects and materials, is mainly achieved via the mobile cobot.

With respect to the cognitive design of the *FELICE* adaptive workstation, the user interface design can adopt the principles of UCD for the part that corresponds to the specific task demands and user preferences regarding performance on the workpiece. The respective information can be derived by the planned methods of focus groups and interviews with end-users (see also section 8) and can be implemented with respect to information related to DIM 3 (visual accessibility/ visual display of information) and DIM 6 (acoustic signals with a semantic component). However, it is expected that the assignment of the cobot to the workstation will predominately change the nature of the assembly task towards teamwork and joint activity principles. Furthermore, the integration of additional, higher-level and tightly interdependent components that will co-define the workflow (e.g. orchestrator) is expected to increase the complexity of the work domain and the information needed by the user. The two approaches (UCD and EID) are not mutually exclusive but should rather be viewed as complementary and be combined (c.f. [382]). Thus, it is reasonable to also include EID principles in the considerations for the interface design of the adaptive workstation (also in potential relation to DIM 7). In this manner, the end-user will be "in the (information) loop" regarding the overall work domain and potential interactions of system components that are beyond the narrow scope of the specific assembly task of the individual user but, if not attended, can affect performance in unanticipated ways.

7 Human robot communication

7.1 Overview

In this section we will focus on the definition of technologies and tools to be employed for the communication between the human and the robot through a command-based interaction. In particular, we will explore two main directions, based on two complementary sources of information, namely based on gesture recognition by visual analysis and voice command detection and recognition by audio analysis. Note that we will talk in this section about the robot, but the same type of interaction also applies to the adaptive workstation.

7.2 Relationship with FELICE project

An important contribution to the Human-Robot Collaboration (HRC) framework includes user friendly human machine-interfaces, and the smart integration of sensors and devices for natural human-robot communication in industrial settings to facilitate fluid and safe collaboration. Bidirectional interfaces are necessary to allow workers to control the robot or other parts of the adaptive workstation. Interpreting voice commands supports hand-free interaction, though is challenging in a noisy industrial environment and can be further augmented by gesture-based analysis to resolve potential ambiguities while supporting natural communication. Still, for discriminating between work-related intentional and non-intentional human gestures, we will assume a constrained set of gestures relying on existing toolboxes. The complementarity of the different streams of information will aim to resolve potential ambiguities in capturing input information from the worker, but also to interact with the worker even if the hands are not visible to the camera. Figure 7 summarises the different methods for information input and output we will explore in the FELICE project.

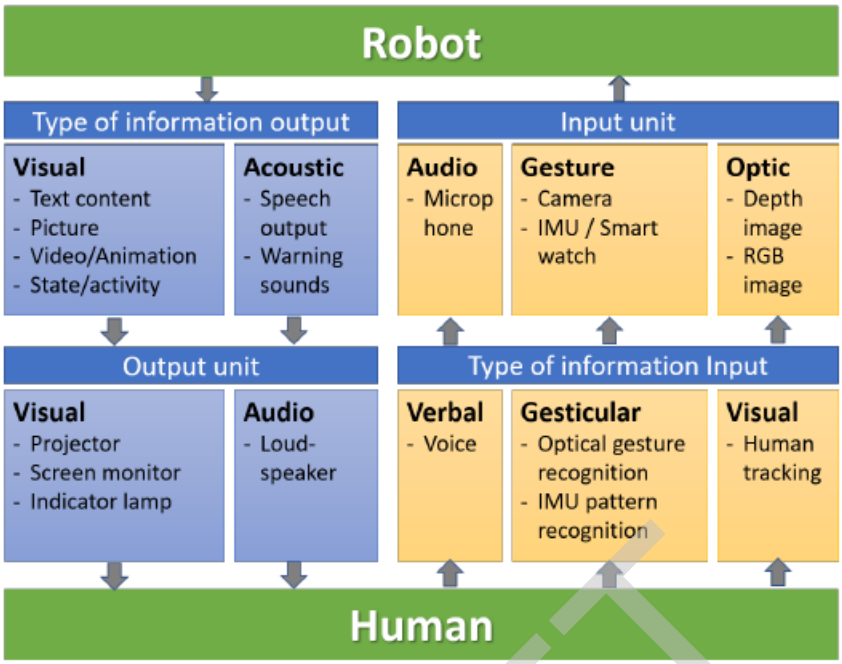


Figure 7: Methods and units of information I/O.

In particular, within this section, we will focus on the state of the art, tools and technologies available for the verbal (through voice) and gesture-based (through visual gesture recognition) communications. Thus, the sensors we will consider are the microphone (for audio analysis) and the camera (for gesture analysis).

7.3 Speech-command interaction

Nowadays, Speech-Command interaction has become the main feature of many industrial products, as Amazon Alexa and voice assistants in general. A speech-command interaction is mainly composed of two AI modules (see Figure 8): Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU). In the first case, ASR is the process of translating or transcribing an audio signal into a written text. The NLU module is responsible for obtaining a semantic interpretation from the (previously) transcribed text, i.e. understand the interlocutor's intents and the involved entities.

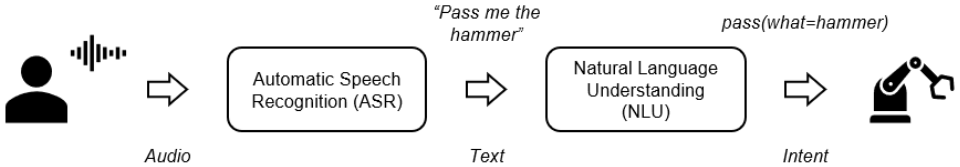


Figure 8: A common pipeline for speech-command interaction. It contains an ASR module to transcribe voice signals into text and an NLU model to understand the related semantics.

Recently, there has been a strong push in both the academic and industrial world to transition from hybrid ASR models to end-to-end (E2E) models. In particular, there

are three promising E2E methods: recurrent neural network transducer (RNN-T), RNN attention-based encoder-decoder (AED), and Transformer-AED [265]. All these models have an acoustic encoder, that generates a high-level representation of the voice audio signal, and a decoder, which is responsible for generating tokens (characters, phonemes, words) in the linguistic domain. The main difference between the discussed E2E models consists of the decoder. The RNN-T independently predicts the next label based, on one hand, on the previously predicted label (language model) and, on the other hand, on the current acoustic representation (acoustic model). The two models are pre-trained independently and then fine-tuned in combination with a *joint network* that combines the two outputs to predict the final label. In AED models, instead, the language model is responsible for predicting the final label based on the previous one and the attention-filtered acoustic features. Finally, RNN-AED and Transformer-AED differ at the realization of encoder and decoder using LSTM-RNN and Transformer, respectively. ASR models can be also categorized in streaming and off-line systems based on the fact that the prediction is synchronized respectively with the audio frame or the sentence.

The authors of [265] perform a large-scale benchmark of these methods, in both streaming and off-line modalities, over 13 test sets containing 1.8 M words. They report that the Transformer-AED architecture outperforms others obtaining a Word Error Rate of 9.16 and 7.83 respectively in the streaming and off-line modalities.

Once the user utterance is processed by the ASR module, the transcribed text is given as input to the Natural Language Understanding component which has to interpret the user's intent with a number of slot-value pairs that need to be filled to accomplish that intent. Slots are defined for each intent: for instance, to bring a tool to the worker, the robot needs to know which is the desired tool. According to [278], two neural-based paradigms can be identified in the literature: *independent models* and *joint models*.

Independent models adopt the approach of training a different model for each task. These architectures are commonly composed of different layers, namely an input layer, one or more encoder layers and an output layer. The input layer is responsible for projecting each word in an embedding space. The embedding function can be either pre-trained or trained from scratch. Regarding the encoder layer, it is common to use a RNN architecture that can be combined with a bi-directional encoding to improve the performance on both tasks. Finally, the output layer is responsible for predicting the class of each word (slot-filling, for example the desired tool) or the class related to the overall sentence (intent classification, for example bring something). The layer is usually implemented with a softmax layer applied respectively on all the hidden-states and on the last hidden-state of the encoder layer for the slot-filling and intent classification tasks respectively. In [278] it is highlighted how the use of long-term RNN, like LSTM and GRU, increases the performance for both the tasks on the standard dataset ATIS. Furthermore, incorporating more context information using attention mechanisms or sentence level representations, further improves the performance for the slot-filling task.

Contrary to independent models, joint models try to exploit the relationships existing between the slot-filling and intent classification tasks. Two categories of joint models can be identified: *parameters and state sharing* and *gating mechanism*. Parameters and state sharing methods, as suggested by their name, share information about the two tasks through either weights or state. In the first case, a shared encoder, usually a Bidirectional RNN (Bi-RNN), is trained by using a multi-objective loss function to obtain

a shared representation. Then, to produce the prediction, several approaches have been proposed, for instance, encoder-decoder methods with a different decoder for each task and attention-based seq-to-seq models. More recently, with the large and successfully adoption of transformers in NLP tasks, the scientific community started to fine-tune pre-trained large-scale language models (e.g. BERT) for the two tasks. State sharing methods differ from weights sharing, in that they use two recurrent models that share states. It is worth mentioning that this approach still requires two models to be (jointly) trained. Finally, gating methods explicitly model the dependency between slots values and the detected intent through the use of a gating layer, function of a *slot context vector* (different for each word step) and a *global intent context vector*.

In [278] the authors report how joint models are competitive with independent models with half of the parameters (weights-sharing approach) on reference datasets ATIS and SNIPS. Moreover, in presence of enough computational power, fine-tuning a pre-trained model such as BERT allows to maximize performance of both slot-filling and intent recognition tasks. Finally, hybrid methods, combining parameter and state sharing and intent gating, obtain the best performance with an F1-Score respectively of 98.75% and 98.78% on the two datasets for the slot-filling problem and an accuracy respectively of 99.76% and 98.96%.

In addition to the Speech-Command interaction pipeline, Sound Source Localization (SSL) is an essential piece in the overall robot scheme. It allows to estimate the position of the speaker and, based on that, suppress sounds from a different direction (i.e. environmental and robot noise). Moreover, the SSL module allows the robot to rotate itself in the direction of the speaker allowing eventually the recognition of visual commands.

The first requirement of the Direction of Arrival (DOA) estimation is to use multiple microphones (i.e. a microphone array). The localization process can be divided in two steps, namely *Feature Extraction* and *Feature-to-Location Mapping* [379]. Commonly used features are the Time difference of arrival (TDOA), i.e. the time-difference between two captured signals, and the inter-microphone intensity difference (IID), i.e. the difference of energy between two signals at a given time. Once the features from the audio signals are extracted, a propagation model is used to identify a mapping function between the features and the sound source position. The most popular propagation model used is the free-field/far-field model which assumes a single direct path between the source and a sound wave that can be considered planar. This model is usually used together with the *generalized cross-correlation with phase transform algorithm (GCC-PHAT)* algorithm which is responsible for estimating the TDOA. GCC-PHAT is robust against reverberation and interfering sources in high signal-to-interference ratio circumstances [379]. GCC-PHAT achieved a Mean Absolute Error of 1.0 degree within the first track of the LOCATA (SSL) Challenge[143]. The feature-to-location mapping step can be also approached with machine learning methods. In this case, a propagation model is not required, instead, it can be learned directly from data; nevertheless, it suffers from generalization issues related to data-driven algorithms. Furthermore, a learning-based model is trained on a particular array geometry and, therefore, it cannot be directly used with a different microphone array.

7.3.1 Baseline technologies and tools

In this section the available technologies and tools publicly available are reported divided by task, i.e. ASR, NLU and SSL.

In Table 4 several ASR technologies have been identified for our languages of interest, i.e. English (en) and Italian (it). Cloud-based systems are firstly identified; no information about the models implemented in these services is publicly available and all of them require an active Internet connection. Almost all the services are available in both the streaming and offline modalities. Then, libraries which do not require the Internet connection are reported. Among these libraries, SpeechBrain, Nvidia NEMO and Vosk API support more recent architectures based on the attention mechanism and Transformers architectures. Nvidia NEMO also includes pre-trained models optimized for embedded systems. Finally, cloud-based services and Vosk API support out-of-vocabulary words; this feature is crucial for application-based vocabulary to recognize.

Table 4: Publicly available ASR libraries and Cloud APIs for English and Italian.

Tool	Internet	Free	Stream / Offline	Available models
Google Speech Recognition API	Yes	Yes	Offline	en , it
Google Cloud Speech API [a]	Yes	No	Both	en , it
Microsoft Azure Speech [b]	Yes	No	Both	en , it
IBM Speech to Text [c]	Yes	No	Both	en , it
CMU Sphinx [d]	No	Yes	Offline	en , it
Mozilla Deep Speech [e]	No	Yes	Offline	en , it
SpeechBrain [f]	No	Yes	Offline	en , it
Nvidia NEMO[g]	No	Yes	Both	en , it
Vosk API[h]	No	Yes	Both	en , it

[a] https://cloud.google.com/speech-to-text
[b] https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/
[c] https://www.ibm.com/cloud/watson-speech-to-text
[d] https://cmusphinx.github.io/
[e] https://deepspeech.readthedocs.io/en/latest/DeepSpeech.html
[f] https://speechbrain.github.io/
[g] https://developer.nvidia.com/nvidia-nemo
[h] https://alphacephei.com/vosk/

As stated for ASR, also in the case of NLU services no information is available about cloud-based frameworks like Microsoft LUIS, DialogFlow and Amazon Alexa. These services require an Internet connection and work with both the languages of interest. Regarding open-source solutions, the Snips NLU library and the RASA framework contain pre-trained models for both the languages; on the other hand, DeepPavlov and Nvidia NEMO need to train new models to handle italian sentences. The RASA framework allows to integrate the HuggingFace and Spacy libraries which contain several state-of-the-art NLP models, like BERT and GPT2. Finally, the RASA framework supports conversation recording for continuous learning of NLU models.

Table 5: Publicly available NLU libraries and Cloud APIs.

Tool	Internet	Free	Available models
Microsoft LUIS [a]	Yes	No	en , it
DialogFlow[b]	Yes	No	en , it
Amazon Alexa [c]	Yes	No	en , it
Snips NLU [d]	No	Yes	en , it
RASA [e]	No	Yes	en , it
DeepPavlov [f]	No	Yes	en
Nvidia NEMO [g]	No	Yes	en

[a]https://www.luis.ai/
[b]https://cloud.google.com/dialogflow
[c]https://developer.amazon.com/alexa
[d]https://github.com/snipsco/snips-nlu
[e]https://rasa.com/
[f]https://deeppavlov.ai/
[g]https://developer.nvidia.com/nvidia-nemo

Table 6 reports two open source libraries for SSL. Both of them contain the implementation of the GCC-PHAT algorithm. The BTK 2.0 library also contains several algorithms for noise suppression and acoustic echo cancellation. On the other hand, the ODAS library is optimized for embedded systems.

Table 6: Publicly available SSL libraries.

Tool	DOA	Noise Suppression
BTK 2.0 [a]	GCC-PHAT	Yes
ODAS [b]	GCC-PHAT	No

[a]https://distantspeechrecognition.sourceforge.io/btk20_documentation/user_docs/index.html
[b]https://github.com/introlab/odas

It is worth reporting that an IoT microphone device optimized for speech acquisition, namely ReSpeaker Mic Array V2.0 from SeedStudio[4] is available on the market. The microphone array executes on board the DOA algorithm combined with Beamforming, Noise Suppression and Acoustic Echo Cancellation algorithms. In this way, it outputs a single-channel signal optimized for ASR systems. The microphone allows to capture and identify also far-field voice activities.

7.3.2 Discussion

Based on what has been reported so far, in this section, the most promising methods and frameworks are identified.

[4]https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/

Regarding the ASR module, the main candidate solution is represented by the Nvidia NEMO framework which contains state-of-the-art models optimized from both the points of view of accuracy and speed. An alternative solution is represented by cloud-services that comes with two main advantages: (i) the speech recognition models are continuously updated and, therefore, their accuracy improves over time; (ii) using a cloud service the computational load of on-robot embedded system is alleviated. The latter is not a negligible feature when several deep learning methods are used to perform complex tasks. The main drawback of this choice is represented by the live Internet requirement and the related additive latency. In the case that out-of-vocabulary words have to be recognized and an Internet connection is not available, the Vosk API is the library that satisfies these requirements. Finally, offline approaches are preferred with respect to streaming ones due to the particular application, which requires short sentences (commands) and high accuracy.

Differently from the ASR module, there are not fully pre-trained models for customized speech command recognition. Trivially, this is due to the fact that the commands to understand have not been defined yet. For this reason, it is important to adopt a fine-tuning strategy from pre-trained state-of-the-art methods and continuously update the obtained model to improve performance and increase robustness. Defined these requirements, the most suitable tool seems to be RASA that satisfies both.

Finally, considering what has been said for ASR and computational power of the embedded system, in the case of SSL a promising solution is the adoption of the ReSpeaker Mic Array V2.0 microphone. In fact, it allows to identify the speech direction of arrival (even when the speaker is far from the microphone) and to suppress background noise (environmental and ego-noise), while decentralizing the computation. Moreover, the identified hardware solution is optimized for speech signal, differently from off-the-shelf libraries which are general-purpose.

7.4 Gesture recognition

Gesture is a form of non-verbal communication that involves the use of one or more parts of the body (in most cases the hands and the head) and can be used both alone and together with speech. After speech, the gesture is the most used form of communication. Moreover, it allows communication at high distance and also in presence of a noisy environment. This last aspect is very important especially in an industrial environment, in which there are many different sources of noise that disturb verbal communication and are difficult to eliminate. For this reason, the gesture may assume a key role in HRC in the industrial field, and in this project as well.

Even if a definitive categorization is not possible, a possible taxonomy for gesture recognition algorithms takes into consideration the sensors we want to use, identifying *contact recognition* methods (based on wearable sensors) and *vision-based gesture recognition* methods (based on RGB or depth cameras). The second possible taxonomy takes into consideration the task to be performed, identifying *isolated gesture recognition* and *continuous gesture recognition*. In the first case (isolated gesture), we have a classification task in which we have already identified starting and ending point of a given gesture; in the second case (continuous gesture) we want to carry out both the detection, to identify starting and ending point, and the classification of the specific gesture. It is important to highlight that most of the methods in the literature focus on the problem of isolated gesture recognition, which is definitively more simple, and most of the dataset available are collected for this specific task. Anyway, as evident, within the project we have to deal with both detection and classification of the gesture, since the input of the module will be the video stream acquired by a camera sensor. Thus, the starting and ending point of the gestures can not be a priori known. Also, in this section we will only focus on vision-based gesture recognition methods, considering as input data for the system the sequence of images acquired through a vision device.

Gesture recognition is a very complex task and, especially when working in the wild, different challenges need to be addressed:

- Encoding temporal information. Temporal information is essential since most of the gestures are dynamic, and temporal information can radically change the meaning of a set of frames (e.g. the action of clockwise or anti-clockwise rotation provide the same set of frames, but arranged in different temporal order and then the meaning is different).

- Small and no-specific training data. Models capable of performing this task (with temporal analysis) typically require large amounts of data not available in the literature and which hardly contain the required set of gestures. Also, most of the datasets are for isolated gesture recognition. The most widely used datasets are shown in Table 7; among them, the most interesting for our purpose could be 20BN-Jester [299] and ChaLearn LAP ConGD [468]. The first contains hand gestures for human-machine interaction, such as alt, thumb up, swiping left and right, sliding two fingers down and up, clockwise and anti-clock wise rotation, and so on. This dataset has no time localization so it can only be used for isolated gesture recognition. The second dataset can be used for continuous gesture recognition but it is about 4 times smaller.

Table 7: Main gesture recognition datasets.

Dataset	Type	Years	Instances	Videos	Instance /video	Classes	Subjects	Scenes	View	Modalitites	Resolution	FPS
ChaLearn LAP IsoGD	Iso	2016	47,933	47,933	1.0	249	21	15	3rd	RGB-D	320 × 240	8.42
NVGesture	Iso	2016	1,532	1,532	1.0	25	20	1	3rd	RGB-D	320 × 240	30
20BN-JESTER	Iso	2019	148,092	148,092	1.0	27	1,376	1,376	3rd	RGB	- × 100	12
ChaLearn LAP ConGD	Con	2016	47,933	22,535	2.1	249	21	15	3rd	RGB-D	320 × 240	8.42
Montalbano V2	Con	2014	13,858	13,858	1.0	20	27	-	3rd	RGB-D	640 × 480	20
IPN Hand	Con	2021	4,218	200	21.1	13	50	28	3rd	RGB	640 × 480	30
EgoGesture	Con	2017	24,161	2,081	11.6	83	50	-	1st	RGB-D	640 × 480	30

- Viewpoint variation and occlusion. All the datasets present in the literature are acquired with subjects positioned in front of the camera and without the presence of occlusions, which makes the systems sensitive to such occurrences. This is a very important and not negligible aspect, since the hand doing the gesture need to be entirely visible from the camera.

- Execution rate variation and repetition. The speed of execution of the gesture depends a lot on the age and emotional state of the person and these features are difficult to extrapolate. It is also important to highlight that all the available datasets, with the only exception of 20BN-Jester, do not foresee variability in the age groups.

- Online motion recognition and prediction. There are far fewer works in the literature that address the continuous gesture recognition problem and those available use very big and heavy models, that do not allow online operation, and then real time operation over some embedded platforms.

- Simultaneous exploitation of spatio-temporal-structural information. These three pieces of information are essential and therefore it is necessary to understand how and when to merge them together. The main problem is that to take advantage of all three components, we have to build even more complex and heavy models.

- Embedded Vision. Since the system must work on embedded devices, the models we have to use for achieving this task must be lightweight.

Even if there is not a standard taxonomy, we can identify 6 main phases in vision-based gesture recognition systems, as shown in Figure 9: (i) the raw data are acquired through vision devices (such as RGB or Depth video cameras); (ii) the raw data are pre-processed obtaining more complex data such as optical flow, skeleton data, segmentation masks (for hands), etc. (iii) extraction of spatial (detection of the parts of interest) and temporal (tracking of the part of interest) features. Typically, spatial information is extracted through CNN, while temporal information can be extracted at multiple levels, such as at feature level through 3D-CNN or RNN, or decision level using 2D-CNN and fusion strategies (more details will be provided in the following); (iv) detection of beginning and end of each gesture in the video. Typically it is treated as a classification task in which each frame is classified as “gesture” or “no-gesture”. This phase can be also associated with the next; (v) classification of the gesture, with one or

more frames, according to the predefined classes; (vi) mapping of the detected gesture in a command to the robot.

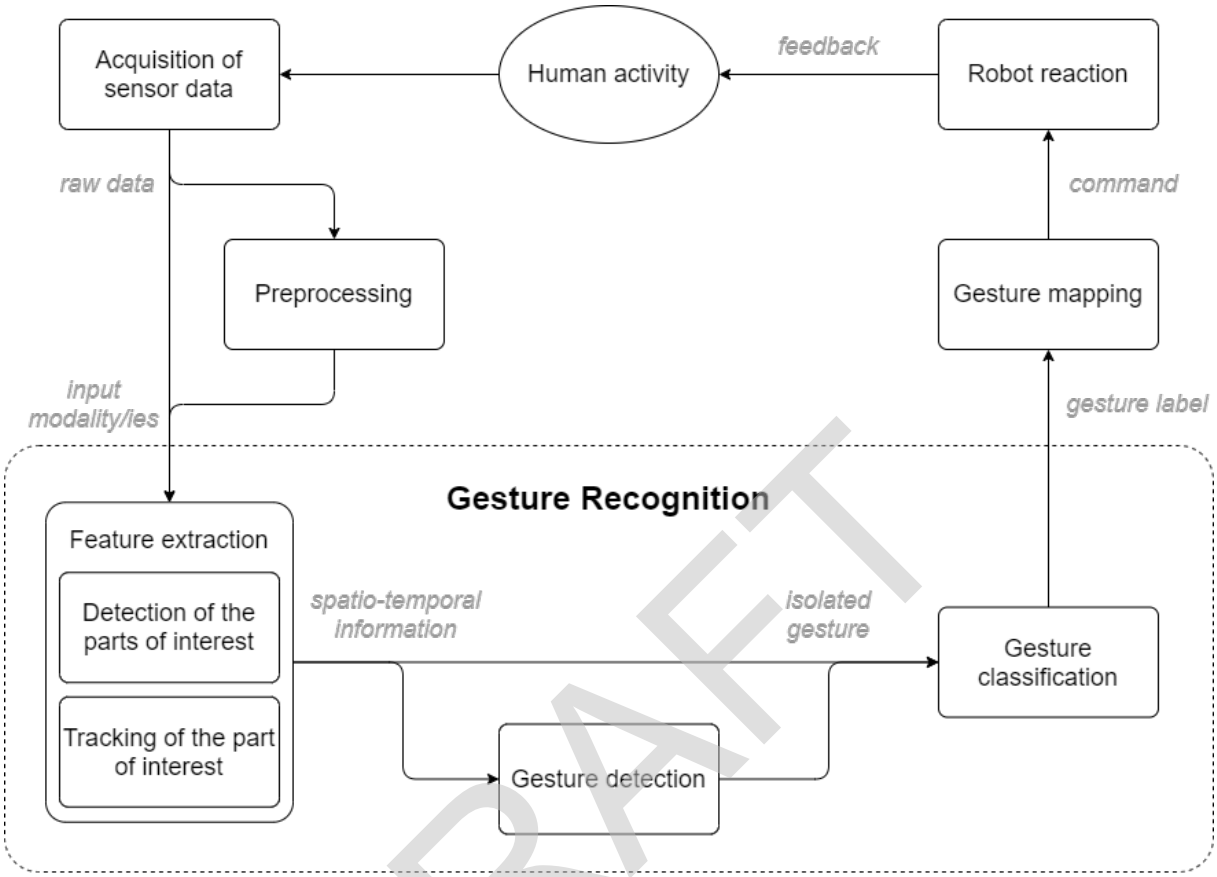
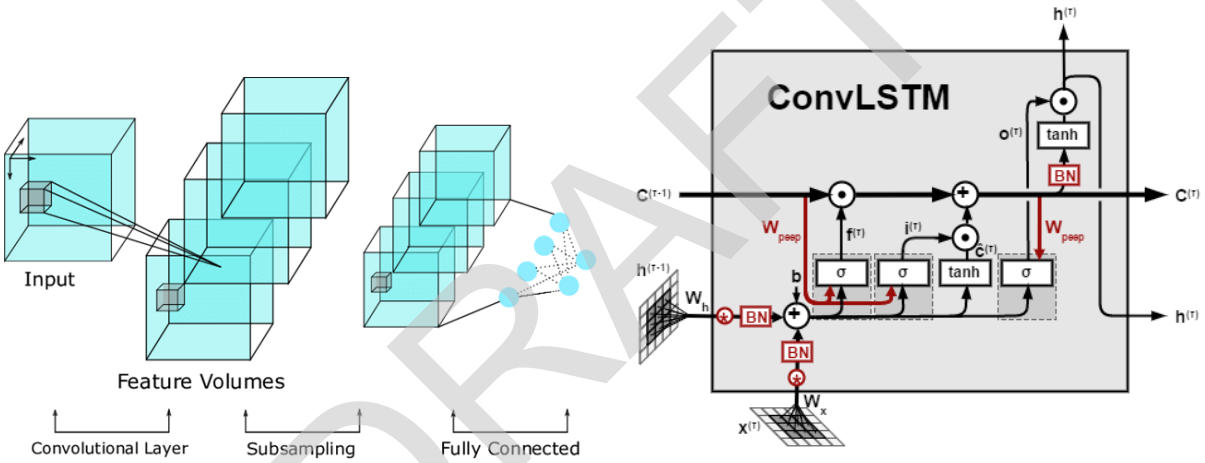


Figure 9: Gesture recognition system workflow.

As we can see from Figure 9, the Gesture Recognition module takes one or more modalities as input. The modality represents the data that is used in input to the module, which can work with one or more of them. The most used modalities are: RGB data, depth data, skeleton data and dense optical flow. The first two modalities use raw data, acquired from traditional camera sensors and range sensors, respectively. Viceversa, the last two modalities are obtained using algorithms and offer more elaborate information by introducing a computational cost. Typically these modalities are not used individually, but together and this implies the necessity to combine the information of the different modalities. The fusion can take place at one or more levels of the following: data, features, or decision level.

The other steps of the system are rarely divided, in most cases some of them are grouped together by subsystems that carry them out simultaneously. The detection and the tracking of the parts of interest are the main phases of the system since they are used to capture, respectively, the spatial and temporal information. Both information will then be used by the next steps. The extraction of spatial information is carried out either by using skeleton data, and therefore algorithms capable of extrapolating the information of specific points in the image [312]; or by exploiting the Convolutional Neural Networks (CNNs) by giving them the stream video frame by frame (the stream can come from RGB, Depth or Optical flow mode) [324]. The most used networks

are the classic 2D-CNN used for tasks such as Object Recognition. The extraction of temporal information is typically carried out using Recurrent Neural Networks (RNNs) and, in particular, the Long Short-Term Memory (LSTM) networks, since they alleviate some problems such as explosion or vanish of the gradient. Although these networks are used for gesture recognition [312], they capture temporal information by losing the previously obtained spatial information. To overcome this problem, two other architectures have been introduced: 3D-CNN and ConvLSTM. 3D-CNNs (see Figure 10a) use three-dimensional kernels in convolutional and pooling levels to consider both spatial and temporal dimensions [496, 477]. The introduction of the third dimension exponentially increases the number of parameters and this causes a trade-off between the spatial and temporal dimension which leads to the ability of networks to extrapolate only short-term temporal information, in addition to the spatial one. ConvLSTM (see Figure 10b) cells modify the classic structure of LSTM cells by introducing convolution both in the input gate and in the forget gate, thus preserving the spatial information allowing the cells to analyze the temporal information in specific regions [508, 509].



(a) 3D-CNN (b) ConvLSTM cell

Figure 10: Architectures capable of capturing spatio-temporal information.

The phases of gesture detection and classification take as input the embeddings extracted in the two previous phases and can be carried out both by a single network with a one stage approach, and by two different networks with a two stage approach. Systems that use the one stage approach perform a double classification, i.e. they classify every instant of the input signal both as a gesture or non-gesture, and by assigning it to a specific class. To do this, a time window is generally applied that flows over the entire video and the classification is made on the instant on which the window is centered based on the instants included in it. In this way, the system outputs vectors of belonging probability which are fused through an appropriate consensus policy, typically applying the average, or selecting the maximum value. This approach can be used both by applying 3D-CNN on RGB video streams, depth, and optical flow [85, 364, 315, 325], and by applying LSTM on skeleton data [469]. Systems using the two stage approach [192, 509] perform video segmentation and classification at different times. This division is very important since a myriad of approaches are available in the literature to carry out classification only. Therefore, it is possible to concentrate only on the devel-

opment of the gesture detection network and use one of the approaches available for classification obtaining slightly better results than the one stage technique. Obviously, at the price of a slightly longer computation time.
The phases of gesture detection and classification are never done by a separate network, but a classifier is mounted on the embeddings extraction networks. The most used classifier is an MLP layer with softmax activation function [98, 312, 508], but Sparse Fusion Network [325], SVM [497] and clustering techniques [477] are also used.

