Chapter 4

Multimodal Human Machine Interactions in Industrial Environments

By Rubén Alonso, Nino Cauli and Diego Reforgiato Recupero

DOI: 10.1561/9781680838770.ch4

Published in *Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production* by John Soldatos and Dimosthenis Kyriazis (eds.). 2021. ISBN 978-1-68083-876-3. E-ISBN 978-1-68083-877-0.

Suggested citation: Rubén Alonso, Nino Cauli and Diego Reforgiato Recupero. 2021. "Multimodal Human Machine Interactions in Industrial Environments" in *Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production*. Edited by John Soldatos and Dimosthenis Kyriazis. pp. 73–92. Now Publishers. DOI: 10.1561/9781680838770.ch4.

now
the essence of knowledge

This chapter will present a review of Human Machine Interaction techniques for industrial applications. A set of recent HMI techniques will be provided with emphasis on multimodal interaction with industrial machines and robots. This list will include Natural Language Processing techniques and others that make use of various complementary interfaces: audio, visual, haptic or gestural, to achieve a more natural human-machine interaction. This chapter will also focus on providing examples and use cases in fields related to multimodal interaction in manufacturing, such as augmented reality. Accordingly, the chapter will present the use of Artificial Intelligence and Multimodal Human Machine Interaction in the context of STAR applications.

4.1 Introduction

Since the beginning of the 20th century, automation played a fundamental role in the manufacturing industry (Wang, 2019). Starting from the sixties, robots were introduced in factories speeding up the manufacturing process. Initially there were strict boundaries between robots' and humans' work-spaces. In order to avoid injuries, workers were not allowed to enter in the robots' working space. Unfortunately, this rigid organization has its limitations. Both robots and humans excel in different areas and a proper collaboration between them can result in a more efficient assembling process. Robots are faster, stronger and more precise in repetitive assembling tasks, while humans are better in decision making and they can easily adapt to unexpected situations. The exponential improvements achieved in the 21st century in AI, perception algorithms and robot control, gradually allowed for a shared work-space between human workers and robots.

Robots use on-board and external sensors to be aware of their surrounding environment. The data output of sensors range from simple single dimension data (contact sensors, ultrasonic distance sensors) to complex high dimensions data (microphones, lidar sensors, RGB cameras, depth cameras). In order to have a better interaction with human workers and other machines in the factory, robots need to merge the information received by every kind of sensor available. This multimodal interaction exists in both ways: while interacting with robots, human workers must not be limited to a restricted group of modalities and devices (keyboards, mouse, screen), but they should be able to use all the modalities made available by their bodies (speech, vision, gestures, touch).

The goal of this chapter is to present the various types of multimodal interaction in industrial environments. After introducing the problem of multimodal interaction we will present some examples of modalities for a natural interaction between human workers and robots/machines such as speech (intended also as Natural Language Processing of text obtained using speech-to-text tools) and vision. We will then make a step further to the idea of multimodal interaction introducing the concept of Extended Reality (XR), where a human is able to remotely control a robot sharing its sensory stimuli.

More specifically, the remainder of this chapter is organized as it follows. Section 4.2 includes all the possible kinds of multimodal interaction between humans and machines. Section 4.3 describes how NLP techniques can be employed within the manufacturing domain. Section 4.4 illustrates human motion recognition and prediction for human robot interaction in manufacturing industry. Section 4.5 illustrates XR technologies which include augmented, mixed and virtual reality. Moreover, a use case showing the application of virtual reality to remote-control a humanoid robot within the manufacturing domain is presented as well. Finally, Section 4.6 concludes the paper.

4.2 Multimodal Interaction

Since the presentation of the famous Put-That-There (Bolt, 1980), innumerable papers have been written about the advantages and disadvantages, problems and solutions aroused from the natural interaction between humans and machines. Multimodal Interaction discipline is based on the idea that human communication is multimodal. Thus, if hoping to interact with machines in the same way as it is done with humans, the interaction must not be limited to a group of modalities and devices, as it has been done until now, using mainly keyboard and mouse as data input and graphical representations as data output.

Some authors (Waibel *et al.*, 1996) point out that it is not advisable to reduce the interaction exclusively to human ↔ machine. They classify the multimodal interaction interfaces in five different classes:

- Human → Machine: in an unidirectional way, as data input mode. For example, a user dictating a text to the computer or giving orders to a robot (without receiving any complex feedback).
- Human ↔ Machine: in a bidirectional and interactive way between the human and the machine, like, for example, in a route planner.
- Human ↔ Multimedia Data: as the extraction of data from multimedia information. For example, the extraction of meaningful images and the transcription of text from video-recorded news, for the subsequent search by a human.
- Human ↔ Machine ↔ Human: where the machine mediates in the interaction between two humans that do not have the same knowledge, lack part of the context or simply because they are far from each other and cannot interact directly.
- Human ↔ Human (observed and assisted by machine): it is not mediated by a machine, but there exists one for assisting the user. For example, a system that records and transcribes meetings which can be searched later looking for actions defined in previous meetings.

