Fusing Components for an Attentive and Emotionally Expressive Companion Robot: Meet ZENIT

Christian Felix Purps[1] and Matthias Wölfel[2]

Fig. 1: **Impressions of ZENIT:** Bodily and Facial Expression Examples, Hardware Setup and Human-Robot Interaction.

***Abstract*— This paper introduces a new companion robot as an open research platform for studying human-robot interaction: The ZENIT Enabling Natural Interaction Technology. Open-source solutions for companion robots are often complicated to build or not thought of holistically. ZENIT is based on a few low-cost components, namely a smartphone and a stationary robot arm in conjunction with an average computer and a separate camera. Our platform aims to enable easy use of the multimodality of human communication channels such as body language, spatial behavior, facial expressions and speech so that ZENIT can understand its interlocutor and express itself appropriately. The fusion of various open-source software components in an adaptable, lightweight and resource-efficient distributed system simplifies its replication. The system provides sensory processing and artificial intelligence models to enable a responsive user experience. The system's response time (speech: 2.69s, facial expressions: 0.55s, bodily movements: 0.32s) for perception and expression combined and its reliability proved in a public testing showed considerably adequate results. Evaluation of the robot's expressive capabilities in an online survey (N=23) revealed users were able to understand six of eight emotions conveyed by ZENIT using facial expressions and/or body language (except for contempt and fear) and perceived its non-verbal feedback as significantly more expressive when facial expressions were accompanied with bodily movements.**

I. Introduction

Growing efforts to employ robots as social actors in various areas of application such as knowledge transfer, care support, customer services or entertainment [18], [39], [20] are driven by the idea to use their co-presence to directly create a social situation that enables natural face-to-face communication [32]. For these (social) robots to be accepted as credible social actors, they must exhibit social behavior that is understandable to humans. This requires an artificial social intelligence (ASI) to recognize and analyze verbal language and non-verbal signals (NVS) during a social interaction and then, coded in the response, reflects comprehensible signals back to the conversation partner [23]. An ideal social robot should comprehend all levels of information encoded in human-emitted messages — factual content, self-disclosure, relationship and appeal — and express itself accordingly in order to avoid misunderstandings and ensure functioning communication [35]. However, due to the complex cosmos of human social behavior, only small aspects of behavior can be broken down, which still significantly limits the performance of current systems.

As a subtype of social robots, companion robots with an animal-, human-like or mechanical appearance, are designed for natural and intuitive human-robot interaction (HRI) in non-expert settings [5]. The appearance of companion robots is currently tending towards more abstract embodiments, both human- and animal-like (e.g. egg-shaped, black and white) [31]. This is explained by a higher acceptance of the majority of users and a reduction in technical complexity (e.g. number of electric motors) [7], [24]. The quality of the social signals emitted is obviously tied to the technical capabilities (e.g. size, mobility, visual appearance, number and strength of electric engines, degrees-of-freedom of joints etc.) of the robotic embodiment itself.

To date, a large number of social robots have flushed onto the market and disappeared again. Possible causes for a missing widespread commercial success include high prices and the complexity of the hardware, false user expectations, normative beliefs, a lack of helpful and accepted use cases and underdeveloped social skills, which can lead to user rejection rather than acceptance [37], [33], [16]. Consequently, many approaches to assistive systems tend

[1]Christian Felix Purps is with Institute for Intelligent Interaction and Immersive Experience, Karlsruhe University of Applied Sciences, Karlsruhe, Germany christian_felix.purps@h-ka.de

[2]Matthias Wölfel is with Institute for Intelligent Interaction and Immersive Experience, Karlsruhe University of Applied Sciences, Karlsruhe, Germany, and University of Hohenheim, Stuttgart, Germany matthias.woelfel@h-ka.de

to forego embodiment and social behavior entirely [4], [6]. This challenge is recognized by developers of more complex social robots with sufficient perceptual, expressive and speech capabilities that are usually offered commercially as embedded systems. However, more sophisticated commercial systems usually have the disadvantage of not being freely customizable or programmable and therefore limited to the application programming interface (API) provided by the manufacturer [15], [27], [25] or requiring further tools to facilitate their usage [14]. This represents an obstacle for tinkering or necessary adaptations to the respective robot and its features, which tend to often be required in research. Free, non-commercial approaches, on the other hand, have no restrictions of this kind. However, do-it-yourself (DIY) solutions for social robots — due to the complexity of sophisticated systems — are rarely thought of holistically. Mostly, specific aspects of the development of social robots are brought into focus. These include ASI modules/frameworks on the software side, electronic approaches for the development of embodiment and movements, as well as design aspects [11], [17], [38], [22].

Considering the basic characteristics for social robots, there are open-source soft- and hardware solutions for all essential partials which can be sensibly combined to holistically create a new companion robot. By "holistically" we mean that all relevant aspects and components were thought of and integrated as interrelated parts of a whole including embodiment, perceptive and expressive capabilities, and ASI. This involves the use of sensors for visual and acoustic perception and the subsequent processing steps for these signals in terms of speech, emotion and pose/gesture recognition, person localization, etc.. Sensor inputs are fused to generate verbal or non-verbal responses. This was implemented using a dialog system in conjunction with dynamic behavioral rules, that control the robot body and its emitted signals (verbal and body language, facial expressions). As an important aspect of non-verbal communication (NVC), we enabled our robot to express itself emotionally by default. For the software, we wanted to rely exclusively on free and open-source components (e.g. for emotion and speech recognition or tools to facilitate usage of perceptive sensors). Additionally, our system should not be dependent on an internet connection or external services, as this can pose challenges in certain experiment settings or environments. Our aim is to facilitate the replication of our system (partially or as a whole) and thus promote further development and experiments based on it. Our approach shows how it is possible to assemble a simple companion robot that has the necessary perceptive, intelligent and expressive capabilities with relatively simple methods, few assembly parts and low financial outlay (see Fig. 1). The proposed embodiment, consisting of robot arm and display device, can be realized in a variety of different combinations depending on requirements (e.g. size, price) without having a major impact on the rest of the components. The robot's behavioral logic, which is largely based on dialog system, allows the development of new interaction scenarios by simply adapting this subsystem. Following, will describe the basic idea, concept, and hard- and software setup and evaluate and discuss the resulting system and its potential.