7.4.1 Baseline technologies and tools

The literature has shown that the skeleton and optical flow modalities are widely used, for this reason many libraries have been developed to simplify and optimize their calculation. The main libraries to obtain skeleton data are:

- *Microsoft Kinect SDK*. The Kinect for Windows Software Development Kit (SDK) 2.0 enables developers to create applications that support gesture and voice recognition, using Kinect sensor technology on computers running Windows 8, Windows 8.1, and Windows Embedded Standard 8 [311]. This library can extract the body skeleton or return the hand position and its pose between: open, close, lasso, unknow. The lasso hand state is a closed hand with the middle and index fingers both up, it is like a pointer hand, but with both the index and middle finger.
- *OpenNI*. The Open Natural Interaction (OpenNI) is a multi-language, cross-platform framework that defines APIs for writing applications utilizing Natural Interaction [369]. The APIs provide support for: voice and voice command recognition, hand gestures, and body motion tracking. The library can extract skeleton only for the whole body, while for the hand it return only one point.
- *OpenPose* [419]. It is a real-time multi-person system to jointly detect human body, hand, facial, and foot keypoints available in C++ and Python. The library has 4 different detectors: body, foot, face, and hand. Each of them can be enabled or disabled.

About the dense optical flow there are two main libraries that offer the code for its calculation: OpenCV [317] and NVIDIA Optical Flow SDK [333].
The most used architectures for gesture recognition are based on 3D-CNN and LSTM. The first type is more common, in fact publicly available pre-trained models have been developed:

- *Video-Caffe: Caffe with C3D implementation and video reader* [336]. There is one of the first 3D-CNNs that was developed, called C3D, pre-trained on UCF-101 dataset [423] (actions dataset).
- *Efficient-3DCNNs* [243]. The most used 2D-CNN architectures have been adapted by introducing convolutional 3D kernels to analyze both spatial and temporal information. The adapted networks are: ShuffleNet (v1 and v2), MobileNet (v1 and v2), SqueezeNet, ResNet (18, 50, and 101), and ResNeXT. The datasets on which they have been pre-trained are: Kinetics-600 [91] (actions dataset), Jester [299] (gestures dataset), and UCF-101 [423] (actions dataset).

7.4.2 Discussion

After a careful analysis of the relevant literature, we concluded that the current state of the art leans towards the use of a multi-stream network architecture that takes input at least the RGB and depth modalities. If performance needs to be improved, the dense optical flow, calculated on both streams, could also be used as an additional input.

Given the complexity of the task and the environment in which it will be performed, we first aim at recognizing static gestures and, then, dynamic gestures. This first step is necessary because in literature the various systems are tested in indoor environments with the performer of the gesture placed in front of the camera and without other moving subjects in the background, and this kind of scenario is as evident definitively less challenging than the industrial one.

The dataset to be used will be defined once the gesture will be definitively chosen. According to the analysis, the two most interesting datasets that could be used to pre-train systems are: (i) 20BN-JESTER, which is the largest dataset available in the literature for Isolated Gesture Recognition and offers a set of gestures that can be used for Human-Machine interaction; (ii) ChaLearn LAP ConGD as it has a much larger set of gestures (even if more general) and can be used for Continuous Gesture Recognition. Obviously, given the complexity of the industrial environment, it will be necessary to acquire a dataset to carry out the final training and performance assessment of the system.

8 Cognitive ergonomics for enhanced human-robot dyads

8.1 Overview

This chapter describes basic principles of cognitive ergonomics that support human information processing and effective interaction between humans and robots. Relevant tools and methods will be outlined and discussed within the context of the *FELICE* project.

8.2 Relationship with FELICE project

Our goal in *FELICE* is not to design autonomous robots with human characteristics, but rather an intelligent system based on the principles of joint activity: a dynamic configuration of a mutually informing, monitoring, predictable and directable human-robot dyad that accommodates/enables adaptive interaction and supports joint goal-driven performance and system resilience. Following this premise, the planned artificial cognitive agent should be able to engage in collaboration by means of awareness and representation of the specific characteristics and states of the human users as well as through mirroring the sociocognitive characteristics of goal-driven interaction. In cognitive ergonomics, models from cognitive psychology are exploited to model cognitive processes of perception, attention, memory, decision making, action preparation and motor coordination as well as socio-cognitive aspects of interaction activity. In addition, theoretical approaches from industrial psychology are used to model and assess mental stress and strain levels and drive the design of human-robot collaboration based on cognitive ergonomic principles that support human information processing and effective interaction. Many of the models also have formal and algorithmic approaches and can be used to model and represent the human counterpart on the cognitive robot side.

8.3 Cognitive ergonomics

In modern working environments, effective task performance strongly relies on cognitive functions of the human operator, i.e., the cognitive processes related to perception, attention, working memory, decision-making, action preparation, and execution. The introduction of technological innovations and automated systems at work contributes to more efficient work processes, productivity, and system reliability. However, the implementation of new technologies also introduces changes in work practices, tools, and processes as well as skills and task-related information that are necessary for task performance. Although technological innovations at work are mostly implemented with the objective of workload reduction and thus optimized system reliability, they can negatively interfere with existing expertise and skills and pose additional and even novel cognitive demands for the human worker, if they fail to consider the respective aspects and characteristics in their design.

This is even more prominent with the integration of intelligent artificial systems in the working context. Aspects such as optimal information flow (in terms of timing and amount of necessary information), information display modes as well as knowledge

about the characteristics and functions of artificial cognitive agents as well as the overall transparency of intelligent systems with respect to shared work processes can create an array of new problems. Particularly, comprehensibility, predictability, interruptions, and information overload can impair task performance resulting in decreased task accuracy [212, 116, 381]. It is therefore essential to consider user-centered requirements for innovative design of work that go beyond classical bioergonomics in order to reduce the negative consequences of such cognitive load and improve working conditions, especially during interaction with new technical systems. In addition to the latter, as technology advances at a fast pace and autonomous artificial agents increasingly share the working space with workers and participate in a rather dynamic task allocation with the human element, it is important to include specific aspects of human teamwork into the considerations and the design of the respective interaction between human and technology [321].

Effective teamwork is defined by the demonstration of core emergent states and process, such as mutual performance monitoring, back-up behaviour, and adaptability/resilience between the team members and influencing factors such as shared mental models, communication, and collective orientation/mutual trust [395, 82]. These aspects also accommodate the principles of joint activity between interacting agents, such as common ground, predictability, and directability [236]. Although the integration of teamwork principles in human-technology collaboration is still in its infancy, considerations of ways to operationalize these provide a human-centered conceptual basis for an implementation of dynamic collaboration and adaptive team performance between human and artificial agents [321, 82].

Designing according to principles of cognitive ergonomics serves these purposes. Cognitive ergonomics is a sub-discipline of ergonomics that is primarily concerned with the performance and resilience of human information processing when dealing with machines. In particular, cognitive ergonomics seeks to understand basic cognitive processes during interaction with a technical system and aims to optimize the overall human-technology system [396]. In this context, optimization means enhancing positive effects and preventing detrimental effects for the human, respectively. Hence, the cognitive functioning of the human worker displays the main factor in cognitive ergonomics and in designing respective work systems and processes. Furthermore, methods of cognitive ergonomics provide guidance and allow for a pre-emptive layout of requirements and challenges for dynamic joint task performance of humans and technology.

8.3.1 Baseline technologies and tools

As mentioned above, human information processing represents the central component of the system to be designed and the respective work processes. Additional factors such as motivation, emotion, and age are also taken into account as they have been shown to modulate the influence on information processing. In order to investigate these factors, cognitive ergonomics relies on findings and methods from cognitive psychology, work psychology, and ergonomics. Particularly, Human Factors methods are of special interest for the *FELICE* project, as they describe a wide range of methods for human interaction with technical devices as well as for designing and evaluating systems [427]. The

following section will outline several methods and techniques that may be particularly suitable in relation to the underlying project.

8.3.1.1 Data collection methods The design of novel systems requires the collection of information regarding the system, the activities as well as information regarding the people that will interact with the system. The following methods are commonly used in this context:

- **Observations**. Observational methods comprise the observation of an individual or group of people when performing a specific work activity. They have to be carefully planned and executed [427] using video and audio software. Importantly, the gathered behavioral data allow for information about movement sequences, posture, and facial expression.
- **Interviews**. Interviews allow for the collection of a wide range of information regarding a specific subject. They provide information on user perceptions, system usability, work requirements, cognitive task analysis (see section 8.3.1.2), and errors [427].

  Moreover, they can be used to determine general relationships between cognitive load and mental stress in the field and laboratory. A special form of the interview is represented by group discussions in the form of focus groups. This interview type offers the opportunity to obtain opinions from a small group of people (approximately 5-10 people) with similar backgrounds and characteristics (such as age, common interest, geography, etc.). Open-ended questions allow a discussion on a specific topic (e.g., a particular product or task scenario) and therefore enable a detailed and deep analysis of a certain aspect.
- **Questionnaires**. The use of questionnaires offers an easy and quick way to gather data from a larger population. Relevant information with respect to user satisfaction, usability, error, user attitudes, preferences as well as the evaluation of system designs can be collected [427]. Typically, validated questionnaire inventories such as the Copenhagen Psychosocial Questionnaire (COPSOQ) are used [330]. These inventories comprise cross-occupational questionnaires on psychosocial job characteristics and are based on cause-effect relationships between external influences on the work situation and individual effects on the human worker [330].

8.3.1.2 Task Analysis Methods In order to describe and analyze work activities and scenarios in more detail, task analysis methods represent a well-established approach. In general, they describe actions of an operator or a team of operators that are required to achieve an overall system goal [427]. As shown in Figure 11, task analysis starts with the definition of a specific task or scenario from which the respective goal and subgoals are derived. Essentially, required operations and actions for achieving each subgoal are defined including the respective execution plan (i.e., the order in which the operations have to take place).

Hierarchical Task Analysis (HTA) represents the most commonly used method of task analysis. It hierarchically breaks down each task into goals, subgoals, physical operations, and plans, thus affording a complete detailed picture. Notably, it sets the starting point for other Human Factors methods such as workload assessment, evaluation,

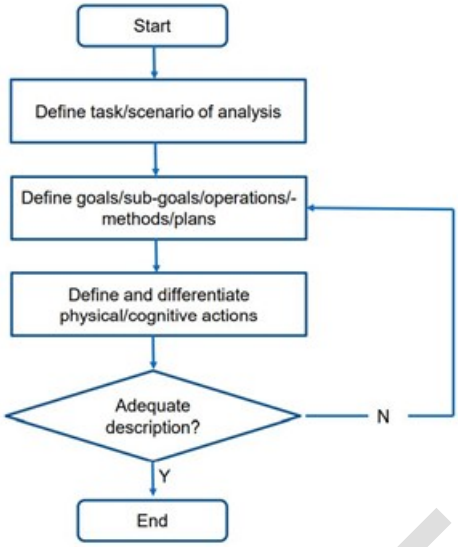


Figure 11: Procedure of task analysis.

and human error identification [9] and can be extended by a Cognitive Task Analysis (CTA). While the HTA provides a physical description of a task, the CTA defines relevant cognitive processes that are engaged in task execution. In particular, demands on attention, perception, and working memory are analyzed to highlight salient information about task performance processes (start/adjust/conclude steps) and decision-making. The CTA mainly relies on data collected from interviews and observations and is highly relevant for the specification of requirements to facilitate human-machine interaction.

8.3.1.3 Usability- and user-experience methods Usability describes the extent to which assistance systems aid in achieving the employees goal effectively and efficiently (e.g., ISO 92419), while user experience expands the focus of usability by including emotional and aesthetic factors. This involves the investigation of attitudes, trust, well-being, and expectations of the user when interacting with technical systems (e.g., [447]]). These methods form the transition between survey and experimental methods and enable the prediction of the acceptance of assistance systems based on user emotions.

8.3.1.4 Experimental methods To assess cognitive processes underlying task performance, experimental methods using paradigms of cognitive psychology are essential. Although such laboratory settings do not resemble real-life scenarios, they provide a highly controlled environment necessary to gain an understanding of basic cognitive mechanisms. At the behavioral level, experimental methods offer information about response times, force, posture, and movement (e.g., movement activity, grasping, and gait analysis). Moreover, physiological parameters (e.g., muscle-, heart-, skin-, eye-, and brain activity) can be assessed using electroencephalography (EEG), electrocardiogram (EKG), or electromyogram (EMG) in order to record physical and psychological stress [167, 50, 384].

8.3.2 Discussion

In sum, principles of cognitive ergonomics are inevitable when aiming for a human-robot dyad based on joint activity and awareness of the human user's state. In terms of a human-centered approach, relevant cognitive processes during execution of a task need to be identified. To this end, the following tools and methods from cognitive psychology, work psychology, and ergonomics will be particularly considered in the *FELICE* project:

1. The **HTA** will form the starting point and provide descriptive information about the task and the required actions. Importantly, the HTA can be expanded by defining not only the human but also the operations of the robot. This allows for a detailed description of the task and illustrates different possible task sequences that result from an interaction with the robot. By this, relevant principles of joint activity for an effective and efficient human-robot collaboration will be identified. Moreover, potential cues (e.g., auditory or visual) for the human worker can be derived from the analysis.

2. In a following step, the HTA will be extended by a **CTA** to identify underlying cognitive processes during task performance, such as perception, attention, working memory, and action coordination. To obtain appropriate data for this analysis, **focus groups** will be conducted in two ways: first, partners from the consortium will be invited to discuss required actions in a specific task scenario. In a later step, focus groups will be conducted with human workers to provide detailed information about the user's perception and work demands. These analyses will be especially relevant in order to specify requirements for efficient human-robot collaboration.

3. Concurrently, experimental studies are planned to analyze cognitive processes in a controlled environment. In particular, laboratory studies on proxemics will be conducted (cf. for example, [217, 383]).

This variety of methods enables the identification and specification of requirements both on the local layer (human-robot interaction) and on a global layer regarding the global supervisory unit. Importantly, they are further addressing one of the main challenges in efficient human-robot collaboration, i.e., the implementation of joint activity principles between humans and artificial agents.

9 Safe robot operation

9.1 Overview

This section gives a brief preliminary overview of robot safety in industrial environments from essentially two different perspectives. The first perspective is a regulatory one resulting from the normative standards and directives. The second originates from academia and includes research concepts that have not made their way into the normative. Finally, a concept of safety within the *FELICE* project will be presented.

9.2 Relationship with FELICE project

Safety is one of the most important factors when considering human-robot collaboration in industrial applications [502]. Focusing on *FELICE*, sharing the collaborative robot among workers at different workstations on the factory floor implies that the robot must move from one workstation to another. To do so, the robot should be able to plan its path and navigate along it autonomously using its on-board sensors to avoid obstacles. The robot should also ensure that it does not injure workers or damage surrounding objects. According to the ISO Technical Specification 15066:2016 [301], this can be achieved by integrating safety controls and safety-related features to the robot, which collectively ensure that hazardous situations are prevented from occurring and their impact is minimised if they do occur. More specifically, ISO/TS 15066:2016 dictates that the robot should continuously monitor the speed and separation of obstacles so as to maintain a pre-determined distance from them and employ built-in power or force feedback sensing to detect contact with persons or objects. Speed and separation monitoring imply that the robot should re-plan its path of movement in order to avoid collision and slow down or even stop completely when being near obstacles is unavoidable [250]. Navigation will be supported by the ubiquitous sensing of Pillar I by synergistically integrating information on the scene geometry, moving objects/humans.

9.3 Normative requirements

In the following we give an overview of the applicable norms and standards that result from the *FELICE* use cases. ISO/TS 15066:2016 specifies safety requirements for collaborative industrial robot systems and the work environment, and supplements the requirements and guidance on collaborative industrial robot operation given in ISO 102181 and ISO 102182. ISO/TS 15066:2016 applies to industrial robot systems as described in ISO 102181 and ISO 102182. It does not apply to non-industrial robots, although the safety principles presented can be useful to other areas of robotics. Pertains to all trucks and their systems except: a) trucks solely guided by mechanical means (rails, guides, etc); b) trucks operating in areas open to persons unaware of the hazards.

ISO 10218-1:2006 specifies requirements and guidelines for the inherent safe design, protective measures, and information for use of industrial robots. It describes basic hazards associated with robots, and provides requirements to eliminate or adequately reduce the risks associated with these hazards. ISO 10218-1:2006 does not apply to non-industrial robots although the safety principles established in ISO 10218 may be

utilized for these other robots. Examples of non-industrial robot applications include, but are not limited to: undersea, military and space robots; tele-operated manipulators; prosthetics and other aids for the physically impaired; micro-robots (displacement ¡1 mm); surgery or healthcare; and service or consumer products. ISO 10218-2:2011 specifies safety requirements for the integration of industrial robots and industrial robot systems as defined in ISO 10218-1, and industrial robot cell(s). The integration includes the following: the design, manufacturing, installation, operation, maintenance and decommissioning of the industrial robot system or cell; necessary information for the design, manufacturing, installation, operation, maintenance and decommissioning of the industrial robot system or cell; component devices of the industrial robot system or cell.

ISO 10218-2:2011 describes the basic hazards and hazardous situations identified with these systems, and provides requirements to eliminate or adequately reduce the risks associated with these hazards. ISO 10218-2:2011 also specifies requirements for the industrial robot system as part of an integrated manufacturing system. ISO 10218-2:2011 does not deal specifically with hazards associated with processes (e.g. laser radiation, ejected chips, welding smoke). Other standards can be applicable to these process hazards.

ISO 3691-7:2010 gives regional requirements specific to the countries within the European Community (EC) and European Economic Area (EEA) for the types of industrial trucks specified in the scopes of ISO 36911, ISO 36912, ISO 3691-3, ISO 3691-4, ISO 3691-5 and ISO 3691-6, respectively. It is intended to be used in conjunction with each of those parts of ISO 3691. This document gives regional requirements for specific countries outside the European Community (EC) and European Economic Area (EEA) for the types of industrial trucks specified in the scopes of ISO 3691-1, ISO 3691-2, ISO 3691-3, ISO 3691-4 and ISO 3691-6.

ISO 13855:2010 establishes the positioning of safeguards with respect to the approach speeds of parts of the human body. It specifies parameters based on values for approach speeds of parts of the human body and provides a methodology to determine the minimum distances to a hazard zone from the detection zone or from actuating devices of safeguards. The values for approach speeds (walking speed and upper limb movement) in ISO 13855:2010 are time tested and proven in practical experience. ISO 13855:2010 gives guidance for typical approaches. Other types of approach, for example running, jumping or falling, are not considered in ISO 13855:2010.

Safeguards considered in ISO 13855:2010 include: - electro-sensitive protective equipment, including light curtains and light grids (AOPDs), and laser scanners (AOPDDRs) and two-dimensional vision systems; - pressure-sensitive protective equipment, especially pressure-sensitive mats; - two-hand control devices; - interlocking guards without guard locking.

ISO 13849-1:2015 provides safety requirements and guidance on the principles for the design and integration of safety-related parts of control systems (SRP/CS), including the design of software. For these parts of SRP/CS, it specifies characteristics that include the performance level required for carrying out safety functions. It applies to SRP/CS for high demand and continuous mode, regardless of the type of technology and energy used (electrical, hydraulic, pneumatic, mechanical, etc.), for all kinds of machinery. It does not specify the safety functions or performance levels that are to be used in a particular case. This part of ISO 13849 provides specific requirements for SRP/CS

using programmable electronic system(s). It does not give specific requirements for the design of products which are parts of SRP/CS. Nevertheless, the principles given, such as categories or performance levels, can be used.

ISO 13849-2:2012 specifies the procedures and conditions to be followed for the validation by analysis and testing of the specified safety functions, the category achieved, and the performance level achieved by the safety-related parts of a control system (SRP/CS) designed in accordance with ISO 13849-1. IEC 60204-1:2016 is available as IEC 60204-1:2016 RLV which contains the International Standard and its Redline version, showing all changes of the technical content compared to the previous edition.

IEC 60204-1:2016 applies to electrical, electronic and programmable electronic equipment and systems to machines not portable by hand while working, including a group of machines working together in a coordinated manner. The equipment covered by this part of IEC 60204 commences at the point of connection of the supply to the electrical equipment of the machine.

ISO 11161:2007 is not intended to cover safety aspects of individual machines and equipment that may be covered by standards specific to those machines and equipment. Therefore it deals only with those safety aspects that are important for the safety-relevant interconnection of the machines and components. Where machines and equipment of an integrated manufacturing system are operated separately or individually, and while the protective effects of the safeguards provided for production mode are muted or suspended, the relevant safety standards for these machines and equipment apply.

ISO 13854:2017 enables the user (e.g. standard makers, designers of machinery) to avoid hazards from crushing zones. It specifies minimum gaps relative to parts of the human body and is applicable when adequate safety can be achieved by this method.

ISO 13854:2017 is applicable to risks from crushing hazards only and is not applicable to other possible hazards, e.g. impact, shearing, drawing-in.

ISO 14118:2017 applies to unexpected start-up from all types of energy source, i.e.: - power supply, e.g. electrical, hydraulic, pneumatic; - stored energy due to, e.g. gravity, compressed springs; - external influences, e.g. from wind.

9.4 Research projects targeting robot safety

The importance of safety in human robot interaction [456] motivated other European projects e.g. PHRIDOM [40], PHRIENDS[5] and SAPHARI[6], and currently COVR[7]. De Santis et al. formulate an atlas of physical human robot interaction (pHRI) with special focus on safety and dependability [125]. The following guidelines should be observed during the design of a robot for pHRI: In order to reduce the potentially catastrophic consequences of the robot colliding with a human, the inertia of the moving parts should be kept as low as possible by means of lightweight design by locating the drives in the robot base and transmitting the mechanical power to the joints using cable or hydraulic/pneumatic actuation. The links should be compliant and the robot surface should be covered with soft material; no sharp elements should be present on the robot

[5] https://cordis.europa.eu/project/id/045359
[6] http://www.saphari.eu/
[7] https://www.safearoundrobots.com/home

surface. Safe limits for the maximum moving weight and velocities of the arm have to be found by simulation and ensured on the robotic system [[181, 207]. The drives should be torque-limited and back-drivable so that the user may change the robot position just by touching and moving it with bare hands. It should be impossible for the human fingers, hands, clothing parts etc. to be clamped in between joints, cables or any other protruding elements. The user should be protected against electric shock; ideally, no voltage higher than 48V DC should be present in the robot. It is obligatory that the robot recognizes the interaction forces exerted by the surrounding humans or objects by its very surface and not only at the end effector or at the joints; adequate sensorised skin is recommended [87].

9.5 FELICE approach to safety

Despite considerable efforts in research devoted specifically to robot safety, and the well elaborated results, methods, guidelines, etc., safe robot manipulators suitable for our application do not yet exist. *FELICE* builds upon the above research and will contribute to their actual technical realization. As it can also be seen, some of the most important standards mentioned above are not yet released. This is partly because of a typical deadlock situation: In order to release a standard, the authorities need proofs of concept; and the robot makers need standards to develop the products. *FELICE* will contribute by providing the proof of concept. Until recently, the common approach to safety was amending safety systems to existing robots (industrial) by means of additional sensors and software/control modifications. Nevertheless, sensors combined with control laws can improve but never guarantee by themselves the safe behavior of a robot. The goal must be to design robots so they are intrinsically/mechanically safe, i.e. avoiding hazards instead of controlling them. This approach, however, requires that safety is "engineered into" electromechanical design at the earliest stage possible. *FELICE* will go beyond the current state of the art in safe robotic manipulator development by developing a robotic manipulator of a workspace comparable to that of a human arm, with overall weight of less than 5 kg and payload capacity of 3 kg, additionally fulfilling the following requirements:

- actuator velocity, power, torque and stiffness shall be limited to the lowest values necessary to carry out the task
- the actuators shall be backdrivable and therefore high gear reduction ratios shall be avoided
- the manipulators shall be passively gravity compensated
- mechanical "fuses" shall be applied disabling the drive propulsion in case of overload during impact or unintentional collision
- mechanical singularities shall be designed out of the workspace in order to prevent uncontrolled motion in case of control or numerical failures
- all wires and electrical connection shall be well protected from wrenching, cutting etc.; wires shall be integrated in the manipulator or into the body

- it shall be impossible to clamp any part of the human body or clothing in between moving parts of the robot
- the joints will be multi-actuated using parallel or differential kinematic structures, so that an uncontrolled behaviour of one motor will not be able to produce dangerous motion of the link.

10 Robot programming

10.1 Overview

This section briefly describes various programming paradigms that enable the configuration of a robotic system for a particular task at varying degrees of task abstraction. Special focus in this section is put on task level programming that enables configuring a robotic task. The task level programming is linked with automated planning framework(s) to deal with dynamic changes in the environment. Finally, some remarks on the work on robot programming in *FELICE* project are outlined to show the aspects related to the advancement of the state of the art.

10.2 Relationship with FELICE project

Ease of programming is one of the principal advantages of cobots (collaborative robots) over traditional industrial robots [461]. A popular paradigm is *task-level programming*, where goals for the positions of objects are specified, instead of the robot motions needed to achieve these goals. A task-level specification is completely robot-independent and no positions or paths dependent on the robot geometry or kinematics need to be specified. Task-level programming requires complete geometric descriptions of the robot surroundings, which in the case of *FELICE* are provided by the sensing of Pillar I (see also section 3).

10.3 Robot programming

In the research community there has been a growing interest to solve the challenge of programming robots for executing tasks in complex real-world environments [230, 122, 49]. In the literature, different methods for programming/configuring agents by human instructors are described. Common to all methods is their aim to reduce programming effort. However, the problem of reducing the programming effort required by an expert using natural modes of communication is still an open issue [353]. When viewed from a broader perspective, robot programming approaches can be divided as follows:

- Programming by advice
- Skill based programming
- Programming by demonstration
- Programming by interaction

Each approach is discussed next.

Programming by advice: This is made possible by the use of natural forms of communication. The authors in [286] developed a method whereby advice is given to the learning agent (Reinforcement Learning).

Skill based programming: The approaches falling under this category use the **task-level programming paradigm** for easy and quick re-programming by non-experts. The

task-level programming paradigm [327] is built upon on a set of actions, where the actions have the capability to alter the current world state.

Programming by demonstration: This is a popular paradigm by which agents learn by physically presenting the task through the human being [68]. The work in [49] provides a detailed overview of relevant methods.

Programming by interaction: This is an interesting approach where learning arises from interaction with the environment. Recent trend points towards applying Reinforcement Learning (RL) [435] for this purpose. The agent (robot) learns how to behave (i.e., which actions to perform when) in order to complete a task in the given environment.

10.4 Task-level programming

Task-level programming is based on lower level entities, usually called robot skills, that instantiate actions. Various representations of robot skills that would be suitable for task-level programming have been pursued during the last decades (e.g. [70]). A systematic survey of the rich relevant literature reveals how fragmented the concept of skills is, lacking a widely accepted, strict definition [39, 353]. What most of the aforementioned attempts have in common, is that robot skills are, in turn, composed of primitive sets of robot motions, called action primitives. Skill primitives require parameterization that is acquired from demonstration.

The general idea is to store the ability to perform elementary robot actions in reusable primitives that allow mobility, coordination, control and supervision of specific tasks (e.g. manipulation tasks [247]). The primitives can incorporate advanced task specifications, necessary control, and sensing capabilities (cf. Sec. 3), which allow a skill to handle uncertainties during execution. As all of this information is encapsulated, the programmer (or a *higher level task planner*) does not need to worry about the details and can utilize the robot by assembling predefined skill primitives.

10.4.1 High level task planning

Within the AI community, there has been a long-standing focus on planning in discrete domains, generally with very large state spaces, but made tractable by using representations and algorithms that exploit underlying regularities in the structure of the domain. Ghallab et al. [166] provide a comprehensive discussion of task planning from the AI perspective, and Karpas and Magazzeni [226] survey task planning for robotics.

The simplest formalization of AI planning is to specify a set of states (the state space) S, a set of transitions $T \subseteq SXS$ that describe permissible changes to the state, an initial state $s_0 \in S$, and a set of goal states $S \subseteq S$. Each directed transition $t = \langle s, \hat{s} \rangle \in T$ moves the system from state s to state $\hat{s}$. The objective for a planner is to find a plan , i.e. a sequence of transitions, that advances the initial state s_0 into a goal state $s \in S$. This problem can be reduced to a graph traversal problem, where the vertices are states and directed edges are transitions, and solved using standard graph-search algorithms.

One focus of AI planning has been to define languages for specifying planning problems. The most widely-used formalism is the planning domain definition language

(PDDL) [304], which can be seen as a transition system where state variables are Boolean facts. The AI planning community has developed domain-independent algorithms that can operate on any problem written in a planning language, without any additional information about the problem. A factored planning representation enables efficient algorithms for solving relaxed problems, simplified versions of the original problem, and using their solutions to estimate the distance to a goal state [187].

There are several extensions to the basic task planning formalism [151], [138], [152]. One of these is numeric planning, which involves planning with real-valued variables such as time, fuel, or battery charge. Recent approaches support planning with convex dynamics [81] and non-convex dynamics by discretizing time [112].

In task-level planning involving robotic systems, the robot actions are specified by their interaction with objects. Often the final goal is known, e.g, "put the object from the box to the table". The goal of the task level planning in this case is to find a sequence of actions that a robot has to perform to modify the environment from the initial state to the desired goal state [288]. For example, the solution for the earlier example will consist of the following actions: 1) "open the box", 2) "grasp the object", 3) "move the object to the table", 4) "release the object". Every single action is then planned with a domain dependent planner. Cao et al. [88] proposed a net called AND/OR used for reasoning about geometrical task constraints. The general idea is to map the proposed net to a Petri net. Then, the solution search is performed by building a reachability tree from the Petri net. Later, Chien et al. [104] [105] proposed an efficient way to incorporate the domain information into the planner for the indoor robot scenario. The data is represented with an object-oriented model. This model includes relation between objects, categories and physical laws. The case study that the solver is capable to solve was to put all metal parts to the bench in the assembly room. This type of planning ignores the kinematics and collision planning. Another interesting example where a robot has to fetch an empty bottle from a customer's hand and then bring it to a counter is implemented using behaviour trees (BTs) in [114].

Task-level planning was also applied to mobile robotics [158] where the hierarchical planning technique was proposed to reduce the computational overhead. For a more detailed overview on robot task level planning, the reader is referred to [41].

10.4.2 Uncertainity in task planning

A critical issue when acting in the real world is *uncertainty*. In the presence of future state uncertainty, a planning algorithm might need to take into account multiple possible outcomes of an action and ensure that there are actions it can take in response, to avoid unlikely but disastrous outcomes. More difficult, but pervasive, is uncertainty about the present state. In this case, the problem can be treated as a belief-space planning problem, in which the planner reasons explicitly about the agent's state of information about the world and takes actions both to gain information and to drive the world into a desired belief state. Several approaches for deterministic observable task planning have been extended to handle these challenges, e.g. [220], [182], [362], [164].

