A Multimodal Attention Tracking in Human-Robot Interaction in Industrial Robots for Manufacturing Tasks

LI, Chen; Kaszowska, Aleksandra; Chrysostomou, Dimitrios

Published in:
ICAC 2023 - 28th International Conference on Automation and Computing

DOI (link to publication from Publisher):
10.1109/ICAC57885.2023.10275168

Publication date:
2023

Document Version
Accepted author manuscript, peer reviewed version

A Multimodal Attention Tracking in Human-Robot Interaction in Industrial Robots for Manufacturing Tasks

Chen Li[1,*], Aleksandra Kaszowska[2], and Dimitrios Chrysostomou[3]
[1,3]Department of Materials and Production, Aalborg University, DK-9220, Aalborg East, Denmark
[2]Department of Electronic Systems, Aalborg University, DK-9220, Aalborg East, Denmark
cl@mp.aau.dk, kaszowska@es.aau.dk, dimi@mp.aau.dk
*Corresponding author: Chen Li(cl@mp.aau.dk)

***Abstract*—The field of human-robot interaction has seen tremendous growth in recent years, and the use of robots in manufacturing tasks has become increasingly common. However, the success of human-robot interaction is highly dependent on the ability of the robot to understand and adapt to the human operator's actions and attention. In this paper, we propose a novel approach that uses a context-aware natural language interface and position tracker to track the operator's attention and improve interaction with the robot. The system integrates multimodal inputs such as head pose estimation and intent recognition to accurately predict the operator's attention and adjust the robot's behaviors. The proposed approach is evaluated in a manufacturing logistic scenario, and the results show a significant improvement in collaboration and a reduction in errors in task completion. The approach is expected to have broad applicability in industrial manufacturing settings, where it can enhance productivity and efficiency by improving human-robot interaction.**

***Index Terms*—Human-robot interaction, natural language interface, Speech recognition, face recognition, mobile robots**

I. Introduction

In recent years, robots have grown more prevalent in the industrial companies as a result of their efficacy, accuracy, and cost-effectiveness. In hazardous and time-consuming industrial conditions, robots may assist human to decrease error rates and improving production. Ineffective communication between humans and robots may lead to production errors, delays, and even accidents during the human-robot interaction (HRI) in manufacturing tasks [1], [2]. Therefore, it is critical to develop methods for robots to detect whether humans are paying attention and respond accordingly.

Many state-of-the-art (SOTA) attention monitoring methods have been proposed for HRI [3]–[6]. These systems employ high-tech sensors such as cameras, microphones, and eye-tracking devices to capture and sensor data in real time on various sorts of human focus. Some systems employ gaze tracking to determine where a person's attention is focused [7], [8], while others use voice recognition and natural language processing to interpret the human's intention [9]. However, most of the above methods are often employed in the HRI for social robots to interpret human attention. Hence, it is necessary to investigate whether we can leverage such technologies to assess many types of human attention, e.g., gaze, speech, and body language, during HRI for industrial robots in industrial setting.

Furthermore, despite advancements in attention monitoring methods, tracking attention during the manufacturing tasks still confronts a number of challenges, e.g., feature representation, unavailable modality. For instance, it may be challenging to merge multimodalities (e.g., speech, vision) into a single feature representation. Gaze tracking may help us estimate where a user's attention is focused, but it may not perform as effectively in low-light or dark environments, highly reflective surfaces or when the individual is wearing helmets or goggles.

Moreover, developing systems that can adapt to the dynamic changes of human attention is another challenging. The industrial robot's behavior will have to account for the human intent to rapidly switch between multiple tasks. To be useful, the robot must be able to interpret human emotions and predict their future actions.

In this paper, we present a multimodal attention monitoring method and integrated it with our industrial virtual assistant (VA), Max [10], [11], to overcome these barriers. Multiple sensors are integrated with the latest iteration of our autonomous industrial mobile manipulator, "Little Helper (LH) [12]," such as cameras and microphones, to monitor and analyze human attention during the internal logistic tasks. The proposed method is not only able to assist VA to hold a task-related conversation, but can also track the dialogue history and determine if the operator's intent changed during the conversation, inferring the operator's attention loss, as well as tracking the real-time head pose.