II. Related Robotic Systems

There are many approaches to social robot creation, however holistically thought free solutions are rare. Approaches such as NELSON [12] or ROSBOT [13] have been developed to provide cost-effective open solutions for a combination of hardware and software to enable easy setup and replication of the system. However, both systems have different shortcomings in terms of their social capabilities. Despite sophisticated perception units, mobility and its voice interface, the expressive capabilities of ROSBOT are barely addressed. NELSON, on the other hand, has good expression capabilities through embodiment, but is severely limited at the perception level and no longer meets today's standards. The Poppy Project [26] is an open-source platform for the creation, use, and sharing of interactive 3D printed robots. It includes hardware and software components and was used for various applications including research, education, and art. However building a Poppy robot can be time-consuming and requires technical expertise, particularly in 3D printing and assembly. A holistically conceived and applicable option is offered by the FLEXI-Kit [1]. The kit was designed to address the challenges of short lifespans and limited applicability in the social robotics market. It enables unlimited customization of embodiment, making it suitable for a wide range of use cases. It also features an open-source end-user programming interface and provides access to materials and a fabrication tutorial, to make it easily accessible. Despite the many advantages of FLEXI, the effort required for self-assembly remains comparatively high which is something we address in our own approach.

III. The ZENIT Platform

The design of our companion robot was inspired by Pixar's Luxo Jr. and Disney's animation style. Our robot follows suit in its small size, always perceived as a subordinate entity, avoiding the uncanny valley by embracing a deliberately mechanical appearance, yet organic motion and behavior (we created a virtual animated prototype before hardware implementation, see Fig. 2). Affordability and ease of replication are central to our design ethos. We achieved this through the use of few, low-cost, and widely available hardware components and open-source/free software, enabling wider access and adoption.

Studies on social robots emphasize the significance of adapting NVS, like eye contact, facial mimicry, and body posture, based on interaction contexts. These signals enhance the robot's attentiveness and emotional connection with users, leading to improved feelings of social closeness [28]. Robots with contingent NVS are perceived as easier to understand, more trustworthy and comfortable to interact with than those lacking such features [30]. To fulfill these requirements we created a social robot based on a robotic arm and a smartphone horizontally attached to it using a car phone holder. It serves as a "robotic face" running a

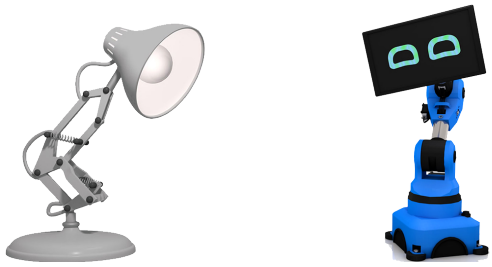

Fig. 2: **Pixar's Luxo Jr., (left) and Virtual ZENIT Prototype (right).** The development of ZENIT was inspired by Luxo Jr.[2], Pixar's virtual anthropomorphic desk lamp character. Before we started tinkering with physical robot components we created a purely virtual prototype.

3D application that displays two abstractly stylized eyes. These eyes can morph into different shapes depicting an abstraction of facial expressions respectively emotions. If the robot performs vocal utterances or speaks it plays a sine wave as a stylized mouth. The display can additionally be used to show other information (videos, text messages etc.). Periphery devices for sensation and diverse software components running in a distributed system are employed to control the robotic arm movements and smartphone display contents. Each hardware unit runs different applications that are interconnected with each other using matching network protocols (see Fig. 4). *ZENIT Face*, *ZENIT Body* and *ZENIT Brain* are the main components of our system[1] and are connected to the wrapped open-source tools employed for perception and speech processing that we have customized and packaged. We employed Node.js to seamlessly network components, essential for e.g. connecting to Python-based AI services and efficiently integrating business logic. For the development of the robotic face and related screen renderings we used Unity (C#) for an easy deployment on mobile devices. Our distributed system encapsulates components with pre-configured network APIs, fostering connectivity for future additions and enables outsourcing of computationally expensive algorithms.

A. Hardware Selection

We explored integrating robotic arms with tablets or smartphones to select suitable hardware. We chose the hardware that best suits our requirements, but any other combination of robotic arm and smartphone should also be possible (e.g. a stronger arm and a tablet computer instead of a smartphone). Our approach enables the use of a wide variety of robotic arms and display devices for the embodiment — a key strength of the concept. Elephant Robotics *MechArm 270-Pi*[3] was selected for its range of motion, degrees-of-freedom

[1]ZENIT's code is open-source and freely available at GitHub (https://github.com/Snagga86/ZENIT).