10.4.3 Baseline technologies and tools

10.4.3.1 Automated planning The objectives of this subsection are twofold: First, the identification of a suitable reference PDDL (Planning Domain Definition Language) domain that might provide a basis for the planning tasks in the *FELICE* use case(s). Second, the identification and evaluation of suitable planners. We extend the scope for pure classical planning with dynamic planning, i.e. the ability to monitor the state of the world, and to update the knowledge base for automated planning with world state information.

We identified the *tidybot* domain as a possible reference domain; the *tidybot* domain has been used in the International Planning Competitions 2011 and 2014 [270]. The *tidybot* domain was introduced with a different motivation, the increasing interest in re-approaching the fields of AI planning and autonomous robotics. State-of-the art planners fail to address problems with large state spaces, like the motion planning problems typically addressed in robotics. Humans are able to quickly find feasible solutions in such domains, because they seem to be able to decompose the problem into separate parts and make use of the geometrical structure. The *tidybot* domain is intended to exercise the ability of planners to find and exploit structure in large but mostly unconstrained problems. Optimal reasoning in such problems is challenging for humans as well, and a secondary motivation for the domain is to test the ability to do optimal reasoning in geometrically structured worlds.

10.4.3.2 Planners The following planners / planning frameworks are investigated for further development in the *FELICE* project:

- Pddl4j [357]
- ROSplan [96]
- Fast Downward planner [188]

Pddl4j is an open source java library for classical planning, released under the Lesser General Public License (LGPL) licence model. The ROSplan framework provides a collection of tools for automated planning in the context of robotic applications, providing integration with ROS. The licence model for ROSplan is based on the MIT licence. Fast Downward (FD) is an award winning planner that has taken part in International Planning Competitions. FD has been in development since 2003, with development still ongoing. It is licensed under the General Public License (GPL).

The XRob software framework The XRob software framework of Profactor [363] enables the creation of complex robot applications within a short time. It builds on unique, easy-to-use features that significantly speed up commissioning and make the operation more cost-efficient and flexible than common programming methods. The special software architecture allows easy and intuitive creation of processes and configuration of the components of a robot system via a single user interface.

Robot interfaces: To facilitate communication with the robotic system, the XRob framework provides a uniform communication interface, which can be extended in a plug-in like fashion to support robotic systems from different vendors.

Application development: The XRob software framework provides an intuitive user interface for application development, which includes an interactive programming environment, and software modules to simulate and visualize robotic movement paths as well as data acquisition via sensors.

Interactive programming environment: The interactive programming is used to obtain all information needed to execute the intended process. The system understands the following basic information: robot-pose, 2D image and 3D image. Additional algorithm specific properties can be changed for every algorithm instance. The basic workflow is to configure the system in an initial state. If the system is to be reconfigured, either all information has to be re gathered or the initial state has to be restored.

10.4.4 Discussion

The idea of breaking down the given robotic task into a set of primitive robotic actions has gathered a lot of attention and several tools are commercially available (for e.g, Drag&Bot, Artiminds) that allow a user to program a robotic task in this fashion. However, these tools require the action primitive parameters to be known a priori for a successful execution and cannot be easily modified during execution. Given the dynamic nature of the environment, the action primitive parameters should adapt according to the given situation. The variations could include:

- Inability to grasp the object due to change in object position (due to sensor or other disturbances)
- Change in user requirements (for e.g., in case of object handover, the robot has to adjust according to the height of the user)
- Slipping of object (from the gripper) during manipulation

The *FELICE* project aims to improve the existing task-based programming tools by enabling them to deal with the dynamic changes like those mentioned above. The strategy is to build on the existing XROB task-based programming framework [39] and exploit multi-modal sensing (verbal, human tracking, object recognition, gestures) to deal with dynamic changes. The coordination between the XROB tool together with the high-level cognitive system in the *FELICE* project will be the key in achieving this goal.

11 Synchronization of the human-robot dyad in taskable pipelines

11.1 Overview

As already discussed in Section 10, task-level programming of industrial and assistance robots is a promising approach towards reducing the software complexity and therefore the development effort for various applications. Effective cooperation between two parties presupposes that both are accustomed to the task and to each other, so as to coordinate their actions. Moreover, besides the ability to alter their plans and actions appropriately and dynamically, their timing should be precise and efficient, resulting in a well-synchronized meshing of their actions.

With the aim of using robots to support the automotive human workforce, *FELICE* is interested in how robotic teammates could perform fluently with human workers. A fluent teammate evokes appreciation and condence. If robotic teammates are to be widely integrated in assembly workplaces to collaborate with humans, their acceptance may depend on the fluent coordination of their actions with that of their human counterparts.

This section examines complementary aspects of human-robot collaboration that range from distributing relevant work to each party after considering their state at the given time, to enhancing real-time human-robot interaction that is expected to enhance human trust and increase the acceptance of using robots in the assembly line.

11.2 Relationship with FELICE project

Human-Robot Interaction (HRI) has been the topic of several survey papers [498, 189, 417]. Broadly speaking, HRI can be roughly separated into two areas considering the characteristics of the application domain. On the one hand, there is human-robot interaction in occupational environments in which robot skills remain rather focused on the needs of the particular operating context. In this context it is quite often that the interaction between the two sides routinely repeats, following a known behavior-interaction pattern. The role of robot is critical for the effective and timely progress of the planned work, and humans expect very high standards from robot performance.

On the other hand, there are personal service robots targeting the much more open-ended social human-robot interaction. This application domain often includes robots to provide entertainment, teaching, and assistance for children and elderly, autistic, and handicapped persons, with the interaction between the two sides being much more broad and unconstrained. In this environment, humans are more tolerant of possible robot errors, as long as human physical integrity is not compromised.

The work implemented in *FELICE* is falls in the first category, so in the following paragraphs we focus on the analysis of human-robot interaction in occupation environments.

11.3 Human-robot interaction taxonomy

Before proceeding with the technical characteristics of the interaction in human-robot dyads, it is important to comment on the different levels of synergetic interaction that can be developed between the two entities [239, 337]. Considering the organization of team work and how the common goal is accomplished, different levels of human-robot interaction are identified [403, 58]:

- At the lowest level we have complete spatial and/or temporal separation of human and robots. Fences, barriers, and other safety environments ensure that no human could get harmed as the robot proceeds in isolation with the implementation of its predetermined tasks.
- In the next level the two entities share a common workspace. They act at the same time (i.e. work in parallel), but they work individually to accomplish different aims of the composite task. The HRI at this level is often referred as Human-Robot Coexistence (shared worktime, workspace).
- When additionally humans and robots are working with the same aim then we speak about Human-Robot Cooperation (shared worktime, workspace, aims).
- Finally, when there is the possibility of contact (physical, auditory, visual) between the two agents the interaction is labeled as a Human-Robot Collaboration. The latter, more enhanced type of interaction is the topic investigated in *FELICE*.

It is important to note that the development of robots working alongside humans should aim not only at task efficiency, but also at human-robot fluency. While previous sections have mostly considered robot efficiency, Section 11 focuses on the fluency of the interaction. To improve Human-Robot collaboration, a number of metrics have been developed to evaluate the level of fluency in human-robot shared-location teamwork, which aim to codify subjective and objective human-robot fluency [198]. These metrics will be considered in *FELICE*, to reach a high level of coordination between human workers and their robotic teammates.

11.4 Timing in human-robot synergies

Several works have considered the notion of time in planning solo robot behavior in the form of action sequences, frequently with the use of PDDL that uses first-order predicates to describe plan transitions [95], or NDDL (New Domain Definition Language) that considers a “timeline” representation to describe sequences of temporal conditions and desired values for given state variables [372] also adopted by the EUROPA Planning Library [378, 57] and its subsequent advancement that considers the description of hierarchical plans [47]. Opportunistic planning provides an alternative view for scheduling long-horizon action sequences [94]. The use of hierarchical plans is additionally considered in [430], focusing on the unification of sub-plans to improve implementation efficiency. Moreover, the high-level Timeline-based Representation Framework provides a structured library for managing operational modes and the synchronization among events [97], or with the use of the forward search temporal planner POPF [93].

Extensions of this framework have been used among others in industrial human-robot collaboration [356, 452] to ensure controllability.

To implement tasks involving multi-agent collaboration, planning algorithms often rely on constraints which provide ordering between the independently implemented activities having the role of prerequisites to one another [319, 411, 421, 318]. Existing approaches explore the controllability of alternative strategies, to identify plans that successfully schedule the required activities in a way that satisfies constraints until the final completion of the goal [108]. Moreover, possible adaptations on the timing of the activities sequences are considered.

To manage temporal constraints, Distance Graphs with Uncertainty (DGUs) are frequently used as a means to represent and study the given problem. Checking a DGU for negative cycles provides information on the consistency of a candidate plan. The non-existence of negative cycles in the DGU indicates that the action sequence is dispatchable, meaning that (i) there are no temporal conflicts and (ii) there is enough time for all events to occur. Following this formulation, previous works have considered back propagation rules to adapt the timing of forthcoming activities and thus dynamically preserve the dispatchability of plans [411, 318, 296].

Interestingly, relevant works consider the use of time in full isolation, without the ability to blend time with other quantities for the time-inclusive multi-criteria evaluation of plans. For example, time-labeled Petri-nets have been used to accomplish fluent resource management and turn-taking in human-robot collaboration focusing mainly on dyadic teams [99]. In a different work, time has been sequentially combined with space to minimize annoyance among participating agents [173].

Other works follow a multi-criteria optimization problem formulation, to accomplish time-aware human-robot cooperation. The objective function is derived from the preference values of participating agents and the temporal relations between entities are mapped on the constraints of the problem [482]. More recent works follow basically the same formulation, representing time in the set of constraints that confine available solutions [172]. Besides the fact that criteria such as the workload and the user preferences can be addressed with these approaches, time is largely kept separate form other quantities, thus not used for the formulation of time-informed multi-criteria objectives. Moreover, the works mentioned above do not consider predictive estimates on the performance of interacting agents and the expected release of constraints among tasks.

Recently, decentralized approaches are used for multi-robot coordination, which work on the basis of auctions. For example, [307] considers scenarios in which tasks have to be completed within a specified time window, but without allowing overlap between time windows. Modern approaches are targeting this issue with particularly successful results in simulation environments [332, 331]. In other similar problems, the routing of working parts is assigned to the most suitable transportation agent through an auction-based mechanism associated to a multi-objective function [90]. However, the relevant approaches assume auctions to proceed on an agent-centered point of view which does not consider the capacities and special skills of other team members. Therefore, it is hard to maximize the usability of all members for the benefit of the team (i.e. it might be beneficial for the team if the second optimal agent undertakes a given task).

Complementary to the above, the Daisy Planner (DP) [295, 294, 210] relies on the daisy representation of tasks and adopts time-inclusive multi-criteria ranking of alter-

native plans. DP operates under the assumption of pursuing immediate, locally optimal assignment of tasks to agents. This is in contrast to previous works on scheduling multi-agent interaction that typically prepare long plans of agents' activities for all future moments [173, 205, 108] under the risk of frequent re-scheduling, due to external disturbances that may render current plans infeasible. In such cases, re-scheduling may take up to a few tenths of seconds [356]. DP effectively operates as a lightweight process which minimizes the chances for re-planning in the case of unexpected events [210].

11.5 Task distribution

In recent years, the distribution of tasks to the human or the robot especially for assembly lines has been focally investigated. A number of parameters have been identified to affect the distribution of work, including (i) the physical characteristics of the processed components, (ii) the special skills that are needed in order to use the components in the right way (e.g. mounting, placing) (iii) the way the components are provided to the assembly line (iv) the safety issues that arise when either the human or the robot undertakes a task (v) the potential fastening of tools that may require more than one hand to complete [290, 238].

The method of task-distribution in HRC starts by decomposing an assembly operation into tasks and identifying their automation potential. In most cases, execution by robot is assumed to be slower than execution by human, while collaborative execution is assumed faster. To estimate the benefit of robot deployment to the workstations of an automotive assembly line, [439, 404] developed an approach to assess the human-robot collaboration potential of workstations. Based on existing standardised work descriptions, the suitability for human-robot collaboration is derived and thus an evaluation and comparison of the whole assembly with and without the robot is achieved.

To take into account the optional collaboration of a human and a robot on tasks, [479] has recently developed a method for the human-robot collaborative assembly line balancing and scheduling problem, after considering the use of a single robot for multiple workstations. Optimization criteria aim to locally improve the station finish times and globally strengthen the collaboration potential of the whole production line. In a similar spirit, [448] considers the assignment of jobs to team members (either humans or robots) considering the average utilization of each member and the average time that a number of jobs spend in a workstation. Other works consider analytical expressions of time expectation and variability, to study process monotonicity and bottleneck identification [218]. Besides the fact that many of the existing works aim to minimize the makespan of tasks (e.g. [73]), this approach is not relevant for the setup assumed in *FELICE*. This is because the assembly time that is spend per door in each work station is predefined and cannot change online (as this would destabilize the whole production line). In fact, the criteria to be considered in the current project focus mostly on how the process will become less tedious for the humans working on the assembly line.

In the majority of existing works, task distribution considers the full implementation of a task either by a human or a robot. Apparently, in agile automotive assembly line we cannot easily adopt this approach because robot skills should be extensively tested to assure that robot actions will not harm the quality and appearance of the end product

(even small errors, may result into the need for disassembling of a large number of components, with significant delays for the production). The approach adopted in the current project assumes that only humans are able to act with and on car parts. The robot adopts a secondary supportive role to the human.

Following the above, humans and robots do not have same skills, and thus they do not equally share tasks. Thus, the task distribution approaches considered so far in the literature assigning complete task implementation either to the human or the robot can only provide sub-optimal solution to the type of problems considered in *FELICE*. For example, the hand-over of tools that is included in the scenarios summarizing HRC in all workstations can not be effectively represented under the assumption of full task implementation. The hand-over of tools and assembly parts is of particular interest to *FELICE* because it is one of the basic interactive elements in the context of close-proximity HRC.

To effectively implement object hand-over, the two involved entities should establish a three-fold agreement which includes the object (what), the handover time (when), and handover location (where). The entity providing the object should additionally consider the comfortability and the convenience of the receiver so that (s)he can easily use it [43]. Human-Robot interaction in handover tasks can be supported be social gestures and the communication of affective information, as suggested in [148, 313] From the robot side, it must be able to estimate the time required by the human to complete the ongoing (current) assembly process, including the effects of potential external and intrinsic factors (e.g., skill level, fatigue, and stress) that can affect the assembly rate [204].

A recent work studied the cycle time, waiting time, and operators subjective preference of a humanrobot collaborative assembly task when three handover prediction models were applied: traditional method-time measurement (MTM), Kalman filter, and trigger sensor approaches [438]. The results revealed that both the Kalman filter prediction model and the trigger sensor method were superior to the MTM fixed-time model in both scenarios in terms of cycle time and subjective preference. The Kalman filter prediction model could adjust the handover timing according to the operators current speed and reduce the waiting time of the robot and operator, thereby improving the subjective preference of the operator. Moreover, the trigger sensor methods inherent flexibility concerning random single interruptions on the operators side earned it the highest scores in the satisfaction assessment.

11.6 Discussion

In order to achieve efficient cooperation between humans and robots, *FELICE* examines the cooperation of the two sides at two different levels of detail. At the highest level, *FELICE* will make decisions about where and how the robot will assist humans in assembly tasks. The relevant decisions will be made by the orchestrator, which will choose the work station where the robot could contribute more depending on the particular needs and states of the employees working on the assembly line. At the lower level, decisions will be made to coordinate task activities between humans and robots so that the two parties can work together cooperatively, fluently and efficiently. The coordination of the human and robot actions will take into account the evolution of the assembly tasks, the

state of the employee and the assembly actions that (s)he currently carries out, as well as the commands that (s)he will address to the robot, asking for tools or any other type of assistance.

12 Prescriptive analytics in production system diagnosis, monitoring, and control

12.1 Overview

Prescriptive analytics is part of the domain of business analytics (BA) which uses new technologies and methods to analyse and control complex processes. It is composed of four stages, namely descriptive, diagnostic, predictive and prescriptive analytics. As summarized in [126], each of the stages is attributed to a specific BA related question: What happened?" (descriptive), "Why did it happen?" (diagnostic), "What will happen?" (predictive) and "What should I do?" (prescriptive) (also see Figure 12 for an overview). These four stages provide a general template whereas the implementation itself greatly depends on the specific domain and is a continuous and hierarchical process as each of the stages depends on the implementation and maturity of the previous ones. The stages can further be categorized with respect to whether they are retrospective (descriptive and diagnostic) or prospective (predictive and prescriptive), meaning they either allow to review and analyse past events or can be used to predict future events and to provide a recommendation of actions in adherence to given optimization criteria (e.g. product quality, productivity, machine utilization, etc.). The resulting set of recommended actions and their predicted outcome can both be used in an automatic and autonomous fashion or act as a guideline for decision-makers, optimizing processes according to potential future demands and conditions.

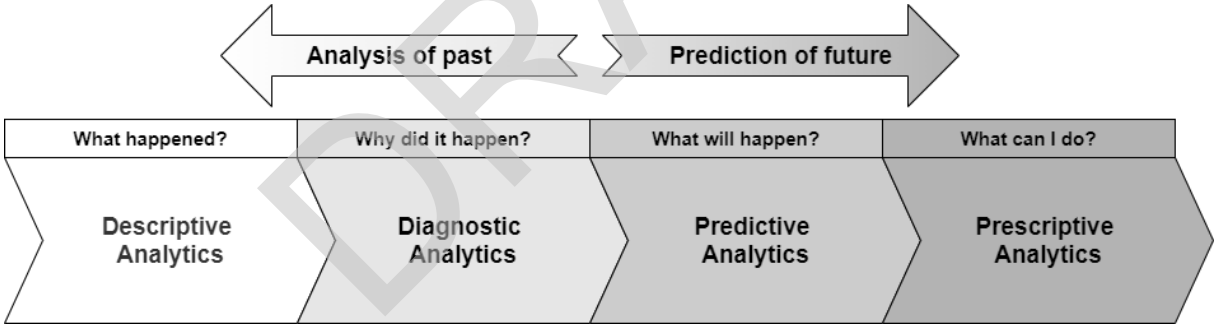


Figure 12: Overview of descriptive, diagnostic, predictive and prescriptive analytics based on [126].

12.2 Relationship with FELICE project

A prescriptive analytics system in the context of human/robot assembly lines must be able to decide, often within a short time window, on the (re)allocation of robots to workstations, speed of the line, and balancing of tasks. This poses challenges in formulating appropriate models of the decision situation and solving such models continuously in real-time and in reaction to an incoming stream of data. Continued adaption of predictive models and their parameters is necessary so that gradual process changes can be detected and reacted to. Optimisation agents that formulate decisions for controlling assembly lines, task assignment, and robot/operator allocation need to continuously

adapt to the environment and align their policies with the actual reward that they receive. A further challenge is the integration of prescriptive and predictive analytics in a single advanced control system. *FELICE* will develop analytics models to predict e.g., performances of human operators and assembly errors. A prescriptive analytics system will be developed, tested, and validated that decides, for instance, on the (re)allocation of cognitive robots, dynamic scheduling of tasks in the assembly line, allocation of tasks in human/robot collaborative workflows, and control of assembly line parameters such as line speed. With regard to the different tasks and research questions defined in the *FELICE* project, all four stages from descriptive to prescriptive analytics must be implemented to achieve these goals.

12.3 Prescriptive analytics

The research field of prescriptive analytics gained additional attention in recent years, mostly due to the maturity of the enabling technologies and new trends such as the Internet of Things [51], 5G [72] and the fourth industrial revolution, Industry 4.0 [254], promising additional business value for companies and therefore creating a huge incentive for further research. Besides the maturity of enabling technologies, another important aspect is the ongoing collection of data in almost all industrial fields, leading to the advent of big data [393]. Gaining additional insights of the available data through different means of data analysis previously lead to a strategic advantage but is considered standard since the fourth industrial revolution.

12.3.1 Baseline technologies and tools

Prescriptive analytics solutions are currently deployed in a variety of different fields and applications, such as e-commerce [453, 224], behaviourism [380], smart manufacturing [259], recommendation systems [36], health care [159, 426], aircraft traffic management [52], employees recruitment [361], logistics [329] and infrastructure planning [79]. Current reviews in the field of prescriptive analytics are analyzing and structuring the advancements in terms of applications [366], specific fields such as healthcare [276] or smart manufacturing [458], and current methodologies and research challenges [154, 257, 258]. Of particular interest are the currently used methodologies as surveyed by [258] and illustrated in Figure 13. The authors reviewed and classified 56 prescriptive analytics papers, of which 23 used Mathematical Programming methods, 16 used Logic-based Models, 7 used Machine Learning methods, 7 used Simulation methods, 3 used Evolutionary Computation methods, and 2 used Probabilistic Models (please note that some papers used multiple methods). Most of the review papers were written in either the application domain of manufacturing (12) or sales/marketing (14).

Prescriptive Analytics Methods					
Probabilistic Models	**Machine Learning / Data Mining**	**Mathematical Programming**	**Evolutionary Computation**	**Simulation**	**Logic-based Models**
Bayesian Network	K-means clustering	Mixed Integer Programming	Genetic algorithm	Simulation over Random Fores	Association rules
Markov Chain Monte Carlo	Reinforcement Learning	Linear Programming	Evolutionary Optimization	Risk Assessment	Decision rules
Hidden Markov Model	Privacy preservation	Binary Quadratic Programming	Greedy algorithm	Stochastic simulation	Criteria-based rules
	Boltzmann Machine	Non-Linear Programming	Particle Swarm Optimization	What-if scenarios	Fuzzy rules
	Nadaraya-Watson estimator	Binary Linear Integer Programming			Distributed rules
	Artificial Neural Networks	Stochastic Optimization			Benchmark rules
		Conditional Stochastic Optimization			Desirability function
		Constrained Bayesian Optimization			Graph-based recommendation
		Fuzzy Linear Optimization			5W1H
		Robust and Adaptive Optimization			
		Dynamic Programming			
		Optimal searcher path			

Figure 13: Overview of prescriptive analytics methods as defined in [258].

Another important aspect are the available prescriptive analytics tools and frameworks (see Figure 14), which were reviewed by [258]. The authors argue that currently no single tool exists, which is suitable to represent and optimize prescriptive analytics tasks, instead a combination of different tools must be used. Therefore the authors additionally analysed so-called business analytics (BA) suites, which are classified as an ecosystem which integrates and provides access to multiple tools. These suits still exhibit three disadvantages which are explicitly stated, namely that most BA suites typically only support procedural programming languages, which requires sufficient knowledge in software engineering, their insufficient ability to express or formalize prescriptive analytics workflow tasks (as the main goal of the reviewed BA suites is not primarily prescriptive analytics) and their often limited capability of distributed computation (scaling with multiple computation nodes). So to summarize the state of currently available prescriptive analytics systems: No out of the box solution or automation for prescriptive analytics exists and a combination of multiple tools and expert knowledge is necessary in order to tackle prescriptive analytics tasks.

	System Class	**Key Representative Systems**	**Describtive Analytics**	**Predictive Analytics**	**Prescriptive Analytics**
BA Tools	Reporting and spreadsheet tools Data Mining&ML libraries Data Mining&ML GUI Tools Online ML cloud services Mathematical optimization tools Computer algebra tools Sytem modeling tools	Excel, Google Sheets Spark MLlib, Mahout Weka, Hugin Watson, Azure ML Gurobi, CPLEX, OptaPlanner Mathematica, Mathcad Dymola, Simulink	Advanced Basic Intermediate Intermediate Basic Intermediate Intermediate	Basic Basic Intermediate Intermediate Basic Intermediate Intermediate	Basic Basic Basic Basic Basic Intermediate Intermediate
BA Suites	Statistical computing suites Statistical GUI suites	MATLAB, R, Julia SAS, SPSS	Advanced Advanced	Advanced Advanced	Intermediate Intermediate

Figure 14: Overview of business analytics tools and their maturity concerning descriptive, diagnostic, predictive and prescriptive analytics as defined in [154]. Please note that the authors omitted diagnostic analytics, therefore only three stages in contrast to the four stages defined previously are present in this illustration. The maturity is classified as either basic, intermediate or advanced.

12.3.2 Discussion

To fully implement a prescriptive analytics methodology in the *FELICE* project, all of the previously defined stages have to be addressed.

- **Descriptive analytics**: Main goal in this first stage is the introduction of sensors to collect relevant data of the production process and environment parameters. This data has to be structured, preprocessed [161] and stored in an suitable storage system. This database should be continually extended as new data becomes available and therefore contains all historical data.

- **Diagnostic analytics**: Based on the data from the first stage, models which represent the current state of the production system must be created. These models can be used to describe and analyse the interaction and impact of different parameters according to specific optimization criteria (such as production line speed, makespan, etc.). The introduction of a digital twin (online) is the final goal in this stage and represents the smooth transition towards the next stage.

- **Predictive analytics**: As an extension to the online digital twin, offline simulation enables the creation of "what-if" scenarios, effectively exploiting the ability to forecast certain variations in production parameters and different settings and assumptions (worst case, best case). As the sensor data and simulations both reside in the time-domain, they can be represented as time series data for which a variety of different general [124] and machine learning-specific [38, 74] forecasting-echniques exist. We foresee both methods, and especially in combination, as a suitable way to implement the predictive analytics stage.

- **Prescriptive analytics**: Given a sufficient degree of maturity of the previous stages, it is possible to access historic and current data, analyse the current state by means of an online digital twin/model and to forecast and evaluate potential future scenarios and conditions. The decision on how to react to the uncertainty of the future is still in the hands of the human operator and therefore prone to human error. The prescriptive analytics stage now aims to provide a recommendation of the (near) optimal operation configuration. Although a number of recommendations is suggested, the final decision still remains in the hands of the human operator who can overrule the recommendation according to given experience or additional information, which is only available to him/her.

As indicated by the literature review, a multitude of different frameworks coexist and a variety of methodologies are currently exploited towards prescriptive analytics tasks. According to the presented reviews, this plurality of approaches provides no single method which is currently proven superior in the field of prescriptive analytics. In addition to the presented technologies and frameworks, we have already successfully utilized HeuristicLab [467] in real world predictive maintenance scenarios [504]. Although HeuristicLab hasn't been mentioned in the presented surveys, we do consider it a suitable approach to create a prescriptive analytics solution as it contains a variety of different preprocessing and modeling techniques and can be easily extended. Due to our experience and the extendability of HeuristicLab, we aim to improve existing predictive analytics methods towards prescriptive analytics. In addition to the functionality

provided by HeuristicLab, we will also consider Python and the feature rich scikit-learn library [355] for a multitude of tasks (preprocessing, modeling, forecasting etc.). Depending on several criteria such as deployed operating systems, available computation power and memory, software architecture, communication protocol between the different modules in the *FELICE* project (which are currently subject to change due to the ongoing process of defining/designing the final architecture of the *FELICE* project) and the optimization criterion, the chosen methods and frameworks might change slightly. In summary, independently of the specific architectural decisions, the presented state of the art includes a broad enough spectrum of technologies, frameworks and methods to cover all four stages, from descriptive up to prescriptive analytics.

12.4 Assembly line balancing

Prescriptive analytics is dependent on a certain problem definition with one or multiple optimization criteria (either single or multi objective optimization). In terms of problem definition, the assembly line balancing problem (ALBP) is a well-established topic in operations research and can be used to formulate the problem.

12.4.1 Baseline technologies and tools

One of the earliest and simplest ALBP definition is the simple assembly line balancing problem (SALBP) [61]. Although the assumptions of this problem definition are thoroughly defined [61], for a better understanding a very condensed and simplified overview is as follows: A set of tasks with specific task times and a (partial) precedence between the tasks is specified. A predefined number of workstations in sequential order is available, creating a paced assembly line. Each of these workstations is capable of performing an arbitrary subset of the given tasks, with the task time being independent of the specific workstation. As constraints, each of the tasks has to be executed, the precedence must be adhered to and the sum of all task times must not exceed the cycle time. Based on this simple problem definition, several criteria can be optimized, such as assuming a fixed cycle time and minimizing the number of workstations (which is known as SALBP-1) or vice a versa, assuming a fixed number of workstations and minimizing the cycle time (known as SALBP-2). As a generalization to the rather strict definition of the SALBP, the generalized assembly line balancing problem (GALBP) was introduced, relaxing most of the given assumptions. Since the early definition of the SALBP and GALBP, many derivatives of the original problem definition and methods for their optimization have emerged [62] and were classified in [76]. Of particular interest is current research which includes collaborative robots (cobots) [120, 480, 478], the skillset of operators [119] or the efficiency of the operators [119] in the design of ALBPs.

12.4.2 Discussion

Due to the long history and maturity of research in the field of assembly line balancing, we propose to use ALBP to formalize the problem definition, which should be optimized

in the context of the prescriptive analytics task. The reviewed methodologies [62] used to solve ALBPs overlap with the methodologies used in the context of prescriptive analytics [154, 257, 258], promising great compatibility. Due to their mathematical definition, ALBPs can be easily customized, providing great flexibility given the iterative development cycles in the *FELICE* project. The first definition may only be a SALBP and that will be continually extended as the project advances.

13 AI-driven digital twins and digital operators

13.1 Overview

Human-robot collaboration is one of the fastest-growing focus areas for digital twins, with research studies focusing on assembly lines, predictability of robot motions or smart wearable devices [46]. Current human-robot collaboration (HRC) assembly systems lack sufficient adaptability to update their strategy quickly when assembly environment changes. At the same time, the perception ability towards environmental data is weak and the assembly system lacks strong cognitive ability. The Digital Twin-enabled HRC assembly system is an appropriate way to meet the flexible assembly requirements.

13.2 Relationship with FELICE project

Digital twins harness the capabilities of physics-based simulations and data analytics to create new insights in a fully virtual environment. To this end, continuously monitoring the assembly line and forecasting the performance using AI-driven digital twins enables swift reaction supporting dynamic and flexible assembly lines [281, 410]. The deployment of robots becomes more dynamic as well as the scheduling of tasks in flexible lines where a task can be performed on several workstations. Sophisticated, interactive simulation-based digital twin models will provide a digital learning environment for human operators, but also for integrated AI methods. Challenges are in the real-time synchronisation of the AI-driven digital twin and the real-world system. Within *FELICE*, we aim to construct such digital twins of both machine/equipment and human operators with the AI methods integrated to achieve a fully digital model of complex assembly processes. AI driven Digital Twins will be able to provide both online (real-time) and off-line (what-if analysis) decision support. The AI driven Digital Twin is a "service-oriented knowledge-aware expert system" which is fully synchronized with the physical system and capable to operate on it. It will be designed and implemented as a holistic software platform, which leverages on the latest 4.0 paradigms and offers a set of comprehensive services for both the assembly line and the digital operator.