We summarize our contributions as follows:

- **Multimodal Attention Predictor**. We proposed a pipe line architecture based attention predictor based on two modalities, speech and vision. We fine-tuned a pre-trained Bidirectional Encoder Representations from Transformers (BERT) [13] model on our dialogue dataset for understanding human intent with a high inference accuracy. An intent switch detection algorithm is designed for helping robot to analyze the different contexts of the conversation based on detected intent. It allows the robot to adapt to different scenarios and provides more personalized and effective assistance to the human. Head pose estimation has been used in many applications, such as gaming or virtual reality, its use in industrial tasks is relatively new. This is because industrial tasks often involve complex environments, where HRI can be unpredictable and require a high degree of precision and accuracy. By tracking

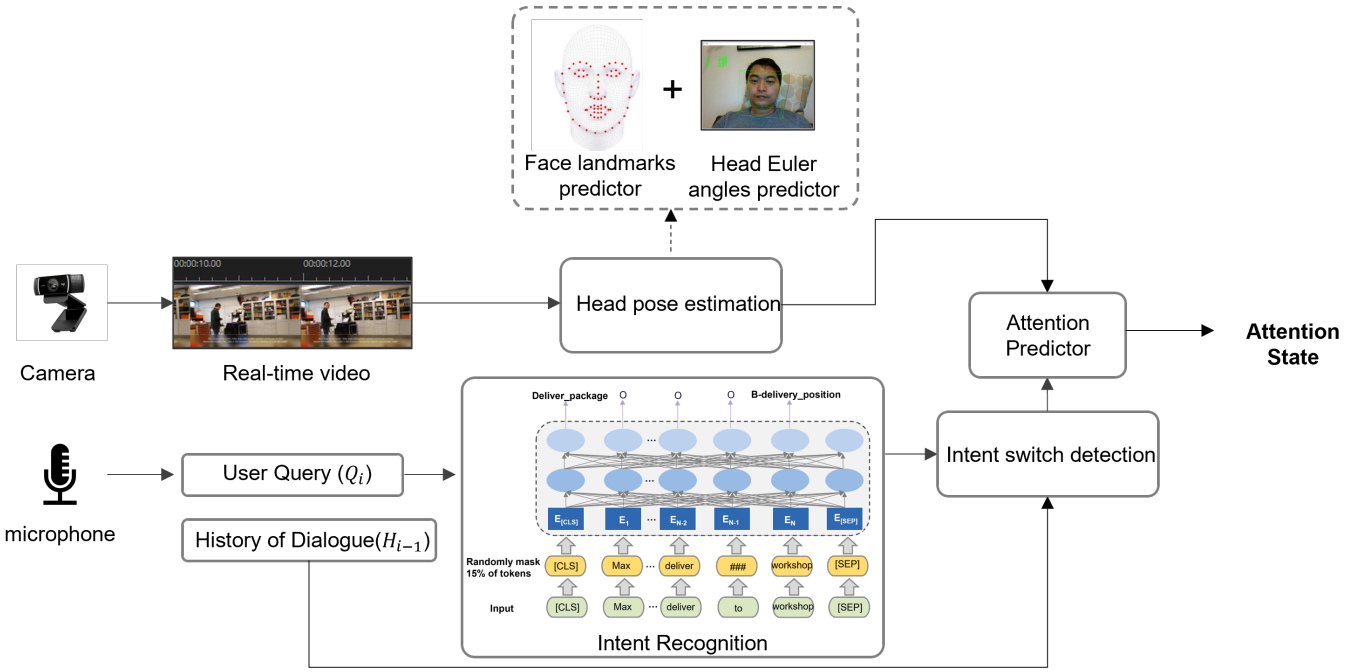

Fig. 1. The proposed multi-modal attention recognition approach

the human's head movements, the robot can adapt its behavior and anticipate the human's needs, leading to a more natural and efficient interaction.In conjunction with the above intent recognition, our algorithm provides a more comprehensive understanding of the human's attention and intentions. This can lead to more intuitive and efficient HRI, where the robot can anticipate the human's needs and respond in a timely and appropriate manner.

- **Industrial Experiments** Multiple industrial experiments were conducted using LH to validate proposed multi-modal attention tracking method performance in a real manufacturing environment. The results demonstrate that the solution effectively increases HRI efficiency. The outcomes of this research promise well for the development of human-robot collaborations in manufacturing and other industries.

In Section II of this paper, we describe the proposed methods for multimodal attention tracking including its design specifications, architecture, and core components. We present the experiments and evaluate method performance in Section III, and we finalize the paper with reflections and concluding remarks in Section IV.