[2]Pixar's Luxo Jr.: https://en.wikipedia.org/wiki/Luxo_Jr._(character)

[3]MechArm 270-Pi is a lightweight and compact 6-axis robotic arm manufactured by Elephant Robotics using Raspberry Pi as controller, with a payload of 250g, which is sufficient to lift an average mobile phone. (www.shop.elephantrobotics.com/en-de/collections/mecharm/products/mecharm)

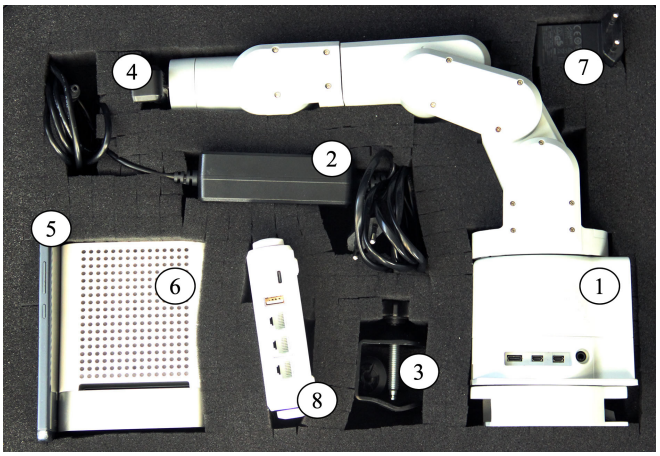

Fig. 3: **Hardware Components of ZENIT** 1) MechArm PI, 2) Power Supply and 3) Table Clamp, 4) Smartphone Car Holder, 5) Smartphone, 6) MS Azure Kinect and 7) Power Supply, 8) Router. Ethernet and USB-C cables not included.

(DOF), payload and price, and mounting options. We argue that five DOF are a decent number for a stationary robot as three of them are required to cover the expression of the most important and powerful head gestures consisting of head nodding, turning and tilting (or combinations of those) [19]. Another one to two DOF would be required for head thrusts or head pulls (two DOF are making the movement look more organic and natural). Speed and acceleration specifications are important for the creation of more organic natural movements for the robot. Accuracy and repeatability do not play a crucial role for our purpose, since small variations from the proposed position can be interpreted as natural behavioral derivatives. Motor noise should be taken into account, as excessively loud motors can greatly reduce the user's enjoyment of interacting with the robot.

For the robot screen face we used a Samsung Galaxy A41 as it offers a relatively large $6.1''$ display and is light weight ($152g$) so robot's arm motion speed or acceleration is not negatively influenced. It also integrates a sufficiently good front camera for computer vision purposes (Full HD at $30fps$). Moreover, we used a Microsoft Azure Kinect to monitor the situation from a different angle utilizing it's depth camera. In order to be able to set up the distributed system with all its components quickly, we also used a preconfigured router. We made sure all components were small and portable (see Fig. 3). As our main computation unit served a regular Laptop (Windows 11, AMD Ryzen 7 5800H, 32GB RAM, NVIDIA GeForce 3070 Laptop GPU).

B. Perceptive System

Our perception system adapts the human senses of vision (camera systems) and hearing (microphones). Vision enables our ASI to interpret the NVS emitted by the conversation partner [42]. In our system, this includes body language (poses and gestures), spatial behavior and facial expressions. Hearing enables our ASI to interpret verbal language and vocal utterances. The fusion and further processing of the

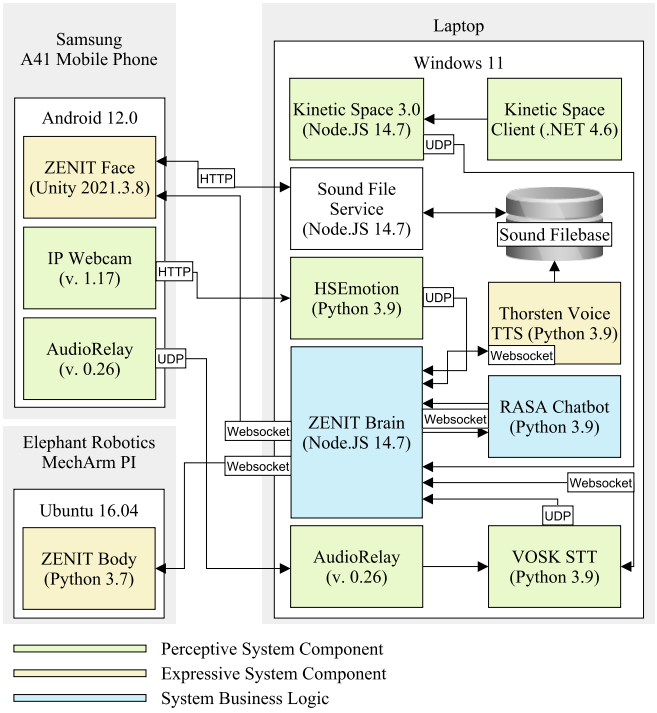

Fig. 4: **ZENIT Distributed System Components.** The architecture includes three hardware devices (robotic arm, smartphone and laptop). Versatile perceptive system components supply the system business logic with recognized NVS. The rule based business logic then may trigger the robot to communicate verbally or non-verbally with its interlocutors.

perceived information takes not place at perception but on business logic level.

1) Body Language and Proximity: Perception of body language and spacial behavior are crucial for the robot to react adequately to human behavior [2], [42]. We used Microsoft's Azure Kinect due to its promising results in recognizing shapes, matching skeletons, and gestures. The Kinect comes with a full *Body Tracking SDK*[4] that allows extraction of joint information in real-time from a depth image. To access data from an Azure Kinect, we use a self-developed middleware that seamlessly integrates the *Kinetic Space*[5] gesture-recognition module. It enables the recognition of gestures from just few examples while the normalization of skeleton data ensures recognition accuracy by removing the influence of individual features, body orientation or localization. In our version, perception supports the recognition of proximity (and is thus able to always focus the user, holding "eye contact") and some basic gestures (e.g. waving, stopping, power pose, requesting, etc.), which we can extend easily. Kinetic Space consists of two applications (C# and Nodes.js) where the C# Client applications enables the Azure Kinect hardware interfacing and detections and the Node.js application fuses the information and routes them to the further processing unit (ZENIT Brain in this case).

