CoMeSy: Multimodal Interaction with a Situated Cobot for Collaborative Tasks

Sven Milde, Alexander Jost, Rainer Blum, Jan-Torsten Milde,
Marius Schultheis, Johannes Weyel, Tobias Müller, Thies Beinke
Niklas Schreiner, Julian Heumüller, Dennis Möller, Frank Hartmann
Department of Computer Science
Fulda University of Applied Sciences
Fulda, Germany
e-mail: (sven.milde, alexander.jost, rainer.blum)@cs.hs-fulda.de, milde@hs-fulda.de,
(marius.schultheis, johannes.weyel, tobias.mueller, thies.beinke, niklas.schreiner)@et.hs-fulda.de,
(julian.heumueller, dennis.möller, frank.hartmann)@alpaka-innovation

***Abstract*—The CoMeSy project is developing a system for multimodal interaction between humans and cobots, where the cobot acts as an intelligent assistant. The system uses speech and gestures as input, and responds with speech, sounds, actions, and visual feedback. A key challenge is dynamically creating action plans based on human input, world knowledge, and visual perception. The system integrates several technologies, including speech recognition and synthesis, image processing, object detection, hand tracking, and acoustic feedback. Currently in development, the project aims to address intelligent communication, situational understanding, dynamic planning, reactive behavior, and robust handling of interruptions, with plans for empirical evaluation.**

Keywords-Multimodal interaction; Human-Robot Interaction (HRI); Collaborative robotics; Cobot as an intelligent assistant.

I. Introduction

Collaborative robotics, especially the interaction between humans and robots (HRI), has become a central research area in robotics [1], [2]. The CoMeSy (CoMeSy: *Co*bot *Me*nsch *S*ymbiose, German for Cobot Human Symbiosis) project focuses on multimodal, situationally embedded interaction with a cobot with the aim of jointly carrying out action processes. The cobot should take on the role of an intelligent, supportive assistant; the human takes the lead in the process. The interaction between human and cobot is realized through multimodal inputs (speech and gestures) [3]. The cobot also reacts multimodally via speech and acoustic signals, can perform situation-adapted actions and give visual feedback via a table projection. A central challenge consists in the creation of dynamic action plans, which are based on the instructions of the human, the inherent world knowledge of the current domain as well as the (visual) perception of the current state of the world.

The rest of the paper ist structerd as follows: Section II will first provide an overview of related work, followed by an explanation of the underlying research agenda in Section III. Section IV will elaborate on the basic technical scenario, for which the situated action control will be described in Section V. Section VI concludes the paper and summarizes the central findings.

II. Related Work

The development of multimodal human-robot interaction (HRI) has seen significant progress in recent years, particularly with regard to the integration of speech, gestures and other modalities. The objective of this integration is to establish seamless and intuitive communication between humans and machines.

A comprehensive overview of human-robot collaboration is provided in [4]–[6]. The impact of robot movements on human-robot collaboration is analysed in [7] and is pertinent to the planning of robot movements to avoid collisions with the user.

The significance of multimodal interfaces is highlighted in [8], with a primary focus on communication through speech and gesture. In particular, it is emphasised that the content of the communication is more important than the method, which is facilitated by the use of multimodality. This assertion is further substantiated by comparative analyses of unimodal and multimodal user interfaces in [9], [10]. The findings of these studies demonstrate that multimodality serves to reduce cognitive load. This is considered to be of crucial importance for the system concept under discussion, with the result that the user can concentrate primarily on the work task and act intuitively with the robot. In a similar vein, the study by Turk [11] corroborates the notion that multimodality fosters enhanced user acceptance. Common misconceptions in the domain of multimodal interaction and its advantages are illustrated by Oviatt [12], while Baltruvsaitis et.al. [13] examines the relationships between the modalities and describes various forms of fusion, as well as the challenges that arise when evaluating multimodal interaction. The challenges of evaluating multimodal interaction are also considered in the context of this work, as it is essential to interpret the various modalities correctly and to integrate them to form a comprehensive understanding.