II. Methods

The framework of the proposed method is depicted in Fig 1. Following the pipeline architecture, the framework includes three core components, head pose estimation, intent switch detection, and attention predictor. A robot control component is also provided for control the robot's behavior, such as mark position and pause the movement. The attention predictor provides the attention estimation results for robot control component adjusting the robot behavior.

A. Head Pose Estimation

Head pose estimation provides a way to track where a shop floor worker is looking and what they are paying attention to. This is particularly useful in situations where a person is working alongside a robot or other automated machinery, as it allows the robot to adapt its behavior and respond appropriately to the person's actions and needs.

In this module, two algorithms are leveraged for *face landmarks prediction* and *head Euler angles prediction* (see Fig. 1).

- Face Landmarks Predictor. An ensemble of regression trees is introduced to estimate the face's landmark positions from realtime camera input [14]. The positions of these landmarks can be used to calculate the position and orientation of the head. The pre-trained model is used for this task [1].
- Head Euler Angles Predictor. Mapping a correspondence between the 2D image points and the 3D world coordinates of the face is performed once the facial landmarks have been detected. This is done by using a 3D model of the face [2]. The 3D model is then used to estimate the 3D coordinates of the facial landmarks based on their 2D image positions. Head Euler angles are introduced to describe the rotation of the head in 3D space. It measured as pitch, yaw, and roll, which correspond to the rotation around the x, y, and z axes, respectively. To estimate

[1]https://github.com/italojs/facial-landmarks-recognition
[2]https://github.com/lincolnhard/head-pose-estimation

the head Euler angles, the 3D coordinates of the facial landmarks are used to calculate the relative positions of the head in 3D space. The Euler angles can then be computed from these relative positions.

B. Intent Switch Detection

Intent recognition during the human-robot dialog is used to detect the level of attention and engagement of the shop floor worker. For example, if the worker is experiencing difficulty, confusion with a task or lost attention, the robot can use intent recognition to identify the problem and offer additional guidance or assistance.

In this module, a pipe line style dialog flow is used to identify the worker's intent, track the dialog history, and generate the appropriate response. It includes three sub-modules, spoken language understanding, dialog state tracker and response generator.

- Spoken language understanding. In this module, the focus was on fine-tuning the pre-trained BERT model to improve its accuracy in predicting the operator's intent and key slots from their utterances. The fine-tuned BERT model offers contextualized sentence representation by taking into account the surrounding words and context. This is particularly useful for understanding the meaning of words in a given context, which is crucial for effective communication in human-robot interaction. The model was trained and validated on a dialog dataset that contains task-related conversations on industrial robots (i.e., MiR200 [3]). This dataset was carefully curated to include a wide range of scenarios and conversations that are relevant to industrial logistic tasks [4]. By training the model on this dataset, it was able to learn the general conversation pattern regarding the logistic tasks (e.g., mark location, package delivery), leading to better performance in predicting the intent and key slots from operator utterances. The results of the model prediction is with an intent accuracy of 97.7% and an F1 score of 96.8% for slots.
- Dialog state tracker. The shop floor worker may need to engage in multiple rounds of dialogue with a robot in order to provide sufficient information to complete a manufacturing operation. For example, a worker may request the Mobile Industrial Robot (MiR) to mark a position on a digital map by saying, *"Hey robot, I need you to mark this position on the map."* One of the possible responses from the MiR can be *"Sure, but I need you to give me a name for this position."* As the consequence, the worker might give a name *"Ok, name it as storage room."*. Therefore, a manufacturing task can involve multiple stages, each of which may require different information to be collected. In our method, we use a pre-defined JSON file that specifies intents and all the necessary slots for completing the job. These intents and slots are essentially pieces of information that the robot needs to collect from the user in order to complete the manufacturing task. The dialog state tracker will verify the current intent of the conversation is align with previous dialog history and then updates it as the user provides the necessary information.
- Response generator. In our method, we use Template-based strategy for response generation. The response templates are created based pre-defined intent categories. For example, robot may response *"Sure, I will create position with the #position-name on the map."* when worker requests to mark the position with name, "storage room", on the digital map. In this case, *#position-name* is the placeholder for dynamic part of the response that will be filled in based on the user's input. The "storage room" will replace this placeholder with the actual position name that the user requests. Once the dialog is verified by the dialog state tracker, including the intent matching and slots filling, the system can generate responses by selecting the appropriate template and filling in the placeholders based on the user input.