2) Facial Expressions: Crucial to NVC is the recognition of facial expressions [10]. To make our robot socially aware of its interlocutors emotions we used the smartphone build-in-camera. We used the Android App *IP Webcam*[6] to establish a webserver on the smartphone that creates camera shots continuously. These shots are requested at $10fps$ by a python service wrapping an instance of *HSEmotions*[7] (model: enet_b0_8_best_afew). The service buffers inputs and streams the recognized emotion (one of the seven basic emotions or neutral) as results to ZENIT Brain.

3) Speech and Vocal Utterances: There is a multitude of free speech recognizers and transcribers available [36]. We prioritized using *VOSK STT*[8] because of its offline language support and low recognition latency. To access the smartphones microphone we used the Android App *AudioRelay*[9]. We created a python service that wraps VOSK and uses a virtual remote microphone established with AudioRelay as speech input. Using VOSK's KaldiRecognizer (model: vosk-model-de-tuda-0.6-900k, 4.4 GB, Tuda-DE Project) we streamed only the final transcription result to the ZENIT Brain for further processing. Additionally our VOSK STT service can be suspended in case listening is inadequate (e.g. while the robot itself is uttering).

C. Expressive System

The embodiment, consisting of a robotic arm and a smartphone, aims to mimic the human way of verbal (speech and vocal utterances) and non-verbal (facial expressions and body language) expression in an abstracted form. In addition to real-time speech synthesis, it must be ensured that the abstract, NVS emitted by the embodiment are also clearly and unambiguously understandable for the user. As there exists a multitude of possible NVS we limited ourselves to enable the robot to express the seven basic facial expressions of emotion (using its eyes and body, see Fig. 5) [10].

1) Head and Face: A user's perception of a robot can be strongly influenced by its facial appearance [8]. A robot's face establishes a robot's identity as a subject and performs together with the body emotional and affective expression. A large percentage of contemporary social robots are equipped

[4]Body Tracking SDK for Azure Kinect enables segmentation of exposed instances and both observed and estimated 3D joints and landmarks for fully articulated, uniquely identified body tracking of skeletons. (http://www.azure.microsoft.com/en-us/services/kinect-dk)

[5]Kinetic Space is a self-developed tool to enable training, analysis, and recognition of individual gestures with a depth camera like Microsoft's Kinect family [41].

[6]IP Webcam is a free universal network camera adapter that can be used with a variety of protocols, cameras with MJPG output or static images (https://ip-webcam.appspot.com/)

[7]HSEmotions (High-Speed face Emotion recognition) is an open source tool that returns a string value of predicted emotions (Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, or Surprise) and scores at the output of the last layer based on an input image (https://github.com/av-savchenko/hsemotion) [34].

[8]VOSK is a versatile speech recognition toolkit offering support for different languages, offline functionality on lightweight devices, streaming API for smooth user experience, speaker identification capabilities etc. (https://alphacephei.com/vosk/).

[9]AudioRelay is software that supports wirelessly streaming audio from an android device to a PC, turning the Android device into a remote microphone for the PC (https://audiorelay.net/).

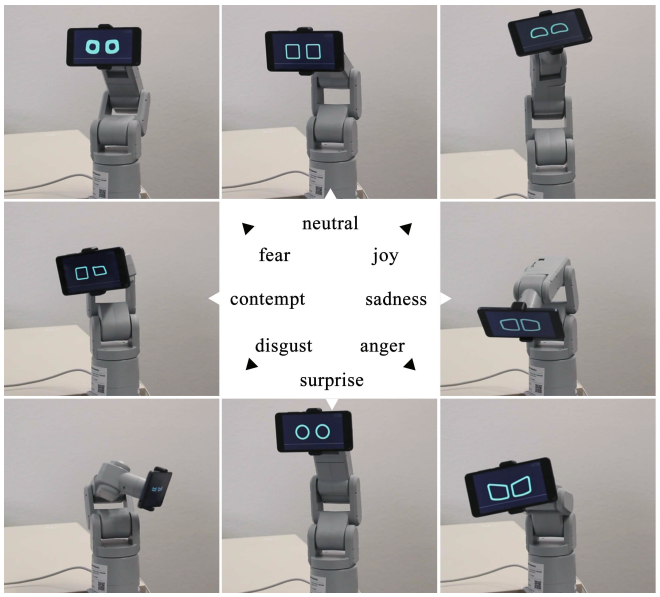

Fig. 5: **Basic emotions provided by ZENIT.** The figures show the combination of facial expressions and bodily movements to express the respective emotion.

with screen faces [21]. Many come with a face only consisting of eyes without displaying a mouth. There is indication that screen faces without mouths are perceived as more mature, yet less human-like and intelligent [31]. Eventually, however, these are design decisions and also characterized by individual preferences. We opted for a representation without a mouth and instead animated a sine wave while the robot is uttering. The usage of classical blendshapes enable an easy addition of new facial expressions. ZENIT Face receives control commands from ZENIT Brain. These include morphing the eyes (7 emotions), displaying further information (text, videos) and playing audio. Audio information is requested via a *Sound File Service* component and then streamed.