The integration of Large Language Models (LLMs) in robot planning is investigated in [14], wherein the capacity of LLMs to generate Behaviour Trees for robot task planning is analysed. An interactive planning approach with LLMs that enables agents to handle multiple tasks in open environments

is presented in Wang et.al. [15]. In this work, however, the LLM will be used to transition from a voice command to a series of granular actions. Additionally, Ao et al. [16] adopt an approach integrating speech and gestures into a planning method with an LLM, with the objective of generating a sequence of actions for a robot. Another LLM-driven approach to robust task planning and execution is presented in [17], where mainly visual information is used to obtain a worldview and plan tasks in it, whereas in this work a multimodal input is used.

The subject of computer-based hand gesture recognition for HRI is addressed in depth in the publication by Qi et al. [18]. The utilisation of direct gestures for communication with robotic systems is addressed in [19]. A context-aware robotic assistance system based on pointing-based gestures to support humans is described in [20] and demonstrates the significance of the robot's world knowledge for this work, ensuring the accurate interpretation of gestures.

A comprehensive overview of Behaviour Trees (BTs) in robotics and artificial intelligence is given in [21], [22], highlighting the application of BTs to control robot behaviour as it will be used in this work. A key benefit of behavior trees is their capacity for dynamic response to alterations during runtime.

III. Research agenda: an intelligent, collaborative system

The primary goal of the project is the development of an intelligent, collaborative system that can be controlled by natural language communication and intuitive gestures. CoMeSy should be able to perform complex tasks in dynamic environments and flexibly adapt to unexpected situations. The research agenda includes the following points:

1) *Intelligent communicative behavior within the framework of collaborative action:* the system should be able to understand and generate natural language and effectively integrate it into the cooperative work process. To do this, the system must be able to exchange relevant information, give and receive instructions, provide and process feedback, and coordinate its own actions with the human actor.
2) *Situational interpretation of linguistic instructions and gestures*: CoMeSy should interpret linguistic instructions and gestures situationally, i.e., grasp the meaning of language and gestures not in isolation, but in the context of the respective situation. This includes resolving ambiguities, processing ellipses, deictic expressions, interjections, and prosodic markings, accepting the pragmatic use of language, and ideally recognizing the intention of the communication partner.
3) *Dynamic action planning based on a linguistic target specification*: the system should be able to create dynamic action plans based on a linguistic target specification. This requires the ability to break down complex tasks into subtasks, develop suitable action strategies, and adapt them flexibly to changing world states
4) *Integration of reactive behavior control into action execution*: the system is able to react quickly and appropriately to unexpected events and changes in the environment and, if necessary, adapt action plans and develop alternative strategies.
5) *Robust system behavior in case of interruptions or dynamic changes of the world state*: CoMeSy should demonstrate robust system behavior in case of interruptions (such as communication breakdowns) or dynamic changes in the world state.

IV. Scenario: collaborative work

CoMeSy is currently in its first construction phase. In addition to the physical construction of a work cell, the components for sensor data preprocessing, in particular, have been implemented:

- Speech recognition and speech synthesis (whisper.cpp, coqui/XTTS-2)
- Reading out the current camera images (opencv)
- Visual object detection and classification (yolo11m with fine-tuning via LabelStudio)
- Hand detection and posture analysis (mediapipe and tensorflow-based ANN (Artificial Neural Network)
- Synthesis of acoustic feedback signals (SuperCollider)

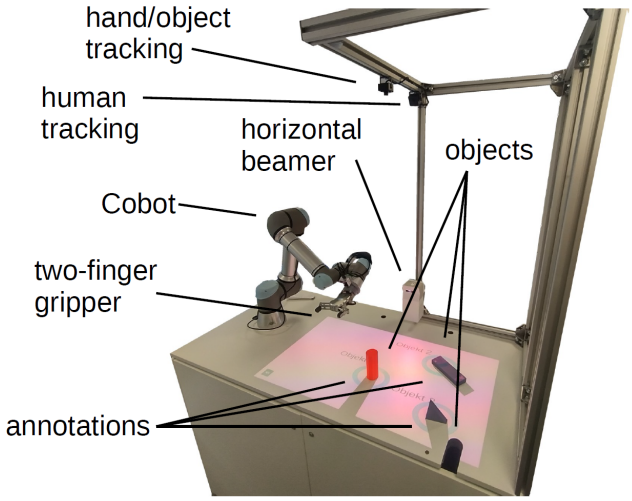

Figure 1. The work cell: Cobot and human share a workspace. Two cameras capture the workspace and the human. Visual information can be displayed on the work surface via a horizontal beamer.