13.3 Digital twin

13.3.1 Baseline technologies and tools

A recent keynote paper of CIRP (Society of Production Engineering) pointed towards the importance of digital twins for HRC production systems (Wang et al., 2019 [472]). Grieves (among the earliest researchers in the field of Digital Twin) has also presented the potential value of digital twins for the development of the field of cobotics. The significance of digital twins for HRC assembly systems has been highlighted, e.g., by Bilberg & Malik (2019) [67] where the relevance of digital twins in relation to human-robot collaboration was discussed. It was argued that digital twins can support HRC systems, however, the study remained limited only to the HRC challenges in the operational phase without a lifecycle approach. The new "lifecycle" approach is the concept of a digital twin [102] an intelligent digital representation of a physical system enabled

by the advancement in virtualization, sensing technologies and computing power [428]. Two comprehensive works on digital twins have been by Qi et al. (2019) [374] and Lu et al. (2020) [281]. However, these studies do not cover the human-robot interaction aspects specifically.

Malik & Brem (2021) [291] provide an overview of the domains where the digital twin can help in systems engineering of an HRC production system:

- the *design phase*;
- the *integration phase*;
- the *operation phase*.

The functions of monitoring, prediction and optimization of the digital twin are compatible for HRC assembly. The prediction module is used to predict the assembly task and the state of assembly system. According to the collected data, the optimization module will provide an optimal solution for each assembly state. For example, in the case of a dynamic market demand, Lv et al. (2021) [283] proposed a digital twin-based HRC assembly system integrating all kinds of data from digital twin spaces to improve the assembly efficiency, safety and accuracy while reducing the workload of the operator. Due to the functions of monitoring, prediction and optimization in digital twins, they can provide an effective cooperation strategy for HRC assembly. Malik & Bilberg (2018) [289] used the digital part which was continuously mirroring the physical part to simulate the assembly plan. The proposed digital twin framework could carry out on-line or off-line experiments, avoiding any economic loss and personal injury in the actual production. In this sense, simulation-in-the-loop digital twin is used extensively in HRC assembly industry to test and develop HRC concepts and setups. Kousi et al. (2019) [244] used the digital modeling technology in production system, making the system reconfiguration realized through shared environment and process awareness. The suggested digital world model infrastructure involves three main functionalities:

a. Virtual representation of the shopfloor, combining multiple sensor data and CAD models. The digital shopfloor is rendered in the 3D environment exploiting the capabilities provided by Robot Operating System (ROS) framework.

b. Semantic representation of the world through the implementation of a unified data model for representing the geometrical as well as the workload state.

c. Dynamic update of the digital twin based on real time sensor and resource data coming from the actual shopfloor.

They argue that future work will involve integration of the Digital Twin with:

a. the physical robotic set up to validate its performance and,

b. high-level decision-making mechanisms allowing the reconfiguration of the system in shopfloor level through task re-allocation based on the real time production needs (new product variants etc.).

Bilberg & Malik (2019) [67] have established a corresponding digital twin of a flexible assembly cell coordinated with a robot to perform assembly tasks alongside humans in which the use of an object-oriented event-driven simulation model is extended to:

- Rapid skills-based workload balancing and task distribution between human and robot for product variety.
- Dynamic workload monitoring during operation to account for human factors.
- Online optimization of robot trajectory generation of robot control program.

Digital twin technology has also been proven useful for HRC workplace layout design. A method for HRC workplace design and task planning was described by Tsarouchi et al. (2017) [449] that also comprised virtual facility layout and evaluation of alternative designs. Both human and robot were modelled in a unified simulation and the model is used for task allocation problem. For designing an HRC workstation, the major challenges for a digital twin are:

- skill-based tasks distribution between robot and human;
- quick adaption and validation of workstation layout;
- virtual commissioning of the designed production system, and
- ensuring safe working conditions for fellow human worker(s).

13.3.2 Discussion

After a careful analysis of the literature, different gaps and open questions emerged. The role of digital twin technology in human-robot collaborative systems is still uncertain and under development. The main points to discuss are the following:

1. A strong modelling effort is still needed to design and validate a digital twin of HRC assembly systems. There are no specific tools or software that are able to automate and speed up the process, which still relies on the modellers capabilities.
2. Simulation-based digital twins started to emerge but only as standalone applications. Interoperability with manufacturing execution systems, robot operating systems and other shop-floor level applications are highly desirable in order to develop context-aware digital twins. Furthermore, the services offered by the digital twin are usually separate (e.g. the task allocation strategy is usually separated from the planning work of robot motion path), therefore a unified approach is not available yet in the literature.
3. The current HRC assembly system lacks adaptability to update strategy quickly when assembly environment changes. This is crucial in the modern dynamic market environment. As a consequence, the perception ability towards environmental data is weak and the assembly system lacks strong cognitive ability. Digital twins of the future must be designed to perceive the context, understand and learn from it, in order to support decision-making and scenario testing.
4. Humans are a fundamental component of assembly systems, not only in human-robot collaborative environments. However, they are often out-of the-loop due to the difficulty of modelling human behaviour. Considerable modelling efforts are needed to create a digital twin that reflects human facets and include human behavioural aspects.

5. Virtual Reality (VR) has not yet been clearly established as a design and validation support tool of a HRC system design. Future digital twin solutions should consider the use of Virtual and Augmented Reality (AR) to immerse humans in the cyber-physical world and support them to explore and exploit fused data.

6. A lifecycle approach is required to enable a continuous and real-time synchronization between the real and virtual spaces in the shortest feasible time (e.g. constant communication and data transmission between the DT, the robot and the adaptive workstation is a crucial aspect). However, data exchange among heterogeneous systems is a major challenge.

14 Orchestration of adaptive assembly lines

14.1 Overview

Orchestration of flexible human-robot based assembly lines raises a clear need for more flexible and reconfigurable manufacturing systems. There are various reconfiguration aspects including hard (physical) and soft (logical) reconfiguration [139].

The Plug&Produce concept presented in [48] is a promising approach to manage the complex and flexible layouts of production systems, i.e. a hard or physical reconfiguration of hardware components. The concepts of skill-based engineering presented in [350, 227, 446, 147] could solve the challenges of soft or logical reconfiguration. Although several promising domain-specific modeling approaches to manage the flexibility of production control systems have been proposed, the general acceptance in industrial applications is generally low [320]. The reasons for this are either that some approaches focus on a very specific application domain, or in general, that many of the proposed tools do not take the end-user in the companies into account [465]. Therefore Plug&Produce and skill-based automation must become more intuitive and quickly manageable for the end-user.

14.2 Relationship with FELICE project

In *FELICE*, an assembly line is to be modernized and prepared for flexible, agile manufacturing. For this purpose, a variety of different technologies and systems are used, which should make it possible to completely model, simulate, validate and optimize the workflow of an employee and, if necessary, replace individual steps by a mobile robot. To make this possible, a central orchestration layer is required that monitors ongoing processes in the assembly line and passes orders on to said mobile robots. In addition, the goal is for the orchestration system to continuously learn from previous process data and optimize the sequence or execution of individual work steps. Various sensor data from e.g. the adaptive workplace as well as indirect or direct interaction with the employee have an influence on decisions made by the orchestration system.

To ensure the versatile use and flexibility of the assembly system, it must be possible to integrate relevant subsystems via standardized interfaces or data structures. The concept of skill-based programming and engineering is used here, since this allows components, machines and systems to be addressed via their abstracted functionalities (skills) without having to know the respective implementation in detail. In addition, this enables the simple replacement of individual components without having to make changes to the generic process execution in the orchestration system.

14.3 Assembly line orchestration

In order to enable flexible and eventually real-time orchestration of adaptive assembly lines, we focus on a two-stage approach. In the first stage, a workflow-based orchestration process will be developed to enable the specification of assembly tasks including their sequence and all required skills to fulfil the production process without assuming any platform-specific constraints. At the second stage, a universal communication

interface between the high-level orchestration unit, featuring a workflow-based runtime system, and the low-level modules is utilized that supports the execution of the orchestrated workflows.

This approach is supported by the WORM tool [271, 405] (see Figure 15) which is to be integrated into the developed modules. WORM will transform task-level specifications to robot-level specifications for task-level programming. The underlying skill-based engineering mechanism using OPC UA [34] allows the on-the-fly mapping of required skills with available production skills in the manufacturing system. Here, we expect to contribute to a new standard that defines in what way robotic manufactures represent and provide such skills.

As a result, it is possible to orchestrate generic workflows including all required skills, which could automatically be executed on any manufacturing system with suitable skills. This implies that all skills and their order may be dynamically re-allocated by the prescriptive analytics system, as described above.

Another aspect of the orchestration is path planning. The dynamic assembly line introduced in *FELICE* necessitates a topological representation able to adapt to changes in the environment, which then serves as basis for mobile robot routing decisions. A topology generator will be developed that fuses geographic and semantic information of the environment into a common topology representation, allowing for high flexibility. A global path planner will provide optimal routes, satisfying spatial constraints on separation of robots and workers.

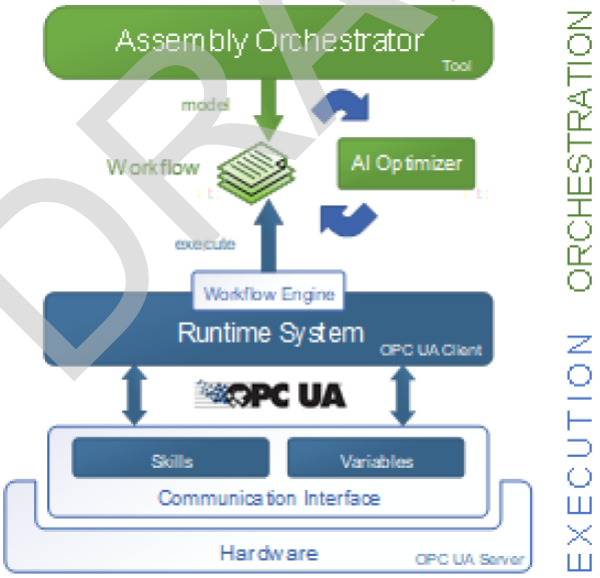


Figure 15: Orchestration overview.

14.3.1 Baseline technologies and tools

Current research topics focus on solving several technical challenges stemming from the upcoming mass customization desired by customers. All systems of a production system must be flexible enough to be capable of a quick change in the assembly workflow and should be able to handle the increasing variability of product configurations. The concept of skill-based engineering tries to manage these challenges of flexible (robotic)

production systems. Instead of coding static and specific robot programs, the developers can define skills to execute different tasks. The skill definition should be used like instantiating generic black boxes that accomplish different work tasks. All following works being presented use the base concept of a skill-based engineering process or a generic approach for modeling (collaborative) assembly tasks i.e. automation systems.

Focusing on sequence-oriented (discrete) automation and production processes, common design/modeling approaches are workflow or finite-state-machine based, like Sequential Function Charts in IEC 61131-3 [32, 348], Function Block Diagrams in IEC 61499 [33] or Open Platform Communications Unified Architecture (OPC UA) based skills i.e. OPC UA methods or programs [129, 34]. Taking a look at more high-level description languages, e.g. the Activity Diagrams of the Unified Modeling Language (UML) [388] or the Business Process Model and Notation (BPMN) [177] are common approaches to define specific processes in a user-oriented and abstract way [135].

Thomas et al. [446] present a new skill-based robot programming language and introduce a domain specific language (DSL) called *LightRocks* (Light Weight Robot Coding for Skills). They use state-charts of the Unified Modeling Language for Programming (UML/P) to describe different levels of detail during defining assembly processes for e.g. industrial workers or robotics experts. Prähofer et al. introduce a domain specific language for programming event-based, reactive automation solutions [368]. Pedersen et al. focus on skills especially designed for manufacturing systems and propose for example a method for an intuitive programming method for industrial mobile robots [351, 352, 354]. They combine robot skills, a graphical user interface and human gesture recognition. In addition, Pedersen et al. present robot skills designed for transformable manufacturing systems to perform a variety of tasks enabled by simple task-level programming methods [350].

Brandenbourger et al. introduce a metamodel which allows an efficient and flexible engineering of automation components [78]. They use skills describing reusable production steps of the automation components. The automatic device discovery in the context of Plug&Produce is in the focus of the work presented by Profanter et al. [370]. They use OPC UA Local Discovery Services with Multicast Extension (LDS-ME) and define services (i.e. basic actions) of devices with the help of skills. Dorofeev et al. work on a device adapter concept, that enables Plug&Produce production environments [128] and suggest a skill-based engineering approach with the help of OPC UA programs [129]. As a result, it is possible to use a device adapter for wrapping device functionality as service and hide the low-level skill implementation. The corresponding skills can be executed and triggered with OPC UA programs.

Danny et al. add a decisional attribute to an existing skill concept [121]. In addition, the authors introduce a way to define task execution tables and corresponding Relationships between the Product, Process and Resource (PPR) domains. Steinmetz and Weitschat present in [429] a new software architecture for robot skills and introduce basic demands on the parameter setting of skills. These speed up the engineering process and make the process more intuitive for the user.

Weichhart et al. analyze different modeling approaches to represent tasks that are shared between human beings and robots [481]. Michalos et al. work on methods for planning human robot shared tasks and try to extract assembly sequences of a product out of CAD models [310]. Keddis et al. examine new ways of modeling production

workflows in the era of mass customization and present a metamodel for these workflows [227].

Tsarouchi et al. introduce a decision making framework for the layout generation of a Human Robot Collaboration (HRC) workplace design to decrease the reconfiguration or set-up time of a workcell [450]. The authors in [309] suggest design considerations for safe HRC workplaces. Stöhr et al. focus on applications with elderly people and those with disabilities [432]. They work on user-centered work instructions shown by multi-modal user interfaces. Zor et al. present a proposal to extend BPMN especially for the manufacturing domain [510].

Most approaches presented above focus either on the engineering process of orchestrating assembly processes with skills for devices or robots but not on applying skill-based orchestration for both.

14.3.2 Discussion

In the context of the *FELICE* project, the following questions need to be further addressed for full implementation and correct execution of the orchestration system.

- **Metamodel design:** Individual generic components of the Asset-Decision-Action-Property-RelaTionship (ADAPT) metamodel [271] must be adapted or extended for the *FELICE* project in order to enable correct use. The resulting set of rules must be compatible with all relevant production systems.

- **Workflow design:** Different options for modeling process workflows must be weighed and evaluated against existing technologies, interfaces, and subsystems within the project. It needs to be determined whether workflows are created using, for example, predefined structures, templates or sub-workflows and who designs them, or whether it is even possible to generate the entire workflow completely and fully purely on the basis of existing data points in the knowledge base.

- **Workflow modification:** Workflows should be continuously optimized based on historical data records. To make this possible, interfaces to the relevant systems must be defined and the exchange of workflow-relevant data must be established. In addition, it must be clarified how workflows are modified. In concrete terms, this means whether, for example, only individual steps can be changed in the sequence or whether the content of entire workflows can be changed.

- **Workflow deployment:** Suitable deployment strategies for the modified workflows to the executing runtime systems must be defined and evaluated (e.g. queuing of workflow instances). Interactions between the local runtime systems and the global orchestration layer must also be clearly defined and prepared for any potential errors and communication problems. This includes e.g. whether the local system proactively requests the supervising layer for new workflows or whether these are sent out by the orchestration system to subordinate systems.

- **Skill-based programming:** The concept of skill-based programming is used to execute individual workflow steps. The systems in question provide abstracted functionalities in the form of skills via a standardized interface. OPC UA is proposed

as the communication protocol for this purpose. This interface must be defined in detail for the *FELICE* project, with the contents from [425] being proposed for the general structure.

- **Workflow performance:** For the heuristic optimization procedures for the ongoing adaptation of workflows, it is necessary to record performance metrics such as execution times of individual workflow steps, store them and make them available to the optimization system. A suitable file format or data interface must be defined for this purpose.

15 Computing infrastructure

15.1 Overview

In this section we will describe technologies and tools, which *FELICE* will take advantage of, in order to communicate information generated or simulated by different components. In particular, components such as messaging buses with sensory inputs, robotic operating systems, and simulation engines will be considered in the computing infrastructure.

15.2 Relationship with FELICE project

Machine-to-machine (M2M) and machine-to-human (M2H) communication is a natural requirement in an industrial environment when the reduction of human errors and manual labor and the overall increase in efficiency both in terms of time and money is sought. Supporting real-time, adaptive and effective human-robot collaboration will require significant amounts of data to be collected, aggregated and shared in a meaningful way. Internet of Things or **IoT** are data-rich frameworks which facilitate the increase of the level of automation in sensors collection and workflow processing by extensive use of API between computing and communication components. In the context of industrial environments where humans and machine collaborate, the latency of decision control systems is of paramount importance, hence an increased automation is required. *FELICE* will require an **Industrial IoT** (**IIoT**) platform where machine sensors and controls are integrated with industrial level accuracy of a set of smart edge devices which can exchange and process data closer to the locally-deployed sensors/actuators. Sensors and robots deployed at a local layer have limited capabilities but modules that running at the global layer will be the drivers of an innovative intelligent orchestration of *FELICE*, which must have adequate processing power to fuel the underlying AI-based processes.

Therefore, a powerful computing infrastructure is required for complex computation of data collection analytics and Machine Learning (ML) models in the contex of *FELICE*.

15.3 Industrial IoT platforms

Realistic any usable Industry IoT applications should support the creation of smart assembly line surveillance environment, with the interconnection and networking of a large number of heterogeneous smart objects and IoT solutions, covering different communication technologies. Various standardized “IoT communication protocols” have been proposed such as the MQ Telemetry Transport (MQTT [55]), the Devices Profile for Web Services (DPWS [131]), and the Constrained Application Protocol (CoAP [415]). In Industry IoT platforms, the business intelligence of the application should be distributed among the cloud and the edge devices. It is important for *FELICE* to aggregate sensor information from a generic industrial shop floor but also to incorporate robotic environments.

Aiming to alleviate the interoperability issues, various commercial and free IoT platforms are emerging, such as EdgeX, Azure IoT Edge, Amazon Greengrass, which in-

corporate raw events as *topics* in order to achieve a Publish-Subscribe platform where signals are shared among a limited set of participants. A service layer consumes signals (i.e. topics) to construct a business oriented workflow. Those service layer are commonly domain-specific (e.g. UniversAAL [29], Home-Assistant [8] for Home Automation or in Smart City environments, the FIWARE [445]). FIWARE has evolved as an open source and open standards initiative, defining a layered set of standards for context data management to facilitate the development of solutions for different domains [9].

Since 2018, FIWARE has been selected by the European Commission, as a *Connecting Europe Facility (CEF) Building Block* [10], This means that the EC officially recommends public administrations as well as industrial players within the European Union (EU) member states to adopt this technology in order to foster the development of digital services, which can be replicated (ported) across the EU.

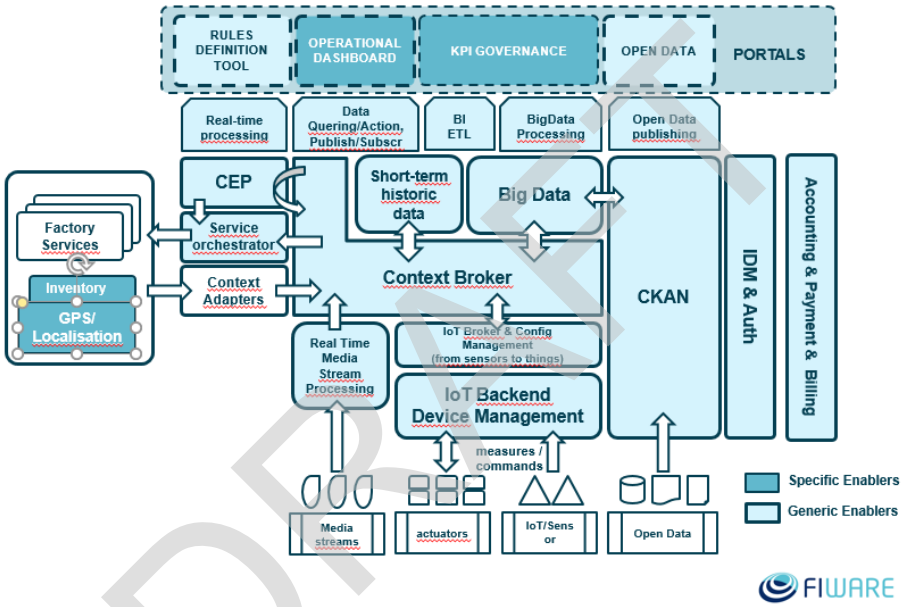


Figure 16: FIWARE components overview [445]

The Orion Context broker [11], retrieves the essential information from the ambient space and manages the interactions between the different systems and components of the smart space in a context-aware manner. A variety of sensors and wireless communication technologies can be orchestrated in the FIWARE IoT platform to support the smart assembly environment and gather relevant contextual information from the shop floor and worker's actions and objects in the vicinity of the robot.

[8] https://www.home-assistant.io/docs/
[9] https://marketplace.fiware.org/pages/solutions
[10] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Digital+Home
[11] https://fiware-orion.readthedocs.io/en/master/

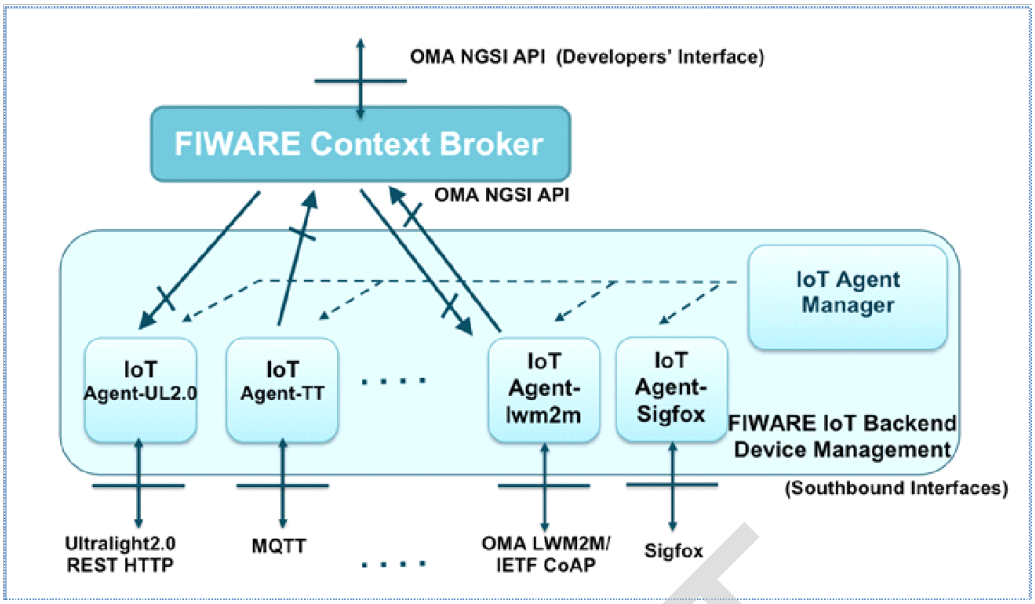


Figure 17: FIWARE IoT technologies [445]

The functionality of FIWARE is partitioned into a set of general-purpose platform functions available through APIs called *Generic Enablers (GEs)*, in terms of software modules. The set of the open and royalty free GEs constitute the FIWARE Reference Architecture [12]. Domain specific implementation of FIWARE platform is realized from the superposition of GE functionality along with Specific Enabler (SE) functionality. SEs provide market differentiation and may not be royalty-free.

FIWARE was the first IoT-ready cloud to support real-time communication (through a middleware -Fast RTPS) between robots and the Orion Context Broker [7], [4]. In future-proof industrial environments [30] the OPC UA [287] [13] (Open Platform Communications United Architecture) is the only recommended modelling solution. Therefore, FIWARE via its OPC UA IoT Agent is able to interconnect Industrial IoT Data in motion streams coming from manufacturing plants with other heterogeneous data sources and datasets. Furthermore, FIWARE has developed functionalities [14] for robotics.

In the context of *FELICE*, FIWARE has been initially selected as the message exchange middleware, as shown in Figure 18.

[12] https://www.fiware.org/developers/catalogue/
[13] https://opcfoundation.org/
[14] https://github.com/Fiware/catalogue/tree/master/robotics

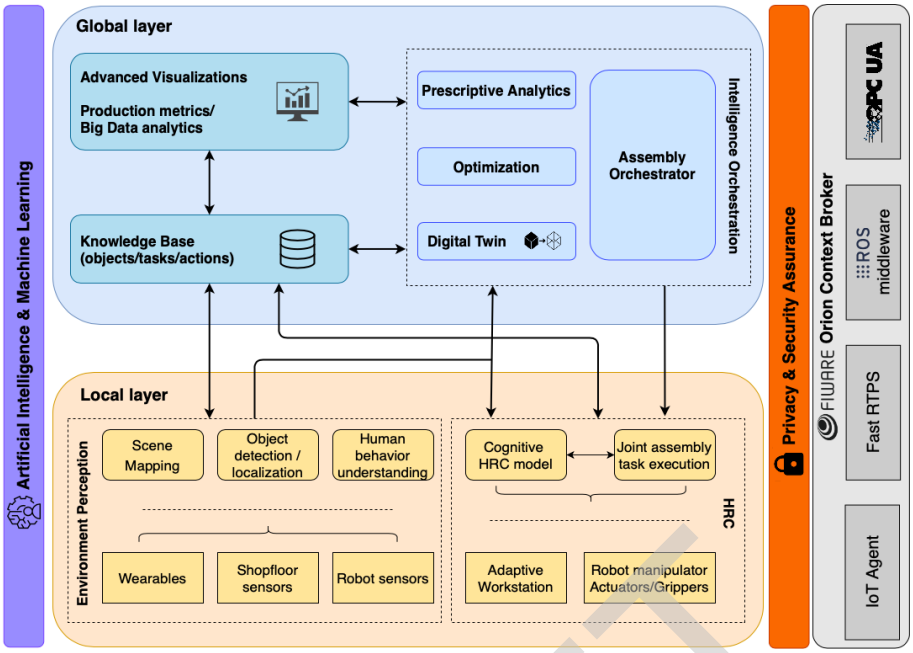


Figure 18: *FELICE* components overview

Apart from removing interoperability barriers, FIWARE has developed modules for security and privacy utilizing the eXtensible Access control Markup Language (XACML [342]) used in FIWARE-identity manager [15] [5] as the de facto standard for specifying and evaluating access control policies [272]. FIWARE may be operated in cloud native environment Openstack [17] with Identity management and Policy Enforcement point (PEP) access control [5] for tight access control of resources (i.e. signals, images) which is necessary in symbiotic and privacy-cautious environments.

In tight-latency (i.e. $<$10ms) domain-specific FIWARE deployments, with a need for distributed computing, there is a partition of the computing infrastructure between cloud and edge devices as shown in Figure 19 in order to meet the stringent time constraints.

[15] https://fiware-idm.readthedocs.io/en/latest/

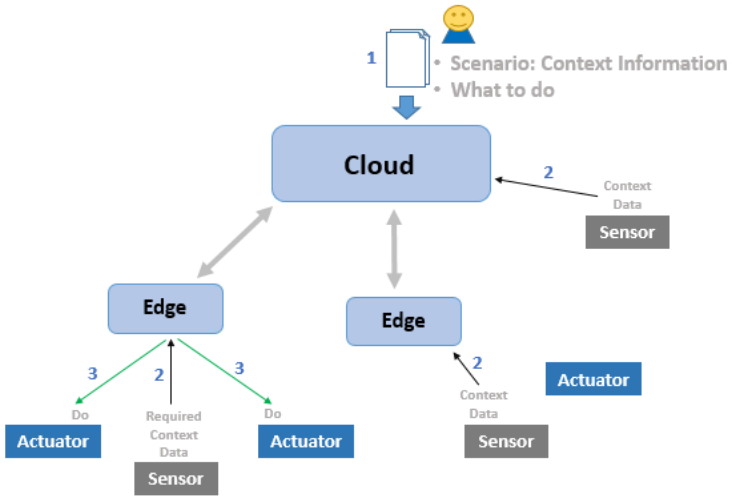


Figure 19: Distributed edge-cloud execution environment [8]

In distributed cloud-edge IoT environments, sensors from the edge devices are exploited locally to minimize latency and to increase distributed intelligence. In complex workflow environments, the overall computation occurs through a dynamic *cloud* orchestration mechanism (to differentiate it from the previously mentioned robot orchestration).

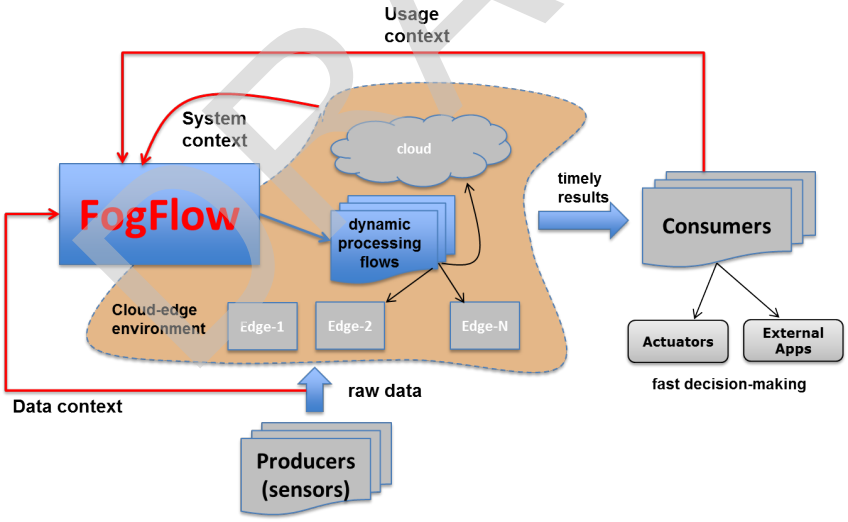


Figure 20: FogFlow high level view [8]

A typical distributed cloud-edge FIWARE environment is the Fogflow [8], which allows service execution objectives to match latency criteria as defined by a user and system context [16]and shown in Figure 20.

[16]https://fogflow.readthedocs.io/en/latest/introduction.html

15.4 Robotic software platforms

The collaboration of humans and robot is a first class objective in *FELICE*, thus the project is foreseen to encompass extensive research in robotics. The diversity of robotic research areas along with the complex requirements of hardware and software for robotic systems have always presented a challenge for system developers. Previous robot control platforms were complex, expensive, and not very user friendly. The Robot Operating System (ROS)[17] is a flexible framework, providing various tools and libraries for developing robotics software. It started in 2007, with the name Switchyard, as part of the Stanford STAIR[18] robot project. Nowadays, it offers several powerful features to help developers in tasks such as message passing, distributing computing, code reuse and implementation of state-of-the-art algorithms for robotic applications. Many researchers select ROS over other robotic platforms, such as Player [19], YARP [20], Orocos [21], OpenRTM [22], MRPT [23], MOOS [24] because it offers a vibrant community [25], supports high-end sensors, i.e. kinect[26], simultaneous Localization and Mapping (SLAM)[27], Adaptive Monte Carlo Localization (AMCL)[28], motion planning[29] of robot manipulators. ROS provides debugging, visualization[30], simulation-interfacing tools[31] with popular simulators i.e. Gazebo[32]. The software modules come in the form of packages bundled in a distribution (http://wiki.ros.org/Distributions). Distributions evolve around the normal operating systems they target on. The current distribution i.e. the Noetic [33] is the 13th one, which is primarily targeted at the Ubuntu 20.04 (Focal) release among others [34].

