Article Identifier
DOI: 10.22032/dbt.58931
URN: urn:nbn:de:gbv:ilm1-2023isc-135:5

60th ILMENAU SCIENTIFIC COLLOQUIUM
Technische Universität Ilmenau, 4 – 8 September 2023
URN: urn:nbn:de:gbv:ilm1-2023isc:1

On Contextual Perception of Workers in Complex Production Environments

Thorsten Hempel and Ayoub Al-Hamadi

Neuro-Information Technology (NIT),
Otto von Guericke University, Magdeburg, Germany

ABSTRACT

In this work we focus on the challenges of perceiving and coordinating spatial actions between humans and robots in production systems. We address the fundamental questions of how the affective states of individuals in the production process can be visually captured and interpreted in order to facilitate intuitive interactions without explicit commands. Additionally, we investigate methods to analyze the environment and action context in a semantic scene to anticipate user and action intentions. Lastly, we formulate decision approaches to derive appropriate interaction strategies based on affective user states and intentions in the scene context to improve productive collaboration between humans and robots in production environments. By addressing these challenges, this work aims to improve the efficiency of productive teaming processes in production systems.

Index Terms - Productive Teaming, Human-Robot Interaction, Contextual Perception, Affective Anticipation, Interaction Strategies, Semantic Mapping

INTRODUCTION

Collaborative robots (Cobots) are considered the new, advanced generation of robots that are intended to replace stationary and inflexible industrial robots. Cobots are mobile and versatile, suitable for various tasks, equipped with the ability to learn and to adapt to increasing and changing demands. These capabilities enable them to safely break the barrier between human and robotic workspaces to unite the abilities of humans and robots in symbiotic collaboration. Yet, the path towards this kind of collaboration with humans in a shared and dynamic workspace is still paved with technical challenges. These challenges can be categorized into three main areas: capturing and processing human actions, capturing and processing the corresponding workspace, and generating suitable interaction strategies based on information from both humans and the workspace. In this paper we present a conceptual approach to address these tasks in a flexible interaction system to increase adaptivity and efficiency in productive teaming processes [1].

METHODS

The development of neural networks in the recent years have led to significant performance leaps for Convolutional Neural Networks (CNNs) in image-based detection and classification. However, their objectives are often limited to individual tasks, based solely on visual features and independent of the scene context. Methods such as person detection [2], identification [3], head pose prediction [4], gaze direction estimation [5], emotion recognition [6], and eye contact detection are popular approaches to tackle human understanding, but each method alone is insufficient to infer more complex affective states (See Fig. 1). Instead, they need to be combined in a meaningful way to create a holistic scene understanding and to anticipate states like willingness to interact, need for help, or refusal.