A. The work cell

Human and Cobot (here Universal Robot UR-5, or IGUS Rebel) work together in a work cell (see Fig. 1) and share a common work surface. The UR-5 is equipped with a 2-finger gripper that allows it to grasp and place objects. The 2-finger gripper is also equipped with a force-torque sensor that enables precise measurements of the forces and torques acting on the gripper.

A depth image camera, which is mounted above the Tool Center Point (TCP) of the robot, captures both the objects to be gripped and the current position of the 2-finger gripper. The captured objects are initially designed in simple geometric shapes, such as cuboids, cylinders, prisms and cubes, whereby they can be in different colors.

The work surface itself is captured by a vertically aligned camera. This camera is able to identify the position and orientation of the objects on the work surface. In addition, it has the ability to identify the hands of the human, to distinguish between the left and right hand and to determine the respective hand posture (see Fig. 2).

The upper body of the human is captured by an inclined camera to obtain additional information about the posture and movements of the user. A short-range beamer projects an image onto the work surface to provide visual information or instructions.

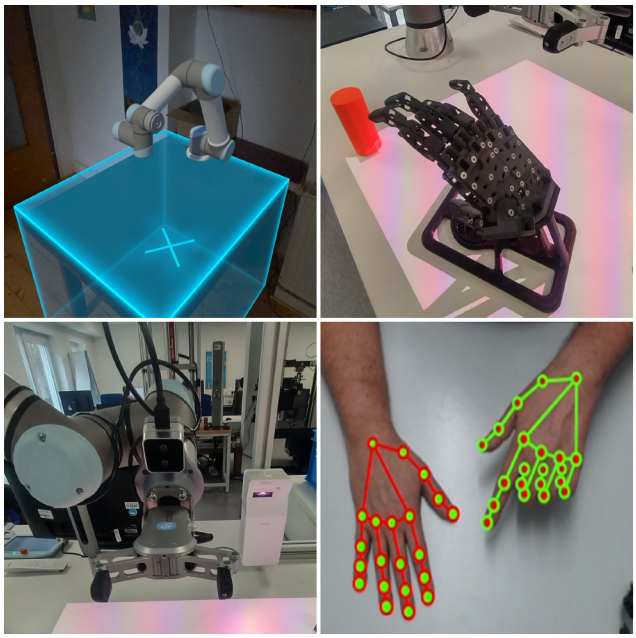

Figure 2. Impressions: a): The prototype of the AR application for the UR-5. b): Prototype 5-finger hand c): View of the 2-finger gripper, depth camera and short-range beamer. d): Hand posture recognition

The multimodal interaction is completed by acoustic inputs via a 4-channel directional microphone and acoustic outputs via a loudspeaker. This setup enables a comprehensive recording of the interactions between human and cobot as well as an intuitive and versatile communication interface.

B. System architecture

The system architecture (see Fig. 3) consists of two major subcomponents. The first consists of ROS2 (Robot Operating System version 2) nodes, which capture the various sensor data and, in particular, perform visual pre-processing and recognition. There are also nodes for controlling the overall system, which relate to different modalities, such as speech, audio, or movement. The Blackboard serves as a memory in this subcomponent and records all sensor data and intermediate results. The BehaviorTree [22] is the central control of the system, which ensures a reactive and dynamic behavior of the robot.

The second sub-component consists of various parallel processes, some of which are particularly computationally intensive and therefore have been outsourced to a corresponding AI server. This includes, above all, audio processing with speech recognition, keyword spotting, and audio synthesis. The integration of a Large-Language Model (LLM) to evaluate speech commands is also implemented as a seperate process. For this purpose, llama3.3 is run locally in Ollama [23]. Information about the internal state of the system, such as recognized objects or understood commands, is visualized on the work surface via a short-range beamer. This visualization is also implemented as a seperate process. The communication between the two sub-components is implemented as a REST API via FastAPI [24].