In (Alonso and Torres, 2010) the authors extended the list to support a new category: **Human ↔ Multiple Machines**, where the user interacts in a multimodal way with a group of programmable machines, such as robots, using different media and devices, and collaborates with all of them.

This theoretical classification, essential to understand multimodal interaction, is somewhat diluted in practice, especially after the emergence of XR technologies. Anyway, the six classes are relevant to the manufacturing industry, and all have been addressed in a certain degree in the literature related to Artificial Intelligence (AI) and Human Machine Interaction (HMI) in recent years.

For example, Roitberg *et al.* (Roitberg *et al.*, 2015) present an interesting approach for improving the efficiency of Human-Robot interaction. This approach is based on multimodal interfaces, and is focused on the industrial environment. Their research is based on monitoring and interpreting human operations, using video depth information provided by different sensors. They use Microsoft Kinect v2 for skeleton tracking, Asus Xtion PRO for object tracking and Leap Motion for hand and finger pose tracking.

Liu *et al.* (Liu *et al.*, 2018) focus on multimodal human ↔ robot collaboration, especially in repetitive and dangerous tasks. They suggest that the more modalities are included and fused, the more robust the collaboration will be. For this purpose, they present an architecture and a use case for operator-robot collaboration in which body motion recognition, hand motion recognition and speech commands recognition are combined.

Concerning the use of multimodal interaction for operator training, (Vélaz *et al.*, 2014) analysed the influence of four interaction technologies and modalities (including mouse, haptic systems and 2D and 3D position capture) for the learning of a procedural assembly task. Among its conclusions it is worth noting that the results showed that the differences between the training performed with these interaction technologies were not significantly different from the traditional training performed by the operators.

Another significant example of multimodal interaction with multiple machines that could be extrapolated to the manufacturing sector is the coordination of multiple unmanned aerial vehicles. Several authors (e.g.: Cacace *et al.*, 2016b, Cacace *et al.*, 2016a) are working on the coordination of machines, using the information obtained through different modalities to solve interaction and coordination problems.

The improvement of recognition thanks to multimodal interaction has been proven in many studies (e.g.: Kettebekov *et al.*, 2002, Oviatt *et al.*, 2003) where the benefits of multimodal HMI were demonstrated for completing the available information and improve the recognition ratio using supporting modalities.

4.3 Employment of Natural Language Processing Within Manufacturing

Natural Language Processing (NLP) is a subset of AI that helps identifying key elements from human instructions, extract relevant information and process them in a manner that machines can understand. Integrating NLP technologies into the system helps machines understand human language and mimic human behaviour. For example, Amazon's Echo, Microsoft's Cortana and Apple's Siri make an extensive use of NLP technologies to interact with the users.

NLP technologies speed up the operation of a whole system cutting down the response time. Imagine a scenario where a manufacturing company hires a data scientist to collect and analyse all the machine readings, reporting any sort of problems. One disadvantage to this scheme is that by the time the management reads the report one problem might have happened causing damage to the entire process. If a robot with sensors and NLP technologies embedded is employed, this might remotely access the machines and detect in real time any change or problem providing an action to be executed. The robot might even communicate with users and accept input in natural language. Therefore, by leveraging NLP technologies, the middleman can be cut out while at the same time keeping the system effective.

Within the manufacturing industry the NLP might be adopted for the following tasks:

- **Process Automation:** The use of NLP technologies in the manufacturing process allows the automatic execution of repetitive tasks like paperwork and report analysis (e.g., Cristian *et al.*, 2019). Besides, it benefits the workflow of the entire process as each employee can be focused on tasks which require human intervention and capabilities. Authors in (Kang *et al.*, 2019) developed the feedback generation method based on Constraint-based Modeling (CBM) coupled with NLP and domain ontology, designed to support formal manufacturing rule extraction. In detail, the developed method identifies the necessity of input text validation based on the predefined constraints and provides the relevant feedback to help the user modify the input text, so that the desired rule can be extracted.
- **Inventory Management**[1]**:** Analysing data about the sales of certain products is essential to assess the correct decisions for a company to optimize and maximize profits. By leveraging NLP technologies the resulting benefits are: (1) the entire process becomes more comprehensive; (2) it is more difficult to incur errors related to the analysis of sales; (3) it is easier to analyse the manufactured products and discard those with low quality without affecting the supply chain and sales. On a different level, authors in (Vicari and Gaspari, 2020, Carta *et al.*, 2021) have employed NLP and Machine Learning techniques to automatically identify patterns, sentiment or other elements within a text which might be correlated to the stock variation.