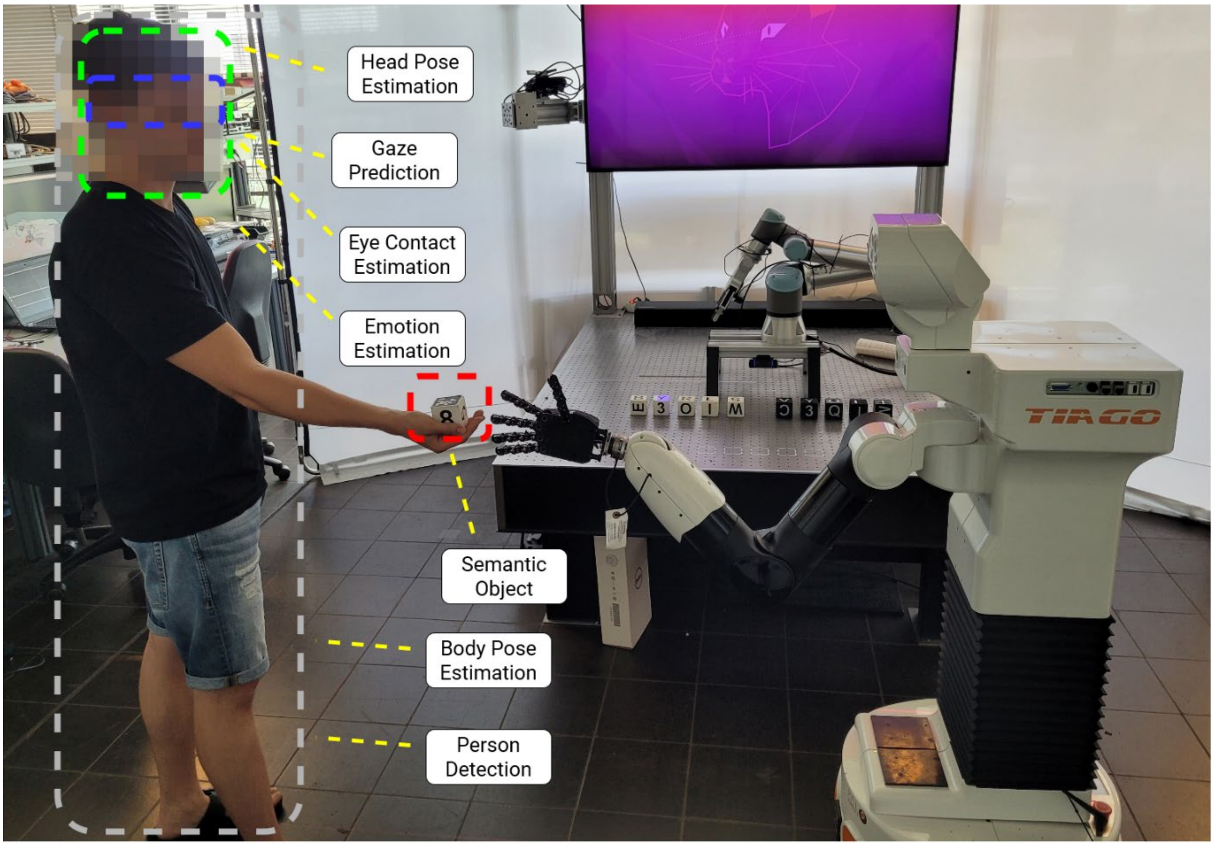

Figure 1: Example situation of a team interaction scenario consisting of a mobile robot and human interaction partner. The robot is capable to capture multiple kinds of information, but it requires further processing algorithms to meaningful combine the extracted information into a holistic scene understanding.

We propose structuring the functional scope of a cobot, designed for deployment as a human assistant in a dynamic workspace, into three categories: Affective Human Anticipation, Semantic Mapping, and Interaction Plan Generation. Figure 2 shows an overview of this classification and its dependencies. In the following, we will delve into the requirements and approaches for each category in more details.

Affective Human Anticipation refers to the ability of a cobot to anticipate and understand the emotional states or intentions of humans with whom it interacts. It involves the robot's capacity to recognize and interpret human emotions, behaviors, and non-verbal cues to predict the emotional or cognitive states of individuals. This enables the cobot to engage in more empathetic, responsive, and effective interactions with humans in various contexts, such as healthcare, social robotics, or customer service.

Semantic mapping involves the process of creating a representation of the environment that incorporates not only spatial information but also semantic or meaningful understanding of the objects, structures, and concepts within the environment. It goes beyond traditional mapping techniques that focus solely on geometric or spatial data. Semantic mapping allows cobots to not only navigate and localize themselves within an environment but also understand the purpose and meaning of different regions or objects. It involves techniques such as object recognition, scene understanding, and semantic segmentation to identify and classify objects, surfaces, and other relevant elements in the environment.

Interaction plan generation refers to the process of generating a structured and coordinated sequence of actions for the robot to effectively interact with humans within an environment represented by semantic maps.
This involves leveraging the semantic understanding of the environment to guide the cobot's behavior and decision-making during interactions. The interaction plan takes into account both the goals of the robot and the intentions or requests of the human user, while considering the semantic information present in the environment.
By utilizing the semantic maps, the cobot can better comprehend the spatial layout, objects, and context of the environment. This understanding enables the cobot to generate interaction plans that account for the semantic attributes of objects, their relationships, and their relevance to the current task or interaction.

