Emotion-awareness for intelligent Vehicle Assistants: a research agenda

Hans-Jörg Vögel, Christian Süß, Thomas Hubregtsen, Viviane Ghaderi, Ronee Chadowitz
BMW Group Research
hans-joerg.voegel@bmwgroup.com

Elisabeth André, Nicholas Cummins, Björn Schuller
Universität Augsburg
andre@informatik.uni-augsburg.de

Jérôme Härri, Raphaël Troncy, Benoit Huet, Melek Önen, Adlen Ksentini
Eurecom
jerome.haerri@eurecom.fr

Jörg Conradt
Technische Universität München
conradt@tum.de

Asaf Adi, Alexander Zadorojniy
IBM Haifa Research Lab
adi@il.ibm.com

Jacques Terken
Technische Universiteit Eindhoven
j.m.b.terken@tue.nl

Jonas Beskow
KTH Royal Institute of Technology
beskow@kth.se

Ann Morrison
University of Southern Queensland
ann.morrison@usq.edu.au

Kynan Eng
iniVation
kynan.eng@inivation.com

Florian Eyben
audEERING
fe@audeering.com

Samer Al Moubayed
Furhat Robotics
samer@furhatrobotics.com

Susanne Müller
Coaching & Consulting
kontakt@susamueller.de

Figure 1. Emotion awareness: crucial for vehicle interaction with passengers and road users.

Abstract

EVA[1] is describing a new class of emotion-aware autonomous systems delivering intelligent personal assistant functionalities. EVA requires a multi-disciplinary approach, combining a number of critical building blocks into a cybernetics systems/software architecture: emotion aware systems and algorithms, multimodal interaction design, cognitive modelling, decision making and recommender systems, emotion sensing as feedback for learning, and distributed (edge) computing delivering cognitive services.

Affective Computing **•Emotion Awareness**; Sensing and Sensor Fusion; Emotion Recognition/Analysis **•Cognitive Systems**; Cybernetic Architecture; User Models; Recommendations **•Human Factors**; Interaction Design; Human Mind; Ethics; **•Privacy**

Keywords Intelligent Assistants, Emotion Awareness, Multi-Modal Interaction Design, Emotional State Analysis, Neuromorphic Emotion Sensing, Cognitive Models and Proactive Recommendations, IoT, Privacy Preserving Machine Learning

1 Introduction

Intelligent personal assistants (aka virtual assistants), smart robots and conversational bots are rapidly finding their way into our daily lives: they control our smart homes, they answer our questions, and they assist us with shopping or finding local

[1]EVA – Emotion-aware Vehicle Assistant.

SEFAIAS'18, May 28, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5739-5/18/05…$15.00
https://doi.org/10.1145/3194085.3194094

services. Future generations of these devices are expected to go beyond simply being a multi-use remote control activated via voice commands. Intelligent devices should be able to learn from our behavior, deduce our preferences and intentions, and based on this knowledge, make decisions and interact with us in a natural manner. Extending this functionality to the car will allow for an optimal driving experience, since hands-free, natural interaction allows for entertainment and productivity without compromising safe vehicle operation.

Some of today's cars already provide more intuitive voice interaction and gesture commands, and visionary prototype transportation devices display advanced capabilities to detect passengers' emotions and intents during guided conversations [1,2]. As depicted in Figure 1, future vehicle interiors are foreseen to become much more spacious than today's, making it on the one hand impossible to apply current control paradigms focused on direct haptic interaction. Quite plainly, this may not be possible any more, as buttons and touchscreens will not be conveniently reachable from a respective seating position. With higher degrees of automation, it may also not be necessary any more, as vehicles drive in fully autonomous mode. What will be of much higher relevance then is that the vehicle displays empathic capabilities to ensure passengers' wellbeing. Also depicted in Figure 1 is the need for these autonomous vehicles to communicate to and interact with vulnerable road users outside the vehicle. Although the picture being an artistic interpretation of this need, it clearly illustrates the difficulties addressing it. As in the vehicle's interior, also on the exterior a multi-modal approach will be necessary for interacting with other road users – just that here, the possibilities are much more limited, due to traffic regulations, environmental conditions, user's attention spans and emission restrictions. Whether visual information projected from the vehicle's exterior parts is preferable for this purpose remains to be investigated.

