(Preprint) July 2020 – Digital Human Modeling (DHM 2020)

Multimodality on the Road

Towards Evidence-Based Cognitive Modelling of Human Interactions in Everyday Roadside Situations

Vasiliki KONDYLI – Mehul BHATT
Örebro University, Sweden
CoDesign Lab > Cognitive Vision
www.codesign-lab.org/cognitive-vision

Abstract We propose an evidence based methodology for the systematic analysis and cognitive characterisation of multimodal interactions in naturalistic roadside situations such as driving, crossing a street etc. Founded on basic human modalities of embodied (inter)action, the proposed methodology utilises three key characteristics crucial to roadside interactions, namely: explicit and implicit mode of interaction, formal and informal means of signalling, and levels of context-specific (visual) attention. Driven by the fine-grained modelling of human behaviour in naturalistic virtual environments, we present an application of the proposed model with examples from a work-in-progress dataset consisting of baseline multimodal interaction scenarios and variations built therefrom with a particular emphasis on joint attention and diversity of modalities employed. Our research aims to open up an interdisciplinary frontier for the human-centred design and evaluation of artificial cognitive technologies (e.g., autonomous vehicles, robotics) where embodied (multimodal) human interaction and normative compliance are of central significance.

Keywords. multimodal interaction, interpersonal communication, naturalistic perception, joint attention, virtual reality, autonomous driving

1. Introduction

Interpersonal communication and interactions are vital for safe and effective coordination of actions in everyday roadside engagements: walking around, driving, riding a bike etc. Failure in interpersonal communication leads to a lack of mutual understanding of a situation and it is responsible for a great number of roadside accidents [24, 28]. With further strides in the autonomous vehicles industry and the present impetus on high-level visual intelligence technologies [3, 30], it will therefore be necessary to account for the role of interpersonal communication on the street and articulate human-centred performance benchmarks, e.g., from the viewpoint of training, testing and validation as part of statutory compliance measures.

Evidence-Based Design Issues pertaining to human-centred design and human-machine interaction have been barely addressed in the autonomous driving sector.

Presently, autonomous vehicles do not have the capacity to communicate intentions, or to anticipate or predict interactions based on deep semantic analysis of observations of behavioural patterns. Considering how ambiguous interpersonal communication is within the context of driving, the interpretation is not trivial for contemporary systems, especially if we take into consideration socio-cultural normative behaviour, environmental and situational context. For this reason, people-centred datasets for training and testing the capabilities of systems will be necessary [16, 31]. These datasets should be based on behavioural analysis from real-world situations, and they should incorporate diverse real-world cases of interactions as they accrue amongst roadside stakeholders.

Interpersonal Communication by Multimodal Interaction Understanding how interpersonal communication develops is about understanding how emotions and intentions are expressed, and how gestures, facial expressions and body posture support, complement, and occasionally override verbal communication. Interpersonal communication involves single or multicomponent signals, together with informative cues and feedback [14].[1] Often communication takes place with the use of different modalities and often these phenomena occur in synchrony. Multimodality refers to a person's way of communicating by using more than one modalities at the same time, or from a perceptual approach, it refers to more than one modality being received based on the receiver's perception of the signal [21, 27].

Multimodality in Embodied Human Interaction The reason why and the point when people utilise multimodal signals for interaction is a question that has been explored from different perspectives [12, 14, 21]. It is primarily related to the circumstances (e.g. environment, events, participants), as well as to efficacy issues of signal structure and permanence to environmental noise, all of which are (most) likely to differ across modalities [13, 14]. For instance, signals expressed in different modalities have different transmission distances and different permanence to environmental noise; hence, by combining modalities the signal has a better chance to get transmitted. Moreover, examining correlations between cognitive load and multimodal communication shows that people respond to dynamic changes in their cognitive load by shifting to multimodal communication when load increases due to task difficulty or communication complexity [20].