[1] https://cmr.berkeley.edu/2021/01/managing-supply-chain-risk/

- **Emotional Mapping:** Sentiment analysis and emotion detection (Atzeni *et al.*, 2018, Atzeni and Recupero, 2020) are one of the most exciting features of NLP. Early NLP systems allowed organizations to collect speech-to-text communication without accurately determining its full meaning. Today, NLP approaches can sort and understand the nuances and emotions in human voices and text, giving organizations unparalleled insight. Learning customer expectations is a very important element in manufacturing. NLP technologies permit to identify emotions and opinions of customers (Dridi *et al.*, 2019, Recupero *et al.*, 2015) and provide actions to improve products and the selling process. Knowing the expectations of customers is key to build a longer relationship and create engagement with them.
- **Operation Optimization:** Furthermore, NLP technologies can be employed to trace the performance of equipment, identifying potential inefficiency. This enables a detailed monitoring of the machinery and taking measures to improve the overall system operability. A review of machine learning approaches for the optimization of production processes covers the majority of relevant literature from 2008 to 2018 dealing with machine learning and optimization approaches for product quality or process improvement in the manufacturing industry (Weichert *et al.*, 2019).

4.4 Human Motion Recognition and Prediction for Human Robot Interaction in Manufacturing

In order to safely interact with humans, robots need to understand human intentions and predict their movements. With the ability to recognise and to predict human actions, industrial robots are able to avoid dangerous collisions and to improve collaborative work anticipating some actions (i.e. passing to the worker the proper tool based on the predicted worker's action).

4.4.1 Video Action Recognition and Prediction

Human action recognition is a complex task that needs as much information as possible about the subject performing the action. RGB and depth cameras are the most suitable sensors for this task: a video sequence of a human performing an action carries information about his visual appearance, the context of the action and the motion of his body.

In order to recognise human actions from images, two steps are needed: action representation and action classification (Kong and Fu, 2018). Traditionally, handcrafted features are used to represent the actions (Jia and Yeung, 2008, Yuan *et al.*, 2016), and standard classifiers are used to recognise the action (e.g. SVN, k-means). The representation of the actions can vary from low level features (edges, corners) to high level ones (body shape, skeletal information). Choosing the optimal handcrafted features that best suit the task of action recognition can be tricky. Automatically extracted features are often more robust and achieve better performances. The recent increase in computational power brought to the rise of Convolutional Neural Networks (CNNs). CNNs are a type of Deep Artificial Neural Networks (DNNs) where for each of the several layers is applied a convolution between 2D weights kernels and the 2D channels of the previous layer. The output of each layer are 2D feature maps extracted from the previous layer (low level features for the initial layers and high level ones for the last layers). With their deep structure and with enough training data, CNNs are able to generate features for action recognition that outperform handcrafted ones. CNNs are frequently used to extract features to represent actions, achieving state-of-the-art results (Kong and Fu, 2018, Özyer *et al.*, 2021).

CNNs are data driven models and one of their drawbacks is the need of big labelled datasets with high quality images. The following are some examples of popular datasets for video action recognition, for a more exhaustive list please refer to (Kong and Fu, 2018, Özyer *et al.*, 2021):

- **UCF-101 (Soomro *et al.*, 2012):** One of the most used datasets for video action recognition. UCF-101 is a large dataset with 13,320 different YouTube videos from 101 categories. This dataset has high variability in camera angles, actors and backgrounds.
- **YouTube-8M (Abu-El-Haija *et al.*, 2016):** This is a very large multi-label video classification dataset (8 million videos for a total of 500K hours). The videos are extracted from YouTube and they are annotated with 4800 machine-generated labels.
- **The Kinetics Human Action Video Dataset (Kay *et al.*, 2017):** This dataset contains 306,245 YouTube clips of 10s each. The clips are grouped in 400 human action classes and are taken from different YouTube videos.
- **Moments in Time (Monfort *et al.*, 2019):** A large-scale human annotated dataset with one million videos of 3 seconds corresponding to dynamic events. Each video is labeled with one among 339 different classes.

While it is possible to recognize action from static images, they lack information about the motion during time. CNNs need to be extended in order to use the time information of video sequences. The most common approaches are the followings:

- **3D CNNs:** These networks are a particular type of CNNs composed by multiple layers of 3D convolutions obtained using 3D kernels. Receiving as input a sequence of frames stacked in one dimension, 3D CNNs are able to extract features related both to space and time. S. Ji *et al.* (Ji *et al.*, 2012) used 3D CNNs to recognize human actions in the real-world environment of airport surveillance videos. The authors compared their model with the state-of-the-art algorithms at the time achieving superior performance.

- **Multi-stream networks:** This type of architecture classifies its input merging together the output of several CNNs. Each CNN receives a different type of input. K. Simonyan and A. Zisserman (Simonyan and Zisserman, 2014) proposed a two-stream CNN for action recognition. The first stream received as input a single RGB frame, while the second stream received as input the multi-frame optical flow, carrying temporal information of the action. The authors tested the network on the UCF-101 dataset obtaining state-of-the-art results.
- **Recurrent neural networks (RNNs):** RNNs are special artificial neural network with internal loops in the connection between layers. Their special structure makes them able to keep a memory of the past and to generate an output based on the sequence of the most recent inputs received. J. Yue-Hei Ng *et al.* (Yue-Hei Ng *et al.*, 2015) introduced an hybrid network that joins together CNNs with RNNs. Their model is composed by GoogLeNet convolutional layers followed by 5 LSTM layers. In the paper the authors perform several ablation studies on a video recognition task showing advantages and disadvantages of using recurrent layers.