However, before achieving truly natural, intuitive interaction, intelligent systems have to overcome several challenges. First, they have to learn to understand the different aspects and subtleties of human communication (non-verbal cues from gestures, facial expressions, tonality, gaze, etc.) and to be aware of the user's current emotional state.

The intelligent system has to be able to understand context, in order to react to a simple question such as "What was that?" This question, simple as it may seem at first glance, bears the contextual challenge in a nutshell. It may refer to a number of events or real-world entities, such as objects, sounds, movement, or sights. It may refer to the immediate context of the car and the activities within (such as unusual sounds, flashing lights, warnings, etc.). It may refer to its vicinity and the traffic situation unfolding around it (such as another car driving by, unusual behavior of other road users, or just the rare pothole). Or it may yet refer further to the greater environment such as buildings, billboards, or spectacular sights. To answer this question, humans usually intuitively infer from a very rich context.

This context has to be extracted from relevant environmental information (sights and objects, sound, car sensor input, intonation voice and sentiment of language, direction of gaze, and gestures amongst others) as well as known and learnt information about the subject's experience, knowledge, usage history, and from other intelligent devices.

To increase relevance, the intelligent assistant needs to become emotionally aware und understand whether the question was uttered in a frightened, curious, or annoyed tonality. Beyond generating answers based on deduced knowledge, maintaining a conversational context and verbalizing actions, the interaction between users and assistant becomes most challenging when there is no verbalization or when verbalization alone just is not sufficient to generate an appropriate reaction. Rightly assessing a potential intent and making recommendations or proactively automating wellbeing by sensing body functions and reading body language beyond pure voice analysis are key functionalities. Hence, EVA will be of prime importance for cars driving fully autonomously, as they might transport passengers who do not hold a driving license, much less understand the least about cars' technology.

2 Challenges and Focus Areas

Pursuing intelligent, emotion-aware systems requires an interdisciplinary approach, by bringing together IoT technologies and protocols with multi-modal user interaction, human-centric emotion-aware design, knowledge base modeling, machine learning and intelligent recommender algorithms, as well as innovative approaches to privacy. Emotion-aware Vehicle Assistants – EVA – will require to develop mechanisms sensing user and context in real-time to facilitate reasoning and decision-making for proactive, emotionally aware intelligent personal assistants. EVA will draw specific capabilities from the in-vehicle situation with all its sensory input.

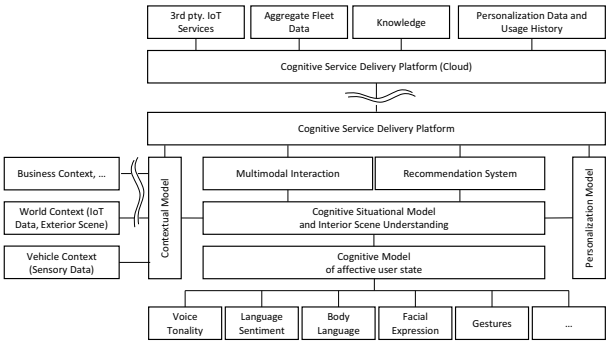

Figure 2. Systems architecture.

The EVA assistant approach defines a multi-layered cognitive systems architecture (Figure 2). Note that it focussed on the vehicle interior scene and interaction with users, leaving out aspects of interaction with exterior users for the sake of simplicity. At its core is the cognitive model of the car users' cognitive state. This model is being fed by an array of ingestion sensors, each sensing aspects of the users' emotional state, such as voice tonality, language sentiment, body language, facial expression, and gestures, among others. These individual sensor feeds need to be fused into a consolidated cognitive model of the users' affective state. Together with contextual information about

vehicle state and events, a world model of the exterior scene and additional IoT data streams available for interpreting it, and other contextual data, the system will be able to fuse a cognitive model for the understanding of the interior scene. Based on this understanding the vehicle will interact with its users through all available modes of interaction, or proactively provide recommendations for the delivery of cognitive services. Affective state information is used also in a feedback loop for the learning mechanisms in the personalization model behind the recommendation engine.