Core Contribution Driven by fine-grained modelling of human behaviour and considerations in human-centred cognitive interaction technology design, we develop a systematic method for the evidence-based modelling of embodied multimodal interactions in an everyday roadside context with a particular focus on cognitive aspects pertaining to visual attention during activities such as driving and crossing a street. We propose a model categorising and characterising the instances of communication based on:

- interaction modalities, e.g. involving gestures, speech, head movement, gaze
- mode of deliverance of the communicative intent, e.g., explicit, implicit
- formalisation and configuration of the message, e.g., formal or informal, body-based, device-based
- levels of (visuoauditory) social attention achieved (e.g. common, mutual, joint) amongst interacting stakeholders such as Drivers, Pedestrians, Cyclists.

[1] A multicomponent signal is different than a multimodal signal and it refers to a number of sensory unimodal elements in a signal. It is considered complex signalling within the same modality.

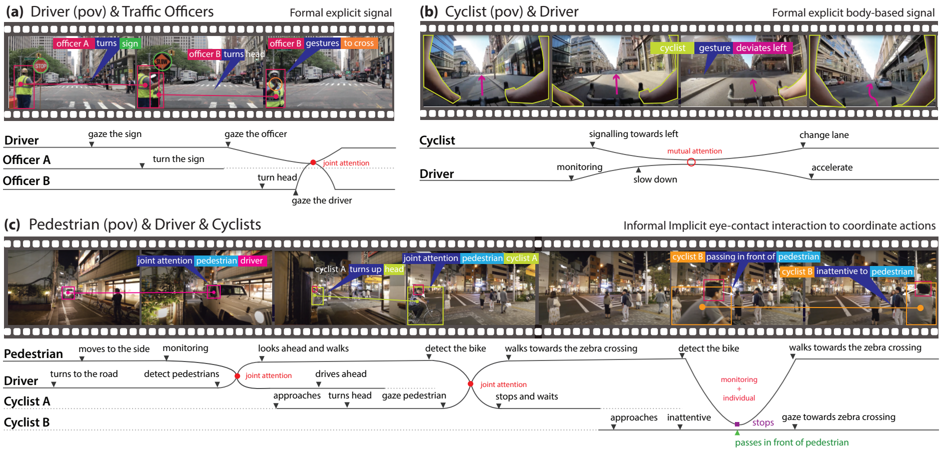

Figure 1. Multimodal interactions in streetscape, characterised based on Table 1; **(a)** From the point of view of the driver, a formal explicit device-based signal from the traffic officers using a sign and gestures (New York); **(b)** From the point of view of the cyclist, a formal explicit gesture indicates change of lane and ask for priority to the driver who follows (Berlin); **(c)** From the point of view of the pedestrian, several interaction episodes back to back involving a driver and two cyclists and focusing on the establishment of joint attention (Tokyo).

Even though real-world instances are the most valuable source for investigating the nature of interpersonal communication, controlling for external factors in naturalistic virtual reality (VR) scenes and projecting the roles of roadside stakeholders to a virtual agent and a participant are a significant addition for the investigation of synchronisation and dynamics between modalities during the course of interaction. To this effect, we also present instances from a work-in-progress dataset of roadside multimodal interactions combining real and VR episodes with the purpose of studying the variations of modalities involved in a sample of scenarios of interactions and the process of achieving joint attention.

2. A Cognitive Characterisation of (Roadside) Multimodal Interactions

Interactions between roadside users are mostly based on non-verbal communication and they are significant in resolving traffic ambiguities considering communication is precarious because of the lack of a homogeneously accepted social set of signals, and their dependance on the circumstantial aspects as situations, country, etc. [22, 29, 39] (Fig. 1). Non-verbal signals serve greatly social functions, creating bonds and shared knowledge, as well as reflecting attitudes, mood, and emotions [10]. In this context non-verbal communication can be characterised as *Spontaneous* (e.g. yawning, scratching their head, stretching their muscles), *Symbolic* (sign language, body movements or facial expressions), and *Pseudo-spontaneous* (performing an action that looks spontaneous) [6]. Focusing on the modalities used, the users, and the cognitive nature of the interaction we further characterise the interactions in the streetscape as follows (Table 1):

Interaction Modalities Roadside users handle a set of modalities to convey their intentions or to give feedback during an interaction. Each modality conveys a great deal of information regarding one's intentions that can be categorised according to its semantic