Video action recognition is the problem of recognising the action performed by a subject based on a video sequence of the entire movement. The problem of predicting the action performed based only on a video of an initial portion of the action is called action prediction. The most recent action/motion prediction systems tend to use the combination of CNNs and RNNs (Lee *et al.*, 2017), better suited for the analysis of video sequences. In Human Robot Collaboration (HRC) scenarios, the prediction of the type of action performed by the human might not be enough. Often the robot needs to know the full body motion during the next action performed by the human in order to successfully perform the collaborative task. Recently some researchers were able to predict the next frames of a motion based on the action to be performed and past frames (Finn *et al.*, 2016, Jung *et al.*, 2019).

For a Robot interacting with a dynamic environment, it is of primary importance being able to model the surroundings and to predict how the environment evolves through time. With a faithful representation of the environment, the robot is able to detect unexpected behaviours and to correct its actions accordingly. This idea is borrowed from cognitive science: in the Predictive Coding (Rao and Ballard, 1999) cognition theory, the brain is constantly predicting the sensory outcome (top-down process) and comparing it with the actual one. At the same time the error between predicted and actual sensory stimuli is back-propagated to the highest layers (bottom-up process) in order to revise and update the internal predictive models (a similar idea applied to robot control was studied under the name of Expected Perception (Barrera and Laschi, 2010, Cauli *et al.*, 2016)). Jun Tani implemented on robotics platforms several models based on the Predictive Coding paradigm (Tani, 2016). One of the most recent is the Predictive Visuo-Motor Deep Dynamic Neural Network (P-VMDNN) (Hwang *et al.*, 2018). This Deep-RNN model can be used both to predict the next RGB frames and encoders values during a motion, and to recognise an action performed by a human placed in front of the robot.

4.4.2 Video Action Recognition and Prediction for HRC in Manufacturing

In recent years we are seeing a gradual introduction of shared spaces and collaborative tasks between humans and robots in factories. Human and robotic workers can collaborate during the assembly process of specific components. In these scenarios, the robot must predict the human coworker action in order to plan its own motion. The application of video action recognition models to HRC in manufacturing is still a relatively new topic (Wang, 2019).

The most straightforward approaches use handcrafted features to represent the actions. E. Coupeté *et al.* (Coupeté *et al.*, 2019) extract the skeletal representation of the upper-torso of a worker from depth images. The sequence of skeletal position during a motion is given as input to an Hidden Markov Model in order to recognise the performed gesture. The model is tested in an assembly scenario where a worker and a robot collaborate to mount a mechanical piece.

A different approach is to automatically extract the best features using a CNN. P. Wang *et al.* (Wang *et al.*, 2018) use AlexNet to recognise specific gestures from a video of a worker assembling an engine. The convolutional layers extract the features while 3 fully connected layers classify the gesture. The architecture based the classification only on single frames.

We already mentioned that single images lack information of the temporal evolution of the action. Using both RGB images and optical flow as inputs solves the problem. Q. Xiong *et al.* (Xiong *et al.*, 2020) use the two-streams network proposed by (Simonyan and Zisserman, 2014) to recognise the actions from closeup videos of workers assembling engines' parts. The network has 2 CNN branches, one receiving as input RGB images and the other optical flow images. Due to the small size of the engine block assembly dataset used in the experiment, the authors apply transfer learning. They first pretrain the entire network on a bigger generic action dataset and then they finetune the last layers on the engine block assembly dataset.

RNNs are other models able to keep temporal information of the recently seen frames. Z. Liu *et al.* (Liu *et al.*, 2019) developed a system able to predict the next action performed by a worker while assembling a computer. A robot passes the worker the proper tool based on the predicted action. The authors use a CNN as feature extractor followed by an LSTM layer and a fully connected layer to classify the next action. The input of the system are the images from a top-down camera mounted above the working table.

While a fair amount of work on action recognition in manufacturing already exists, the problem of human motion prediction in HRC needs to be studied in more details. A robot able to predict in each instant where the body of the human co-worker will be, can easily avoid collision, spot mistakes and make recovering actions.

It is clear that CNNs are the most reliable tool for features extraction from videos. CNNs need a big amount of data to learn properly and be able to generalise. Unfortunately, not many datasets for video action recognition in factory assembly scenario exist (Kong and Fu, 2018, Özyer *et al.*, 2021). New specific video datasets are difficult to generate and the labelling process is highly time consuming. Domain transfer and simulated datasets are a valid solution to the problem. M. Fabbri *et al.* (Fabbri *et al.*, 2018) generated a big dataset for Multi-People Tracking using the Grand Teft Auto V game engine. Generating a simulated dataset is faster than collecting a real one and labelling is automatic. An action recognition model trained on a simulated dataset with high variability and realism is able to transfer the knowledge learned in simulation to the real world.