Natural interaction and emotion awareness need to be realized by leveraging sensor semantics and processing them with advanced machine intelligence algorithms capable of interpreting the sensor data in real-time combined with background knowledge. Machine learning technologies are required to monitor user behavior, discriminate relevant patterns and treating the learnt knowledge with appropriate privacy mechanisms. Privacy preserving machine learning will have to be further developed to accommodate the diverse set of data sources and usage scenarios. For sheer volume, data will have to be treated in a federated learning approach. Combining local per-vehicle onboard learning with centralized cloud-based learning both at the user, vehicle, or fleet level will require distributed (edge) processing and analytics approaches.

Overall, the system will need to act – and interact with the passenger – in a way such as to provide full transparency about what it is doing and why it is doing it. Thereby, passengers can grow sufficient trust to rely on the assistant confidently. The situation of (multiple) passengers sharing rides in autonomous vehicles require particular attention.

The research challenge EVA poses is towards emotional awareness in autonomous systems with particular focus on in-vehicle applications. The ultimate vision is an autonomous car, complemented with an intelligent assistant that naturally interacts with the user. It should have affective skills for emotional awareness and personalized pro-active behavior. It has to be capable of extending autonomously, based on a rich situative context analysis of vehicular sensor and world information from any IoT- and social media data stream. This requires a highly interdisciplinary approach. The vehicular domain is providing innovative applications and non-functional requirements driving novel approaches such as hybrid architectures with cognitive edge computing. We divide this research field into the categories briefly described hereafter.

2.1 Emotion in Multi-Modal Interaction Design

EVA should provide applications that drivers/occupants will integrate into their everyday driving context (both for manual and autonomous driving). To identify factors that govern this integration and adoption of emotion-aware assistants, methods need to be enhanced to collect user opinions in early stages of the system and software design process, building on the existing Co-Constructing Stories method and related methods. Models predicting the adoption by target users will have to be developed and validated through evidence-based frameworks. These models direct the iterative development of actual applications.

EVA will learn from the interactions with users and adapt to them in real-time. It will exploit dialogue management and behavior generation based on reinforcement learning. Assistive dialogue systems have already employed such techniques. [12] However, in many scenarios the rapidly increasing state space makes a straightforward learning approach computationally impractical. Studies show that a permanent stream of requests upsets users and distracts them from their current activities. Hence, novel dialog system design will have to explore techniques for policy optimization based on deep reinforcement learning, taking into account user's emotions as implicit reinforcement signals that help the assistant learn users' preferences and enable it to adapt the dialog's content, form and modalities to the user's current needs.

Novel system design approaches are required with respect to active learning, which – in the context of data annotation – is to select from large pools of unlabeled data those instances that are the most informative ones for the task being modeled, and subsequently query a human or machine annotator for labeling. Previous research focused on desktop-based settings processing previously recorded data offline. Further work is to investigate how to make use of active learning in an interactive setting with the automotive assistant. The EVA assistant has to decide when it is a good time to interrupt the user and afterwards to update the model on the fly. Furthermore, users do not like to serve as a pure data provider. Hence, the EVA approach should develop probabilistic methods to minimize the costs for collecting data, visualization and explanation strategies that make the inner workings of the machine learning process more transparent, and federate machine-learning techniques that keep all training data on the user's private device.

Smart interaction techniques need to be devised that allow for rich, multi-modal interaction between the user and the system and that allow easy switching between the focus and periphery of attention, within the constraints imposed by the context. While (embodied) conversational agents have long been considered a promising ally for the vehicular context, from an interaction perspective such conversational agents raise serious concerns along a number of dimensions (e.g., social, ease of use, obtrusiveness), so that extending the concept for EVA must not narrowly focus on conversational agents but draw from multiple modalities for exchanging information. A model integrating task, context and user characteristics is established, predicting the suitability of different interaction techniques and giving direction to user interface design for different applications.