ROS is more than a development framework, however, it is a meta-operating system, as it does not only offer tools and libraries but even OS-like functions, such as hardware abstraction, package management, and a developer toolchain. Like a real operating system, ROS files are organized in a particular manner, as illustrated in Figure 21:

[17] https://www.ros.org/about-ros/
[18] http://stair.stanford.edu/
[19] http://playerstage.sourceforge.net/
[20] http://www.yarp.it/git-master/
[21] https://orocos.org/
[22] https://www.openrtm.org/openrtm/
[23] https://www.mrpt.org/
[24] https://www.robots.ox.ac.uk/~mobile/MOOS/wiki/
[25] (http://answers.ros.org)
[26] http://wiki.ros.org/kinect
[27] slam toolbox http://wiki.ros.org/slam_toolbox
[28] amcl http://wiki.ros.org/amcl
[29] moveit https://moveit.ros.org/
[30] Rviz http://wiki.ros.org/rviz
[31] gazebo ros pkgs, http://wiki.ros.org/gazebo_ros
[32] Gazebo simulator, http://gazebosim.org
[33] http://wiki.ros.org/noetic
[34] https://www.ros.org/reps/rep-0003.html

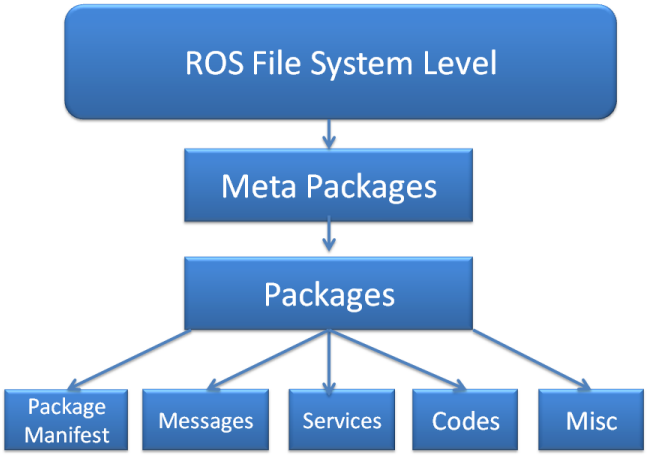


Figure 21: ROS high level view

The ROS packages are the basic units of the ROS system containing one or more ROS programs (nodes), libraries, configuration files, which are organized together as a single (software) unit. The ROS messages (.msg) are a type of information that is sent from one ROS process to the other. The specification of a message resides inside the message folder of the package. Similarly a ROS service is a kind of request/reply interaction between processes. The specification of the interaction resides inside the srv folder of the package.

In a typical **IIoT** environment operating with low latency requirements, robotic platforms may need to utilize the Fast RTPS, or DDS FIROS 2 [6] to share their data. DDS (Data Distribution Service) of the OMG (Object Management Group) is a RTPS (Real Time Publish Subscribe) protocol standard, which provides publisher-subscriber communications over unreliable transports such as the User Datagram Protocol (UDP), typically used in communication between ROS modules. Those functionalities have been developed commonly by the FIWARE and robotic platform communities.

15.5 Simulation of robotic environments and workflows

Simulators play an important role in robotics as tools for testing the efficiency, safety, and robustness of new algorithms. This is of particular importance in scenarios that require robots to closely interact with humans i.e. *assistive environments*, collaborative robotics. A typical workflow scenario in *FELICE* originates after receiving a user-specified task plan, where a corresponding end-effector pose is calculated. Via the simulator, a trajectory of robotic plan is calculated, using its robotics manipulation platform, so that both the simulated robot itself and the object carried by the robot do not collide with the environment. If the user approves the simulated trajectory plan, the robot will be notified with an approval message to convert the approved trajectory plan into robot control commands [475]. Testing the robustness of the performance of a robotic design by modifying only specific environmental conditions is one of the benefits of simulation software. Consequently, with regard to simulation rendering accuracy, physics simulation accuracy is essential. For that reason, Gazebo supports four different engines:

"ODE - Open Dynamics Engine" [35], "Bullet Physics" [36], "DART - Dynamic Animation and Robotics Toolkit" [37], and Simbody [38]. This wide choice offers robustness and flexibility, as it is possible to select the best-suited engine for each specific task in a project. Bullet Physics, in particular, achieves state-of-the-art performance and accuracy, a reason for which it is commonly used in reinforcement learning research [359, 223]. Regarding rendering quality, Gazebo uses the OGRE [39] rendering engine. Its capabilities are not on par with of state-of-the-art photorealistic engines such as *Unreal Engine*, *Unity3D* or *Nvidia Omniverse*. ROS provides bridges, i.e. ROS# [40], to interconnect with non ROS native technology engines such as Unity3D. Nvidia is developing a simulator within its Isaac framework [41], based on Nvidia multi-GPU Omniverse [42] real-time simulation platform for 3D production pipelines based on Universal Scene Description ray tracing technology to produce extremely high photorealistic quality. Furthermore, IsaacSim uses Unity3D as the simulation environment for Isaac robotics, providing an infinite stream of procedurally generated, fully annotated training data for machine learning with emulated sensor hardware, robot base models, scene randomization, and scenario management.

Simulated environments require a **Digital Model (DM)**, a digital version of a pre-existing or planned physical environment with objects. Examples of a *DM* could be, but not limited to, plans for buildings, product designs etc. Robotic simulation engines offer *Unified Robot Description Format*[43], an XML specification, for models containing information about robot mechanical, kinematic and dynamic description, visual representation, and collision model. If a virtual model represents the physical model only, with one-way data flow, this is considered to be a **Digital Shadow (DS)** [221], as shown in Figure 22. Once a *DM* is created, an evolutionary change in the environment, i.e due to human movement, has no representation on the *DM*, hence there is need of a virtual copy of the *(DM)* of any physical entity (physical twin) to the simulated world via data exchange in real time [326].

[35] https://www.ode.org/
[36] https://pybullet.org/wordpress/
[37] http://dartsim.github.io/
[38] https://simtk.org/projects/simbody/
[39] https://www.ogre3d.org/
[40] https://github.com/siemens/ros-sharp/
[41] https://developer.nvidia.com/isaac-sdk
[42] https://developer.nvidia.com/nvidia-omniverse-platform
[43] www.ros.org/urdf

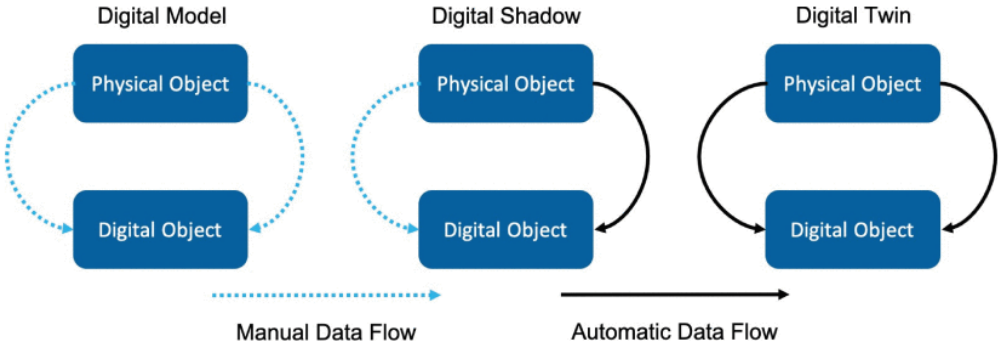


Figure 22: Digital model vs. digital twin [157]

This constitutes the foundation of a *Digital Twin (DT)*,a virtualisation counterpart of a simulated robotic environment of digital objects which can be used for predictive analytics [136], in hypothetical situations [326]. Hence, in order to predict the CPS actions, the DTs virtually replicate physical world assumptions in order to interconnect both worlds.This is typical Cyber-Physical System (CPS) as shown in Figure 23, where CPS modelling is essential to accurately virtualise the operations of the physical world. Not only do DTs virtualise processes, but also generate high-value data for production efficiency [298].

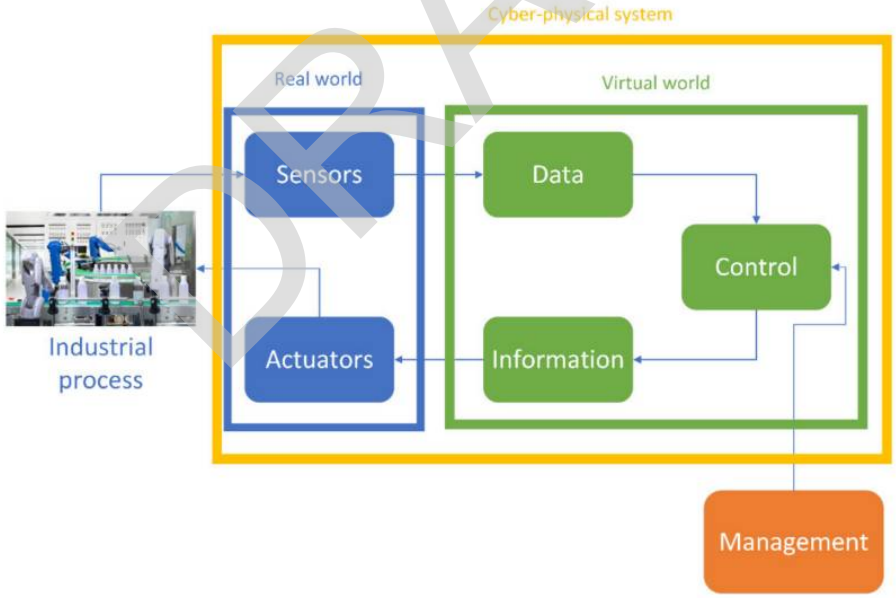


Figure 23: Robotic cyber-physical system

In more general aspect *Industry 4.0* manufacturing systems are equipped with Cyber-Physical Systems that are characterized by a strong interlinkage between the real world and the digital one: actions in one world have an impact on the other. In this paradigm, Digital Twins (DT) are defined as simulation models that are both getting data from the field and triggering actions on the physical equipment. More details regarding digital twins can be found in Section 13 of this document.

15.6 Discussion

Most of the aforementioned tools and components will be considered while selecting the necessary building blocks in *FELICE*. Although FIWARE constitutes a mature software component, there is still the need to verify its latency performance in the production environment of the automotive industry.

16 Data privacy, vulnerability management, and security assurance

16.1 Overview

In this section we will describe technologies and tools aimed at protecting person-identifiable information recorded by different *FELICE* components. In particular, implementations of anonymization and encryption algorithms, identification of vulnerabilities, authentication, authorization and accounting will take place at different levels.

16.2 Relationship with FELICE project

The General Data Protection Regulation (GDPR) is a key consideration in *FELICE*, hence effort addressing relevant ethical, legal and privacy concerns is allocated to a specific task (i.e., T1.3) which will continuously run during the whole project timeline. To ensure GDPR compliance, *FELICE* will "sanitize" sensitive information before its further processing, with privacy-achieving guarantees reinforced and ensured by a well-defined strategy for ethics, privacy and data protection compliance. At the local layer, devices and robots deployed in the work environment will record person identifiable information, such as voices and faces. Nevertheless, none of this information must reach the components of the global layer, therefore *FELICE* plans to protect personal data in two ways: (i) anonymization and encryption algorithms for human identifiable information (voice, body, face characteristics) will be deployed during process/robot monitoring at the local layer so that only extracted features without identifiers are uploaded to the cloud for global training of the ML models; and (ii) data processing, decision making and human-robot collaboration will be mostly executed in real-time at the local layer (edge) (WP4, WP5) so as to entirely avoid the need of information flow to the cloud. Constant monitoring of robot actions (T3.2) will also allow quick identification of vulnerabilities and will strengthen automatic vulnerability management imposed by the safety protocols implemented in the robotic hardware (T5.1). Regarding security, *FELICE* will seek a solid implementation of authentication, authorisation, and accounting at all levels. The required protection mechanisms will span across all communication interfaces and data handling modules. Access control and identity management delegation to all layers of the platform will allow for auditing of data access and support conditional access to data and generated analytics.

16.3 Anonymization, authentication, authorization and vulnerability scanning

Anonymization can be utilized to achieve privacy preservation, by removing personal identifiers, both direct and indirect, that may lead to an individual being identified. Anonymized data must have two properties. First, they should be irreversible and second, it should be extremely impractical or even impossible to identify the data subject. One anonymization technique is k-Anonymity. Trying to hide personal information of an individual in a dataset among at least k-1 others with respect to quasi-identifiers,

k-Anonymity is a model for protecting privacy and was introduced by Samarati and Sweeney [397]. Achieving optimal k-anonymity is an NP-hard problem as proved by Meyerson and Williams [308]. In practice, many different algorithms have been proposed towards achieving optimal k-anonymization in datasets by Bayardo et al. [60], Park et al. [343] and Kenig et al [229]. l-Diversity was proposed by Machanavajjhala et al. [285] as an extension to the k-Anonymity model and its main concept is to promote the diversity of the sensitive values within an anonymized group. Just like in the k-Anonymity model, it is equally hard to achieve optimal l-Diversity in practice, as stated by Xiao et al [489]. t-Closeness (proposed by Li et al. [267]) is a further refinement of the l-Diversity model that reduces the granularity of the individuals data. This reduction is a trade-off that results in some loss of effectiveness of the data in order to gain some privacy. Differential privacy [137] is the most recently proposed approach that has been used over the last few years. Differential privacy states that any possible outcome of an analysis should be almost equally likely, independent of whether any individual is included or removed from the data set. Consequently, the data of any specific individual can never seriously affect the result of the analysis.

Under the concept of preventive security services lay different tools in various forms that try to prevent a security incident from happening by detecting vulnerable system areas. For example, vulnerability scanning tools allow for detection of vulnerabilities in different system parts and kinds. Static (code) analysis tools can find code bugs potentially exploitable by attackers. Audit tools can spot well-known rootkits, Trojans and backdoors unveiling hidden processes and sockets. Finally, antivirus tools can detect viruses which either attempt to infect or have already infected the underlying operating system (OS). Although malware and antivirus software is well-adopted, it seems that choosing the right vulnerability scanning tool is not a straightforward process due to the great tool diversity and varied coverage. Holm et al. [199] evaluate a set of popular vulnerability scanners, finding significant differences between scans of Linux and Windows hosts and accuracy of the scanners. Fonseca et al. [149] propose a method to evaluate web vulnerability scanners using software fault injection techniques. According to their results, the coverage is low and the percentage of false positives is very high. Deraison et al. [127] introduce the concept of Passive Vulnerability Scanning not as a replacement for active vulnerability scanning but as a complementary tool that produces interesting information about the security profile of a monitored network. Kritikos et al. [246] tries to connect vulnerability management to the application lifecycle, presents the state-of-the-art open-source vulnerability scanning tools and databases, and explores the possibility that combined vulnerability scanning tools can reach higher vulnerability coverage.

Basic authentication schemes can be found in a large number of works in the literature. Lamport [253] describes a basic user-password authentication scheme for the single server environment with a secure one-way encryption function implemented in the user's terminal. Since this approach allows for improvements, extensions of this study are found in the works of Yoon et al. [494], Guo et al. [178], and Shen et al. [416]. An authentication scheme that uses both a password and a smart card is described by Hwang and Li [206]. Device-centric and attribute-based authentication is the main feature of the federated architecture that is presented by Papadamou et al. [339]. Three different user authentication schemes are analyzed by Wang et al. [470], addressing their weaknesses and the related countermeasures taken. Regarding IoT, a survey for

authentication protocols is discussed in the work of Ferrag et al. [145], while an authentication protocol is designed by Amin et al. [44] for IoT devices in cloud computing environments. A robot cloud service system is presented by Chen et al. [100] and the focus is on four crucial issues: cloud platform central control, robot intelligence technology, robot privatization, and communication security. The authors prove that the proposed scheme achieves mutual authentication, and user anonymity. An algorithm for secure key management, and secure communication in an insecure wireless and noisy environment, is provided by Yfantis and Fayed [493]. An offline authentication approach that uses biometric data to authenticate a user on mobile robots, is presented by Haas et al. [180]. Said approach uses expiring passwords and a smart card for the authentication of authorized people.

The simplest approach available in the literature for authorization refers to the Identification Based Access Control (IBAC) mechanism [401], where permission to use is linked to the user identity. On the other hand, permission to use is linked to roles in the Role Based Access Control (RBAC) approach, introduced by Ferraiolo and Kuhn [146]. Extensions of these approaches can be found by Karp et al. [225], where an authoriZation Based Access Control (ZBAC) scheme is presented. Other approaches link access to resources to specific attributes of the user identity (Attribute Based Access Control) [202]. Regarding Cyber-Physical Systems, a multi-authority access control scheme is proposed by Sciancalepore et al. [406]. The state-of-the-art of access control solutions in IoT domain is presented by Ouaddah et al. [338], who highlight challenges and opportunities.

16.3.1 Baseline technologies and tools

Anonymization tools that have been created as a result of research include the followings: i) UTD Anonymization Toolbox, ii) Cornell Anonymization Toolkit, iii) TIAMAT, iv) Anamnesia, and v) SECRETA. The disadvantage of those usually is the narrow scope regarding data transformation models. sdcMicro and μ-Argus are two tool examples coming from the statistics community, supporting a wider variety of methods for measuring risks, transforming data, and analyzing the usefulness of output data. The ARX Data Anonymization Tool is an open source software, achieving at the same time a wide range of anonymization techniques.

Regarding vulnerability, a well-known schema to rate the severity of a vulnerability is the common vulnerability scoring schema (CVSS) [1]. The produced numerical score can then be also translated into a qualitative representation (low, medium, high, and critical) helping towards the assessment and prioritization of the vulnerability management processes. Additionally, there are IoT Web platforms that maintain lists of vulnerabilities, such as the common vulnerability and exposures (CVE) database from MITRE [2] and the national vulnerability database (NVD) [11]. There is a great number of vulnerability scanning tools that, taking advantage of the online vulnerability platforms, manage to spot known vulnerabilities. OpenVAS [14] offers unauthenticated and authenticated testing, high level and low level Internet and industrial protocols, performance tuning, and internal programming language to implement any type of vulnerability test. The OpenSCAP [16] ecosystem provides multiple tools to assist with assessment, measurement, and enforcement of security baselines. Its goal is to the en-

force Security Content Automation Protocol (SCAP), a U.S. standard maintained by the National Institute of Standards and Technology (NIST). OWASP ZAP [19] is an open-source web application security scanner. It creates a proxy between the client and a website, capturing all actions.

There is a plethora of open standards and market solutions, as far as authentication and authorization are concerned, which should be mentioned. Fast Identity Online Alliances goal is to address the end user problem of creating and remembering multiple credentials [3]. The OpenID Foundation allows individuals to give their credentials to only the identity provider, and that provider then confirms their identity to the websites they visit. In that way, the whole authentication mechanism is outsourced to the identity provider [15]. OAuth [13] and its evolution OAuth 2.0 [183] is an open protocol to allow secure authorization in a simple and standard method from web, mobile and desktop applications. Being safer than asking users to log in with passwords, is the industry-standard protocol for authorization. Another open standard for access control is XACML [12].

Moreover, regarding free identity management and access control solutions, the following choices can be mentioned. Idemix [84] is an anonymous credential system developed at IBM Research that enables both strong authentication and privacy. The user can apply transformation on their attested information based on the information to be disclosed. KeyCloak [9] and WSO2 [31] are open source identity management and access control solutions. Finally, OpenUnison [18] is an open source identity management, highly customizable solution.

16.3.2 Discussion

Any of the proposed anonymization techniques present in the literature, or even a combination of them, could help us achieving data privacy in *FELICE*. The choice depends on the data collected and the individual needs of the use-case consortium participants.

As discussed in the previous subsections, there is a plethora of vulnerability scanning tools that use different vulnerability databases. However, there is a need to combine scanning tools together for various reasons. The first reason is enhancement the vulnerability coverage. Orchestrating scanning tools with different focuses will solve this problem. Next, is the addition of source code analysis that may be missed from vulnerability scanners. Tools like antiviruses could play that role and enhance the overall protection of a system. The third reason is the conflicting factors that may appear in a scanning tool orchestration. Said conflicting factors include complementarity of the participating tools, the integration level, and the user requirements to be addressed. Regarding the user requirements, there is a trade-off between the following properties: scanning time, accuracy and overhead. An orchestration of different vulnerability scanning tools in the *FELICE* project will offer a scanning solution with enhanced vulnerability coverage.

To this end, all the open source technologies mentioned above could be considered as possible solutions regarding the authentication and authorization needs of the project. The general methodology that needs to be followed, allowing for decoupling of both said functionalities, corresponds to the following steps: (i) all logical entities implementing specific algorithms and procedures (e.g. users, services) are authenticated by an Identity Manager; (ii) the Identity Manager releases authentic tokens, storing the

attributes associated to the authenticated entities; (iii) authenticated entities can use these tokens for performing authorization procedures. Management of access policies dynamically should rely on open standards, like XACML.

17 Modular technologies and tool kits for agile production

17.1 Overview

This chapter describes the modular technologies and tool kits for agile production and their state-of-the-art necessary for advanced digital solutions and robotic technologies in industrial production processes.

Since the *FELICE* consortium strives to make the tools developed in the project open and freely available - thus encouraging end-users and robotics solution developers to employ them in multiple application domains beyond the proposed work, existing tools are examined, especially those from Digital Innovation Hubs such as DIH^2 and TRINITY [44]. These hubs have already established networks of potential stakeholders interested to exploit and re-use the developments of the project in various application domains which may be beneficial for the *FELICE* project to tie in, if possible.

17.2 Relationship with FELICE project

FELICE's two-layered architecture is split into the local layer and the global layer. The local layer includes components for perceiving the environment and facilitating human-robot collaboration (HRC). The global layer comprises components for digital twin modelling, assembly orchestration, optimization, and analytics. AI and machine learning algorithms are omnipresent throughout the system. A cloud infrastructure provides components with processing resources and supports the communication among them with appropriate privacy and security mechanisms.

FELICE will seek collaboration with current and future EU robotics Digital Innovation Hubs by developing and annually updating a concrete DIH Networking Action Plan that will guide the relevant activities, thus capitalizing on the technologies and dissemination/exploitation channels already established by the DIH networks. According to the theme that each of the existing hubs focuses on, TRINITY and DIH^2 are well-aligned with the scope and the thematic priorities of *FELICE*. Both TRINITY and DIH^2 aspire to improve the agility of the European manufacturing sector, with TRINITY focusing on the combination of robotics, IoT and cybersecurity and DIH^2 seeking to develop standard robotics solutions. *FELICE* will take advantage of both of these networks to unlock and enhance its collaboration and networking potential. This concerns a two-way collaboration where on the one hand *FELICE* may gain from the potential exploitation of available freemium assets to speed up prototyping, and, on the other hand, *FELICE* will seek to provide its technological innovations as open source modules publicly available for third-party collaborators through the Digital Innovation Hubs.

TRINITY has made available a number of assets to be used and applied in industrial applications. Besides the fact that *FELICE* participants are fully equipped and well experienced with the background technologies involved in the project, some of the available

[44] Respectively http://www.dih-squared.eu/ and https://trinityrobotics.eu/

TRINITY modules can potentially be useful and will be examined/tested during Phase I of the project. These are listed below:

- **Safe Human Detection in a Collaborative Work Cell**
- **ROS Peripheral Interface**
- **Projection-based Interaction Interface for HRC**
- **Kinaesthetic Teaching of Robot Skills**
- **Robot Trajectory Generation Based on Digital Design Content**

Furthermore, *FELICE* will consider providing its developed technologies as compact, well-interfaced open-source packages that can be reused in other application domains beyond the scope of the present project. Along this line, *FELICE* will take advantage of the DIH^2 marketplace which provides an Open Integration Framework for members to develop and offer new services, enabling dynamic networking and interactions among the users. DIH^2 is already connected with the Robotics and Automation Market Place (RAMP) that is powered by a Europe-wide network of Digital Innovation Hubs (DIHs), operating a platform that connects manufacturing companies with providers of the optimal robotics solutions. *FELICE* will participate as a new member of the DIH^2 marketplace that will enhance existing and develop new service offerings. The project will develop a plan and a formal procedure on how to better promote and exploit the developed assets through the established DIH^2 functions that include (i) user customisation, (ii) user communication, (iii) brokerage, i. e., robotics repository, rating and feedback, service exchanges' tracking, (iv) tools such as an investment calculator, a file repository, an active matchmaking, and (v) a training platform. Tools of the DIH^2 project that will possibly be included are:

- **DIH^2 Marketplace**
- **DIH^2 Digital Platform**
- **LER Procedures**
- **Robots and Digitization - Needs for Standardisation**
- **Plan for the Exploitation and Dissemination of Results (PEDR)**

17.3 Agile production

17.3.1 Introduction

Agile production describes the ability of a company to react quickly to changing requirements [107]. These include, among other things, customer requests and changes in the market. Despite this flexibility, costs and quality should not suffer which is the special feature of agile production. The target audience are companies that operate in a highly competitive environment where small fluctuations in quality and slower reactions to changes can make a big difference. Agile production can therefore be seen as

an evolutionary step after lean manufacturing. As the term suggests, the latter is light manufacturing that is not specifically focused on agility. In order to decide which type is right for one's own company, the Consumer Order Cycle (COC) must be considered, according to Martin Christopher (see [107]). Lean production is possible only if the supplier has a short response time. It is also possible to pursue both concepts at the same time, whereby lean manufacturing then leads to avoiding waste and costs as far as possible, unless they are directly necessary for production for the customer.

Agile productivity can mainly be achieved by building a strong supplier network that allows one to quickly negotiate new agreements [171]. In addition, a quick changeover of workstations is important, as well as a variety of cooperative teams working together within the company to deliver products effectively. In addition, further steps are conceivable that aim to fulfil customer requirements quickly, cost-effectively and with high quality. An example of a necessary tool is a common database of parts and products shared by market participants, designers and production personnel. Sharing information about production capacity and problems is also important, especially when minor difficulties can cause major delays.

17.3.2 Relation to FELICE

FELICE starts exactly at this point. Workstations are operated by general robots and human operators. This allows for quick reconfiguration. In addition, the "Resilient Assembly Line" is an explicit part of the project, which is intended to avoid minor inconsistencies in the process from the outset and to correct them if they have already occurred. The entire production process is also driven by algorithms.

17.3.3 Baseline technologies and tools

As mentioned above, both TRINITY and DIH^2 enable a wide variety of tools for agile production which can be used for *FELICE*. The goal of TRINITY is to create a network of multidisciplinary and synergistic local DIHs composed of research centers, companies, and university groups that cover a wide range of topics that can contribute to agile production. The focus of TRINITY is to deploy tools to achieve highly intelligent, agile and re-configurable production which will ensure Europe's welfare in the future. One safety tool is the Safe Human Detection in a Collaborative Work Cell to create a safe collaborative working space for robots and employees in closer proximity. Another tool is the ROS Peripheral Interface which provides a bridge between hardware that is not ROS-compliant.

DIH^2 is a European project funded by Horizon 2020 and aspiring to apply the power of robotics in order to transform the agility of manufacturing in Small and Medium-sized Enterprises (SMEs). The idea is to facilitate the connections that enable agile production in factories where speed and versatility are essential to satisfy customer demand. One concept is the DIH Marketplace which provides service for members and newcomers, to develop and offer new services, enabling dynamic networking and interactions among users.

17.3.3.1 Tools from TRINITY

TRINITY provides about 40 reusable assets or modules that can be used and applied in industrial applications. Some of the modules provided by TRINITY can be potentially useful within the scope of *FELICE* and a relevant selection of these is outlined in the following.

- **Safe Human Detection in a Collaborative Work Cell** presents a flexible and adaptive dynamic safety area based on information from safety approved multi-modal sensors such as laser scanners, microwave radars, RF indoor positioning and panoramic cameras. The broad goal is to enable humans and and robots to collaboratively work in a cell, e. g. in the context of an agile assembly process. The hardware components used in this module are solely off-the-shelf commercial components and thus can easily supplement existing industrial environments [28].

- **ROS hardware and software interface for peripheral elements that are not ROS-compliant** provides a bridge between hardware that is not ROS-compliant and the ROS backbone of the actual system. It therefore provides an interface between the ROS-based system and modular hardware elements typically found frequently in industrial environments. The ROS peripheral interface has been designed as a self-contained unit allowing seaming-less integration of new hardware components [27].

- **Projection-based Interaction Interface for HRC** offers an interface the user can interact with by placing a hand over it. The interface consists of multiple buttons, such as *go*, *stop*, and *confirm* to manually interact with the workspace. The system setup consists of a 3LCD projector, a Microsoft Kinect v2, an UR5 robotic arm as well as a work station. The Kinect v2 sensor installed at the ceiling is utilized for observing the whole workspace [20].

- **Kinaesthetic Teaching of Robot Skills** allows users to intuitively program robots based on simple interactions. The robot has to be equipped with necessary sensing equipment, such as joint-torque sensors or 6D force-torque sensors that allow a gravity compensation. The robot is enabled to acquire new skills guided by human demonstrations that present the desired configurations. The skill acquisition process is guided by a graphical user interface to further reduce complexity [10].

- **Robot Trajectory Generation Based on Digital Design Content** aims at speeding up robot simulation and programming by using digital design data, such as Building Information Model (BIM). Existing digital design data (e. g., from a CAD model) can be utilized to generate trajectories for robotics tasks and thus effectively shorten the design-to-production time. Furthermore, the design data allows for creating AR/VR models in a very time-efficient manner. Possible application areas of this module include training, safety as well as production planning [26].

17.3.3.2 Tools from DIH^2

The following paragraphs describe the tools, modules and analyses which have been made available in the *DIH^2* Digital Innovation Hub.

- **DIH^2 Marketplace**
  The Marketplace is an open platform for manufacturing end-users and robotic solution providers. It offers dynamic networking and interaction between users. On the one hand, manufacturers gain access to robotics digitization technologies to improve their efficiency and productivity. On the other hand, automation suppliers can reach their customers faster and access a larger market. A range of different services and support in robotics and automation is therefore provided for the end-user. This should lead to the reduction of the knowledge and access gap between the end-users and robotics and automation industry [23].

- **DIH^2 Digital Platform**
  The ability of a production line to flexibly adapt to changes in the production context inside (breakdown, delays, strikes) and outside the workshop (supply chain events, new customers) needs to be considered. Therefore, a Digital Platform must be data- and event-driven. It provides a first set of robotic-based open standard enablers for an agile production. The digital platform corresponds to an open and inter-operable platform to integrate physical robotic-based industrial facilities with agile production applications such as, for example, production planning scheduling, infra-logistics optimization or maintenance [21].