2) Body: MechArm 270-PI[2] with its 6 DOF and concealed cabling contributes to our robot's clean design to creates a coherent overall picture. It employs a Raspberry PI 4B running Linux (Ubuntu 14) as its control unit. It runs a Python application (ZENIT Body), which receives control information from ZENIT Brain. This includes the playback of manually preconfigured arm movements (e.g. for 7 emotions, incl. speed adjustments among others), as well as rotation data for focusing the user.

3) Speech: To give our robot speech capabilities we used *Thorsten Voice TTS*[10] (model: Neutral TTS) as our target group is german native speakers. Wrapping pythons TTS module, our service receives text from ZENIT Brain to synthesize. Generated audio data is saved (if not yet available) and can then be streamed by ZENIT Face via our Sound File Service. This procedure allows us to reduce latency, as once generated audio does not have to be recreated.

[10]Thorsten Voice TTS is a high quality (AI), German, artificial TTS/text-to-speech voice that can be generated offline. (http://www.thorsten-voice.de).

D. Decision-making System

The core ASI component is the ZENIT Brain in interaction with a dialog system. This is where the various threads of the perceptual and expressive systems are brought together and behavioral decisions are made for verbal and non-verbal expressions. We used *RASA Chatbot*[11] to enable guided conversations. The range of interaction capabilities of our prototype can therefore be extended simply by adapting the underlying RASA chatbot. Depending on the behavioural mode RASA receives transcribed speech as input, recognizes an intent and generates a text response and if applicable also a suitable NVS. Depending on the context, the results are then forwarded to a behavioral state machine and expressed via the expression system. The underlying behavioral system (ZENIT Brain) also includes post-processing steps for recognized gestures or emotions (e.g. to avoid false triggering of the expressive system) and was architecturally solved with the state pattern. This is necessary because RASA implementations are impractical when many context changes occur. Our state system makes it possible to use e.g. separate complex state machines for system sections in different contexts. In addition, the dialog system must sometimes be suspended (e.g. to restrict the escape from certain program sections). Furthermore, RASA is not created to deal with triggers such as timers, emotions or gestures or to organize the sequential tracking of states. The combination of our own state machine solution in conjunction with RASA therefore allows us the necessary flexibility to freely develop applications for the robot.

IV. Technical Evaluation

We evaluated our system considering the most important aspects which are user comprehension of the emitted NVS, latency and system reliability. For performance metrics of the employed 3rd party hard- and software the respective literature is to be considered.

A. Emotion Expression

A major feature of our robot is the capability to express an array of emotions (seven basic emotions and neutral) by default. It can do so either through bodily movements, facial expressions or both combined. The facial expressions and bodily movements created are naturally influenced by the ideas of the 3D and motion designer and thus have to be evaluated considering their unambiguity/distinguishability and expressiveness. Additionally we were interested how clearly the emotions can be conveyed through body language, facial expressions or using a combination of both.

1) Participants, Questionnaire & Procedure: To evaluate the distinguishability and expressiveness of the robotic emotions we drove a within-subject online video survey. Although the experience of interacting with a real robot is

[11]Rasa is an open-source framework that enables building advanced chatbots and conversational assistants capable of understanding user intents, handling natural language processing, and maintaining context throughout conversations, empowering more natural and contextually relevant interactions. (http://www.rasa.com).

obviously different from watching a robot in a video, it can be assumed that video testing is a valid method for our purposes [40], [3]. A total of 23 participants completed a valid questionnaire. Participants had an average age of 34 years (23 youngest, 69 oldest) and belonged to the Western (German) culture. 11 participants considered themselves male, 11 female (1 did not answer). 7 participants never use AI-Assistance, while 16 use it frequently. 13 participants had at least once interacted with a social robot before.

In the online questionnaire, participants had to answer demographic questions and were shown 22 different videos. For each individual our robot expressed emotions in a randomly generated sequence. Each video showed the robot performing one emotion out of 7 (joy, sadness, anger, surprise, disgust, contempt, fear) in a loop, starting and ending it's emotional expression in a neutral position. Each participant saw these emotions for each condition (only body movements, only facial expressions, both together) and neutral (8th emotion, only displayed once, as it is identical for all conditions). Participants could watch the video loop for as long as they wished. For each video, participants had to answer two questions. First, we asked participants "What emotion is the robot trying to communicate to you?" to measure the *distinguishability* of each emotion. Possible answers had to be given using a 5-point Likert scale consisting of 8 (7+1) items (joy, sadness, anger, surprise, disgust, contempt, fear, and neutral) ranging from "not likely" to "very likely". Secondly, we asked participants "How do you rate the robot's behavior?" to measure the level of perceived *expressiveness* of each emotion. Possible answers had to be given using a 5-point Likert scale consisting of 4 self-developed questions ("The behavior shown was expressive.", "The behavior shown touched me emotionally.", "The behavior shown was evidence of strong feelings.", "The behavior shown was clearly different from inactivity.") ranging from "does not apply" to "fully applies". Reliability analysis to assess the internal consistency for emotional expressiveness yielded an adequate result ($\alpha = 0.87$).