2.2 Sensor Technologies and Emotion Sensing

Sensing human body posture and activity, emotion and cognitive state in vehicles poses a number of complex challenges, such as sensor arrangement in limited space, unclear lightning conditions, and required separation between body and unknown distractor objects, noisy sensing environments, high-dynamic range of lighting conditions and restrictions on active lighting due to driver distraction reasons. Moreover, to effectively sense emotional context and infer emotional states, one needs to observe human signals and sense fast human activity (such as pointing or waving

gestures) with a high spatial and temporal resolution. This is especially true for gestures and mimics, as very casual and transient signals greatly contribute to the understanding of the actual emotional context of a situation [8, 9].

Here, novel event-based vision sensors provide multiple advantages over traditional camera technology for emotional state sensing and for high-speed human position and activity tracking in a driving as well as a passenger context. In particular, high temporal responsiveness, drastically reduced data rates in typical car-indoor scenes and insensitivity over large range of illumination display properties relevant to the challenging vehicular environment. Combined with novel algorithms for robust tracking of facial features from event-based real-time data, and body and limb poses in the constrained car-indoor environment this will yield descriptions of emotional states. These novel event-based neuromorphic algorithms also pose new challenges to software architectures and the signal processing chains.

Interactive learning approaches and non-invasive (in seating, in vehicle interior) monitoring will allow the EVA assistant to learn from positional and embodied interactions with which end-users adapt to situations. It will learn what individuals like and dislike, and adapt content, form and modalities as information garnered from the embodied, positional and gestural information and address the end user's continually changing needs and preferences. The situational context will take into account prior knowledge, purpose of use, gender preferences and will log use and adapt accordingly.

Current state-of-the-art automatic emotion detection systems can achieve near human level of performance when predicting emotions. However, such systems generally rely on resource-heavy deep learning [10, 11]. The convolutional and recurrent neural network topologies generally used in contemporary emotion detection systems can have connection numbers measuring in the billions, and require sizable amounts of memory and energy. While there have been some research efforts into designing emotion detection system capable of running natively in smart or embedded devices [13], the vast majority of research into emotion detection only consider performance metrics relating to system accuracy. One such approach to reduce runtime, and computational resources in emotion detection is to explore spiking-based neuromorphic systems, both at the sensor [14] and analytic [15] phases of emotion detection. Despite offering the potential of being faster, more robust to noise, and more computationally efficient when compared to conventional methods, neuromorphic systems have yet to be realized for emotion detection.

2.3 Emotion-aware Cognitive Systems

The impact of drivers' emotional states on safety and joy are well known. What is less known is how to take information from all type of available sensors, such as biofeedback (e.g., HR, HRV, GSR), visual (e.g., facial images), and car-specific (e.g., steering wheel, acceleration) sensors and use them to design a precise, person-centric controlling system. We consider three main phases in the research: data collection, labeling, results assessment, and metric development; designing a recommendation system with limited data and sensors; and finally, designing a large-scale recommendation system using state-of-the-art machine learning and control algorithms.

The customer acceptance of future automated vehicles will strongly depend on vehicles to drive considering drivers' or passengers' emotions. Early control mechanisms have been mostly studied to reach a so-called 'string stability' between Automated Cruise Controlled ACC vehicles or have been enhanced to accommodate long-term driver profiles [6]. The EVA assistant will monitor and detect the emotional state of drivers and passengers and accordingly need to provide appropriate mitigation strategies. Cooperative trajectory planning adapted to the identified emotional states is expected to be a major mitigation strategy. Integrating long-term driving profiles in ACC or Cooperative-ACC controllers will not be sufficient but need to be capable to adapt to short-term emotions in real-time. Cooperation between various individual EVA assistant profiles will be critical to resolve conflicts and design cooperative emotion-aware trajectories.