4.5 XR in Manufacturing Industry

XR related technologies are facilitating multimodal interaction in Industry 4.0 and thus enabling tangible in-site visualisations and interactions with industrial assets (Simões *et al.*, 2018).

The term XR can be considered as an umbrella for the terms augmented (AR), mixed (MR) and virtual (VR) reality, which differ in how much real and virtual content they display and the level of interactivity. As detailed in Alizadehsalehi *et al.*, 2020 VR is characterised by high virtual content and low interactivity, while AR is characterised by high real content and higher interactivity. MR lies in the middle of both, including higher levels of virtual and real content, and high interactivity.

4.5.1 Related Work of XR in Industry

The use of XR in industry has been suggested since the early 90's, where for example Thomas and David, 1992 proposed the superimposition of certain information on real world objects. Since that point there are hundreds of examples of XR aided manufacturing, Bottani and Vignali present an exhaustive list of them in their article *"Augmented reality technology in the manufacturing industry: A review of the last decade"* (Bottani and Vignali, 2019).

In addition to the Boeing article (Thomas and David, 1992) already mentioned above, for example Karlsson *et al.* (Karlsson *et al.*, 2017) suggest an approach for the presentation of superimposed information, e.g. information on potential bottlenecks, that can help decision making in manufacturing.

Workforce training is another activity where the use of XR is increasing, especially after the rise of robotic systems and complex machines in shopfloors. For example. safety training is another area where multimodal interaction and XR are absolutely worthwhile. As detailed in Doolani *et al.*, 2020, these systems reduce the risks of harm that can be caused by machines as well as damage to them, and offer a platform for learning-by-doing approach that can be used multiple times without worrying about the costs, availability or risks associated with the use of real machines.

The possibility of remote guidance is another advantage of XR systems in the manufacturing environment. For example Fast-Berglund *et al.*, 2018 validated a use case in which the expert uses AR to guide the novice operator in an assembly task and gives directions and corrections in case there is something wrong in the assembling. Their conclusion is that thanks to the AR being able to give instant feedback, it makes it practically impossible to do the assembly wrong and therefore the results are highly positive.

4.5.2 Use Case: Virtual Reality to Remote-control a Robot

In this section we are going to describe the work of authors in Alonso *et al.*, 2021 related to a general-purpose, open-source framework for teleoperating a NAO humanoid robot through a Virtual Reality (VR) headset. As the proposed architecture is general, it would be straightforward to replace the NAO robot with Kuka[2] or Universal Robot,[3] two well known robots used in several production environments around the world. The architecture presented in Alonso *et al.*, 2021 includes a VR

[2] http://www.kuka.com

[3] https://www.universal-robots.com

interface for the Oculus Rift[4] using the Unity game engine to perform robot actions through the VR controllers and exploits the flexibility of the Robot Operating System (ROS) for the control and synchronization of the robot hardware. This work gives ideas on potential architecture that can be employed within the manufacturing domain to allow the robots (e.g. Kuka or Universal Robot, both supported by ROS) to protect workers from repetitive, mundane, and dangerous tasks while also creating more desirable jobs such as engineering, programming, management and equipment maintenance. In the following we will show details of the tools used for their work. Let us first start giving some background information about the Unity, ROS and NAO software platforms.

Unity 3D[5] is a game engine which supports the development of 2D and 3D games, Virtual and Mixed Reality experiences and simulations.

ROS is an open-source framework for robot software whose architecture includes Nodes, Messages, Topics, Services, and Actions. Nodes are processes that carry out a computation. Messages are exchanged by nodes. A node sends a message by posting it on a certain topic. Services are needed by nodes that need to perform remote procedure calls. Actions are used to send a request to a node to perform a certain task for longer time and receive a reply. Then, ROS packages are a collection of code for easy reuse and stacks are a collection of packages that jointly offer some functionalities.

The authors employed NAO as the robotic platform but, as already mentioned, robots such as Kuka or Universal Robots may be employed. The Kuka system software is the operating software containing all the basic functions needed for the deployment of the robot system. Kuka robots come with a control panel with a display and axis control buttons and a 6D mouse which is used to manually move the robot. The control panel allows the users to view and create new and modify existing programs. A rugged computer lies in the control cabinet communicates with the robot system via the Multi Function Card, which controls the real-time servo drive electronics. Servo position feedback is transmitted to the controller through the DSE-Resolver Digital Converter/RDC connection. The software includes two elements running on parallel – the user interface and program storage. Figure 4.1 shows a Kuka robot palletizing food in a bakery. Universal robots consist of industrial collaborative robot arms (cobots), which are six-jointed robot arms with a very low weight (from 11 to 33 kilos) with a lifting ability from 3 to 16 kilos. These cobots can work right alongside personnel with no safety guarding, based on the

[4] https://www.oculus.com/rift/

[5] https://unity.com/

Figure 4.1. A Kuka robot palletizing food in a bakery (taken from Wikipedia).

results of a mandatory risk assessment.[6] The robot arm can run in two operating modes of the safety functions; a normal and a reduced one. A switch between safety settings during the cobot's operation is also possible. Figure 4.2 shows a Universal Robot lifting an object.