2) Distinguishability: We aimed to make the emotions conveyed by the robot easy and unambigous to interpret by its interlocutor. The following formula was used to easily illustrate the distinguishability of the emotions shown to the participants:

$$\text{dist}(e_x, \vec{e}_o) = \frac{7(e_x - 1) - \sum_{i=1}^{i=7}(e_{oi} - 1)}{28}$$

e_x is the given score (5-point Likert scale) to the emotion displayed we expected the participants to interpret. The vector $\vec{e}_o$ consists of the scores for all other 7 emotions we expected the participants not to interpret. Through normalization the formula outputs values between 1 and -1. A result of $\text{dist}(e_x, \vec{e}_o) = 1$ means that the emotion we expected a participant to interpret was rated 5, while all others were rated 1, and thus the theoretically best possible result. A result of $\text{dist}(e_x, \vec{e}_o) = 0$ means that the intended emotion was rated the same as all other emotions and is therefore indistinguishable. A result of $\text{dist}(e_x, \vec{e}_o) < 0$ indicates a tendency for the emotion to be misinterpreted and another, unintended emotion to be given greater consideration. The distinguishability was calculated for every sample of every emotion and condition taken. The aritmetic mean values ($N = 23$) can be seen in Tab. I. We also conducted a one-way ANOVA to analyze the impact of expression performance on distinguishability with significance level set at $p < .05$. There was significant effect of expression performance on distinguishability ($F = 22.43$, $p < .001$) with a medium effect size ($\eta^2 = .082$). Post-hoc analyses indicated that emotions performed only using bodily movements were significantly less distinguishable from each other than emotions performed using facial expressions ($p < .001$) or using both facial & bodily movements ($p < .001$).

TABLE I: **Distinguishability of 7+1 emotions for each performance type.** Green shows a high, yellow a low and orange a contradicting distinguishability.

	Only Body	Only Face	Face & Body
Neutral	0.75	0.75	0.75
Joy	0.54	0.41	0.67
Sadness	0.55	0.63	0.76
Anger	-0.15	0.74	0.66
Surprise	0.25	0.46	0.67
Disgust	0.16	0.11	0.41
Contempt	-0.13	0.13	0.06
Fear	-0.04	0.00	0.03

3) Expressiveness: We hypothized that a fusing of facial expressions and bodily movements yields higher levels of perceived emotional expressiveness compared to only face or bodily movements. A one-way ANOVA was conducted to analyze the perception of expressiveness based on the expression performance (only face, only body, face & body), with significance level set at $p < .05$. A significant effect of expression performance on expressiveness ($F = 48.40$, $p < .001$) was revealed (see Fig. 6). Post-hoc analyses further elucidated the nature of these differences, indicating that emotions performed only using bodily movements were perceived significantly less expressive than emotions performed only using facial expressions ($p < .007$) and emotions performed using facial & bodily movements ($p < .001$). Furthermore, emotions performed using face & body were perceived significantly more expressive than emotions performed only by facial expressions ($p < .001$). A large effect size was indicated by $\eta^2 = .161$, suggesting that the expression performance accounts for a substantial proportion of the variability in perceived expressiveness of an emotion.

B. Latency

As in each human-computer interaction task, system response time is crucial for a functioning interaction with the robot and should not exceed $300ms$ to make users feel like they are in a closed-loop system [9]; a challenging task, especially in sequential speech processing. To evaluate latency of our system we used a $120fps$ high speed camera to film the robot and calculated the time delta between human input, system recognition and robot expression using

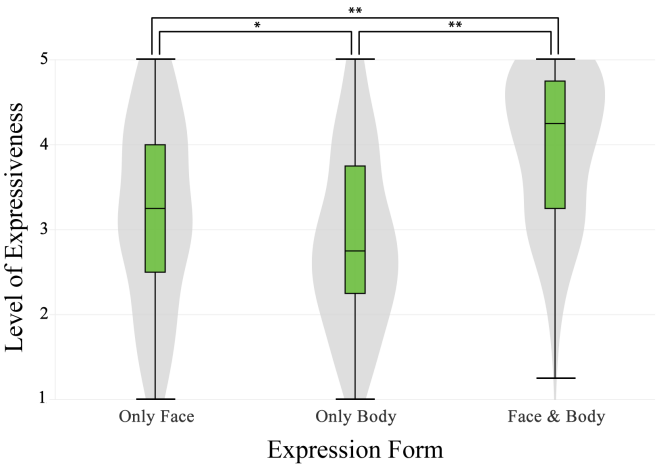

Fig. 6: **Perceived expressiveness of emotional performances across expression modalities**

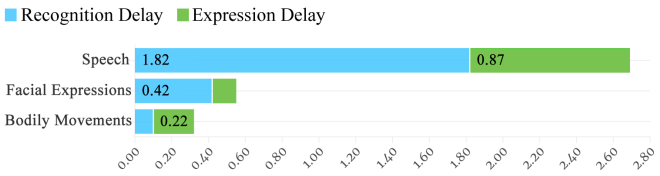

Fig. 7: **Latency (sec.) of input and output modalities.**

the average of 20 samples recorded for each speech, facial expression and bodily movement input randomly selected from interactions with the robot during a simple small talk scenario (see Fig. 7).

C. Stability and Reliability

We tested the basic functionality of our hard- and software setup in a real-world setting under challenging conditions in public space where passerby had the chance to interact with the robot (facial recognition features were not included due to the use-case). At the two day event, our system was running for 8-9 hours per day with a total of 450 passerby encounters taking place [29]. We noted about 3-4 smartphone crashes during the event. At least one was due to overheating, for the rest, memory leaks must be considered. Sometimes the robotic arm did not execute a scheduled movement. We investigated that problem and figured out, that the robotic arm did not perform movements in about 3%-4% of the cases. The issue seems to occur when several movement commands are overwritten in quick succession and could be due to the low baud rate of the robot's serial interface. All other components ran smoothly for the duration of the event.

V. Discussion, Conclusion and Future Work

We developed a new companion robot (ZENIT) based on a small number of low-cost components: A robotic arm with an attached smartphone as embodiment, an optical sensor for perception, and a common computer to run AI and business logic based on freely available software. The robot has perceptual and expressive capabilities for speech and NVS. Our hard- and software architecture as a distributed system allows the system to be easily expanded to include new languages, gestures/poses (Kinetic Space) and behavioral logic (RASA). For our base solution, we gave our robot the ability to express seven basic emotions through its screen face and robotic arm.