- **LER Procedures**
  Local Evangelists in Robotics' (LERs) main mission is to effectively communicate with other DIHs, regional bodies and Small and Medium-sized Enterprises (SMEs). LER procedures cover activities, procedures and training that affect LERs. The main contents consist of a catalogue of services to SMEs, selling skills and techniques to offer those services to SMEs and robot demonstrator exhibition [22].

- **Robots and Digitization: Needs for Standardisation**
  The needs for standardization are presented by giving several issues that could be tackled with the use of standardization. This topic gives an overview of the available body of standards and current issues of companies and Small and Medium-sized Enterprises (SMEs) while focusing on specific issues that are largely encountered by manufacturing companies. Gaps in standardization are distinguished and potential challenges to solve these are described [25].

- **PEDR**
  This report gives an overview of the exploitation and dissemination strategy covering an exploitation plan for the project result, target audience, different channels, e. g. events, social media and communication materials. This report will be further elaborated and updated in the future [24].

17.3.4 Discussion

Any of the existing tools of TRINITY and DIH^2 can be useful during the lifespan of *FELICE*. Initially, the focus will be on the consideration of the "Safe human Detection in a Collaborative Work Cell" of TRINITY. Regarding the networking with DIH^2, the consortium will focus in the first year on the communication with the LER and will

have in mind and contribute to the package “Robots and Digitization: Needs for Standardisation” as well as on the “PEDR” to identify further exploitation and dissemination strategies. Once services and modules are developed by *FELICE*, those can be made available in the market place of DIH^2 and, if possible, will be distributed to the partners of the DIH^2 as well as the TRINITY project.

18 Conclusions

This document has analyzed the tools and technologies that are already available either in the scientific literature or commercially, and can be adopted in specific tasks that have been identified in the context of the *FELICE* project. For each such tool or technology, its advantages and disadvantages have been analyzed and related challenges as well as potential issues have been pointed out. This process has paved the way for the technical developments that will follow in the course of the project. An overview of the baseline technologies and tools selected for the various topics is reported in Table 8, and some additional comments and insights derived from this analysis are summarized in the following paragraphs.

A key ingredient of the *FELICE* project is the collaborative mobile robot which must be able to autonomously build a map and navigate its environment. Thus, SLAM algorithms will be deployed, and in particular online visual SLAM algorithms, such as LSD-SLAM, which is highly resilient in indoor environments. Also, the possibility to improve the localization accuracy by online data fusion, employing sensors such as IMU, will be also considered. Additionally, the robot needs the capability to perceive the objects of interest that are present within the environment in which it is moving, together with their pose. This can be achieved by means of object detection and 6D pose estimation algorithms, such as CosyPose and EPOS, which are able to operate in near real time on RGB images. Apart from objects, *FELICE* needs to extract visual data for characterizing human behavior. This will be supported by an off-board multi-camera system overcoming occlusions.

Another fundamental aspect pertains to the hardware architecture of the robot that will be designed and constructed during the project. The *FELICE* robot is an autonomously moving service platform with an on-board touch screen and a dexterous robotic hand and arm. Despite that existing solutions available on the market meet some of the project requirements, there exists no off-the-shelf robot that fulfills them all. Therefore, the development of the robot will be based on ACC's past designs.

Even if the robot is not required to have a human-like appearance, the basic principles of cognitive ergonomics have to be taken into account, so as to allow engaging the robot in collaboration by means of awareness and representation of the specific characteristics and states of the human users as well as through mirroring the socio-cognitive characteristics of goal-driven interaction. Cognitive ergonomics will rely on hierarchical task analysis, extended by cognitive task analysis.

The robot also has to behave in a safe way with respect to both itself, the environment where it operates and the human workers. The normative standards and directives to be respected in industrial environments are thus analyzed, together with the choices made within the project in order to guarantee safety by design. More specifically, the robot will be designed to be intrinsically/mechanically safe, i.e. avoiding hazards instead of controlling them. This implies that safety is "engineered into" the electromechanical design at the earliest stage possible.

The developed cobot will be complemented by adaptive workstations that will be based on the evolution of existing TUD designs and will combine user centered and ecological interface design principles. The purpose of an adaptive workstation is to allow individual customization of the workplace to improve physical and cognitive ergonomics, productivity/efficiency, and work quality for the worker. Consequently, since

the workstation needs to adapt itself according to a user's individual needs, the dimensions and options of its user-adaptivity are discussed and analyzed.

Both the robot and the workstation have to interact with the human worker, using a command based interaction approach. This is realized by using two complementary sources of information, namely gesture recognition by visual analysis and voice command recognition by audio analysis. Concerning audio analysis, the algorithms and tools available in the literature, all of them based on deep learning, were analyzed and discussed, for both Automatic Speech Recognition (i.e. NVIDIA NEMO framework) and Natural Language Understanding (i.e. RASA framework). With respect to gesture recognition, the different possible input modalities were analyzed as well (i.e. RGB, RGBD, optical flow), together with the different possible deep neural network architectures (3D-CNN, ConvLSTM etc). In both gesture and voice command recognition, the analysis was conducted by taking into account accuracy and resource requirements, since it is expected that such components have to run in real time on embedded devices.

Another important aspect taken into account pertains to the various programming paradigms, that enable the configuration of a robotic system for a particular task. Among the different robot programming techniques (i.e. programming by advice, skill based programming, programming by demonstration and programming by interaction), special focus has been given to skill based programming, and in particular to task level programming, which has the advantage of being easily and quickly re-programmable by non-expert users. The strategy is to build on PROF's XROB task-based programming framework and exploit multi-modal sensing to deal with dynamic environment changes.

Effective cooperation between human and robot presupposes that both are accustomed to the task and to each other, to coordinate their actions. Thus, the timing should be precise and efficient, resulting in a well-synchronized meshing of their actions. This aspect, very important for the *FELICE* project, has been analyzed and discussed as well.

Human-robot cooperation occurs at two levels of detail in *FELICE*. At the highest level (global layer), the orchestrator will monitor ongoing processes in the assembly line and decide on where and how the robot will assist human workers in assembly tasks. A workflow-based orchestration process is outlined and discussed in this report, aiming at enabling the specification of assembly tasks including their sequence and all required skills to fulfil the production process without assuming any platform-specific constraints. The WORM tool has been identified as being suitable for transforming task-level specifications to robot-level specifications for task-level programming. At the lower level, decisions will be made to coordinate task activities between humans and robots so that the two parties can work fluently together.

Another important aspect of the project relates to constructing a digital twin of both machine/equipment and human operators. Indeed, AI driven Digital Twins will be designed and developed as a knowledge-aware expert system which is fully synchronized with the physical system and capable of being operated upon. Digital twins will be coupled with four-stage analytics, namely descriptive, diagnostic, predictive and prescriptive.

The protection of person- identifiable information recorded by different FELICE components is highlighted. Hence, anonymization and encryption techniques (such as the ARX data anonymization tool), identification of vulnerabilities, authentication, authorization and accounting have been analyzed and discussed.

Table 8: Overview of candidate algorithms, technologies and tools to be used within *FELICE*.

Topics	**Methods, Technologies and Tools**
Scene and object perception	*Simultaneous localization and mapping*: LSD-SLAM, ORB-SLAM, SVO *Object detection and pose estimation*: CNN on RGB/RGB-D data (CosyPose, EPOS)
Human behavior monitoring in assembly task execution	Graph Neural Network on skeleton data. Deep Multi-Task Learning.
Robotic hardware	Robotic manipulator: RAMCIP-based mobile robot, ARIA robotic arm
Adaptive workstation	Ecological Interface Design, and User Centered Design.
Human robot communication	*Speech-command interaction*: NVIDIA Nemo, VOSK, RASA, ReSpeaker. *Gesture Recognition*: CNN on RGB-D or optical flow data, LSTM on skeleton data.
Cognitive ergonomics for enhanced human-robot dyads	Human Factors methods (Data collection, Task analysis, Usability- and User experience, Experimental methods).
Safe robot operation	Normative requirements (ISO standards).
Robot programming	*Planners*: PDDL4j, ROSplan, FastDOwnward planner. *Framework*: XRob.
Synchronization of the human-robot dyad in taskable pipelines	Not applicable.
Prescriptive analytics in production system diagnosis, monitoring, and control	*Prescriptive analytics*: HeuristicLab. *Assembly line balancing*: SALBP, GALBP.
AI-driven digital twins and digital operators	Not applicable.
Orchestration of adaptive assembly lines	Sequential Function Charts (IEC 61131-3), Functional Block Diagrams (IEC 61499), Open Platform Communications Unified Architecture.
Computing infrastructure	FIWARE, ROS.
Data privacy, vulnerability management, and security assurance	*Anonymization*: ARX data anonymization tool. *Authentication and Authorization*: Fast Identity Online Alliance, OpenID Foundation, OAuth, XACML, Idemix, KeyCloak, WSO2, OpenUnison. *Vulnerability scanning*: OpenVAS, OpenSCAP, OWASP ZAP.
Modular technologies and tool kits for agile production	Tools from TRINITY and DIH^2.

Computing infrastructures for the deployment of components, the execution of simulations and the exchange of messages between components were also discussed.

Finally, offerings by relevant Digital Innovation Hubs that can be of usefulness in *FELICE* were examined. Furthermore, the dissemination of *FELICE* developments that can be of interest to a wider audience has been considered.

References

[1] Common vulnerability scoring system. https://www.first.org/cvss/. Accessed: 2021-05-17.

[2] Cve program. https://cve.mitre.org/. Accessed: 2021-05-17.

[3] Fido alliance. https://fidoalliance.org/. Accessed: 2021-05-17.

[4] Fiware-cloud. https://www.fiware.org/2015/07/29/cloud-robotics-when-robots-got-smart/.

[5] Fiware-idm.

[6] Fiware-ros. https://www.rosin-project.eu/ftp/ros2-integration-service.

[7] Fiware-success. https://cordis.europa.eu/docs/projects/cnect/3/632893/080/deliverables/001-D1162ReportonFIWARESuccessStories.pdf.

[8] Fogflow. https://github.com/smartfog/fogflow.

[9] keycloak. https://www.keycloak.org/. Accessed: 2021-05-17.

[10] Kinesthetic teaching of robot skills. https://trinityrobotics.eu/wp-content/uploads/2019/11/TRINITY-Modules-JSI-kinesthetic-teaching.pdf. Accessed: 2021-06-15.

[11] National vulnerability database. https://www.nist.gov/programs-projects/national-vulnerability-database-nvd. Accessed: 2021-05-17.

[12] Oasis extensible access control markup language. https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml. Accessed: 2021-05-17.

[13] Oauth. http://oauth.net. Accessed: 2021-05-17.

[14] Open vulnerability assessment scanner. https://www.openvas.org/. Accessed: 2021-05-17.

[15] Openid. http://openid.net/. Accessed: 2021-05-17.

[16] Openscap ecosystem. https://www.open-scap.org/. Accessed: 2021-05-17.

[17] Openstack. https://www.openstack.org.

[18] Openunison. https://github.com/TremoloSecurity/OpenUnison. Accessed: 2021-05-17.

[19] Owasp zap. https://owasp.org/www-project-zap/. Accessed: 2021-05-17.

[20] Projection-based interaction interface for HRC. https://trinityrobotics.eu/modules/projection-based-interaction-interface-for-hrc/. Accessed: 2021-06-15.

[21] Report on the DIH^2 digital platform. http://www.dih-squared.eu/sites/default/files/D1.3.pdf. Accessed: 2021-06-15.

[22] Report on the DIH^2 LER procedures. http://www.dih-squared.eu/sites/default/files/D2.1.pdf. Accessed: 2021-06-15.

[23] Report on the DIH^2 marketplace. http://www.dih-squared.eu/sites/default/files/D1.2.pdf. Accessed: 2021-06-15.

[24] Report on the DIH^2 plan for the exploitation and dissemination of results. http://www.dih-squared.eu/sites/default/files/D6.1.pdf. Accessed: 2021-06-15.

[25] Report on the DIH^2 standardization analyses. http://www.dih-squared.eu/sites/default/files/D2.2.pdf. Accessed: 2021-06-15.

[26] Robot trajectory generation based on digital design content. https://trinityrobotics.eu/modules/robot-trajectory-generation-based-on-digital-design-content-2/. Accessed: 2021-06-15.

[27] ROS hardware and software interface for peripheral elements that are not ROS-compliant. https://trinityrobotics.eu/modules/ros-hardware-and-software-interface-for-peripheral-elements-that-are-not-ros-compliant/. Accessed: 2021-06-15.

[28] Safe human detection in a collaborative work cell. https://trinityrobotics.eu/modules/safe-human-detection-in-a-collaborative-work-cell-2/. Accessed: 2021-06-15.

[29] Universaal. http://www.universaal.info.

[30] upc-ua. https://www.spotlightmetal.com/iot-basics-what-is-opc-ua-a-842878/.

[31] Wso2 identity server. https://wso2.com/library/articles/2017/08/what-is-wso2-identity-server/. Accessed: 2021-05-17.

[32] IEC 61131 programmable controllers—part 3: Programming languages, 1993.

[33] Function blocks for industrial process measurement and control systems. part i: Architecture, 2005.

[34] OPC UA Specification, 2015.

[35] A. Abobakr, D. Nahavandi, J. Iskander, M. Hossny, S. Nahavandi, and M. Smets. A kinect-based workplace postural analysis system using deep residual networks. In *2017 IEEE International Systems Engineering Symposium (ISSE)*, pages 1–6, 2017.

[36] K. Achmad, L. Nugroho, and A. Djunaedi. Context-aware based restaurant recommender system: a prescriptive analytics. *J Eng Sci Technol*, 14(5):2847–2864, 2019.

[37] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer vision and image understanding*, 73(3):428–440, 1999.

[38] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny. An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, 29(5-6):594–621, 2010.

[39] S. C. Akkaladevi, A. Pichler, M. Plasch, M. Ikeda, and M. Hofmann. Skill-based programming of complex robotic assembly tasks for industrial application. *e & i Elektrotechnik und Informationstechnik*, 136(7):326–333, 2019.

[40] R. Alami, A. Albu-Schaeffer, A. Bicchi, R. Bischoff, R. Chatila, A. De Luca, A. De Santis, G. Giralt, J. Guiochet, G. Hirzinger, F. Ingrand, V. Lippiello, R. Mattone, D. Powell, S. Sen, B. Siciliano, G. Tonietti, and L. Villani. Safe and dependable physical human-robot interaction in anthropic domains: State of the art and challenges. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–16, 2006.

[41] S. Alatartsev, S. Stellmacher, and F. Ortmeier. Robotic task sequencing problem: A survey. *Journal of intelligent & robotic systems*, 80(2):279–298, 2015.

[42] G. Alenyà, S. Foix, and C. Torras. Using ToF and RGBD cameras for 3D robot perception and manipulation in human environments. *Intelligent Service Robotics*, 7(4):211–220, 2014.

[43] J. Aleotti, V. Micelli, and S. Caselli. An affordance sensitive system for robot to human object handover. *Int. J. Soc. Robotics*, 6(4):653–666, 2014.

[44] R. Amin, N. Kumar, G. Biswas, R. Iqbal, and V. Chang. A light weight authentication protocol for iot-enabled devices in distributed cloud computing environment. *Future Generation Computer Systems*, 78:1005–1019, 2018.

[45] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[46] L. Ante. Digital twin technology for smart manufacturing and industry 4.0: A bibliometric analysis of the intellectual structure of the research discourse. *Manufacturing Letters*, 27:96–102, 2021.

[47] G. Antonelli, S. Chiaverini, N. Sarkar, and M. West. Adaptive control of an autonomous underwater vehicle: experimental results on odin. *IEEE Transactions on Control Systems Technology*, 9(5):756–765, 2001.

[48] T. Arai, Y. Aiyama, Y. Maeda, M. Sugi, and J. Ota. Agile assembly system by plug and produce. *CIRP Annals-Manufacturing Technology*, 49(1):1–4, 2000.

[49] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[50] S. Arnau, T. Möckel, G. Rinkenauer, and E. Wascher. The interconnection of mental fatigue and aging: An EEG study. *International Journal of Psychophysiology*, 117:17–25, 2017.

[51] L. Atzori, A. Iera, and G. Morabito. The internet of things: A survey. *Computer Networks*, 54(15):2787–2805, Oct. 2010.

[52] S. Ayhan, P. Costas, and H. Samet. Prescriptive analytics system for long-range aircraft conflict detection and resolution. In *Proceedings of the 26th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 239–248, 2018.

[53] R. Azzam, T. Taha, S. Huang, and Y. Zweiri. Feature-based visual simultaneous localization and mapping: A survey. *SN Applied Sciences*, 2(2):1–24, 2020.

[54] K. Bacharidis and A. Argyros. Extracting action hierarchies from action labels and their use in deep action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 339–346. IEEE, 2021.

[55] A. Banks and R. Gupta. MQTT Version 3.1.1, 2014.

[56] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori. Object level visual reasoning in videos. In *ECCV*, 2018.

[57] J. Barreiro, M. Boyce, M. Do, J. Frank, M. Iatauro, T. Kichkaylo, P. Morris, J. Ong, E. Remolina, T. Smith, and D. Smith. EUROPA: A platform for AI planning, scheduling, constraint programming, and optimization. 2012.

[58] W. Bauer, M. Bender, M. Braun, P. Rally, and O. Scholtz. Leichtbauroboter in der manuellen montage–einfach einfach anfangen. *IRB Mediendienstleistungen, Stuttgart*, 2016.

[59] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417. Springer, 2006.

[60] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *21st International conference on data engineering (ICDE'05)*, pages 217–228. IEEE, 2005.

[61] I. Baybars. A survey of exact algorithms for the simple assembly line balancing problem. *Management science*, 32(8):909–932, 1986.

[62] C. Becker and A. Scholl. A survey on problems and methods in generalized assembly line balancing. *European journal of operational research*, 168(3):694–715, 2006.

[63] D. Beltran and L. Basañez. A comparison between active and passive 3D vision sensors: Bumblebeexb3 and microsoft kinect. In *Robot2013: First iberian robotics conference*, pages 725–734. Springer, 2014.

[64] K. B. Bennet and J. Flach. *Display and Interface Design: Subtle Science, Exact Art*. CRC Press, Boca Raton, 2011.

[65] K. B. Bennett, J. M. Flach, C. Edman, J. Holt, and P. Lee. Ecological interface design: A selective overview. 2015.

[66] P. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[67] A. Bilberg and A. A. Malik. Digital twin driven humanrobot collaborative assembly. *CIRP Annals*, 68(1):499–502, 2019.

[68] A. Billard, S. Calinon, R. Dillmann, and S. Schaal. Survey: Robot programming by demonstration. Technical report, Springrer, 2008.

[69] A. Billard and D. Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446), 2019.

[70] A. Björkelund, L. Edström, M. Haage, J. Malec, K. Nilsson, P. Nugues, S. G. Robertz, D. Störkle, A. Blomdell, R. Johansson, et al. On the integration of skilled robot motions for productivity in manufacturing. In *2011 IEEE International Symposium on Assembly and Manufacturing (ISAM)*, pages 1–9. IEEE, 2011.

[71] J. W. Blake Hannaford. *Actuator Properties and Movement Controls: Biological and Technological Models*. Springer, https://link.springer.com/content/pdf/10.10071990.

[72] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski. Five disruptive technology directions for 5g. *IEEE Communications Magazine*, 52(2):74–80, Feb. 2014.

[73] K. Bogner, U. Pferschy, R. Unterberger, and H. Zeiner. Optimised scheduling in humanrobot collaboration a use case in the assembly of printed circuit boards. *International Journal of Production Research*, 56(16):5522–5540, 2018.

[74] G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne. Machine learning strategies for time series forecasting. In *European business intelligence summer school*, pages 62–77. Springer, 2012.

[75] M. Bortolini, M. Faccio, F. G. Galizia, M. Gamberi, and F. Pilati. Adaptive automation assembly systems in the industry 4.0 era: A reference framework and fullscale prototype. *Applied Sciences*, 11(3), 2021.

[76] N. Boysen, M. Fliedner, and A. Scholl. A classification of assembly line balancing problems. *European journal of operational research*, 183(2):674–693, 2007.

[77] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6D object pose estimation using 3D object coordinates. In *European conference on computer vision*, pages 536–551. Springer, 2014.

[78] B. Brandenbourger, M. Vathoopan, and A. Zoitl. Engineering of automation systems using a metamodel implemented in AutomationML. In *Industrial Informatics (INDIN), 2016 IEEE 14th International Conference on*, pages 363–370. IEEE, 2016.

[79] T. Brandt, S. Wagner, and D. Neumann. Prescriptive analytics in public-sector decision-making: A framework and insights from charging infrastructure planning. *European Journal of Operational Research*, 291(1):379–393, 2021.

[80] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser. Simultaneous localization and mapping: A survey of current trends in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2(3):194–220, Sept. 2017.

[81] D. Bryce, S. Gao, D. Musliner, and R. Goldman. Smt-based nonlinear pddl+ planning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[82] C. S. Burke, K. C. Stagl, E. Salas, L. Pierce, and D. Kendall. Understanding team adaptation: a conceptual analysis and model. *Journal of Applied Psychology*, 91(6):1189, 2006.

[83] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

[84] J. Camenisch and E. Van Herreweghen. Design and implementation of the idemix anonymous credential system. In *Proceedings of the 9th ACM Conference on Computer and Communications Security*, pages 21–30, 2002.

[85] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. Using convolutional 3d neural networks for user-independent continuous gesture recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 49–54, 2016.

[86] C. Campos, R. Elvira, J. J. G. Rodrguez, J. M. M. Montiel, and J. D. Tards. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. arXiv, 2021. 2007.11898v2.

[87] G. Cannata, M. Maggiali, G. Metta, and G. Sandini. An embedded artificial skin for humanoid robots. In *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 434–438, 2008.

[88] T. Cao and A. C. Sanderson. Task sequence planning in a robot workcell using and/or nets. 1991.

[89] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[90] E. Carpanzano, A. Cesta, A. Orlandini, R. Rasconi, M. Suriano, A. Umbrico, and A. Valente. Design and implementation of a distributed part-routing algorithm for reconfigurable transportation systems. *International Journal of Computer Integrated Manufacturing*, 29(12):1317–1334, 2016.

[91] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. A short note about kinetics-600, 2018.

[92] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[93] M. Cashmore, M. Fox, T. Larkworthy, D. Long, and D. Magazzeni. Auv mission control via temporal planning. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6535–6541, 2014.

[94] M. Cashmore, M. Fox, D. Long, D. Magazzeni, and B. Ridder. Opportunistic planning in autonomous underwater missions. *IEEE Transactions on Automation Science and Engineering*, 15(2):519–530, 2018.

[95] M. Cashmore, M. Fox, D. Long, D. Magazzeni, B. Ridder, A. Carrera, N. Palomeras, N. Hurtós, and M. Carreras. Rosplan: Planning in the robot operating system. In *ICAPS*, 2015.

[96] M. Cashmore, M. Fox, D. Long, D. Magazzeni, B. Ridder, A. Carreraa, N. Palomeras, N. Hurtós, and M. Carrerasa. Rosplan: Planning in the robot operating system. In *Proceedings of the Twenty-Fifth International Conference on International Conference on Automated Planning and Scheduling*, ICAPS'15, page 333341. AAAI Press, 2015.

[97] A. Cesta, G. Cortellessa, S. Fratini, and A. Oddi. Developing an end-to-end planning application from a timeline representation framework. In *Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence*, 01 2009.

[98] T. Chalasani and A. Smolic. Simultaneous segmentation and recognition: Towards more accurate ego gesture recognition. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4367–4375, 2019.

[99] C. Chao and A. Thomaz. Timed petri nets for fluent turn-taking over multimodal interaction resources in human-robot collaboration. *The International Journal of Robotics Research*, 35(11):1330–1353, 2016.

[100] C.-L. Chen, Y.-T. Li, Y.-Y. Deng, and C.-T. Li. Robot identification and authentication in a robot cloud service system. *IEEE Access*, 6:56488–56503, 2018.

[101] L. Chen, H. Wei, and J. Ferryman. A survey of human motion analysis using depth imagery. *Pattern Recognition Letters*, 34(15):1995–2006, 2013.

[102] J. Cheng, H. Zhang, F. Tao, and C.-F. Juang. Dt-ii:digital twin enhanced industrial internet reference framework towards smart manufacturing. *Robotics and Computer-Integrated Manufacturing*, 62:101881, 2020.

[103] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[104] S. Y. Chien, L. Q. Xue, and M. Palakal. Task planning for a mobile robot in an indoor environment using object-oriented domain information. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 27(6):1007–1016, 1997.

[105] Y. Chien, A. Hudli, and M. Palakal. Using many-sorted logic in the object-oriented data model for fast robot task planning. *Journal of Intelligent and Robotic Systems*, 23(1):1–25, 1998.

[106] Y. G. Cho. Awesome SLAM datasets. https://github.com/younggguncho/awesome-slam-datasets, 2021.

[107] M. Christopher. Logistics and supply chain management: Strategies for reducing cost and improving service financial times: Pitman publishing. london, 1998 isbn 0 273 63049 0 (hardback) 294+ 1× pp., 1999.

[108] A. Cimatti, A. Micheli, and M. Roveri. Dynamic controllability of disjunctive temporal networks: validation and synthesis of executable strategies. In *in proc. 30th AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[109] J. Civera and S. H. Lee. RGB-D odometry and SLAM. In *RGB-D Image Analysis and Processing*, pages 117–144. Springer, 2019.

[110] F. Clement, K. Shah, and D. Pancholi. A review of methods for textureless object recognition. *arXiv preprint arXiv:1910.14255*, 2019.

[111] CO-ADAPT. Adaptive environments and conversational agent based approaches for healthy ageing and work ability. https://cordis.europa.eu/project/id/826266, 2018-2020.

[112] A. J. Coles, A. I. Coles, M. Fox, and D. Long. Colin: Planning with continuous linear numeric change. *Journal of Artificial Intelligence Research*, 44:1–96, 2012.

[113] A. Colim, C. Faria, J. Cunha, J. Oliveira, N. Sousa, and L. A. Rocha. Physical ergonomic improvement and safe design of an assembly workstation through collaborative robotics. *Safety*, 7(1), 2021.

[114] M. Colledanchise and L. Natale. On the implementation of behavior trees in robotics. *IEEE Robotics and Automation Letters*, 6(3):5929–5936, 2021.

[115] R. S. Consensus. A paradigm for model fitting with applications to image analysis and automated cartography. *MA Fischler, RC Bolles*, 6:381–395, 1981.

[116] C. Couffe and G. A. Michael. Failures due to interruptions or distractions: A review and a new framework. *American journal of psychology*, 130(2):163–181, 2017.

[117] M. Cramer, J. Cramer, K. Kellens, and E. Demeester. Towards robust intention estimation based on object affordance enabling natural human-robot collaboration in assembly tasks. *Procedia CIRP*, 78:255–260, 2018.

[118] A. Crivellaro, M. Rad, Y. Verdie, K. M. Yi, P. Fua, and V. Lepetit. Robust 3D object tracking from monocular images using stable parts. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1465–1479, 2017.

[119] M. Dalle Mura and G. Dini. Worker skills and equipment optimization in assembly line balancing by a genetic approach. *Procedia CIRP*, 44:102–107, 2016.

[120] M. Dalle Mura and G. Dini. Designing assembly lines with humans and collaborative robots: A genetic approach. *CIRP Annals*, 68(1):1–4, 2019.

[121] P. Danny, P. Ferreira, N. Lohse, and K. Dorofeev. An event-based AutomationML model for the process execution of 'plug-and-produce' assembly systems. In *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*, pages 49–54. IEEE, 2018.

[122] K. Dautenhahn. Methodology & themes of human-robot interaction: A growing research field. *International Journal of Advanced Robotic Systems*, 4(1):15, 2007.

[123] A. J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proceedings Ninth IEEE International Conference on Computer Vision*, volume 2, pages 1403–1410, 2003.

[124] J. G. De Gooijer and R. J. Hyndman. 25 years of time series forecasting. *International journal of forecasting*, 22(3):443–473, 2006.

[125] A. De Santis, B. Siciliano, A. De Luca, and A. Bicchi. An atlas of physical humanrobot interaction. *Mechanism and Machine Theory*, 43(3):253–270, 2008.

[126] D. Delen and S. Ram. Research challenges and opportunities in business analytics. *Journal of Business Analytics*, 1(1):2–12, Jan. 2018.

[127] R. Deraison, R. Gula, and T. Hayton. Passive vulnerability scanning: Introduction to nevo. *Revision*, 9(1-13):7, 2003.

[128] K. Dorofeev, C.-H. Cheng, M. Guedes, P. Ferreira, S. Profanter, and A. Zoitl. Device adapter concept towards enabling plug&produce production environments. In *2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, sep 2017.

[129] K. Dorofeev and A. Zoitl. Skill-based engineering approach using OPC UA programs. In *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*. IEEE, jul 2018.

[130] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[131] D. Driscoll, A. Mensch, T. Nixon, and A. Regnier. Devices profile for web services, version 1.1, 2009.

[132] B. Drost, M. Ulrich, P. Bergmann, P. Hartinger, and C. Steger. Introducing MVTec ITODD - a dataset for 3D object recognition in industry. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2200–2208, 2017.

[133] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3D object recognition. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 998–1005. Ieee, 2010.

[134] G. Du, K. Wang, S. Lian, and K. Zhao. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artificial Intelligence Review*, pages 1–58, 2020.

[135] M. Dumas and A. H. M. ter Hofstede. UML activity diagrams as a workflow specification language. In *≪UML≫ 2001 — The Unified Modeling Language. Modeling Languages, Concepts, and Tools*, pages 76–90. Springer Berlin Heidelberg, 2001.

[136] L. F. C. S. Durão, S. Haag, R. Anderl, K. Schützer, and E. Zancul. Digital twin requirements in the context of industry 4.0. In P. Chiabert, A. Bouras, F. Noël, and J. Ríos, editors, *Product Lifecycle Management to Support Industry 4.0*, pages 204–214, Cham, 2018. Springer International Publishing.

[137] C. Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[138] S. Edelkamp and J. Hoffmann. Pddl2. 2: The language for the classical part of the 4th international planning competition. Technical report, Technical Report 195, University of Freiburg, 2004.

[139] H. A. ElMaraghy. Flexible and reconfigurable manufacturing systems paradigms. *International Journal of Flexible Manufacturing Systems*, 17(4):261–276, Oct 2005.

[140] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard. 3-D mapping with an RGB-D camera. *IEEE Transactions on Robotics*, 30(1):177–187, 2014.

[141] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):611–625, 2018.

[142] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *Computer Vision – ECCV 2014*, pages 834–849, Cham, 2014. Springer International Publishing.

[143] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann. The locata challenge: Acoustic source localization and tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1620–1643, 2020.

[144] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, 2019.

[145] M. A. Ferrag, L. A. Maglaras, H. Janicke, J. Jiang, and L. Shu. Authentication protocols for internet of things: a comprehensive survey. *Security and Communication Networks*, 2017, 2017.