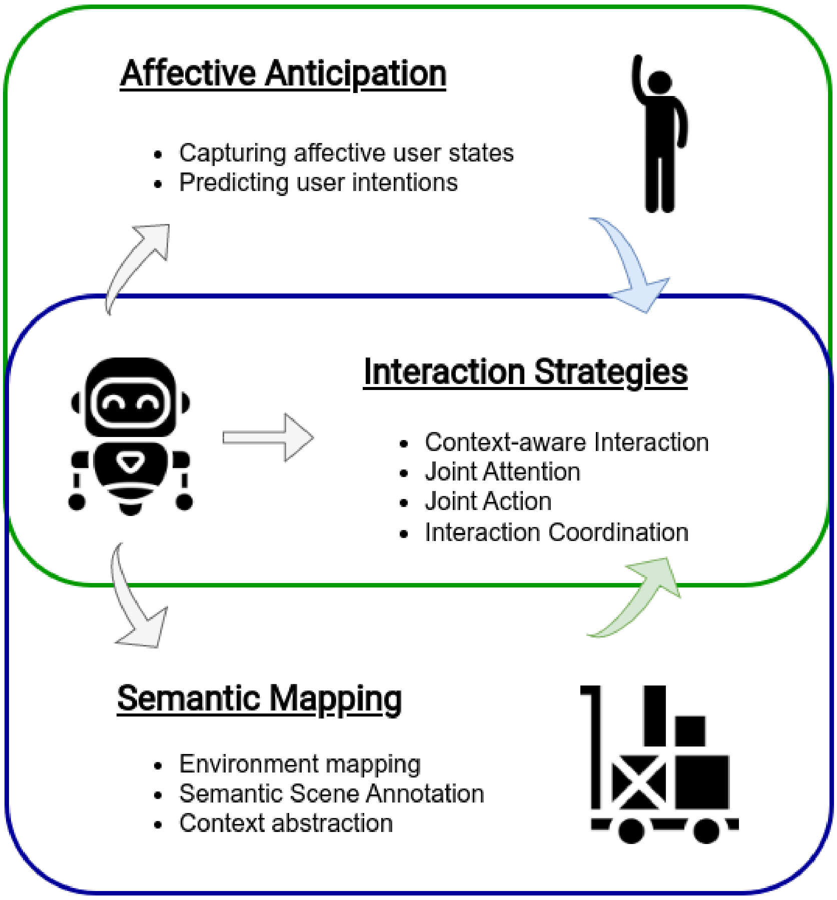

Figure 2: Conceptual overview of the functional structure of a cobot. (Icons from Flaticon.com).

1.1 Affective Human Anticipation

Based on our former work [7], we propose using a simplified Partially Observable Markov Decision Process (POMDP) for this purpose.

A POMDP is a mathematical framework used to model probabilistic decision-making in situations where the underlying state is not directly observable but is instead inferred from partial observations.
It can formally be described as a 7-tuple

$$P =< S, A, T, R, Z, O, y >,$$

where S is a set of partially observable states, A is a set of actions, T is a set of conditional transition probabilities with $T(s' \mid s, a)$ for the state transition $s \rightarrow s'$ conditioned on the taken action. $R: S \times A \rightarrow R$ is the reward function, O is a set of observations, Z is the observation function with $Z(o \mid s', a)$ conditioned on the reached state and the taken action, and $y \in [0,1]$ is the discount factor.

For our task, we focus on set O, which consists of the individual predicted features. These include body pose, head pose, gaze direction, facial expressions (emotions), and eye contact. These observations provide insights into the not directly observable states "Willingness to interact" (s_w), "Need for help" (s_n), and "Refusal"(s_r)

$$S = \{s_w, s_n, s_r\}$$

Now, iteratively sampling new observations for each time t we try to find the policy π^* which maximizes the total expected discounted reward

$$\pi^* = arg\ max\ E\ \left[\sum_{t=0}^{T} y_t R(s_t, a_t)\right]$$

To solve the equation, the Bayesian belief update function approach can be utilized. By incorporating multiple measurements (observations) and continuously updating the belief states, the robustness of state estimation can be improved. Additionally, the observations themselves can be augmented with an additional uncertainty factor (based on prediction confidence). In this way, POMDPs provide a robust, reliable and extendable methodology for estimating affective human states for these scenarios where training data is not available and only subtasks can be processed by neural networks.