In their work the authors show how through the remotes and the VR headset the VR interface allows the teleoperation of the NAO and the recording of a movements sequence for later execution. During the former, the user and the robot are not in the same room. Therefore, the user exploits the VR interface as a source of input and for having a visible and understandable representation of the remote robot status. The recording of a movements sequence allows the user to perform a number of tasks and save them in certain collections. Whenever needed, they can play them back.

As the ROS framework allows the development and run on different machines it is easier and more flexible to support both the storing and the playing of recorded actions of the robot.

Figure 4.3 illustrates the architecture of the VR system developed by the authors. It includes three main software components (VR, ROS and Rosbridge) and two hardware devices (Oculus Rift and NAO). The VR Component leverages the Unity

[6] https://www.iso.org/standard/62996.html

Figure 4.2. A Universal Robot lifting an object (taken from https://www.therobotreport.com/voith-robotics-cuts-ties-franka-emika-adds-universal-robots/).

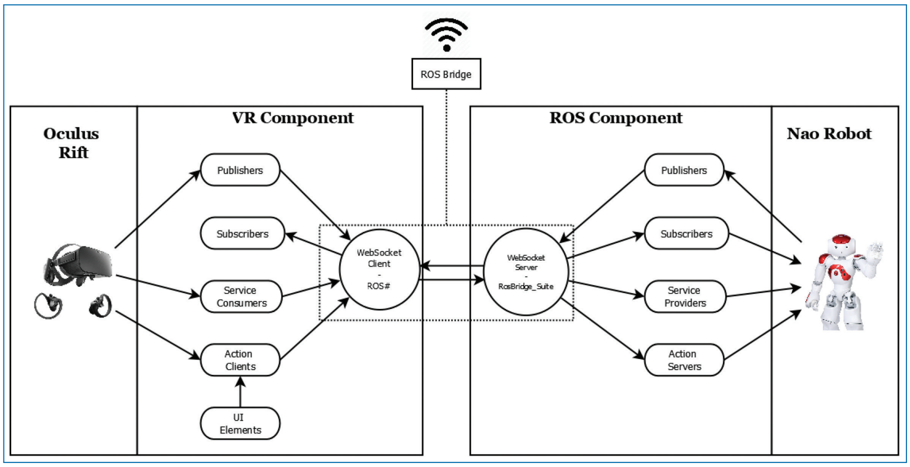

Figure 4.3. Architecture of the Virtual Reality system. Taken from Alonso *et al.*, 2021.

game engine for displaying the interface on the Oculus Rift. Unity has been chosen for the existing Oculus SDK that facilitates the developing process. The ROS component controls the robot through the VR simulation or the management of real hardware. It includes the ROS framework, multiple packages provided by ROS Nao Drivers, custom Publisher, Subscriber, Action Servers and Service Provider that have been implemented for supporting the VR control. The Ros Bridge is the connection between the VR and ROS components. It provides the methods for passing messages between them, for managing the information serialization and deserialization, and the connection and the delivery through WebSockets.

4.6 Conclusions

In this chapter we have presented various types of multimodal interaction within the manufacturing domain. First we have introduced the classification of multimodal interaction interfaces, indicating all the possible ways a user can interact with one or multiple machines. Then we briefly described the NLP research area and how it can be employed to automatically let an independent system (e.g., an agent or robot) to identify relevant information within the manufacturing. Next, we examined the ability of robots of recognising and predicting human actions by using cameras as sensors and deep learning as breakthrough machine learning technology. We continued discussing the XR related technologies (e.g., augmented, mixed, virtual reality) and how they can facilitate multimodal interaction in Industry 4.0. Finally, we showed an architecture of a use case where virtual reality technology has been adopted to remote-control a robot and how this schema can be adapted to be employed within the manufacturing domain.

Secure, safe, reliable AI systems in manufacturing environments, such as those investigated in the STAR project, can benefit from all of these technologies in their goal to make systems more trusted and human-centric. As part of the STAR project, research will continue on Human Robot Interaction and on knowledge systems, that benefit from NLP techniques and are accessible through multimodal interaction.

Acknowledgement

This work has been carried out in the H2020 STAR project, which has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 956573.

References

Abu-El-Haija, S., N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan (2016). “Youtube-8m: A large-scale video classification benchmark”. *arXiv preprint arXiv:1609.08675*.

Alizadehsalehi, S., A. Hadavi, and J. C. Huang (2020). “From BIM to extended reality in AEC industry”. *Automation in Construction*. 116: 103254.

Alonso, R. and M. I. Torres (2010). “Architecture for the Multimodal coordination of semi-autonomous agents”. In: *International Conference on Agents and Artificial Intelligence, ICAA 2010, Valencia, Spain. 22–24 January, 2010*.

Alonso, R., A. Bonini, D. Reforgiato Recupero, and D. Spano (2021). “Exploiting Virtual Reality and the Robot Operating Systemto remote-control a Humanoid Robot”. *Submitted to Multimedia Tools and Applications*.