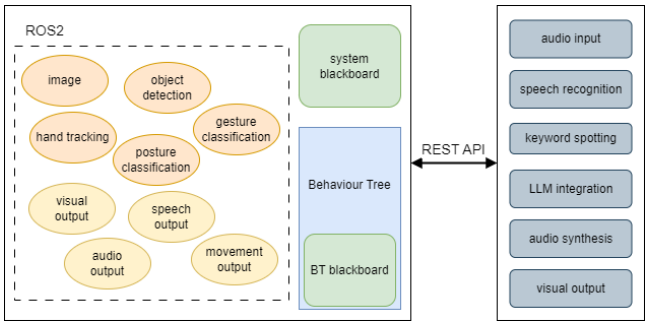

Figure 3. The system architecture: ROS2 nodes control the robot actions. Computationally expensive processes run on the AI (Artificial Intelligence) Server. The two parts communicate via a REST API.

V. Situational Action Control

A. BehaviorTrees

In order for the system to be able to dynamically react to user input and sensor readings at any time, the individual running processes must be fine-grained and interruptible. With the help of BehaviorTrees, this property can be achieved, since BehaviorTrees are run at a defined frequency and the individual nodes are stopped when they receive a corresponding signal from higher levels of the tree. This signal can be generated by other nodes or reactively when data on the Blackboard has been changed."

To solve varying tasks using the BehaviorTree [22], the individual actions that the robot is able to perform are represented in the smallest possible subtrees. Based on user input, a sequence of granular actions can then be defined, which can then be dynamically loaded into the tree as a sequence. This allows different sequences of operations to be put together which are not pre-programmed into the system as predetermined fixed processes. The basic BehaviorTree (see Fig. 4) thus consists of a node that calibrates the entire system when it is started and performs a system check, a higher-level

error handling, and a node that dynamically loads the various subtrees, depending on the command given.

In addition, a parallel node should run as high as possible in the BehaviorTree, ensuring the processing of the sensor data. For this purpose, different topics are subscribed within the ROS network and stored partly directly and partly after prior processing in the Blackboard, so that the BehaviorTree can access them quickly and easily.

The use of the Blackboard has the advantage that during a single tick within the BehaviorTree, the data in the Blackboard is not changed, so that the entire BehaviorTree works with the same consistent values. If individual nodes were to subscribe to the same topics independently, the values could change and the tree could thus work with inconsistent data.

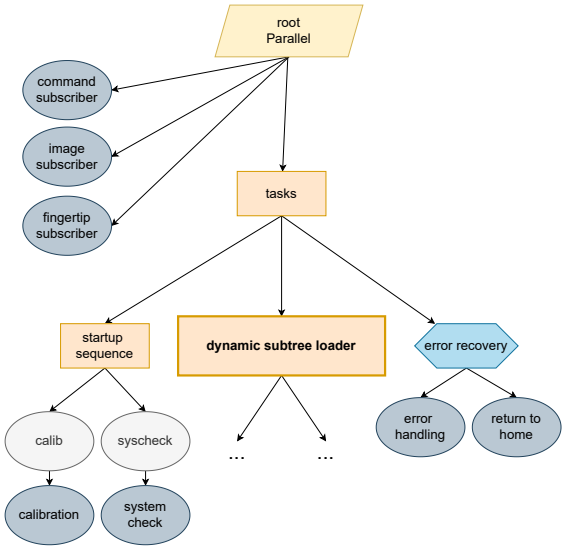

Figure 4. The basic BehaviorTree: At the very top (marked yellow) is the parallel node, which contains the various subscribers and the main task.

B. Multimodal action sequences: examples

The interaction between humans and robots in collaborative robotics requires intuitive and efficient communication strategies. Speech-gesture-control allows humans to interact with the robot in a natural intuitive way and it also allows to coordinate complex tasks. In the context of a work cell, where objects are located on a table, the situation-based formulation of action instructions is crucial for the success of human-robot collaboration.

In the following, 3 exemplary multimodal interaction sequences are outlined to illustrate the complexity of speech-gesture controlled human-robot interaction. We assume the following initial situation: several geometric objects are positioned on the table. The cobot is ready to take on tasks.