1.2 Semantic Mapping

Enriching robot maps with semantic information is an important factor for integrating objects as entities into human-robot interaction. We propose a method based on our previous works [8, 9], where point cloud maps can be augmented in real-time with semantic objects. The foundation is a CNN for 2D object detection. Two challenges arise from this approach: project 2D detection from the image plane into 3D point cloud and data association to assign actual object entities in the map to individual image objects.

Similar to our Affective Human Anticipation method, we pursue a probabilistic approach to suppress measurement errors. We aim to find the centroid of a 3D object represented by $X = (X, Y, Z)^T$, which is projected onto the image plane at point $x = f(X)$ using the projection function [10] f. To account for inaccuracies in both image capture and object prediction, we add measurement noise to this observation model: $x = f(X) + \eta$ with $\eta \sim N(0, \Sigma)$.

For additional robustness, we consider multiple detections of an object, which results in the following probability distribution:

$$p(x_{1:t}|X) = \frac{exp\left(-\frac{1}{2}(f(X) - x_{1:t})^T \Sigma^{-1} - 1(f(X) - x_{1:t})\right)}{\sqrt{(2\pi)^2|\Sigma|}}$$

The posterior distribution of X, given the measurements $x_{1:t}$, can be obtained using Bayes' rule:

$$p(X|x_{1:t}) = \frac{p(x_{1:t}|X)P(X)}{p(x_{1:t})}$$

Assuming a uniform prior distribution and independent measurements, we obtain the following factor representation:

$$p(X|x_{1:t}) = \prod_{t=1}^{T} p(x_t|X),$$

To determine the 3D position X, we search for the X that maximizes the posterior probability:

$$X^* = arg\,max(x)\,p(X|x_{1:t})$$

Having the position and the predicted class of an object, we are left with the challenge of data association. This problem can be simplified to the question: “Is this observed object not mapped yet, or to which mapped object do I have to refer it?”. If the object is already registered in the map, but is assumed not to be, it will be placed twice in the map and can lead to confusion in later data association steps and for the object interaction itself.

Common solution approaches are also based of probabilistic methods, but they tend to make the processing pipeline very computational costly, especially in environment with large amounts of objects. As alternative, we apply a nearest-neighbor search to find the nearest mapped objects with the same class and calculate their distance towards the object candidate. If the distance is below a threshold, it will be referred to the object in map. Otherwise, it will be assumed as a new object and transferred into the map. The threshold is determined dynamically depended on the size of the target object.

Figure 3 shows an illustration of a colored point cloud map with added semantic objects. By comparing the 2D detection rectangle with the appearing point clouds cluster in the corresponding 3D area, one also receives a rough indication of the object’s dimensions.

In summary, this solution provides fast results and works well with objects of different categories. However, if multiple objects of the same class are close together, the data association tends to introduce error in the referencing.