C. Attention Predictor

The attention predictor takes the results from head pose estimation and intent switch recognition as inputs to predict the final attention state. When the head is facing towards the speaker and the conversation topic is consistent, the person is likely to be paying attention. When the head is facing away from the workspace or the conversation topic changes frequently, the person's attention may be divided or distracted. The thresholds are set for head pose and dialog intent switch to determine when attention levels are high, medium, or low. For example, if the head is facing towards the workspace for during the task and the dialog intent switches less than 2 times per task, the attention level may be considered high. If the head is facing outwards the workspace during the task or the dialog intent switches more than 2 times per task, the attention level may be considered low. By using the detected attention levels to provide feedback to the speaker (e.g., slow down the robot operation speed) or adjust the conversation (e.g., reminding worker of current task through the pre-defined response templates) accordingly. For example, if the attention level is low, the speaker may need to pause the current task and wait for the worker's attention back.

Table I shows the pre-defined attention prediction rules. For head pose estimation, we pre-defines the threshold with the value of (-25, 25) of rotation of the x, y and z of the worker's head pose. The threshold of intent switch frequency is set to 2 (i.e., 2 times per task). In our case, the dialog consistency is given high priority than the head pose direction when it comes to predict the attention level. Observation of the human-to-human interaction shows that speaker's attention does not necessary go lower when the speaker looks at other direction as long as the speaker can maintain the consistent conversation. Therefore, we set up the attention prediction value to *medium* even the operator does not look at workspace (e.g., finding a

[3]https://www.mobile-industrial-robots.com/

[4]https://github.com/lcroy/Virtual-Assistant-Max/tree/main/MaxModel/data/mir

TABLE I
PRE-DEFINED RULES OF ATTENTION PREDICTION

Head Pose Estimation rotation(x, y, z)	Intention Switch Frequency times/task	Attention Prediction Low/medium/high
x, y, z $\in$ (-25,25)	$<$ 2 times/task	high
x, y, z $\in$ (-25,25)	$\geq$ 2 times/task	low
x, y, z $\notin$ (-25,25)	$<$ 2 times/task	medium
x, y, z $\notin$ (-25,25)	$\geq$ 2 times/task	low

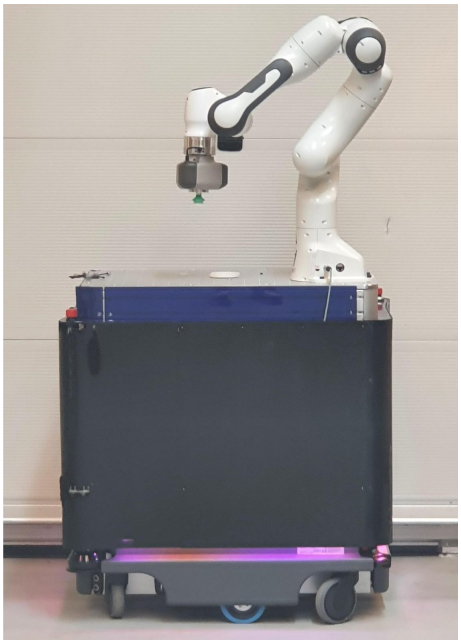

Fig. 2. Robotics Platform: Little Helper

part, checking the machine) as long as the operator's intent switch does not switch more than two times during the task.

III. EXPERIMENTAL SETUP

We conducted the experiments at our Aalborg University learning factory [15]. The industrial robot used in our case is autonomous industrial mobile manipulator, LH. The LH is composed by a MiR and a industrial manipulator, Franka Emika. Fig. 2

Two industrial scenarios, i.e., digitization and localization, are selected for testing the proposed approach. These scenarios explore and test whether the proposed approach can predict the shop floor worker attention level while adjusting the robot's behavior.

In the *digitization* scenario, the operator navigates the LH robot inside the lab to explore the working environment and building a 2D digital map. Task focused in this scenario is *marking position* during the map generation. Those positions are used for building the path between positions for internal logistic. The shop floor worker is asked to switch the intents during the human-robot dialog in this task. The purpose is to testing the whether the system can predict the intent switching and provide appropriate suggestions to the operator. Fig. 3 shows an example of intent switch detection for a position marking task.

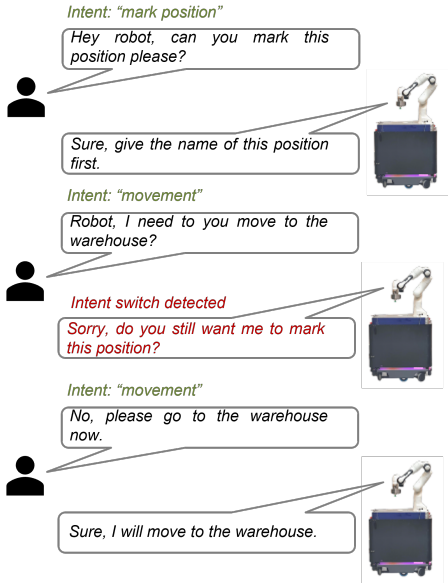

Fig. 3. An example dialogue of Intent Switch for a digitization task.