Humans continuously adapt their speech production to the communicative context, which is one of the keys to efficient and fluent spoken interaction that humans excel at. Speech synthesis, in contrast, is remarkably static. While the output quality of synthetic speech has increased tremendously in the past years, and has reached a point where it in many cases is indistinguishable from a human voice, current TTS systems are essentially oblivious to the context in which they operate. For smooth and efficient spoken in-vehicle interaction, it is highly desirable to have a speech synthesizer that is able to adapt its output to the context. Research into speaking style variation in TTS has focused primarily on synthesis of expressive speech, which in most cases means a fixed set of pre-defined emotions. This is too restrictive for most real-world applications [4]. Hence, leveraging recent improvements in DNN-based synthesis, the EVA assistant design is to add continuous control over several paralinguistic parameters, such as the ones described above, also allowing speech output generation to be directly conditioned on contextual data. Existing datasets may need to be amended with newly recorded data that contains the desired speaking style variations.

EVA's intelligent personal recommender system is based on advanced machine learning models. The models are trained using state of the art algorithms, which – while providing unprecedented performance – produce outputs that are often difficult to explain or interpret. Explicability and interpretability are important for enabling trust of users for a system. Our aim is to research techniques and approaches for providing insight into the models learnt using deep networks in the context of proactive recommendation. Due to the lack of intermediate information or decision in deep architecture, the final result comes without any particular explanation, which is probably fine when the system is accurate or when the produced output follows human intuition but not when an error occurred. In such case, it is important to know how the decision was made, in order to identify what caused the erroneous prediction/classification.

2.4 Cognitive IoT and Service Delivery

Personal assistants aim to generate personalized experience based on a user profile and contextual information. For example, depending on the weather, the location, the time of day or the season in the year, and the user preferences – a personal assistant will respond differently to a simple request such as "please, make a reservation in a near-by restaurant." The intelligent personal assistant should take into account the time, user preferences, who the user is eating with and naturally the set of restaurants nearby the current location, destination or any points in the shortest journey of the user for answering this question.

It is well established that edge cloud may represent a key enabler for low-latency-oriented applications, such as automotive industry and ITS. These types of service require ultra-low latency and reliable data analytics solutions that rely in real-time on heterogeneous data gathered from the ITS network and the vehicle environments (including environment sensors and emotional sensors). [5] Locating the cloud service at the network edge will considerably reduce the latency access to remote applications, like data analytics. Indeed, this may require low-latency access to the data analytic application located in the edge cloud to react to any urgency, such as heart failure. EVA will rely on the edge cloud to guarantee low latency as well as reliability for ITS applications.

EVA will devise mechanisms to deploy distributed analytic applications over the edge, advanced algorithms, based on multi-criteria combinatorial optimization and machine learning to decide the placement of an analytical application. Algorithms, which decide dynamically to duplicate an application to in-car fog, by predicting when the vehicle will lose connection or entering an area with bad connection.

2.5 Privacy and Human Factors

For users to accept the intelligent personal assistant provided by EVA, they need to be given reasonable guarantees that it will not pose significant threats to their privacy; i.e. that the data collected remain confidential and under the control of the users. While privacy-preserving machine learning has been studied in the past in the context of anonymous databases (differential privacy) or public database releases (privacy-preserving data mining), EVA offers significantly new challenges mainly because of the large amount of produced data, the existence of multiple data sources, and the necessity to use users' data by several authorized parties. In EVA, we will investigate customized privacy primitives based on advanced cryptographic techniques such as homomorphic encryption or secure multi-party computation that would enable the processing of the data while being encrypted. In order to integrate practical cryptographic tools, the underlying machine learning algorithm may sometimes be approximated into low degree polynomials. Therefore, the goal is to achieve a high degree of privacy (thus complying with the upcoming General Data Protection Regulation GDPR) without sacrificing utility (accuracy) too much.