We evaluated our robot in terms of the distinguishability and expressiveness of displayed emotions, as well as system latency, stability and reliability. The results of our user study show that contempt and fear were not clearly recognizable for the users. However, expressed through facial expressions or a combination of facial expressions and body movements emotions were significantly better recognized than through body movements alone. Investigating the expressiveness of emotions showed a significant difference between all forms of presentation (body only, face only, face & body), with the interplay of body and face being perceived as the most expressive. It cannot be generalized, but for our system it can be concluded that the extension of facial expressions by body movements does not lead to better recognizability, but if the emotion is recognized, its expressive power is significantly increased. This immediately raises the question of how body movements, which stand in contrast to facial expressions, are perceived in interaction. Further studies could aim to find the critical number of DOF at which the movement is still perceived as expressive but also natural. We are also planning field tests with various interaction scenarios to investigate the likability, trust and competence of the robot.

We found that our system's latency exceeds recommended limits, particularly for speech interactions which should be under $300ms$. Our latency was about $2.69s$, far too long for smooth interaction. While expert users can cope with this (as long as the robot indicates that it is still listening), this latency is problematic for less technically experienced users. Therefore, reducing latency remains a major focus for improvement. This is particularly important because we aim to supplement the natural language understanding system used with a large language model in which the text results are unpredictable. In addition, voice-based interruptions of the talking robot, could enable a more human-like communication in future implementations. We examined the system stability and reliability in a 2-day use case. The errors of smartphone crashes and unexecuted robot arm movements that occurred are secondary for short research deployments of the robot, but must be solved for use in potential long-term studies. Overall the system shows a decent reliability. ZENIT offers a platform for future HRI research with full control over all subprograms and hardware. We see potential applications for example in assisted living (reminder functions, activation, empowerment), as well as for telepresence and entertainment, but above all in the field of research.

Today's social robots are still in their infancy. Progressive developments in the field of AI (especially computer vision and speech processing) could accelerate the development of acceptable systems. There is an urgent need to improve the social skills of these intelligent machines. Eventually though, it may be individual preferences that decide whether and if so which social robot someone wants to use or not.

References

[1] Alves-Oliveira, P., Bavier, M., Malandkar, S., Eldridge, R., Sayigh, J., Björling, E.A., Cakmak, M.: Flexi: A robust and flexible social robot embodiment kit. In: Designing Interactive Systems Conference. pp. 1177–1191 (2022)

[2] Argyle, M.: Bodily communication. Routledge (2013)

[3] Bahadori, S., Cesta, A., Grisetti, G., Iocchi, L., Leone, R.G., Nardi, D., Oddi, A., Pecora, F., Rasconi, R.: Robocare : an integrated robotic system for the domestic care of the elderly (2003)

[4] Barchard, K.A., Lapping-Carr, L., Westfall, R.S., Fink-Armold, A., Banisetty, S.B., Feil-Seifer, D.: Measuring the perceived social intelligence of robots. ACM Transactions on Human-Robot Interaction (THRI) **9**(4), 1–29 (2020)

[5] Dautenhahn, K., Woods, S., Kaouri, C., Walters, M., Koay, K.L., Werry, I.: What is a robot companion - friend, assistant or butler? In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1192–1197 (2005)

[6] De Graaf, M.M., Allouch, S.B., Klamer, T.: Sharing a life with harvey: Exploring the acceptance of and relationship-building with a social robot. Computers in human behavior **43**, 1–14 (2015)

[7] Dereshev, D., Kirk, D.: Form, Function and Etiquette–Potential Users' Perspectives on Social Domestic Robots. Multimodal Technologies and Interaction **1**(2) (2017)

[8] DiSalvo, C.F., Gemperle, F., Forlizzi, J., Kiesler, S.: All robots are not created equal: the design and perception of humanoid robot heads. In: Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques. pp. 321–326 (2002)

[9] Doherty, R.A., Sorenson, P.: Keeping users in the flow: mapping system responsiveness with user experience. Procedia Manufacturing **3**, 4384–4391 (2015)

[10] Ekman, P., Oster, H.: Facial expressions of emotion. Annual review of psychology **30**(1), 527–554 (1979)

[11] Elfaki, A.O., Abduljabbar, M., Ali, L., Alnajjar, F., Mehiar, D., Marei, A.M., Alhmiedat, T., Al-Jumaily, A.: Revolutionizing social robotics: A cloud-based framework for enhancing the intelligence and autonomy of social robots. Robotics **12**(2), 48 (2023)

[12] Ferguson, M., Webb, N., Strzalkowski, T.: Nelson: a low-cost social robot for research and education. In: Proceedings of the 42nd ACM technical symposium on Computer science education. pp. 225–230 (2011)

[13] Fu, G., Zhang, X.: Rosbot: A low-cost autonomous social robot. In: 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM). pp. 1789–1794. IEEE (2015)

[14] Ganal, E., Siol, L., Lugrin, B.: Peput: A unity toolkit for the social robot pepper. In: 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). pp. 1012–1019. IEEE (2023)

[15] Glauser, R., Holm, J., Bender, M., Bürkle, T.: How can social robot use cases in healthcare be pushed-with an interoperable programming interface. BMC Medical Informatics and Decision Making **23**(1), 118 (2023)

[16] de Graaf, M.M., Ben Allouch, S., Van Dijk, J.A.: Why would i use this in my home? a model of domestic social robot acceptance. Human–Computer Interaction **34**(2), 115–173 (2019)