EMBODIED INTERACTIONS	INTERPRETATION
A1. MODE	
Explicit Interaction	Joint Attention - Facial Expressions - Gestures - Speech - Nodding
Implicit Interaction	Body Posture - Head Rotation - Behaviour Changes (pace, direction change) - Gaze Allocation (referential, aversion) - Stigmergy
A2. METHOD	
Formal Device-based	Hazard lights - Turn signal - Honking
Informal - Device-based	Head-light flashing for warning - Head-light blinking for acknowledgement - Honking as social etiquette - Honking for expressing displeasure - Honking for gratitude
Formal Body-based	Cyclist gesture to turn - Hand signals by traffic officer
Informal Body-based	Nodding for encouragement - Gesture to order yield - Gesture as gratitude for yielding - Eye contact to encourage yielding
A3. LEVEL (OF SOCIAL ATTENTION)	
Individual	Individual A attends objects/event X and individual B
Monitoring	Individual A attends to B's attention to objects/event X
Common	Individual A attends to B's attention to X and him/herself
Mutual	Individual A attends to B who attends to X and him/herself characterised by non-communicative eye contact
Joint (Shared)	Individual A attends to B who attends to X and him/herself characterised by communicative eye contact and/or other bi-directional communication
MODALITIES	EXAMPLES
HEAD MOVEMENTS (HM)	Turn towards the street - Tilt to a direction - Nod for disapproval - Slide for notice - Protrusion for warning
FACIAL EXPRESSIONS (FE)	Smiles - Frowns - Wrinkle - Eye Rolling - Cut Eye - Eyebrows Raising - Lips Movement - Mouth Movement
GESTURES (GE)	Emblematic (thumbs up, hitchhiking, stop) - Iconic (direction of movement) - Deictic (pointing) - Beat (irritation, gratitude)
BODY POSTURES (BP)	Crossing arms - Idle - Stand with the back to the street - Lean towards the car/ a kid's stroller - Stand besides a car/bike
GAZE (GZ)	Eye contact - Seek attention - Follow other's gaze - Follow a moving object - Aversion - Point towards a direction(Referential) - Look the traffic light
AUDITORY CUES (AU)	Honking - Car engine - Traffic light sound - Brakes - Siren - Voice
SPEECH (SP)	Ask - Warn - Shout - Scold - Give directions

Table 1. A Cognitive Characterisation of Roadside Interactions and the Modalities Involved.

functions. The classification is based on measurable properties of the modality such as direction, intensity, angle, fluidity, which can provide details for fine-grained modelling of interactions. For instance, manual gestures are classified with respect to their semantic function by McNeill [18] into *emblematic* (bare conventionalised meaning e.g. "thumbs up"), *iconic* (convey the shape of an object, direction of movement), *metaphoric* (resemble abstract concepts e.g. shape hands into a heart), *deictic* (point out locations in space), *beat* (keep the rhythm of speech with no semantic content). Head gestures vary in their exact of kinematic realisations (angles, extent), as well as overlap with other movements. However, the main categorisation includes *tilt, nod, turn, slide and protrusion*, and the relevant measurable properties include the pitch rotation in the up-down direction, roll in X axis, yaw and translation in X and Y axes [35].

A1. Mode of Interaction Signals of interpersonal communication can be expressed explicitly, e.g. via a handwave (Fig. 1a), or a gesture (Fig. 1b); however an implicit mode of signal deliverance is more common in streetscape scenarios, such as eye-contact (Fig. 1c). In implicit interactions, intending any practical action primarily aimed to reach a practical goal, can also lead to achieving a communicative purpose, without any predetermined (conventional or innate) specialised meaning. For instance, changing the speed of a vehicle indicates driver's intention to give or take priority. There are several steps in the scale between pure action and direct communication, with the general principle that the message is based on observation and it exploits simple side effects of acts and the agent's natural disposition to observe and interpret the behaviour of others [1].

A2. Method of Interaction The role and the tools that different roadside users have at their disposal also indicate the types of modalities they use and the nature of interaction they get involved in. Pedestrians and cyclists use their body as a communication tool (e.g. hand gesture - Fig. 1b; eye gaze and subtle movement towards the side of the road - Fig. 1c). Drivers also use body-based configurations and additionally the available technological device such as hazard lights, or horn. With or without equipment roadside users produce a range of formal and informal signals that they integrate into their interactions [23]. Formal signals refer to established traffic rules such as gesturing before changing lane for a cyclist (Fig. 1b), or traffic officer's sign (Fig. 1a), while informal signals vary widely, as they are highly context and culture dependant (e.g. gesture for gratitude, gesture to give priority).