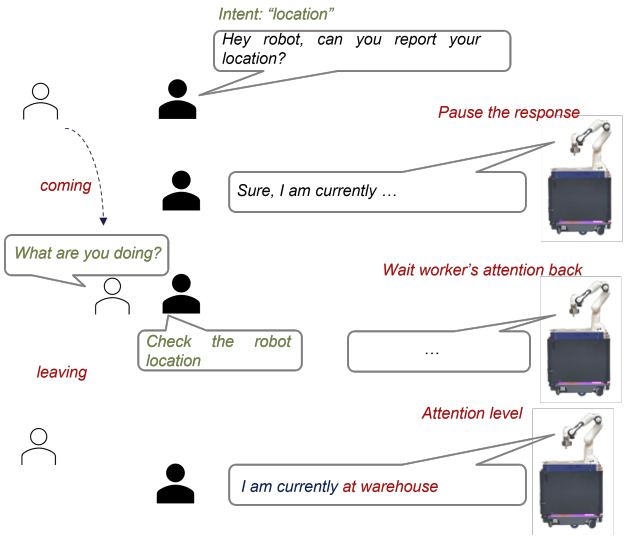

Fig. 4. An example of head pose estimation for a localization task

In the *Localization* scenario, the operator needs to check the LH's current position in a digital map. Two shop floor workers are involved in this scenario. One worker is asked to operate the LH moving in the lab and requests the LH to report the current position on the digital map. The other worker needs to interrupt the above interaction by asking a random question. The purpose is to testing the whether the system can predict the attention lapses through the head pose estimation. Fig. 4 shows an example of head pose estimation for a location checking task.

The performed tasks and respective prediction accuracy are presented in Table II 30 times experiments are conducted on each scenario to verify the accuracy of the attention prediction and validate its overall performance. The sample experiment

TABLE II
PERFORMANCE OF THE ATTENTION PREDICTOR IN SELECTED SCENARIOS

Scenario ID	Scenario description	Attention Prediction Accuracy
1	digitization	0.74
2	Localization	0.80

videos [5] [6] are available.

IV. DISCUSSION AND CONCLUSIONS

The proposed multimodal attention prediction method benefits from both vision (i.e., head pose estimation) and speech (i.e., dialog consistency) inputs, enabling high accuracy attention tracking in HRI scenarios in industrial robots. The advantages of the proposed methods are threefold. Firstly, it is non-invasive as it does not require physical contact with the shop floor worker being monitored, which can be essential in industrial settings where safety is a concern. Secondly, it provides real-time monitoring that can provide real-time information on a worker's attention level, important for adjusting the robot's behavior to optimize task performance and safety. Thirdly, it is robust, as the method can be robust to changes in lighting conditions and other environmental factors, making it suitable for use in industrial settings. However, there are also observed limitations, such as accuracy, limited modality, and privacy. For example, the pre-defined rules for attention prediction are relatively simple. Although the current prediction accuracy is high, it may not always accurately reflect a person's true attention level. For instance, a person may be looking at their workspace but thinking about something else. Furthermore, the current two modalities do not capture all aspects of a person's attention level, such as emotional state or level of engagement in a task. Additionally, using cameras and other sensors to monitor worker behavior can raise privacy concerns, particularly in industrial settings where sensitive information may be involved.

In the future, we plan to include more modalities, such as eye-gazing and body position, with the current approach and provide a more comprehensive representation of the attention level. Furthermore, we will investigate industrial standards (e.g., ISO 10218 [7]) and design strategies [16] [17] to build a safer environment for human-robot collaboration in industrial settings. We are also reaching out to our industrial partners to identify potential industrial cases where human attention needs to be tracked and monitored for further validation.

ACKNOWLEDGMENT

The authors would like to acknowledge support by Aalborg University Bridging Project (A Multimodal Attention Tracking in Human-Robot Collaboration for Manufacturing Tasks), EU's SMART EUREKA programme under grant agreement S0218-chARmER and the H2020-WIDESPREAD project no. 857061 "Networking for Research and Development of Human Interactive and Sensitive Robotics Taking Advantage of Additive Manufacturing – R2P2".