While privacy may be receiving much attention as of now, it is essential to keep investigating human factors and ethical questions of any kind, since they will be crucial for the adoption of EVA and critically influence system design [7].

3 Conclusions

In this paper we present a multi-disciplinary approach to the future design of emotion-aware autonomous systems. A combination of multi-modal user interaction design approaches, advanced sensing with high temporal sensitivity, elaborate federated machine learning, distributed edge processing, cognitive modelling and cybernetic system design principles govern that approach. A layered architecture will provide for the delivery of cognitive services based on personalization information gathered from users' interaction with the system. That will allow for a pro-active mobility experience.

Acknowledgments

This work is in its infancy, currently seeking funding.

References

[1] BMW. 2016. BMW i Vision Future Interaction. In: *BMW i. Born Electric. The Official BMW i Channel.* Video Blog, YouTube, (Jan. 2016). https://youtu.be/LqCVfn7mwgw

[2] IBM. 2016. Cognitive Mobility: Olli the self-driving vehicle and Watson the cognitive system. In: *IBM Watson Internet of Things*. Video Blog, YouTube, (Jul. 2016). https://youtu.be/9joEsWiYFEI

[3] S. Skriabine, G. Taverni, F. Corradi, L. Longinitti, K. Eng, T. Delbruck. 2017. Ultra-High Speed Marker-Free Human Tracking using the Dynamic Vision Sensor, Technical Report ETHZ, 2017.

[4] Székely, E., Mendelson, J., Gustafson, J. 2017. Synthesizing uncertainty: the interplay of vocal effort and hesitation disfluencies. In: *Proceedings of INTERSPEECH*. 2017.

[5] S. Nastic et al. 2017. A Serverless Real-Time Data Analytics Platform for Edge Cloud. In: *IEEE Internet Cloud*, Vol. 21, Issue. 4 (2017).

[6] M. Hasenjäger, H. Wersing. 2017. Personalization in Advanced Driver Assistance Systems and Autonomous Vehicles: A Review. In: *Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE (2017).

[7] Susanne Müller. 2017. The Enhanced Human: Will We Need a New Self-Concept Between Automated Cars and a Super Intelligent Environment? In: *Breakout Workshops*, BMW Summer School, Bad Wörishofen (Jul. 2017).

[8] M-H. Yang, S-C. Liu, T. Delbruck. 2015. A Dynamic Vision Sensor with 1% temporal contrast sensitivity and in-pixel asynchronous delta modulator for event encoding, In: *IEEE Journal of Solid-State Circuits*, 50:(9), (2015).

[9] G. Dikov, M. Firouzi, F. Röhrbein, J. Conradt, C. Richter. 2017. Spiking cooperative stereo-matching at 2ms latency with neuromorphic hardware. In: *Proceedings of Living Machines*, (2017).

[10] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, T. S. Huang. 2016. Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge* (AVEC '16). ACM, New York, NY, USA, 97-104. (2016).

[11] S. Chen, Q. Jin, J. Zhao, and S. Wang. 2017. Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition. In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* (AVEC '17). ACM, New York, NY, USA, 19-26. (2017).

[12] V. Rieser, O. Lemon. 2011. *Reinforcement Learning for Adaptive Dialogue Systems: A Data-driven Methodology for Dialogue Management and Natural Language Generation*. Springer. (2011).

[13] E. Marchi, F. Eyben, G. Hagerer, and B. W. Schuller. 2016. *Realtime Tracking of Speakers' Emotions, States, and Traits on Mobile Platforms*. In: Proceedings INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, (San Francisco, CA), pp. 1182–1183, ISCA, ISCA, September 2016.

[14] T. Delbrück, B. Linares-Barranco, E. Culurciello and C. Posch. 2010. *Activity-driven, event-based vision sensors*. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems, Paris, 2010, pp. 2426-2429

[15] Y. Cao, Y. Chen, & D. Khosla. 2015. *Spiking deep convolutional neural networks for energy-efficient object recognition*. International Journal of Computer Vision, 113(1), 54-66, 2015