[146] D. Ferraiolo, D. R. Kuhn, and R. Chandramouli. *Role-based access control*. Artech House, 2003.

[147] P. Ferreira and N. Lohse. Configuration model for evolvable assembly systems. In *4th CIRP Conference On Assembly Technologies And Systems*, 2012.

[148] K. Fischer, L. Jensen, F. Kirstein, S. Stabinger, . Erkent, D. Shukla, and J. Piater. The effects of social gaze in human-robot collaborative assembly. pages 204–213, 10 2015.

[149] J. Fonseca, M. Vieira, and H. Madeira. Testing and comparing web vulnerability scanning tools for sql injection and xss attacks. In *13th Pacific Rim international symposium on dependable computing (PRDC 2007)*, pages 365–372. IEEE, 2007.

[150] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza. SVO: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, 33(2):249–265, 2017.

[151] M. Fox and D. Long. Pddl2. 1: An extension to pddl for expressing temporal planning domains. *Journal of artificial intelligence research*, 20:61–124, 2003.

[152] M. Fox and D. Long. Modelling mixed discrete-continuous domains for planning. *Journal of Artificial Intelligence Research*, 27:235–297, 2006.

[153] F. Fraundorfer and D. Scaramuzza. Visual odometry: Part II – matching, robustness, optimization, and applications. *IEEE Robotics Automation Magazine*, 19(2):78–90, 2012.

[154] D. Frazzetto, T. D. Nielsen, T. B. Pedersen, and L. Šikšnys. Prescriptive analytics: a survey of emerging trends and technologies. *The VLDB Journal*, 28(4):575–595, May 2019.

[155] M. Fu and W. Zhou. DeepHMap++: Combined projection grouping and correspondence learning for full dof pose estimation. *Sensors*, 19(5):1032, 2019.

[156] K. Fukuda, I. G. Ramirez-Alpizar, N. Yamanobe, D. Petit, K. Nagata, and K. Harada. Recognition of assembly tasks based on the actions associated to the manipulated objects. In *2019 IEEE/SICE International Symposium on System Integration (SII)*, pages 193–198. IEEE, 2019.

[157] A. Fuller, Z. Fan, C. Day, and C. Barlow. Digital twin: Enabling technologies, challenges and open research. *IEEE Access*, 8:108952–108971, 2020.

[158] C. Galindo, J.-A. Fernandez-Madrigal, and J. Gonzalez. Improving efficiency in mobile robot task planning through world abstraction. *IEEE Transactions on Robotics*, 20(4):677–690, 2004.

[159] L. Galli, T. Levato, F. Schoen, and L. Tigli. Prescriptive analytics for inventory management in health care. *Journal of the Operational Research Society*, pages 1–14, 2020.

[160] D. Galvez-Lopez and J. D. Tardos. Real-time loop detection with bags of binary words. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 51–58. IEEE, 2011.

[161] S. García, J. Luengo, and F. Herrera. *Data preprocessing in data mining*, volume 72. Springer, 2015.

[162] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[163] S. Garg, N. Sünderhauf, F. Dayoub, D. Morrison, A. Cosgun, G. Carneiro, Q. Wu, T. Chin, I. D. Reid, S. Gould, P. Corke, and M. Milford. Semantics for robotic mapping, perception and interaction: A survey. *CoRR*, abs/2101.00443, 2021.

[164] C. R. Garrett, C. Paxton, T. Lozano-Pérez, L. P. Kaelbling, and D. Fox. Online replanning in belief space for partially observable task and motion problems. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5678–5684. IEEE, 2020.

[165] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer vision and image understanding*, 73(1):82–98, 1999.

[166] M. Ghallab, D. Nau, and P. Traverso. *Automated planning and acting*. Cambridge University Press, 2016.

[167] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 2019.

[168] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

[169] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018.

[170] A. Glover, W. Maddern, M. Warren, S. Reid, M. Milford, and G. Wyeth. Openfabmap: An open source toolbox for appearance-based loop closure detection. In *2012 IEEE International Conference on Robotics and Automation*, pages 4730–4735. IEEE, 2012.

[171] S. L. Goldman, R. N. Nagel, and K. Preiss. *Agile competitors and virtual organizations: strategies for enriching the customer*, volume 8. Van Nostrand Reinhold New York, 1995.

[172] M. Gombolay, A. Bair, C. Huang, and J. Shah. Computational design of mixed-initiative human-robot teaming that considers human factors: situational awareness, workload, and workflow preferences. *The International Journal of Robotics Research*, 36(5-7):597–617, 2017.

[173] M. Gombolay, R. Wilcox, and J. Shah. Fast scheduling of multi-robot teams with temporospatial constrints. In *in Proceedings of the Robots: Science and Systems (RSS)*, 2013.

[174] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, and E.-h. Zahzah. Human pose estimation from monocular images: A comprehensive survey. *Sensors*, 16(12):1966, 2016.

[175] V. Gopinath and K. Johansen. Understanding situational and mode awareness for safe human-robot collaboration: case studies on assembly applications. *Production Engineering*, 13(1):1–9, 2019.

[176] F. Gouidis, A. Vassiliades, T. Patkos, A. A. Argyros, N. Bassiliades, and D. Plexousakis. A review on intelligent object perception methods combining knowledge-based reasoning and machine learning. In *AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice, (AAAI-MAKE), also available at CoRR, arXiv*, March 2020.

[177] O. M. Group. Business Process Model and Notation (BPMN), Version 2.0.2, January 2014.

[178] P. Guo, J. Wang, X. H. Geng, C. S. Kim, and J.-U. Kim. A variable threshold-value authentication architecture for wireless mesh networks. *ŁŁ*, 15(6):929–935, 2014.

[179] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, and N. M. Kwok. A comprehensive performance evaluation of 3D local feature descriptors. *International Journal of Computer Vision*, 116(1):66–89, 2016.

[180] S. Haas, T. Ulz, and C. Steger. Secured offline authentication on industrial mobile robots using biometric data. In *Robot World Cup*, pages 143–155. Springer, 2017.

[181] S. Haddadin, A. Albu-Schaffer, A. De Luca, and G. Hirzinger. Collision detection and reaction: A contribution to safe physical human-robot interaction. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3356–3363, 2008.

[182] D. Hadfield-Menell, E. Groshev, R. Chitnis, and P. Abbeel. Modular task and motion planning in belief space. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4991–4998. IEEE, 2015.

[183] D. Hardt et al. The oauth 2.0 authorization framework, 2012.

[184] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun. PVN3D: A deep point-wise 3D keypoints voting network for 6DoF pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11632–11641, 2020.

[185] Z. He, W. Feng, X. Zhao, and Y. Lv. 6d pose estimation of objects: Recent technologies and challenges. *Applied Sciences*, 11(1):228, 2021.

[186] N. Heidari and A. Iosifidis. Temporal attention-augmented graph convolutional network for efficient skeleton-based human action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7907–7914. IEEE, 2021.

[187] M. Helmert. The fast downward planning system. *Journal of Artificial Intelligence Research*, 26:191–246, 2006.

[188] M. Helmert. The fast downward planning system. *CoRR*, abs/1109.6051, 2011.

[189] A. Hentout, M. Aouache, A. Maoudj, and I. Akli. Humanrobot interaction in industrial collaborative robotics: a literature review of the decade 20082017. *Advanced Robotics*, 33(15-16):764–799, 2019.

[190] S. Herath, M. Harandi, and F. Porikli. Going deeper into action recognition. *Image Vision Comput.*, 60(C):421, Apr. 2017.

[191] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit. Gradient response maps for real-time detection of textureless objects. *IEEE transactions on pattern analysis and machine intelligence*, 34(5):876–888, 2011.

[192] N. N. Hoang, G. Lee, S. Kim, and H. Yang. Continuous hand gesture spotting and classification using 3d finger joints information. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 539–543, 2019.

[193] T. Hodaň, D. Baráth, and J. Matas. EPOS: Estimating 6D pose of objects with symmetries. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[194] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017.

[195] T. Hodan, R. Kouskouridas, T.-K. Kim, F. Tombari, K. Bekris, B. Drost, T. Groueix, K. Walas, V. Lepetit, A. Leonardis, et al. A summary of the 4th international workshop on˜ recovering 6D object pose. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[196] T. Hodaň, M. Sundermeyer, B. Drost, Y. Labbé, E. Brachmann, F. Michel, C. Rother, and J. Matas. BOP challenge 2020 on 6D object localization. In *European Conference on Computer Vision*, pages 577–594. Springer, 2020.

[197] T. Hodaň, X. Zabulis, M. Lourakis, Š. Obdržálek, and J. Matas. Detection and fine 3D pose estimation of texture-less objects in RGB-D images. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4421–4428. IEEE, 2015.

[198] G. Hoffman. Evaluating fluency in humanrobot collaboration. *IEEE Transactions on Human-Machine Systems*, 49:209–218, 2019.

[199] H. Holm, T. Sommestad, J. Almroth, and M. Persson. A quantitative evaluation of vulnerability scanning. *Information Management & Computer Security*, 2011.

[200] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE Journal of selected topics in signal processing*, 6(5):538–552, 2012.

[201] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 495–504, June 2021.

[202] V. C. Hu, D. Ferraiolo, R. Kuhn, A. R. Friedman, A. J. Lang, M. M. Cogdell, A. Schnitzer, K. Sandlin, R. Miller, K. Scarfone, et al. Guide to attribute based access control (abac) definition and considerations (draft). *NIST special publication*, 800(162), 2013.

[203] S. Huang and G. Dissanayake. A critique of current developments in simultaneous localization and mapping. *International Journal of Advanced Robotic Systems*, 13(5):1–13, 2016.

[204] M. Huber, C. Lenz, C. Wendt, B. Färber, A. Knoll, and S. Glasauer. Predictive mechanisms increase efficiency in robot-supported assemblies: An experimental evaluation. 2013.

[205] L. Hunsberger. A faster algorithm for checking the dynamic controllability of simple temporal networks with uncertainty. In *in Proceedings of the 6th International Conference on Agents and Artificial Intelligence (ICAART)*, 2014.

[206] M.-S. Hwang and L.-H. Li. A new remote user authentication scheme using smart cards. *IEEE Transactions on consumer Electronics*, 46(1):28–30, 2000.

[207] K. Ikuta, H. Ishii, and M. Nokata. Safety evaluation method of design and control for human-care robots. *The International Journal of Robotics Research*, 22(5):281–297, 2003.

[208] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele. Arttrack: Articulated multi-person tracking in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6457–6465, 2017.

[209] International Ergonomics Association. Definition, domains of specialization, systemic approach. https://iea.cc/definition-and-domains-of-ergonomics/. Last accessed on 2021-07-20.

[210] S. Isaacson, G. Rice, and J. C. B. Jr. Mad-tn: A tool for measuring fluency in human-robot collaboration, 2019.

[211] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011.

[212] H. Jahncke, S. Hygge, N. Halin, A. M. Green, and K. Dimberg. Open-plan office noise: Cognitive performance and restoration. *Journal of Environmental Psychology*, 31(4):373–382, 2011.

[213] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013.

[214] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020.

[215] J. Jones, G. D. Hager, and S. Khudanpur. Toward computer vision systems that understand real-world assembly processes. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 426–434. IEEE, 2019.

[216] J. D. Jones, C. Cortesa, A. Shelton, B. Landau, S. Khudanpur, and G. D. Hager. Fine-grained activity recognition for assembly videos. *IEEE Robotics and Automation Letters*, 6(2):3728–3735, 2021.

[217] J. Jost, T. Kirks, S. Chapman, and G. Rinkenauer. Examining the effects of height, velocity and emotional representation of a social transport robot and human factors in human-robot collaboration. In *IFIP Conference on Human-Computer Interaction*, pages 517–526. Springer, 2019.

[218] Y. jun Zhang, N. Huang, R. G. Radwin, Z. Wang, and J. Li. Flow time in a human-robot collaborative assembly process: Performance evaluation, system properties, and a case study. *IISE Transactions*, 0(0):1–13, 2021.

[219] A. Kadambi, A. Bhandari, and R. Raskar. *3D Depth Cameras in Vision: Benefits and Limitations of the Hardware*, pages 3–26. Springer International Publishing, Cham, 2014.

[220] L. P. Kaelbling and T. Lozano-Pérez. Integrated task and motion planning in belief space. *The International Journal of Robotics Research*, 32(9-10):1194–1227, 2013.

[221] S. Kaewunruen, P. Rungskunroch, and J. Welsh. A digital-twin evaluation of net zero energy building for existing buildings. *Sustainability*, 11(1), 2019.

[222] K. Kahatapitiya and M. S. Ryoo. Coarse-fine networks for temporal activity detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8385–8394, June 2021.

[223] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, and V. V. et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint*, 1806.10293, 2018.

[224] S. Kandula, S. Krishnamoorthy, and D. Roy. A prescriptive analytics framework for efficient e-commerce order delivery. *Decision Support Systems*, page 113584, 2021.

[225] A. H. Karp, H. Haury, and M. H. Davis. From abac to zbac: the evolution of access control models. *Journal of Information Warfare*, 9(2):38–46, 2010.

[226] E. Karpas and D. Magazzeni. Automated planning for robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:417–439, 2020.

[227] N. Keddis, G. Kainz, A. Zoitl, and A. Knoll. Modeling production workflows in a mass customization era. In *2015 IEEE International Conference on Industrial Technology (ICIT)*, pages 1901–1906, March 2015.

[228] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab. SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again. In *Proceedings of the IEEE international conference on computer vision*, pages 1521–1529, 2017.

[229] B. Kenig and T. Tassa. A practical approximation algorithm for optimal k-anonymity. *Data Mining and Knowledge Discovery*, 25(1):134–168, 2012.

[230] O. Khatib, K. Yokoi, O. Brock, K. Chang, and A. Casal. Robots in human environments: basic autonomous capabilities. *International Journal of Robotics Research*, 18(7):684–696, 1999.

[231] S.-h. Kim and Y. Hwang. A survey on deep learning based methods and datasets for monocular 3D object detection. *Electronics*, 10(4):517, 2021.

[232] T. S. Kim and A. Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 1623–1631. IEEE, 2017.

[233] W. Kim, J. Lee, L. Peternel, N. Tsagarakis, and A. Ajoudani. Anticipatory robot assistance for the prevention of human static joint overloading in humanrobot collaboration. *IEEE Robotics and Automation Letters*, 3(1):68–75, 2018.

[234] W. Kim, J. Lee, N. Tsagarakis, and A. Ajoudani. A real-time and reduced-complexity approach to the detection and monitoring of static joint overloading in humans. In *2017 International Conference on Rehabilitation Robotics (ICORR)*, pages 828–834, 2017.

[235] W. Kim, M. Lorenzini, P. Balatti, P. D. Nguyen, U. Pattacini, V. Tikhanoff, L. Peternel, C. Fantacci, L. Natale, G. Metta, and A. Ajoudani. Adaptable workstations for human-robot collaboration: A reconfigurable framework for improving worker ergonomics and productivity. *IEEE Robotics Automation Magazine*, 26(3):14–26, 2019.

[236] G. Klein, P. J. Feltovich, J. M. Bradshaw, and D. D. Woods. Common ground and coordination in joint activity. *Organizational simulation*, 53:139–184, 2005.

[237] G. Klein and D. Murray. Parallel tracking and mapping on a camera phone. In *8th IEEE International Symposium on Mixed and Augmented Reality*, pages 83–86, 2009.

[238] D. Kofer, C. Bergner, C. Deuerlein, R. Schmidt-Vollus, and P. Hess. Human-robot-collaboration: Innovative processes, from research to series standard. *Procedia CIRP*, 97:98–103, 2021. 8th CIRP Conference of Assembly Technology and Systems.

[239] A. Kolbeinsson, E. Lagerstedt, and J. Lindblom. Foundation for a classification of collaboration levels for human-robot cooperation in manufacturing. *Production & Manufacturing Research*, 7(1):448–471, 2019.

[240] R. König and B. Drost. A hybrid approach for 6dof pose estimation. In *European Conference on Computer Vision*, pages 700–706. Springer, 2020.

[241] K. Konolige and M. Agrawal. FrameSLAM: From bundle adjustment to real-time visual mapping. *IEEE Transactions on Robotics*, 24(5):1066–1077, 2008.

[242] D. Konstantinidis, K. Dimitropoulos, and P. Daras. Towards real-time generalized ergonomic risk assessment for the prevention of musculoskeletal disorders. In *14th ACM International Conference on Pervasive Technologies Related to Assistive Environments Conference (PETRA)*, June-July 2021.

[243] O. Köpüklü, N. Kose, A. Gunduz, and G. Rigoll. Resource efficient 3d convolutional neural networks. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1910–1919. IEEE, 2019.

[244] N. Kousi, C. Gkournelos, S. Aivaliotis, C. Giannoulis, G. Michalos, and S. Makris. Digital twin for adaptation of robots behavior in flexible robotic assembly lines. *Procedia Manufacturing*, 28:121–126, 2019. 7th International conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV2018).

[245] C. Kraft. *User Experience Innovation*. Apress, 2012.

[246] K. Kritikos, K. Magoutis, M. Papoutsakis, and S. Ioannidis. A survey on vulnerability assessment tools and databases for cloud-based web applications. *Array*, 3:100011, 2019.

[247] T. Kröger, B. Finkemeyer, and F. M. Wahl. Manipulation primitives – a universal interface between sensor-based motion control and robot programming. In *Robotic Systems for Handling and Assembly*, pages 293–313. Springer, 2010.

[248] A. Krull, E. Brachmann, F. Michel, M. Y. Yang, S. Gumhold, and C. Rother. Learning analysis-by-synthesis for 6d pose estimation in rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pages 954–962, 2015.

[249] J. Krger and T. D. Nguyen. Automated vision-based live ergonomics analysis in assembly operations. *CIRP Annals*, 64(1):9–12, 2015.

[250] D. Kulić and E. A. Croft. Real-time safety for human-robot interaction. *Robotics and Autonomous Systems*, 54(1):1–12, 2006.

[251] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. G2o: A general framework for graph optimization. In *2011 IEEE International Conference on Robotics and Automation*, pages 3607–3613, 2011.

[252] Y. Labbe, J. Carpentier, M. Aubry, and J. Sivic. CosyPose: Consistent multi-view multi-object 6D pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[253] L. Lamport. Password authentication with insecure communication. *Communications of the ACM*, 24(11):770–772, 1981.

[254] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann. Industry 4.0. *Business & Information Systems Engineering*, 6(4):239–242, June 2014.

[255] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[256] Q. Lei, J.-X. Du, H. Zhang, S. Ye, and D. Chen. A survey of vision-based human action evaluation methods. *Sensors (Basel, Switzerland)*, 19, 2019.

[257] K. Lepenioti, A. Bousdekis, D. Apostolou, and G. Mentzas. Prescriptive analytics: A survey of approaches and methods. In *Business Information Systems Workshops*, pages 449–460. Springer International Publishing, 2019.

[258] K. Lepenioti, A. Bousdekis, D. Apostolou, and G. Mentzas. Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, 50:57–70, Feb. 2020.

[259] K. Lepenioti, M. Pertselakis, A. Bousdekis, A. Louca, F. Lampathaki, D. Apostolou, G. Mentzas, and S. Anastasiou. Machine learning for predictive and prescriptive analytics of operational data in smart manufacturing. In *International Conference on Advanced Information Systems Engineering*, pages 5–16. Springer, 2020.

[260] V. Lepetit, F. Moreno-Noguer, and P. Fua. EPnP: An accurate O(n) solution to the PnP problem. *International journal of computer vision*, 81(2):155, 2009.

[261] B. Li, H. Chen, Y. Chen, Y. Dai, and M. He. Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 613–616, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[262] B. Li, M. He, Y. Dai, X. Cheng, and Y. Chen. 3d skeleton based action recognition by video-domain translation-scale invariant mapping and multi-scale dilated cnn. *Multimedia Tools and Applications*, 77:22901–22921, 2018.

[263] C. Li and S. Lee. *Computer Vision Techniques for Worker Motion Analysis to Reduce Musculoskeletal Disorders in Construction*, pages 380–387.

[264] G. LI and P. BUCKLE. Current techniques for assessing physical exposure to work-related musculoskeletal risks, with emphasis on posture-based methods. *Ergonomics*, 42(5):674–695, 1999.

[265] J. Li, Y. Wu, Y. Gaur, C. Wang, R. Zhao, and S. Liu. On the comparison of popular end-to-end models for large scale speech recognition. In H. Meng, B. Xu, and T. F. Zheng, editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 1–5. ISCA, 2020.

[266] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[267] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[268] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox. DeepIM: Deep iterative matching for 6D pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018.

[269] Z. Li, G. Wang, and X. Ji. CDPN: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7678–7687, 2019.

[270] C. Linares Lpez, S. Jimnez Celorrio, and ngel Garca Olaya. The deterministic part of the seventh international planning competition. *Artificial Intelligence*, 223:82–119, 2015.

[271] R. Lindorfer, R. Froschauer, and G. Schwarz. ADAPT - a decision-model-based approach for modeling collaborative assembly and manufacturing tasks. In *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*, pages 559–564, July 2018.

[272] A. X. Liu, F. Chen, J. Hwang, and T. Xie. Designing fast and scalable XACML policy evaluation engines. *IEEE Transactions on Computers*, 60(12):1802–1817, 2011.

[273] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[274] J. Liu, Z. Zou, X. Ye, X. Tan, E. Ding, F. Xu, and X. Yu. Leaping from 2D detection to efficient 6DoF object pose estimation. In *European Conference on Computer Vision*, pages 707–714. Springer, 2020.

[275] Z. Liu, J. Zhu, J. Bu, and C. Chen. A survey of human pose estimation: the body parts parsing based methods. *Journal of Visual Communication and Image Representation*, 32:10–19, 2015.

[276] J. Lopes, T. Guimarães, and M. F. Santos. Predictive and prescriptive analytics in healthcare: A survey. *Procedia Computer Science*, 170:1029–1034, 2020.

[277] M. I. Lourakis and A. A. Argyros. SBA: A software package for generic sparse bundle adjustment. *ACM Transactions on Mathematical Software*, 36(1):1–30, 2009.

[278] S. Louvan and B. Magnini. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics.

[279] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[280] S. Lowry, N. Snderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2016.

[281] Y. Lu, C. Liu, K. I.-K. Wang, H. Huang, and X. Xu. Digital twin-driven smart manufacturing: Connotation, reference model, applications and research issues. *Robotics and Computer-Integrated Manufacturing*, 61:101837, 2020.

[282] D. C. Luvizon, H. Tabia, and D. Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, 2019.

[283] Q. Lv, R. Zhang, X. Sun, Y. Lu, and J. Bao. A digital twin-driven human-robot collaborative assembly approach in the wake of covid-19. *Journal of Manufacturing Systems*, 2021.

[284] C.-Y. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, and H. P. Graf. Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6800, 2018.

[285] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

[286] R. Maclin and J. W. Shavlik. Creating advice-taking reinforcement learners. *Machine Learning*, 22(1):251–281, 1996.

[287] W. Mahnke, S.-H. Leitner, and M. Damm. *OPC unified architecture*. Springer Science & Business Media, 2009.

[288] O. Maimon. The robot task-sequencing planning problem. *IEEE Transactions on Robotics and Automation*, 6(6):760–765, 1990.

[289] A. A. Malik and A. Bilberg. Digital twins of human robot collaboration in a production setting. *Procedia Manufacturing*, 17:278–285, 2018. 28th International Conference on Flexible Automation and Intelligent Manufacturing (FAIM2018), June 11-14, 2018, Columbus, OH, USAGlobal Integration of Intelligent Manufacturing and Smart Industry for Good of Humanity.

[290] A. A. Malik and A. Bilberg. Collaborative robots in assembly: A practical approach for tasks distribution. volume 81, 06 2019.

[291] A. A. Malik and A. Brem. Digital twins for collaborative robots: A case study in human-robot interaction. *Robotics and Computer-Integrated Manufacturing*, 68:102092, 2021.

[292] J. G. Mangelson, D. Dominic, R. M. Eustice, and R. Vasudevan. Pairwise consistent measurement set maximization for robust multi-robot map merging. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2916–2923. IEEE, 2018.

[293] F. Manhardt, D. M. Arroyo, C. Rupprecht, B. Busam, T. Birdal, N. Navab, and F. Tombari. Explaining the ambiguity of object detection and 6D pose from visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6841–6850, 2019.

[294] M. Maniadakis, E. Aksoy, T. Asfour, and P. Trahanias. Collaboration of heterogeneous agents in time constrained tasks. In *Proc. IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2016.

[295] M. Maniadakis, E. Hourdakis, and P. Trahanias. Time-informed task planning in multi-agent collaboration. *Cognitive Systems Research*, 2016.

[296] M. Maniadakis and P. Trahanias. Time-informed, adaptive multi-robot synchronization. In *From Animals to Animats 14*, pages 232–243, Cham, 2016. Springer International Publishing.

[297] N. A. F. Marc Zupan, Mike F. Ashby. *Actuator Classification and Selection- The Development of a Database*. N.A., https://onlinelibrary.wiley.com/doi/10.1002/adem.200290009, 2002.

[298] A. e. a. Martnez-Gutirrez. Digital twin for automatic transportation in industry 4.0. *PubMed, Sensors (Basel, Switzerland)*, 21, 2021.

[299] J. Materzynska, G. Berger, I. Bax, and R. Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2874–2882, 2019.

[300] J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang, and T. Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020.

[301] B. Matthias. Iso/ts 15066 - collaborative robots - present status, 03 2015.

[302] L. Mcatamney and E. N. Corlett. Rula: a survey method for the investigation of work-related upper limb disorders. *Applied ergonomics*, 24 2:91–9, 1993.

[303] L. McAtamney and S. Hignett. Rapid entire body assessment. In *Handbook of Human Factors and Ergonomics Methods*, pages 97–108. CRC Press, 2004.

[304] D. McDermott. Regression planning. *International Journal of Intelligent Systems*, 6(4):357–416, 1991.

[305] R. Mehrizi, X. Peng, Z. Tang, X. Xu, D. N. Metaxas, and K. Li. Toward marker-free 3d pose estimation in lifting: A deep multi-view solution. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 485–491, 2018.

[306] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017.

[307] J. Melvin, P. Keskinocak, S. Koenig, C. A. Tovey, and B. Y. Ozkaya. Multi-robot routing with rewards and disjoint time windows. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, page 23322337, 2007.

[308] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228, 2004.

[309] G. Michalos, S. Makris, P. Tsarouchi, T. Guasch, D. Kontovrakis, and G. Chryssolouris. Design considerations for safe human-robot collaborative workplaces. *Procedia CIRP*, 37:248 – 253, 2015. CIRPe 2015 - Understanding the life cycle implications of manufacturing.

[310] G. Michalos, J. Spiliotopoulos, S. Makris, and G. Chryssolouris. A method for planning human robot shared tasks. *CIRP Journal of Manufacturing Science and Technology*, 22:76 – 90, 2018.

[311] Microsoft. Microsoft Kinect SDK, 2011.

[312] Y. Min, Y. Zhang, X. Chai, and X. Chen. An efficient pointlstm for point clouds based gesture recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5760–5769, 2020.

[313] R. Mizanoor and W. Yue. Dynamic affection-based motion control of a humanoid robot to collaborate with human in flexible assembly in manufacturing. page V003T40A005, 10 2015.

[314] T. B. Moeslund, A. Hilton, and V. Krger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126, 2006. Special Issue on Modeling People: Vision-based understanding of a persons shape, appearance, movement and behaviour.

[315] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4207–4215, 2016.

[316] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *Eighteenth National Conference on Artificial Intelligence*, pages 593–598, USA, 2002. American Association for Artificial Intelligence.

[317] A. Mordvintsev and A. K. Dense optical flow with OpenCV, 2013.

[318] P. Morris. Dynamic controllability and dispatchability relationships. In *Integration of AI and OR Techniques in Constraint Programming*, Lecture Notes in Computer Science, pages 464–479. Springer International Publishing, 2014.

[319] P. Morris, N. Muscettola, and V. T. Dynamic control of plans with temporal uncertainty. In *Intl. Joint Conf. on AI (IJCAI)*, 2001.

[320] M. Moser, M. Pfeiffer, and J. Pichler. Domain-specific modeling in industrial automation: Challenges and experiences. In *Proceedings of the 1st International Workshop on Modern Software Engineering Methods for Industrial Automation*, MoSEMInA 2014, pages 42–51, New York, NY, USA, 2014. ACM.

[321] K. L. Mosier, U. Fischer, B. K. Burian, and J. A. Kochan. Autonomous, context-sensitive, task management systems and decision support tools i: Human-autonomy teaming fundamentals and state of the art. 2017.

[322] A. I. Mourikis and S. I. Roumeliotis. A multi-state constraint Kalman filter for vision-aided inertial navigation. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3565–3572, 2007.

[323] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, Oct. 2015.

[324] P. Narayana, J. R. Beveridge, and B. A. Draper. Gesture Recognition: Focus on the Hands. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5235–5244, Salt Lake City, UT, June 2018. IEEE.

[325] P. Narayana, J. R. Beveridge, and B. A. Draper. Continuous gesture recognition through selective temporal fusion. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019.

[326] E. Negri, L. Fumagalli, and M. Macchi. A review of the roles of digital twin in cps-based production systems. *Procedia Manuf.*, 11:939948, 2017.

[327] M. N. Nicolescu and M. J. Mataric. Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 241–248, 2003.

[328] R. Niese, A. Al-Hamadi, and B. Michaelis. A novel method for 3D face detection and normalization. *Journal of Multimedia*, 2(5), 2007.

[329] P. M. Notz and R. Pibernik. Prescriptive analytics for flexible capacity management. *Management Science*, 2021.

[330] M. Nübling, U. Stößel, H.-M. Hasselhorn, M. Michaelis, and F. Hofmann. Measuring psychological stress and strain at work-evaluation of the copsoq questionnaire in germany. *GMS Psycho-Social Medicine*, 3, 2006.

[331] E. Nunes and M. Gini. Multi-robot auctions for allocation of tasks with temporal constraints. In *AAAI Conf. on Artificial Intelligence*, page 21102116, 2015.

[332] E. Nunes, M. Nanjanath, and M. Gini. Auctioning robotic tasks with overlapping time windows. In *Intl Conf. on Autonomous Agents and Multi-Agent Systems*, page 12111212, 2012.

[333] NVIDIA. Nvidia optical flow SDK.

[334] E. Occhipinti and D. Colombini. Updating reference values and predictive models of the ocra method in the risk assessment of work-related musculoskeletal disorders of the upper limbs. *Ergonomics*, 50(11):1727–1739, 2007.

[335] T. U. of Applied Sciences and A. of Southern Switzerland. Manufacturing through ergonomic and safe anthropocentric adaptive workplaces for context aware factories in europe. hg. v. cordis. online document. https://cordis.europa.eu/project/id/609073, 2016.

[336] U. of California. Video-caffe: Caffe with c3d implementation and video reader, 2016.

[337] L. Onnasch and E. Roesler. A taxonomy to structure and analyze human-robot interaction. *International Journal of Social Robotics*, 2020.