A3. Levels of Social Attention Gaze has a crucial role in non-verbal communication in the streetscape. Naturalistic studies show that pedestrians often establish eye contact with drivers to make sure they are seen, and drivers also often gaze at the face of other road users to assess their intentions [37]. Gaze in combination with gestures or speech, aims to establish a common ground between the road users and achieve a high level of social attention. The levels of social attention correspond to different degrees of situation awareness, and are defined in a scale from *individual* to *joint (or shared)*[2], where *individual* refer to one person attentional engagement with the environment from a first-person perspective only, while in every additional state of the scale (*monitoring, common, mutual, and joint*) the person's engagement is modified to second or third person perspective in order to acquire common knowledge with others [26]. The ultimate state of interaction is *joint* attention, the state where both participants have awareness of the situation and are also both aware that they are engaged. The different levels can be established with a combination of multiple interaction modalities.

Joint Attention In developmental psychology, the ability to share attention and to coordinate behaviour is defined under the joint attention framework, or visual co-orientation [7]. Joint attention traditionally refers to a triadic relationship between two interacting parties and a shared object or event [9, 19], and it is mostly related to the ability to follow a person's gaze to an objects or event [9, 19]. However, in recent work joint attention is also interpreted as *mental focus* [15] or *shared intentionality* [33]. Even though joint attention has been investigated in more sensory modalities other than vision (such as touch [5]), in the context of interpersonal communication in streetscape the focus is on joint *visual* attention between the roadside users. We address joint attention as the ability of a person to engage with another for the purpose of a common objective, or task, which may not involve explicit gaze following action or specific object involved.

Factors Influencing Roadside Interactions Multimodal interactions highly vary and they can convey very different meanings depending on the users involved (F1-Table 2), their intentions, and activities in the streetscape (Table 2, F2), as well as the environmental and situational context (Table 2, F3). For example, *social* factors refer to differences in behaviour recorded as a result of the group size of users, or the compliance levels to traffic rules; while *demographics* refers to correlations between age or gender groups with attentive behaviour from themselves and a cautious treatment from others [28]. Although we emphasise the importance of these factors to the overall outcome of the interactions, we do not provide further analysis in this paper, however, we address parts of topic on our previous work focusing on visuospatial complexity of naturalistic driving stimuli [16].

[2] The terms *joint attention* and *shared attention* have a very similar meaning however there is not one widely acceptable term for the phenomenon. For this paper we will refer to it as *joint attention*.

ROADISE INTERACTIONS	STAKEHOLDERS – ACTIONS – CONTEXT
F1. Roadside Users	
Pedestrian - Motorcyclist - Cyclist - Driver Kid's stroller - Wheelchair - Truck Driver - Bus Driver Emergency Vehicle Driver - Trailer Driver - Animal Rider - Traffic Control Person - Pedal-cyclist	
F2. Intentions – Activities	
Slow down/Accelerate - Cross - Overtake - Stop/Start - Enter/Exit Point - Turn - Ask - Perform work - Play - Retrieving an object - Warn - Regulate traffic	
F3. Context — Environmental – Situational	
VISUOSPATIAL COMPLEXITY	Spatial Configuration - Street Width - Visibility - Auditory Cues - Clutter - Luminance - Traffic Density - Order - Regularity - Motion - Speed - Direction
DEMOGRAPHICS	Age - Gender - Culture
INDIVIDUAL DIFFERENCES	Psychical Capability - Cognitive Capability - Experience - Emotions - Attitude and beliefs - Personality traits
SOCIAL	Group Size - Social Norm - Law Compliance to Traffic Rules - Behavioural Imitation/Observation - Movement Flow - Informal Best Practices

Table 2. Factors influencing the behaviour of roadside users and the multimodal interactions developed during roadside actions.