Atzeni, M., A. Dridi, and D. R. Recupero (2018). “Using frame-based resources for sentiment analysis within the financial domain”. *Prog. Artif. Intell.* 7(4): 273–294. DOI: 10.1007/s13748- 018-0162-8. URL: https://doi.org/10.1007/s13748-018-0162-8.

Atzeni, M. and D. R. Recupero (2020). “Multi-domain sentiment analysis with mimicked and polarized word embeddings for human-robot interaction”. *Future Gener. Comput. Syst.* 110: 984–999. DOI: 10.1016/j.future.2019.10.012. URL: https://doi.org/10.1016/j.future.2019.10.012.

Barrera, A. and C. Laschi (2010). “Anticipatory visual perception as a bio-inspired mechanism underlying robot locomotion”. In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE. 3206–3209.

Bolt, R. A. (1980). ““Put-that-there” Voice and gesture at the graphics interface”. In: *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. 262–270.

Bottani, E. and G. Vignali (2019). “Augmented reality technology in the manufacturing industry: A review of the last decade”. *IISE Transactions*. 51(3): 284–310.

Cacace, J., A. Finzi, and V. Lippiello (2016a). “Multimodal interaction with multiple co-located drones in search and rescue missions”. *arXiv preprint arXiv:1605.07316*.

Cacace, J., A. Finzi, V. Lippiello, M. Furci, N. Mimmo, and L. Marconi (2016b). “A control architecture for multiple drones operated via multimodal interaction in search & rescue mission”. In: *2016 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE. 233–239.

Carta, S., S. Consoli, L. Piras, A. S. Podda, and D. R. Recupero (2021). “Event detection in finance using hierarchical clustering algorithms on news and tweets”. *PeerJ Comput. Sci.* 7: e438. DOI: 10.7717/peerj-cs.438. URL: https://doi.org/10.7717/peerj-cs.438.

Cauli, N., E. Falotico, A. Bernardino, J. Santos-Victor, and C. Laschi (2016). “Correcting for changes: expected perception-based control for reaching a moving target”. *IEEE Robotics & Automation Magazine*. 23(1): 63–70.

Coupeté, E., F. Moutarde, and S. Manitsaris (2019). “Multi-users online recognition of technical gestures for natural human–robot collaboration in manufacturing”. *Autonomous Robots*. 43(6): 1309–1325.

Cristian, M., S. Christian, and T. Tudor (2019). "A Study in the Automation of Service Ticket Recognition using Natural Language Processing". In: 1–6. DOI: 10.23919/SOFTCOM.2019.8903676.

Doolani, S., C. Wessels, V. Kanal, C. Sevastopoulos, A. Jaiswal, H. Nambiappan, and F. Makedon (2020). "A Review of Extended Reality (XR) Technologies for Manufacturing Training". *Technologies*. 8(4): 77.

Dridi, A., M. Atzeni, and D. Reforgiato Recupero (2019). "FineNews: fine-grained semantic sentiment analysis on financial microblogs and news". *International Journal of Machine Learning and Cybernetics*. 10(8): 2199–2207. DOI: 10.1007/s13042-018-0805-x. URL: https://doi.org/10.1007/s13042-018-0805-x.

Fabbri, M., F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, and R. Cucchiara (2018). "Learning to detect and track visible and occluded body joints in a virtual world". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 430–446.

Fast-Berglund, Å., L. Gong, and D. Li (2018). "Testing and validating Extended Reality (xR) technologies in manufacturing". *Procedia Manufacturing*. 25: 31–38.

Finn, C., I. Goodfellow, and S. Levine (2016). "Unsupervised learning for physical interaction through video prediction". In: *Advances in neural information processing systems*. 64–72.

Hwang, J., J. Kim, A. Ahmadi, M. Choi, and J. Tani (2018). "Dealing with large-scale spatio-temporal patterns in imitative interaction between a robot and a human by using the predictive coding framework". *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Ji, S., W. Xu, M. Yang, and K. Yu (2012). "3D convolutional neural networks for human action recognition". *IEEE transactions on pattern analysis and machine intelligence*. 35(1): 221–231.

Jia, K. and D.-Y. Yeung (2008). "Human action recognition using local spatio-temporal discriminant embedding". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 1–8.

Jung, M., T. Matsumoto, and J. Tani (2019). "Goal-Directed Behavior under Variational Predictive Coding: Dynamic Organization of Visual Attention and Working Memory". *arXiv preprint arXiv:1903.04932*.

Kang, S., L. Patil, A. Rangarajan, A. Moitra, T. Jia, D. Robinson, and D. Dutta (2019). "Automated feedback generation for formal manufacturing rule extraction". *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*. 33(3): 289–301. DOI: 10.1017/S0890060419000027.

Karlsson, I., J. Bernedixen, A. H. Ng, and L. Pehrsson (2017). "Combining augmented reality and simulation-based optimization for decision support in manufacturing". In: *2017 Winter Simulation Conference (WSC)*. IEEE. 3988–3999.