A crucial element for the interaction between humans and robots is *visual perception*. The camera enables the cobot to perceive its environment, recognize objects, and move safely within the space. In the example dialogue in table I, objects are verbally referenced via shape, color, and spatial relation, and identified through the visual channel.

TABLE I. ACTION-SEQUENCE 1: COLLABORATIVE BUILDING

agent	action/utterance
Human	"Hand me the red cube."
Cobot	(Moves camera to the tabletop, identifies the red cube.) "Alright, which arm would you like?"
Human	(Holds out left hand.) "Left, please."
Cobot	(Grasps the cube with the left hand, lifts it, and rotates it so that the human can easily grasp it.)
Human	(Grasps the cube and places it on top of the blue pyramid.)
Cobot	(Observes the action.) "Would you like to continue building?"

TABLE II. ACTION-SEQUENCE 2: FAILED COMMUNICATION

agent	action/utterance
Human	(Points at the red cube and makes a rotating motion with their hand.) "Turn the cube."
Cobot	(Misinterprets the gesture.) "Would you like me to rotate the cube around its own axis?"
Human	(Shakes head.) "No, turn it so that the red side faces upwards."
Cobot	(Re-analyzes the situation.) "Ah, now I understand." (Rotates the cube accordingly.)

In the situational embedding of the cobot, it is assumed that communication and action can *fail*, and in many cases they will. Errors can occur due to misunderstandings (see Table II) or unforeseen events (see Table III). It is, therefore, important that humans have the opportunity to intervene in the work process and make dynamic corrections if necessary.

VI. SITUATION-BASED LANGUAGE CONTROL

Language processing in CoMeSy takes place on two conceptual levels. At the lower level 1, we consider strictly situationally embedded instructions, which often have deictic constructions or are situationally expanded and interpreted as linguistic ellipses. These instructions usually refer to the current running sub-action and allow humans to linguistically control and modify it.

At the higher level 2, the specification of more complex actions takes place; actions that are constructed as a sequence of sub-actions and that need to be planned in a dynamic and situatiated way.

The following are examples of level 1 instructions, explained in detail. We differentiate between *general instructions*, *position-related instructions*, *manipulation-related instructions*, and *situational modifiers*.

A. general instructions

- *Stop, Halt*: Interrupt the current sub-action.
- *Continue*: Resume the previously interrupted sub-action.
- *Yes, No*: Confirm or reject instructions or feedback from the robot.
- *Faster, Slower*: Adjust the robot's action speed (to suit human needs).
- *Correct, Incorrect*: Evaluate the robot's action, allowing for error correction.

B. position-related instructions

- *Left, Right, Forward, Backward, Up, Down*: Description of directions and positions in space, often in combination with prepositional phrases (e.g., "in front of the object").
- *In front of, Behind*: Describe the position of an object in spatial relation to one another.
- *Above, Below*: Also relational, but refer to the vertical positioning of objects.

C. manipulation-related instructions

- *Grasp, Release*: Describe the basic actions of the robot gripper.
- *Turn/Rotate the hand, The other way around, Clockwise, Counterclockwise*: Control of the robot's gripper during rotational movements.

TABLE III. ACTION-SEQUENCE 3: UNFORSEEN EVENTS

agent	action/utterance
Human	"Can you bring the green sphere to the pyramid?"
Cobot	"Sure!" (Moves to the sphere, grasps it.)
-	(Unexpected event): The sphere rolls away and falls off the table.
Cobot	(Searches for the sphere with the camera.) "Oh, the sphere fell. I'll try to get it." (Moves to the floor, grasps the sphere.)
Human	(Points to a new spot on the table.) "Put it there."
Cobot	(Places the sphere in the desired location.)

D. situational modifiers

- *A little more, A little bit further* : Fine-tuning of positions and movements. These instructions have lower precision and may require additional visual or tactile feedback.

VII. CONCLUSIONS

CoMeSy is currently in the technological development phase. Hardware and software components have been identified, a system architecture has been outlined, and in the following weeks, with the completion of the first iteration, integration into a working prototype will take place. Subsequently, the points of the research agenda will be addressed. In addition to carrying out the necessary functional tests, the system will be empirically evaluated via subject tests. Furthermore, the complexity of the scenario is to be gradually increased, for example, by integrating the 5-finger hand and using further objects.