3. Human-Centred Interaction Modelling: From Real-World to Naturalistic Virtual Scenes

To develop evidence-based modelling on frequently encountered multimodal interactions in the streetscape we firstly analyse incidents from real-world dynamic scenes recorded from the perspective of a driver, cyclist, or pedestrian (select scenes in Table 3). The analysis is based on the cognitive categorisation in Table 1, with the aim to examine:

1. What kind of interpersonal communication does take place between roadside users and which modalities are used?
2. What kind of interrelations can be found among the modalities during the course of interaction and how do they vary in similar scenarios?
3. Can properties of the modalities be measured systematically to serve in fine-grained modelling for the purpose of design multimodal interaction in VR?

As a second step, a number of incidents from the chosen scenes are subsequently (re)constructed in a virtual environment with variations on the modalities used for communication and the level of social (visuoauditory) attention established between the participating roadside users. We present two example scenarios with their corresponding variations (Scenarios A and B; Fig. 2-3):

Scenario A. Zebra-Crossing Situation (Fig. 2)

Pedestrian (P) with a kid's stroller is crossing a two-lane road on a zebra crossing while two drivers (D1 and D2) are approaching. P turns the head towards D1 as he approaches a zebra crossing, and establishes eye contact with D1. P then looks straight. D2 approaches the zebra crossing in the second lane without detecting P. Momentarily P turns the head, detects D2, stops and expresses disapproval towards D2 by extending his leg and using frowns and lip movements.

► *Scenario A Analysis based on Table 1.* Pedestrian P performs informal body-based explicit interaction with drivers D1 (via eye contact) and D2 (via body posture and facial expressions). P establishes joint attention with D1, as both parts engage in eye contact and both slow down or stop indicating intentional communication and situation awareness. On the contrary, for the interaction between P and D2 we only annotate monitoring attention for P towards D2. Concerning the interrelations between the modalities used, P uses body posture together with facial expressions instead of gestures (because of his

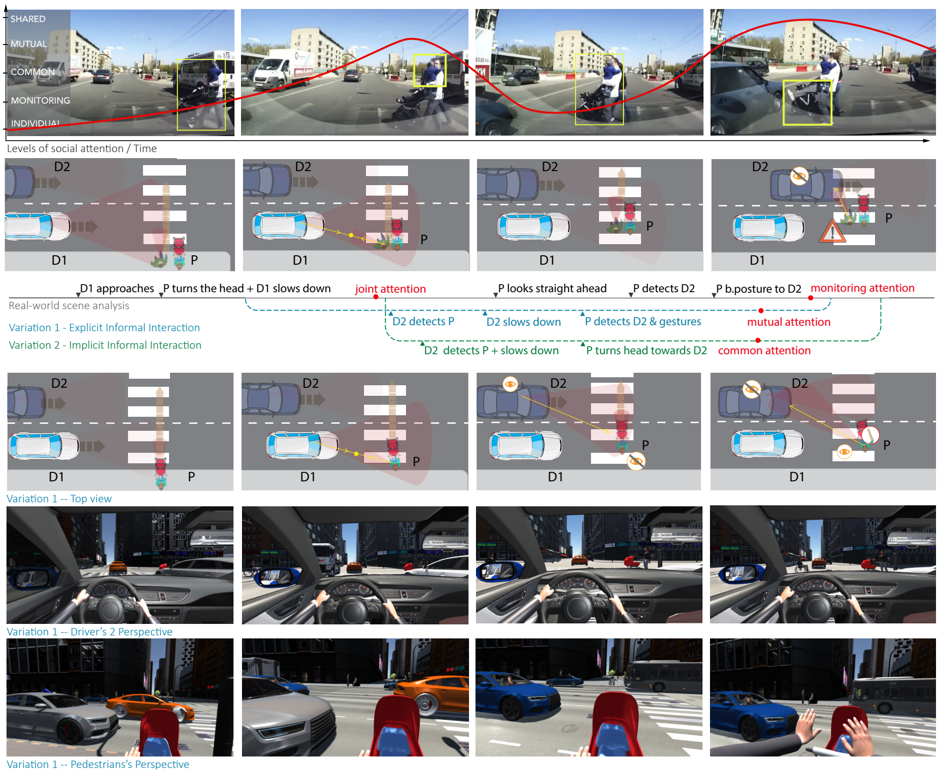
SHARED
MUTUAL
COMMON
MONITORING
INDIVIDUAL
Levels of social attention / Time
D2
D1
P
D1 approaches
P turns the head + D1 slows down
joint attention
P looks straight ahead
P detects D2
P b.posture to D2
monitoring attention
Real-world scene analysis
Variation 1 - Explicit Informal Interaction
Variation 2 - Implicit Informal Interaction
D2 detects P
D2 slows down
P detects D2 & gestures
mutual attention
D2 detects P + slows down
P turns head towards D2
common attention
Variation 1 -- Top view
Variation 1 -- Driver's 2 Perspective
Variation 1 -- Pedestrians's Perspective