[5]https://shorturl.at/knTZ0

[6]https://shorturl.at/epJOZ

[7]https://www.iso.org/standard/51330.html

REFERENCES

[1] C. Jost, B. Le Pévédic, T. Belpaeme, C. Bethel, D. Chrysostomou, N. Crook, M. Grandgeorge, and N. Mirnig, Eds., *Human-robot interaction : evaluation methods and their standardization*. Springer, 2020, vol. 12. [Online]. Available: https://doi.org/10.1007/978-3-030-42307-0

[2] J. F. Buhl, R. Grønhøj, J. K. Jørgensen, G. Mateus, D. Pinto, J. K. Sørensen, S. Bøgh, and D. Chrysostomou, "A dual-arm collaborative robot system for the smart factories of the future," *Procedia manufacturing*, vol. 38, pp. 333–340, 2019. [Online]. Available: https://doi.org/10.1016/j.promfg.2020.01.043

[3] L. Paletta, A. Dini, C. Murko, S. Yahyanejad, M. Schwarz, G. Lodron, S. Ladstätter, G. Paar, and R. Velik, "Towards real-time probabilistic evaluation of situation awareness from human gaze in human-robot interaction," 03 2017, pp. 247–248.

[4] M. Lagomarsino, M. Lorenzini, P. Balatti, E. D. Momi, and A. Ajoudani, "Pick the right co-worker: Online assessment of cognitive ergonomics in human-robot collaborative assembly," *IEEE Transactions on Cognitive and Developmental Systems*, pp. 1–1, 2022.

[5] M. Lagomarsino, M. Lorenzini, E. De Momi, and A. Ajoudani, "Robot trajectory adaptation to optimise the trade-off between human cognitive ergonomics and workplace productivity in collaborative tasks," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 663–669.

[6] M. Lagomarsino, M. Lorenzini, E. De Momi, and A. Ajoudani, "An online framework for cognitive load assessment in industrial tasks," *Robotics and Computer-Integrated Manufacturing*, vol. 78, p. 102380, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736584522000679

[7] E. G. Moreno-Esteva and M. S. Hannula, "Using gaze tracking technology to study student visual attention during teacher's presentation on board," in *CERME 9-Ninth Congress of the European Society for Research in Mathematics Education*, 2015, pp. 1393–1399.

[8] O. Lorenz and U. Thomas, "Real time eye gaze tracking system using cnn-based facial features for human attention measurement." in *VISIGRAPP (5: VISAPP)*, 2019, pp. 598–606.

[9] G. Skantze and J. Gustafson, "Attention and interaction control in a human-human-computer dialogue setting," in *Proceedings of the SIGDIAL 2009 conference*, 2009, pp. 310–313.

[10] C. Li, J. Park, H. Kim, and D. Chrysostomou, "How can i help you? an intelligent virtual assistant for industrial robots," ser. HRI '21 Companion. New York, NY, USA: Association for Computing Machinery, 2021, p. 220–224. [Online]. Available: https://doi.org/10.1145/3434074.3447163

[11] C. Li, D. Chrysostomou, D. Pinto, A. K. Hansen, S. Bøgh, and O. Madsen, "Hey max, can you help me? an intuitive virtual assistant for industrial robots," *Applied Sciences*, vol. 13, no. 1, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/1/205

[12] C. Schou, R. S. Andersen, D. Chrysostomou, S. Bøgh, and O. Madsen, "Skill-based instruction of collaborative robots in industrial settings," *Robotics and Computer-Integrated Manufacturing*, vol. 53, no. June 2016, pp. 72–80, 2018. [Online]. Available: https://doi.org/10.1016/j.rcim.2018.03.008

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[15] M. Nardello, O. Madsen, and C. Møller, "The smart production laboratory: A learning factory for industry 4.0 concepts," in *CEUR Workshop Proceedings*, vol. 1898. CEUR Workshop Proceedings, 2017.

[16] L. Chen, L. Huang, C. Li, L. Wu, and W. Luo, "Design and safety analysis for system architecture: A breeze/adl-based approach," in *2014 IEEE 38th Annual Computer Software and Applications Conference*, 2014, pp. 261–266.

[17] A. Hameed, A. Ordys, J. Możaryn, and A. Sibilska-Mroziewicz, "Control system design and methods for collaborative robots," *Applied Sciences*, vol. 13, no. 1, p. 675, 2023.