[338] A. Ouaddah, H. Mousannif, A. Abou Elkalam, and A. A. Ouahman. Access control in the internet of things: Big challenges and new opportunities. *Computer Networks*, 112:237–262, 2017.

[339] K. Papadamou, S. Zannettou, B. Chifor, S. Teican, G. Gugulea, A. Caponi, A. Recupero, C. Pisa, G. Bianchi, S. Gevers, et al. Killing the password and preserving privacy with device-centric and attribute-based authentication. *IEEE Transactions on Information Forensics and Security*, 15:2183–2193, 2019.

[340] K. Papoutsakis, T. Papadopoulos, M. Maniadakis, M. Lourakis, M. Pateraki, and I. Varlamis. Detection of physical strain and fatigue in industrial environments using visual and non-visual sensors. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*, PETRA 2021, page 270271, New York, NY, USA, 2021. Association for Computing Machinery.

[341] K. E. Papoutsakis and A. A. Argyros. Unsupervised and explainable assessment of video similarity. In *BMVC*, page 151, 2019.

[342] B. Parducci, H. Lockhart, and E. Rissanen. eXtensible Access Control Markup Language (XACML) Version 3.0, 2013.

[343] H. Park and K. Shim. Approximate algorithms for k-anonymity. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 67–78, 2007.

[344] K. Park, T. Patten, and M. Vincze. Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7668–7677, 2019.

[345] B. Parsa and A. G. Banerjee. A multi-task learning approach for human activity segmentation and ergonomics risk assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2352–2362, January 2021.

[346] B. Parsa, A. L. narayanan, and B. Dariush. Spatio-temporal pyramid graph convolutions for human action recognition and postural assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[347] B. Parsa, E. U. Samani, R. Hendrix, C. Devine, S. M. Singh, S. Devasia, and A. G. Banerjee. Toward ergonomic risk prediction via segmentation of indoor object manipulation actions using spatiotemporal convolutional networks. *IEEE Robotics and Automation Letters*, 4(4):3153–3160, 2019.

[348] F. Pauker, I. Ayatollahi, and B. Kittl. Service orchestration for flexible manufacturing systems using sequential functional charts and OPC UA, Sep. 2015.

[349] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis. 6-DoF object pose from semantic keypoints. In *IEEE international conference on robotics and automation (ICRA)*, pages 2011–2018. IEEE, 2017.

[350] M. R. Pedersen. Robot skills for transformable manufacturing systems, 2015.

[351] M. R. Pedersen, D. L. Herzog, and V. Kruger. Intuitive skill-level programming of industrial handling tasks on a mobile manipulator. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, sep 2014.

[352] M. R. Pedersen and V. Krger. Gesture-based extraction of robot skill parameters for intuitive robot programming. *Journal of Intelligent & Robotic Systems*, 80(S1):149–163, feb 2015.

[353] M. R. Pedersen, L. Nalpantidis, R. S. Andersen, C. Schou, S. Bøgh, V. Krüger, and O. Madsen. Robot skills for manufacturing: From concept to industrial deployment. *Robotics and Computer-Integrated Manufacturing*, 37:282–291, 2016.

[354] M. R. Pedersen, L. Nalpantidis, R. S. Andersen, C. Schou, S. Bøgh, V. Krger, and O. Madsen. Robot skills for manufacturing: From concept to industrial deployment. *Robotics and Computer-Integrated Manufacturing*, 37:282–291, feb 2016.

[355] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[356] S. Pellegrinelli, A. Orlandini, N. Pedrocchi, A. Umbrico, and T. Tolio. Motion planning and scheduling for human and industrial-robot collaboration. *CIRP Annals*, 66(1):1–4, 2017.

[357] D. Pellier and H. Fiorino. Totally and partially ordered hierarchical planners in PDDL4J library. *CoRR*, abs/2011.13297, 2020.

[358] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao. PVNet: Pixel-wise voting network for 6DoF pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019.

[359] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (TOG)*, 37(4):114, 2018.

[360] X. Perez-Sala, S. Escalera, C. Angulo, and J. Gonzalez. A survey on model based approaches for 2d and 3d visual human pose recovery. *Sensors*, 14(3):4189–4210, 2014.

[361] D. Pessach, G. Singer, D. Avrahami, H. C. Ben-Gal, E. Shmueli, and I. Ben-Gal. Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, 134:113290, 2020.

[362] C. Phiquepal and M. Toussaint. Combined task and motion planning under partial observability: An optimization-based approach. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9000–9006. IEEE, 2019.

[363] A. Pichler, S. C. Akkaladevi, M. Ikeda, M. Hofmann, M. Plasch, C. Wögerer, and G. Fritz. Towards shared autonomy for robotic tasks in manufacturing. *Procedia Manufacturing*, 11:72–82, 2017.

[364] L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre. Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. *International Journal of Computer Vision*, 126(2-4):430–439, Apr. 2018.

[365] P. Plantard, H. Shum, A.-S. Pierres, and F. Multon. Validation of an ergonomic assessment method using kinect data in real workplace conditions. *Applied Ergonomics*, 65, 11 2016.

[366] S. Poornima and M. Pushpalatha. A survey on various applications of prescriptive analytics. *International Journal of Intelligent Networks*, 1:76–84, 2020.

[367] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[368] H. Prähofer, D. Hurnaus, R. Schatz, C. Wirth, and H. Mössenböck. Monaco: A DSL approach for programming automation systems. In *Software Engineering*, pages 242–256. Citeseer, 2008.

[369] A. PrimeSense, Willow Garage. Open natural-interaction, 2011.

[370] S. Profanter, K. Dorofeev, A. Zoitl, and A. Knoll. OPC UA for plug & produce: Automatic device discovery using LDS-ME. In *Proceedings of the IEEE International Conference on Emerging Technologies And Factory Automation (ETFA)*, 2017.

[371] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021.

[372] F. Py, K. Rajan, and C. McGann. A systematic agent framework for situated autonomous systems. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 2 - Volume 2*, AAMAS '10, page 583590. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

[373] A. Qammaz and A. A. Argyros. Occlusion-tolerant and personalized 3d human pose estimation in rgb images. In *IEEE International Conference on Pattern Recognition (ICPR 2020)*, January 2021.

[374] Q. Qi, F. Tao, T. Hu, N. Anwer, A. Liu, Y. Wei, L. Wang, and A. Nee. Enabling technologies and tools for digital twin. *Journal of Manufacturing Systems*, 58:3–21, 2021. Digital Twin towards Smart Manufacturing and Industry 4.0.

[375] M. Quigley, J. Faust, T. Foote, J. Leibs, et al. ROS: an open-source robot operating system. In *ICRA workshop on open source software*, 2009.

[376] M. Rad and V. Lepetit. BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3828–3836, 2017.

[377] M. M. Rahman, Y. Tan, J. Xue, and K. Lu. Recent advances in 3d object detection in the era of deep neural networks: A survey. *IEEE Transactions on Image Processing*, 29:2947–2962, 2019.

[378] K. Rajan, C. McGann, F. Py, and H. Thomas. Robust mission planning using deliberative autonomy for autonomous underwater vehicles. In *Proc. ICRA workshop on robotics in challenging and hazardous environments*, pages 21–25, 2007.

[379] C. Rascon and I. Meza. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 96:184–210, 2017.

[380] S. J. Raychaudhuri, S. Manjunath, C. P. Srinivasan, N. Swathi, S. Sushma, K. N. N. Bhushan, and C. N. Babu. Prescriptive analytics for impulsive behaviour prevention using real-time biometrics. *Progress in Artificial Intelligence*, 10(2):99–112, Jan. 2021.

[381] J. Rennecker and L. Godwin. Delays and interruptions: A self-perpetuating paradox of communication technology use. *Information and Organization*, 15(3):247–266, 2005. Technology as Organization/Disorganization.

[382] K. Revell, P. Langdon, M. Bradley, I. Politis, J. Brown, and N. Stanton. User centered ecological interface design (UCEID): a novel method applied to the problem of safe and user-friendly interaction between drivers and autonomous vehicles. In *International Conference on Intelligent Human Systems Integration*, pages 495–501. Springer, 2018.

[383] G. Rinkenauer, A. Böckenkamp, and F. Weichert. Man-robot collaboration in the context of industry 4.0: Approach-avoidance tendencies as an indicator for the affective quality of interaction? In *Advances in Ergonomic Design of Systems, Products and Processes*, pages 335–348. Springer, 2017.

[384] G. Rinkenauer and T. Plewan. Geschwindigkeits-genauigkeitsabgleich und körperhaltung: Altersbedingte unterschiede bei der balancekontrolle. In *Soziotechnische Gestaltung des digitalen Wandels–kreativ, innovativ, sinnhaft: 63. Kongress der Gesellschaft für Arbeitswissenschaft, HNW Brugg-Windisch, Schweiz*, volume 15, page 17, 2017.

[385] A. Roitberg, N. Somani, A. Perzylo, M. Rickert, and A. Knoll. Multimodal human activity recognition for industrial manufacturing processes in robotic workcells. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 259266, New York, NY, USA, 2015. Association for Computing Machinery.

[386] A. Rosinol, M. Abate, Y. Chang, and L. Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2020.

[387] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In *International Conference on Computer Vision*, pages 2564–2571, 2011.

[388] J. Rumbaugh, I. Jacobson, and G. Booch. *Unified Modeling Language Reference Manual, The (2nd Edition)*. Pearson Higher Education, 2004.

[389] P. Rupprecht and S. Schlund. Taxonomy for individualized and adaptive human-centered workplace design in industrial site assembly. In *Russo D., Ahram T., Karwowski W., Di Bucchianico G., Taiar R. (eds) Intelligent Human Systems Integration 2021. IHSI 2021. Advances in Intelligent Systems and Computing, vol 1322. Springer, Cham*, pages 576–584. IEEE, 2021.

[390] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3D model acquisition. *ACM Transactions on Graphics (TOG)*, 21(3):438–446, 2002.

[391] K. Rnick. *Analyse des Einflusses unterschiedlicher Individualisierungsgrade eines Montagearbeitsplatzes.* 2020.

[392] K. Rnick, T. Kremer, J. Wakula, S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, and Y. E. Fujita. Evaluation of an adaptive assistance system to optimize physical stress in the assembly. In *Proceedings of the 20th congress of the International Ergonomics Association (IEA 2018), Volume IX Aging, Gender and Work, Anthropometry, Ergonomics for Children an Educational Enviroments*, pages 576–584. IEEE, 2018.

[393] S. Sagiroglu and D. Sinanc. Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, May 2013.

[394] C. Sahin and T.-K. Kim. Recovering 6d object pose: a review and multi-modal analysis. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[395] S. Salas, D. Sims, and C. Burke. Is there a big five in teamwork? small gr res 36: 555–599, 2005.

[396] G. Salvendy. *Handbook of human factors and ergonomics*. John Wiley & Sons, 2012.

[397] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998.

[398] M. R. U. Saputra, A. Markham, and N. Trigoni. Visual SLAM and structure from motion in dynamic environments: A survey. *ACM Comput. Surv.*, 51(2), Feb. 2018.

[399] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1–20, 2016.

[400] H. Sarbolandi, D. Lefloch, and A. Kolb. Kinect range sensing: Structured-light versus time-of-flight kinect. *Computer Vision and Image Understanding*, 139:1–20, 2015.

[401] R. Sargent. Cdsa explained: An indispensable guide to common data security architecture. *The Open Group, Reading, Berkshire, UK*, 1998.

[402] K. Schaub, G. Caragnano, B. Britzke, and R. Bruder. The european assembly worksheet. *Theoretical Issues in Ergonomics Science*, 14(6):616–639, 2013.

[403] J. Schmidtler, V. Knott, C. Hlzel, and K. Bengler. Human centered assistance applications for the working environment of the future. *Occupational Ergonomics*, 12(3):83–95, 2015.

[404] D. Schrter, P. Jaschewski, B. Kuhrke, and A. Verl. Methodology to identify applications for collaborative robots in powertrain assembly. *Procedia CIRP*, 55:12–17, 2016. 5th CIRP Global Web Conference - Research and Innovation for Future Production (CIRPe 2016).

[405] D. Schnberger, R. Lindorfer, and R. Froschauer. Modeling workflows for industrial robots considering human-robot-collaboration. In *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*, pages 400–405, 2018.

[406] S. Sciancalepore, G. Piro, D. Caldarola, G. Boggia, and G. Bianchi. On the design of a decentralized and multiauthority access control scheme in federated and cloud-assisted cyber-physical systems. *IEEE Internet of Things Journal*, 5(6):5190–5204, 2018.

[407] M. Servieres, V. Renaudin, A. Dupuis, and N. Antigny. Visual and visual-inertial SLAM: State of the art, classification, and experimental benchmarking. *Journal of Sensors*, 2021:1–26, Feb. 2021.

[408] R. Sessa. Al centro della filosofia di marchionne, la progettazione di linee di lavoro a zero fatica. *l'Industria Meccanica*, 717:52–54, 2018.

[409] A. Shafti, A. Ataka, B. U. Lazpita, A. Shiva, H. Wurdemann, and K. Althoefer. Real-time robot-assisted ergonomics*. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1975–1981, 2019.

[410] M. Shafto, M. Conroy, R. Doyle, E. Glaessgen, C. Kemp, J. Le, Moigne, and L. Wan. *DRAFT Modeling, Simulation, Information Technology & Processing Roadmap*. National Aeronautics and Space Administration (NASA), Technology Area, 2010.

[411] J. Shah, J. Stedl, B. Williams, and P. Robertson. A fast incremental algorithm for maintaining dispatchability of partially controllable plans. In *Proc. 18th Int Conf on Automated Planning and Scheduling*, pages 296–303, 2007.

[412] D. Shan, J. Geng, M. Shu, and D. F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[413] D. Shao, Y. Zhao, B. Dai, and D. Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[414] N. K. Sharma, M. Tiwari, A. Thakur, and A. K. Ganguli. A systematic review of methodologies and techniques for integrating ergonomics into development and assessment of manually operated equipment. *International Journal of Occupational Safety and Ergonomics*, 0(0):1–13, 2021. PMID: 33308028.

[415] Z. Shelby, K. Hartke, and C. Bormann. The Constrained Application Protocol (CoAP), 2014.

[416] J. Shen, H.-W. Tan, J. Wang, J.-W. Wang, and S.-Y. Lee. A novel routing protocol providing good transmission reliability in underwater sensor networks. *ŁŁ*, 16(1):171–178, 2015.

[417] T. B. Sheridan. Human-robot interaction: Status and challenges. *Human Factors*, 58(4):525–532, 2016. PMID: 27098262.

[418] Y. Shi, J. Huang, X. Xu, Y. Zhang, and K. Xu. StablePose: Learning 6D object poses from geometrically stable patches. *arXiv preprint arXiv:2102.09334*, 2021.

[419] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Openpose, 2016.

[420] R. C. Smith and P. Cheeseman. On the representation and estimation of spatial uncertainty. *The International Journal of Robotics Research*, 5(4):56–68, 1986.

[421] S. Smith, A. Gallagher, T. Zimmerman, L. Barbulescu, and Z. Rubinstein. Distributed management of flexible times schedules. In *Int Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, 2007.

[422] C. Song, J. Song, and Q. Huang. HybridPose: 6D object pose estimation under hybrid representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 431–440, 2020.

[423] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012.

[424] P. Spielholz, B. Silverstein, M. Morgan, H. Checkoway, and J. Kaufman. Comparison of self-report, video observation and direct measurement methods for upper extremity musculoskeletal disorder physical risk factors. *Ergonomics*, 44:588 – 613, 2001.

[425] F. Spitzer, R. Lindorfer, R. Froschauer, M. Hofmann, and M. Ikeda. A generic approach for the industrial application of skill-based engineering using opc ua. In *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, volume 1, pages 1671–1678, 2020.

[426] S. Srinivas and A. R. Ravindran. Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: A prescriptive analytics framework. *Expert Systems with Applications*, 102:245–261, 2018.

[427] N. A. Stanton, P. M. Salmon, G. H. Walker, C. Baber, and D. P. Jenkins. *Human factors methods: a practical guide for engineering and design*. CRC Press, 2017.

[428] R. Stark, C. Fresemann, and K. Lindow. Development and operation of digital twins for technical systems and services. *CIRP Annals*, 68(1):129–132, 2019.

[429] F. Steinmetz and R. Weitschat. Skill parametrization approaches and skill architecture for human-robot interaction. In *2016 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE, aug 2016.

[430] S. Stock, M. Mansouri, F. Pecora, and J. Hertzberg. Online task merging with a hierarchical hybrid task planner for mobile service robots. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6459–6464, 2015.

[431] H. Strasdat, J. M. M. Montiel, and A. J. Davison. Visual SLAM: Why filter? *Image and Vision Computing*, 30(2):65–77, Feb. 2012.

[432] M. Sthr, M. Schneider, and C. Henkel. Adaptive work instructions for people with disabilities in the context of human robot collaboration. In *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*. IEEE, jul 2018.

[433] D. Summers-Stay, C. L. Teo, Y. Yang, C. Fermüller, and Y. Aloimonos. Using a minimal action grammar for activity understanding in the real world. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4104–4111. IEEE, 2012.

[434] M. Sundermeyer, Z.-C. Marton, M. Durner, and R. Triebel. Augmented autoencoders: Implicit 3D orientation learning for 6D object detection. *International Journal of Computer Vision*, 128(3):714–729, 2020.

[435] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[436] H. Tan, L. W. 0003, Q. Zhang, Z. Gao, N. Zheng, and G. Hua. Object affordances graph network for action recognition. In *BMVC*, page 145, 2019.

[437] M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[438] K.-H. Tang, C.-F. Ho, J. Mehlich, and S.-T. Chen. Assessment of handover prediction models in estimation of cycle times for manual assembly tasks in a humanrobot collaborative environment. *Applied Sciences*, 10(2), 2020.

[439] J. Teiwes, T. Bänziger, A. Kunz, and K. Wegener. Identifying the potential of human-robot collaboration in automotive assembly lines using a standardised work description. *2016 22nd International Conference on Automation and Computing (ICAC)*, pages 78–83, 2016.

[440] A. Tejani, D. Tang, R. Kouskouridas, and T.-K. Kim. Latent-class hough forests for 3D object detection and pose estimation. In *European Conference on Computer Vision*, pages 462–477. Springer, 2014.

[441] B. Tekin, F. Bogo, and M. Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2019.

[442] B. Tekin, S. N. Sinha, and P. Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 292–301, 2018.

[443] M. Tenorth, J. Bandouch, and M. Beetz. The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1089–1096, 2009.

[444] G. Terzakis and M. Lourakis. A consistently fast and globally optimal solution to the perspective-n-point problem. In *European Conference on Computer Vision (ECCV)*, pages 1–17. Springer Nature Switzerland, 2020.

[445] The FIWARE Community. Fiware: The open source platform for our smart digital future. https://www.fiware.org/, 2020. Accessed on 20.05.2020.

[446] U. Thomas, G. Hirzinger, B. Rumpe, C. Schulze, and A. Wortmann. A new skill based robot programming language using UML/P statecharts. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 461–466. IEEE, 2013.

[447] M. Thüring and S. Mahlke. Usability, aesthetics and emotions in human–technology interaction. *International journal of psychology*, 42(4):253–264, 2007.

[448] P. Tsarouchi, A.-S. Matthaiakis, S. Makris, and G. Chryssolouris. On a human-robot collaboration in an assembly cell. *International Journal of Computer Integrated Manufacturing*, 30(6):580–589, 2017.

[449] P. Tsarouchi, G. Michalos, S. Makris, T. Athanasatos, K. Dimoulas, and G. Chryssolouris. On a humanrobot workplace design and task allocation system. *International Journal of Computer Integrated Manufacturing*, 30(12):1272–1279, 2017.

[450] P. Tsarouchi, J. Spiliotopoulos, G. Michalos, S. Koukas, A. Athanasatos, S. Makris, and G. Chryssolouris. A decision making framework for human robot collaborative workplace generation. *Procedia CIRP*, 44:228 – 232, 2016. 6th CIRP Conference on Assembly Technologies and Systems (CATS).

[451] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: A survey. *Found. Trends. Comput. Graph. Vis.*, 3(3):177–280, July 2008.

[452] A. Umbrico, A. Cesta, M. Cialdea Mayer, and A. Orlandini. Platinum: A new framework for planning and acting. In F. Esposito, R. Basili, S. Ferilli, and F. A. Lisi, editors, *AI*IA 2017 Advances in Artificial Intelligence*, pages 498–512, 2017.

[453] A. Vakhutinsky, K. Mihic, and S.-M. Wu. A prescriptive analytics approach to markdown pricing for an e-commerce retailer. *Journal of Pattern Recognition Research*, 14(1):1–20, 2019.

[454] E. Valero, A. Sivanathan, F. Bosch, and M. Abdel-Wahab. Musculoskeletal disorders in construction: A review and a novel system for activity tracking with body area network. *Applied Ergonomics*, 54:120–130, 2016.

[455] Various authors. Papers with code - 6D pose estimation using RGB. https://paperswithcode.com/task/6d-pose-estimation, 2021.

[456] M. Vasic and A. Billard. Safety issues in human-robot interactions. In *2013 IEEE International Conference on Robotics and Automation*, pages 197–204, 2013.

[457] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[458] J. Vater, L. Harscheidt, and A. Knoll. Smart manufacturing with prescriptive analytics. In *2019 8th International Conference on Industrial Technology and Management (ICITM)*. IEEE, Mar. 2019.

[459] J. W. a. S. Verena Klaer, Hendrik Groll. Evaluation of different degrees of support in human-robot cooperation at an assembly workstation regarding physiological strainand perceived team fluency. In *Proceedings of IEA 2021, (Publication in preperation)*, page na. TU Darmstadt, 2021.

[460] K. J. Vicente. Ecological interface design: Progress and challenges. *Human factors*, 44(1):62–78, 2002.

[461] F. Vicentini. Collaborative robotics: a survey. *Journal of Mechanical Design*, pages 1–29, Feb. 2020.

[462] J. Vidal, C.-Y. Lin, X. Lladó, and R. Martí. A method for 6D pose estimation of free-form rigid objects using point pair features on range data. *Sensors*, 18(8):2678, 2018.

[463] J. Vidal, C.-Y. Lin, and R. Martí. 6D pose estimation using an improved method based on point pair features. In *2018 4th international conference on control, automation and robotics (iccar)*, pages 405–409. IEEE, 2018.

[464] E. R. Vieira and S. Kumar. Working postures: a literature review. *Journal of occupational rehabilitation*, 14(2):143–159, 2004.

[465] M. J. Villanueva, F. Valverde, and O. Pastor. Involving end-users in the design of a domain-specific language for the genetic domain. In M. José Escalona, G. Aragón, H. Linger, M. Lang, C. Barry, and C. Schneider, editors, *Information System Development*, pages 99–110, Cham, 2014. Springer International Publishing.

[466] N. N. Vo and A. F. Bobick. Sequential interval network for parsing complex structured activity. *Computer Vision and Image Understanding*, 143:147–158, 2016.

[467] S. Wagner, G. Kronberger, A. Beham, M. Kommenda, A. Scheibenpflug, E. Pitzer, S. Vonolfen, M. Kofler, S. Winkler, V. Dorfer, et al. Architecture and design of the heuristiclab optimization environment. In *Advanced methods and applications in computational intelligence*, pages 197–261. Springer, 2014.

[468] J. Wan, S. Z. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 761–769, 2016.

[469] J. Wan, C. Lin, L. Wen, Y. Li, Q. Miao, S. Escalera, G. Anbarjafari, I. Guyon, G. Guo, and S. Z. Li. Chalearn looking at people: IsoGD and ConGD large-scale RGB-D gesture recognition. *IEEE Transactions on Cybernetics*, pages 1–12, 2020.

[470] D. Wang, H. Cheng, D. He, and P. Wang. On the challenges in designing identity-based privacy-preserving authentication schemes for mobile devices. *IEEE Systems Journal*, 12(1):916–925, 2016.

[471] D. Wang, F. Dai, and X. Ning. Risk assessment of work-related musculoskeletal disorders in construction: State-of-the-art review. *Journal of Construction Engineering and Management*, 141(6):04015008, 2015.

[472] L. Wang, R. Gao, J. Vncza, J. Krger, X. Wang, S. Makris, and G. Chryssolouris. Symbiotic human-robot collaborative assembly. *CIRP Annals*, 68(2):701–726, 2019.

[473] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera. Rgb-d-based human motion recognition with deep learning: A survey. *Computer Vision and Image Understanding*, 171:118–139, 2018.

[474] X. Wang, A. Farhadi, and A. Gupta. Actions˜ transformations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2658–2667, 2016.

[475] X. Wang, C.-J. Liang, C. Menassa, and V. Kamat. Real-time process-level digital twin for collaborative human-robot construction work. In H. Osumi, H. Furuya, and K. Tateyama, editors, *Proceedings of the 37th International Symposium on Automation and Robotics in Construction (ISARC)*, pages 1528–1535. International Association for Automation and Robotics in Construction (IAARC), October 2020.

[476] Y. Wang, G. Ajaykumar, and C.-M. Huang. See what i see: Enabling user-centric robotic assistance using first-person demonstrations. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 639–648, 2020.

[477] Z. Wang, Q. She, T. Chalasani, and A. Smolic. Catnet: Class incremental 3d convnets for lifelong egocentric gesture recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 935–944, 2020.

[478] C. Weckenborg, K. Kieckhäfer, C. Müller, M. Grunewald, and T. S. Spengler. Balancing of assembly lines with collaborative robots. *Business Research*, 13(1):93–132, 2020.

[479] C. Weckenborg, K. Kieckhfer, C. Mller, M. Grunewald, and T. S. Spengler. Balancing of assembly lines with collaborative robots. *Business Research*, 13(1):93–132, 2020.

[480] C. Weckenborg and T. S. Spengler. Assembly line balancing with collaborative robots under consideration of ergonomics: a cost-oriented approach. *IFAC-PapersOnLine*, 52(13):1860–1865, 2019.

[481] G. Weichhart, M. Åkerman, S. C. Akkaladevi, M. Plasch, Å. Fast-Berglund, and A. Pichler. Models for interoperable human robot collaboration. *IFAC-PapersOnLine*, 51(11):36 – 41, 2018. 16th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2018.

[482] R. Wilcox, S. Nikolaidis, and S. J. A. Optimization of temporal dynamics for adaptive human-robot interaction in assembly manufacturing. *Robotics*, 441, 2013.

[483] J. Womack. From lean tools to lean management. *Lean Enterprise Institute Email Newsletter*, 21, 2006.

[484] J. Wu, D. Yin, J. Chen, Y. Wu, H. Si, and K. Lin. A survey on monocular 3D object detection algorithms based on deep learning. In *Journal of Physics: Conference Series*, volume 1518, page 012049. IOP Publishing, 2020.

[485] L. Wu, J. Li, T. Lei, and B. Li. Eid vs ucd: A comparative study on user interface design in complex electronics manufacturing systems. In *International Conference on Engineering Psychology and Cognitive Ergonomics*, pages 354–362. Springer, 2016.

[486] L. Xia, J. Cui, R. Shen, X. Xu, Y. Gao, and X. Li. A survey of image semantics-based visual simultaneous localization and mapping: Application-oriented solutions to autonomous navigation of mobile robots. *International Journal of Advanced Robotic Systems*, 17(3), 2020.

[487] Y. Xia. Awesome SLAM. https://github.com/SilenceOverflow/Awesome-SLAM, 2021.

[488] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[489] X. Xiao, K. Yi, and Y. Tao. The hardness and approximation algorithms for l-diversity. In *Proceedings of the 13th International Conference on Extending Database Technology*, pages 135–146, 2010.

[490] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[491] X. Yan, H. Li, C. Wang, J. Seo, H. Zhang, and H. Wang. Development of ergonomic posture recognition technique based on 2d ordinary camera for construction hazard prevention through view-invariant features in 2d skeleton motion. *Advanced Engineering Informatics*, 34:152–163, 2017.

[492] Y. Yang, Y. Aloimonos, C. Fermüller, and E. E. Aksoy. Learning the semantics of manipulation action. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 676–686, Beijing, China, July 2015. Association for Computational Linguistics.

[493] E. A. Yfantis and A. Fayed. Authentication and secure robot communication. *International Journal of Advanced Robotic Systems*, 11(2):10, 2014.

[494] E.-J. Yoon, K.-Y. Yoo, C. Kim, Y.-S. Hong, M. Jo, and H.-H. Chen. A secure and efficient sip authentication scheme for converged voip networks. *Computer Communications*, 33(14):1674–1681, 2010.

[495] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Q. Weinberger. Pseudo-lidar++: Accurate depth for 3D object detection in autonomous driving. *arXiv preprint arXiv:1906.06310*, 2019.

[496] Z. Yu, B. Zhou, J. Wan, P. Wang, H. Chen, X. Liu, S. Li, and G. Zhao. Searching Multi-Rate and Multi-Modal Temporal Enhanced Networks for Gesture Recognition. *arXiv*, 08 2020.

[497] Yunan Li, Qiguang Miao, Kuan Tian, Yingying Fan, Xin Xu, Rui Li, and J. Song. Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 25–30, 2016.

[498] A. Zacharaki, I. Kostavelis, A. Gasteratos, and I. Dokas. Safety bounds in human robot interaction: A survey. *Safety Science*, 127:104667, 2020.

[499] F. Zafari, A. Gkelias, and K. K. Leung. A survey of indoor localization systems and technologies. *IEEE Communications Surveys & Tutorials*, 21(3):2568–2599, 2019.

[500] S. Zakharov, I. Shugurov, and S. Ilic. DPOD: 6D pose object detector and refiner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1941–1950, 2019.

[501] K. Zampogiannis, Y. Yang, C. Fermüller, and Y. Aloimonos. Learning the spatial semantics of manipulation actions through preposition grounding. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1389–1396, 2015.

[502] A. M. Zanchettin, N. M. Ceriani, P. Rocco, H. Ding, and B. Matthias. Safety in human-robot collaborative manufacturing environments: Metrics and control. *IEEE Transactions on Automation Science and Engineering*, 13(2):882–893, 2016.

[503] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1802–1811, 2017.

[504] J. Zenisek, F. Holzinger, and M. Affenzeller. Machine learning based concept drift detection for predictive maintenance. *Computers & Industrial Engineering*, 137:106031, 2019.

[505] C. Zhang, A. Gupta, and A. Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021.

[506] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen. A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5), 2019.

[507] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2117–2126, 2017.

[508] G. Zhu, L. Zhang, P. Shen, and J. Song. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. *IEEE Access*, 5:4517–4524, 2017.

[509] G. Zhu, L. Zhang, P. Shen, J. Song, S. A. A. Shah, and M. Bennamoun. Continuous gesture segmentation and recognition using 3DCNN and convolutional LSTM. *IEEE Transactions on Multimedia*, 21(4):1011–1021, 2019.

[510] S. Zor, F. Leymann, and D. Schumm. A proposal of BPMN extensions for the manufacturing domain. In *Proceedings of 44th CIRP international conference on manufacturing systems*, 2011.