REFERENCES

[1] R. Müller *et al.*, *Handbuch Mensch-Roboter-Kollaboration*. Carl Hanser Verlag GmbH Co KG, 2023.

[2] H. Su *et al.*, "Recent advancements in multimodal human–robot interaction," *Frontiers in Neurorobotics*, vol. 17, p. 1 084 000, 2023.

[3] R. Younes, F. Elisei, D. Pellier, and G. Bailly, "Impact of verbal instructions and deictic gestures of a cobot on the performance of human coworkers," in *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, IEEE, 2024, pp. 1040–1047.

[4] B. Chandrasekaran and J. M. Conrad, "Human-robot collaboration: A survey," in *SoutheastCon 2015*, IEEE, 2015, pp. 1–8.

[5] M. Peshkin and J. E. Colgate, "Cobots," *Industrial Robot: An International Journal*, vol. 26, no. 5, pp. 335–341, 1999.

[6] A. Bauer, D. Wollherr, and M. Buss, "Human–robot collaboration: A survey," *International Journal of Humanoid Robotics*, vol. 5, no. 01, pp. 47–66, 2008.

[7] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of robot motion on human-robot collaboration," in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 2015, pp. 51–58.

[8] D. Perzanowski, A. C. Schultz, W. Adams, E. Marsh, and M. Bugajska, "Building a multimodal human-robot interface," *IEEE intelligent systems*, vol. 16, no. 1, pp. 16–21, 2001.

[9] S. Oviatt, "Advances in robust multimodal interface design," *IEEE computer graphics and applications*, vol. 23, no. 05, pp. 62–68, 2003.

[10] S. Oviatt, R. Coulston, and R. Lunsford, "When do we interact multimodally? cognitive load and multimodal communication patterns," in *Proceedings of the 6th international conference on Multimodal interfaces*, 2004, pp. 129–136.

[11] M. Turk, "Multimodal interaction: A review," *Pattern recognition letters*, vol. 36, pp. 189–195, 2014.

[12] S. Oviatt, "Ten myths of multimodal interaction," *Communications of the ACM*, vol. 42, no. 11, pp. 74–81, 1999.

[13] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[14] Y. Lai *et al.*, "Nmm-hri: Natural multi-modal human-robot interaction with voice and deictic posture via large language model," *arXiv preprint arXiv:2501.00785*, 2025.

[15] Z. Wang *et al.*, "Describe, explain, plan and select: Interactive planning with llms enables open-world multi-task agents," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34 153–34 189, 2023.

[16] J. Ao, F. Wu, Y. Wu, A. Swikir, and S. Haddadin, "Llm as bt-planner: Leveraging llms for behavior tree generation in robot task planning," *arXiv preprint arXiv:2409.10444*, 2024.

[17] C. Rivera *et al.*, "Conceptagent: Llm-driven precondition grounding and tree search for robust task planning and execution," *arXiv preprint arXiv:2410.06108*, 2024.

[18] J. Qi, L. Ma, Z. Cui, and Y. Yu, "Computer vision-based hand gesture recognition for human-robot interaction: A review," *Complex & Intelligent Systems*, vol. 10, no. 1, pp. 1581–1606, 2024.

[19] M. Pascher *et al.*, "Hands-on robotics: Enabling communication through direct gesture control," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 822–827.

[20] S. Drolshagen, M. Pfingsthorn, and A. Hein, "Context-aware robotic assistive system: Robotic pointing gesture-based assistance for people with disabilities in sheltered workshops," *Robotics*, vol. 12, no. 5, p. 132, 2023.

[21] M. Iovino, E. Scukins, J. Styrud, P. Ögren, and C. Smith, "A survey of behavior trees in robotics and ai," *Robotics and Autonomous Systems*, vol. 154, p. 104 096, 2022.

[22] M. Colledanchise and P. Ögren, *Behavior trees in robotics and AI: An introduction*. CRC Press, 2018.

[23] Software: ollama, last access: May, 2025 https://github.com/ollama/ollama.

[24] Software: FastAPI last access: May, 2025 https://github.com/fastapi/fastapi.