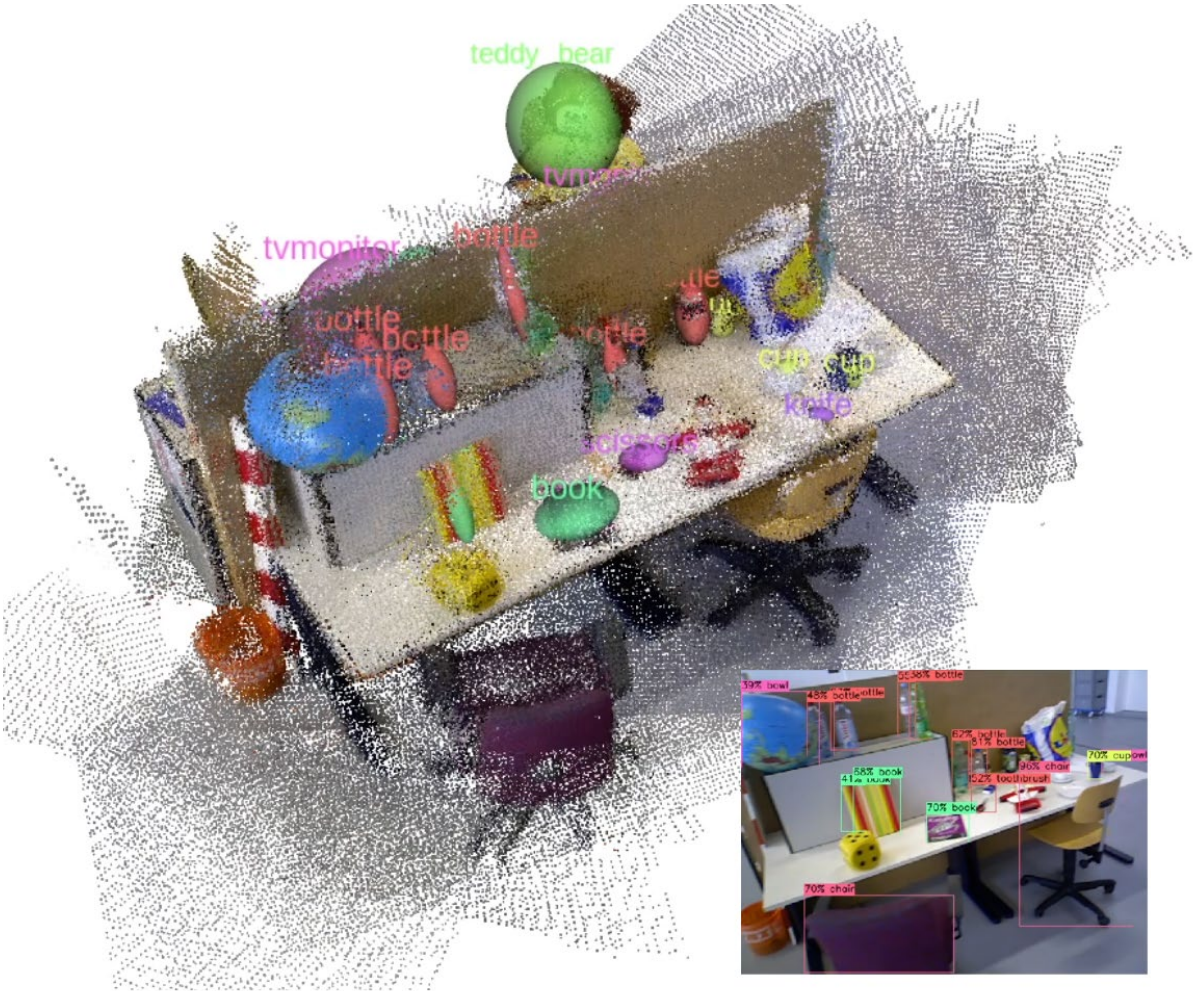

Figure 3: Exemplary image visually capturing the environment as spatial point cloud with addition semantic information.

1.3 Interaction Plan Generation

The interaction plan generation step relies on the condensed information from the affective anticipation, the semantic environment map to interpret interaction goals and to generate appropriate interaction strategies. As these data inputs are attached with uncertainties and inaccuracies in measurements, we address the reasoning probabilistically using a Bayesian network.

Let's consider a scenario where a robot is assisting a human worker in a production environment. The robot's goal is to understand the user's intentions and generate appropriate interaction plans based on semantic mapping and affective user anticipation. The affective user anticipation module may detect signs of "need of help" when the user is struggling to reach a high shelf or observe a smile when they find the mechanical error they were looking for. The semantic mapping module creates a representation of objects, their locations, and their semantic labels in the workspace. The Bayesian network captures the probabilistic dependencies between the user's intentions, the objects in the environment, and the user's affective state.

We will consider three variables: User's Intention (I), Object in the Environment (O), and Affective State (A). Accordingly, the prior probabilities are $P(I)$ for the user's intention, $P(O)$ for the object in the environment, and $P(A)$ for user's affective state.

It follows the conditional probability distributions with $P(I|O)$ for the user's intention given the object in the environment, $P(I|A)$ for the user's intention given the affective state, and $P(O|A)$ for the object in the environment given the affective state. Finally, we define the joint probability $P(I, O, A)$ combining user's intention, object in the environment, and affective state.