[17] Hayosh, D., Liu, X., Lee, K.: Woody: low-cost, open-source humanoid torso robot. In: 2020 17th International Conference on Ubiquitous Robots (UR). pp. 247–252. IEEE (2020)

[18] Hegel, F., Lohse, M., Swadzba, A., Wachsmuth, S., Rohlfing, K., Wrede, B.: Classes of applications for social robots: a user study. In: RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication. pp. 938–943. IEEE (2007)

[19] Heylen, D.: Challenges ahead: Head movements and other social acts in conversations. Virtual Social Agents pp. 45–52 (2005)

[20] Huang, D., Chen, Q., Huang, J., Kong, S., Li, Z.: Customer-robot interactions: Understanding customer experience with service robots. International Journal of Hospitality Management **99** (2021)

[21] Kalegina, A., Schroeder, G., Allchin, A., Berlin, K., Cakmak, M.: Characterizing the design space of rendered robot faces. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction. pp. 96–104 (2018)

[22] Kang, D., Kim, S., Kwak, S.S.: Social human-robot interaction design toolkit. In: HRI 18: Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction (2018)

[23] Kappas, A., Stower, R., Vanman, E.J.: Communicating with robots: What we do wrong and what we do right in artificial social intelligence, and what we need to do better. pp. 233–254. Springer (2020)

[24] Lacey, C., Caudwell, C.B.: The robotic archetype: Character animation and social robotics. In: Social Robotics: 10th International Conference, ICSR 2018, Qingdao, China, November 28-30, 2018, Proceedings 10. pp. 25–34. Springer (2018)

[25] Lambert, A., Norouzi, N., Bruder, G., Welch, G.: A systematic review of ten years of research on human interaction with social robots. International Journal of Human–Computer Interaction **36**(19), 1804–1817 (2020)

[26] Lapeyre, M., Rouanet, P., Grizou, J., Nguyen, S., Depraetre, F., Le Falher, A., Oudeyer, P.Y.: Poppy project: open-source fabrication of 3d printed humanoid robot for science, education and art. In: Digital Intelligence 2014. p. 6 (2014)

[27] Mubin, O., Ahmad, M.I., Kaur, S., Shi, W., Khan, A.: Social robots in public spaces: A meta-review. In: Social Robotics: 10th International Conference, ICSR 2018, Qingdao, China, November 28-30, 2018, Proceedings 10. pp. 213–220. Springer (2018)

[28] Park, H.W., Gelsomini, M., Lee, J.J., Breazeal, C.: Telling Stories to Robots: The Effect of Backchanneling on a Child's Storytelling. In: 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI. pp. 100–108 (Mar 2017)

[29] Purps, C.F., Hettmann, W., Zylowski, T., Sautchuk-Patrício, N., Hepperle, D., Wölfel, M.: Exploring perception and preference in public human-agent interaction: Virtual human vs. social robot. In: International Conference on ArtsIT, Interactivity and Game Creation. pp. 342–358. Springer (2023)

[30] Rae, I., Takayama, L., Mutlu, B.: In-body experiences: embodiment, control, and trust in robot-mediated communication. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1921–1930. CHI '13, Association for Computing Machinery, New York, NY, USA (2013)

[31] Ramírez, V., Deuff, D., Indurkhya, X., Venture, G.: Design Space Survey on Social Robotics in the Market. Journal of Intelligent & Robotic Systems **105**(2), 25 (2022)

[32] Riether, N., Hegel, F., Wrede, B., Horstmann, G.: Social facilitation with social robots? In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction. pp. 41–48. ACM, Boston Massachusetts USA (2012)

[33] Šabanović, S.: Robots in society, society in robots: Mutual shaping of society and technology as a framework for social robot design. International Journal of Social Robotics **2**(4), 439–450 (2010)

[34] Savchenko, A.: Facial expression recognition with adaptive frame rate based on multiple testing correction. In: International Conference on Machine Learning. pp. 30119–30129. PMLR (2023)

[35] Schulz von Thun, F., Ruppel, J., Stratmann, R.: Miteinander reden (2014)

[36] Tivarekar, R.P., Khadye, R.M., Chavande, S.R., Talkatkar, P.S.: Review of deep speech recognizer using transcriber. In: 2023 6th International Conference on Advances in Science and Technology (ICAST). pp. 460–463. IEEE (2023)

[37] Tulli, S., Ambrossio, D.A., Najjar, A., Lera, F.J.R.: Great expectations & aborted business initiatives: The paradox of social robot between research and industry. In: BNAIC/BENELEARN. pp. 1–10 (2019)

[38] Vandevelde, C., Wyffels, F., Vanderborght, B., Saldien, J.: Do-it-yourself design for social robots: An open-source hardware platform to encourage innovation. IEEE Robotics & Automation Magazine **24**(1), 86–94 (2017)

[39] Woo, H., LeTendre, G.K., Pham-Shouse, T., Xiong, Y.: The use of social robots in classrooms: A review of field-based studies. Educational Research Review **33**, 100388 (2021)

[40] Woods, S., Walters, M., Koay, K.L., Dautenhahn, K.: Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In: 9th IEEE International Workshop on Advanced Motion Control. pp. 750–755. IEEE (2006)

[41] Wölfel, M.: Kinetic Space – 3D Gestenerkennung für Dich und Mich. Konturen **32** (2012)

[42] Wölfel, M., Purps, C.F., Percifull, N.: Enabling embodied conversational agents to respond to nonverbal behavior of the communication partner. In: International Conference on Human-Computer Interaction. pp. 591–604. Springer (2022)