Kay, W., J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.* (2017). "The kinetics human action video dataset". *arXiv preprint arXiv:1705.06950*.

Kettebekov, S., M. Yeasin, N. Krahnstoever, and R. Sharma (2002). "Prosody based co-analysis of deictic gestures and speech in weather narration broadcast". In: *Workshop on Multimodal Resources and Multimodal System Evaluation. (LREC 2002), Las Palmas, Spain*. Citeseer.

Kong, Y. and Y. Fu (2018). "Human action recognition and prediction: A survey". *arXiv preprint arXiv:1806.11230*.

Lee, N., W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker (2017). "Desire: Distant future prediction in dynamic scenes with interacting agents". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 336–345.

Liu, H., T. Fang, T. Zhou, and L. Wang (2018). "Towards robust human-robot collaborative manufacturing: Multimodal fusion". *IEEE Access*. 6: 74762–74771.

Liu, Z., Q. Liu, W. Xu, Z. Liu, Z. Zhou, and J. Chen (2019). "Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing". *Procedia CIRP*. 83: 272–278.

Monfort, M., A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, Y. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick, *et al.* (2019). "Moments in time dataset: one million videos for event understanding". *IEEE transactions on pattern analysis and machine intelligence*.

Oviatt, S. *et al.* (2003). "Multimodal interfaces". *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*. 14: 286–304.

Özyer, T., D. S. Ak, and R. Alhajj (2021). "Human action recognition approaches with video datasets—A survey". *Knowledge-Based Systems*: 106995.

Rao, R. P. and D. H. Ballard (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects". *Nature neuroscience*. 2(1): 79.

Recupero, D. R., M. Dragoni, and V. Presutti (2015). "ESWC 15 Challenge on Concept-Level Sentiment Analysis". In: *Semantic Web Evaluation Challenges*. Ed. by F. Gandon, E. Cabrio, M. Stankovic, and A. Zimmermann. Cham: Springer International Publishing. 211–222. ISBN: 978-3-319-25518-7.

Roitberg, A., N. Somani, A. Perzylo, M. Rickert, and A. Knoll (2015). "Multimodal human activity recognition for industrial manufacturing processes in robotic workcells". In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 259–266.

Simões, B., R. De Amicis, I. Barandiaran, and J. Posada (2018). "X-reality system architecture for industry 4.0 processes". *Multimodal Technologies and Interaction*. 2(4): 72.

Simonyan, K. and A. Zisserman (2014). "Two-stream convolutional networks for action recognition in videos". In: *Advances in neural information processing systems*. 568–576.

Soomro, K., A. R. Zamir, and M. Shah (2012). "UCF101: A dataset of 101 human actions classes from videos in the wild". *arXiv preprint arXiv:1212.0402*.

Tani, J. (2016), *Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena*. Oxford University Press.

Thomas, P. and W. David (1992). "Augmented reality: An application of heads-up display technology to manual manufacturing processes". In: *Hawaii international conference on system sciences*. 659–669.

Vélaz, Y., J. Rodríguez Arce, T. Gutiérrez, A. Lozano-Rodero, and A. Suescun (2014). "The influence of interaction technology on the learning of assembly tasks using virtual reality". *Journal of Computing and Information Science in Engineering*. 14(4).

Vicari, M. and M. Gaspari (2020). "Analysis of news sentiments using natural language processing and deep learning". *AI & SOCIETY*. DOI: 10.1007/s00146-020-01111-x. URL: https://doi.org/10.1007/s00146-020-01111-x.

Waibel, A., M. T. Vo, P. Duchnowski, and S. Manke (1996). "Multimodal interfaces". *Artificial Intelligence Review*. 10(3): 299–319.

Wang, L. (2019). "From intelligence science to intelligent manufacturing". *Engineering*. 5(4): 615–618.

Wang, P., H. Liu, L. Wang, and R. X. Gao (2018). "Deep learning-based human motion recognition for predictive context-aware human-robot collaboration". *CIRP annals*. 67(1): 17–20.

Weichert, D., P. Link, A. Stoll, S. Rüping, S. Ihlenfeldt, and S. Wrobel (2019). "A review of machine learning for the optimization of production processes". *The International Journal of Advanced Manufacturing Technology*. 104(5): 1889–1902. DOI: 10.1007/s00170-019-03988-5. URL: https://doi.org/10.1007/s00170-019-03988-5.

Xiong, Q., J. Zhang, P. Wang, D. Liu, and R. X. Gao (2020). "Transferable two-stream convolutional neural network for human action recognition". *Journal of Manufacturing Systems*. 56: 605–614.

Yuan, C., B. Wu, X. Li, W. Hu, S. Maybank, and F. Wang (2016). "Fusing $\mathcal{R}$ Features and Local Features with Context-Aware Kernels for Action ecognition". *International Journal of Computer Vision*. 118(2): 151–171.

Yue-Hei Ng, J., M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici (2015). "Beyond short snippets: Deep networks for video classification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4694–4702.