Using these components, the Bayesian network is defined as follows:

$$P(I, O, A) = P(I|O) * P(O|A) * P(A) * P(I|A) * P(O)$$

This equation represents the joint probability distribution of the variables involved in the human-robot interaction scenario, that captures the dependencies between the variables and their conditional probabilities.

To make inferences and generate interaction plans, we can use the network to compute posterior probabilities given observed evidence using Bayes' rule. By calculating the posterior probabilities, the robot can infer the user's intentions and generate appropriate interaction plans such as approaching the user with caution, offering assistance, or retrieving other tools or fare for the worker.

CONCLUSION

In this work, we addressed the challenge of achieving a holistic scene understanding for cobots, which is crucial for flexible and efficient human-robot interactions in productive teaming processes. We presented a concept for dividing the overall scenario into subtasks to enable the use of individual solutions. For each subtask, namely affective human anticipation, semantic mapping, and interaction plan generation, a robust probabilistic inference solution is proposed.

The concepted presented in this paper will be implemented as part of an interaction system prototype on a mobile robot and evaluated in a user study.

ACKNOWLEDGEMENT

This work is funded and supported by the Federal Ministry of Education and Research of Germany (BMBF) (AutoKoWAT-3DMAt under grant Nr. 13N16336) and German Research Foundation (DFG) under grants Al 638/13-1, Al 638/14-1 and Al 638/15-1.

REFERENCES

[1] M. Johnson and J. M. Bradshaw, “How Interdependence Explains the World of Teamwork,” in *Engineering Artificially Intelligent Systems*: Springer, Cham, 2021, pp. 122–146. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-89385-9_8

[2] M.-A. Fiedler, P. Werner, A. Khalifa, and A. Al-Hamadi, “SFPD: Simultaneous Face and Person Detection in Real-Time for Human-Robot Interaction,” *Sensors (Basel, Switzerland)*, vol. 21, no. 17, 2021, doi: 10.3390/s21175918.

[3] A. Khalifa, A. A. Abdelrahman, D. Strazdas, J. Hintz, T. Hempel, and A. Al-Hamadi, “Face Recognition and Tracking Framework for Human–Robot Interaction,” *Applied Sciences*, vol. 12, no. 11, p. 5568, 2022, doi: 10.3390/app12115568.

[4] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, “6d Rotation Representation For Unconstrained Head Pose Estimation,” in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022.

[5] A. A. Abdelrahman, T. Hempel, A. Khalifa, and A. Al-Hamadi, “L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments,” Mar. 2022. [Online]. Available: http://arxiv.org/pdf/2203.03339v1

[6] S. Handrich, L. Dinges, A. Al-Hamadi, P. Werner, and Z. A. Aghbari, “Simultaneous Prediction of Valence/Arousal and Emotions on AffectNet, Aff-Wild and AFEW-VA,” *Procedia Computer Science*, vol. 170, pp. 634–641, 2020, doi: 10.1016/j.procs.2020.03.134.

[7] T. Hempel, L. Dinges, and A. Al-Hamadi, “Sentiment-based Engagement Strategies for intuitive Human-Robot Interaction,” Jan. 2023. [Online]. Available: http://arxiv.org/pdf/2301.03867v1

[8] H. Thorsten, F. Marc-Andre, K. Aly, A.-H. Ayoub, and D. Laslo, “Semantic-Aware Environment Perception for Mobile Human-Robot Interaction,” in *2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2021.

[9] T. Hempel and A. Al-Hamadi, “An online semantic mapping system for extending and enhancing visual SLAM,” *Engineering Applications of Artificial Intelligence*, vol. 111, p. 104830, 2022, doi: 10.1016/j.engappai.2022.104830.

[10] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision,* 2nd ed. Cambridge: Cambridge University Press, 2003.

CONTACTS

M. Sc Thorsten Hempel
email: thorsten.hempel@ovgu.de
ORCID: https://orcid.org/0000-0002-3621-7194

Prof. Dr.-Ing. habil. Ayoub Al-Hamadi
email: ayoub.al-hamadi@ovgu.de
ORCID: https://orcid.org/0000-0002-3632-2402