Figure 2. Zebra-Crossing Situation (Scenario A) / Real-world scene analysis involving an interaction incident between pedestrian(s) with a kid's stroller and drivers on two-lane zebra crossing. Two variations of this scenario developed in VR, differ from the original scene in terms of the embodied interaction factors (A1-A3) and the combination of modalities involved (Table 1).

occupied hands) to communicate agitation. Social attention levels change three times (represented with the red line on Fig. 2) as a result of P's distraction after interacting with D1, and the head movements that follow. In this scenario measurements of the angle of head rotation in both interaction instances, the synchronisation between the gaze allocation and the reaction time among users are significant for modelling. In variations 1 and 2 we manipulate the series and the number of events, as well as the timing between them, in order to examine via behavioural studies in VR (Section 4) the establishment of lower levels of social attention (e.g. mutual and common).

Scenario B. Cyclist changes lane / turns in front of a car (Fig. 3)
(Motor)Cyclist (C) changes lane / turns in front of a car and the driver (D). C slightly deviates from his lane, D who is following C slows down and starts monitoring C, C listens to the car approaching, performs overhead check and establishes common attention with D, then C looks ahead and changes lane.

▶ *Scenario B Analysis based on Table 1*. (Motor)Cyclist C performs an informal body-based implicit interaction with D by the action of changing direction of movement, and

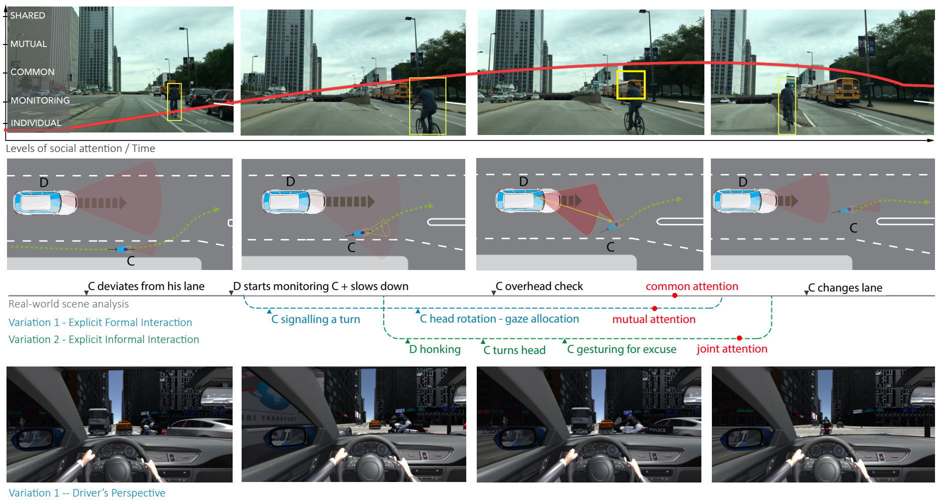
SHARED
MUTUAL
COMMON
MONITORING
INDIVIDUAL
Levels of social attention / Time
D
C
C deviates from his lane
D starts monitoring C + slows down
C overhead check
common attention
C changes lane
Real-world scene analysis
Variation 1 - Explicit Formal Interaction
Variation 2 - Explicit Informal Interaction
C signalling a turn
C head rotation - gaze allocation
mutual attention
D honking
C turns head
C gesturing for excuse
joint attention
Variation 1 -- Driver's Perspective

Figure 3. Cyclist changes lane / turns in front of a car (**Scenario B.**) / Real-world scene analysis involving a driver and a cyclist, and two variations that differ on the formalisation and the configuration of the signal and the modalities involved (based on Table 1). In variation 1 an explicit formal interaction is developed using formal gestures and gaze by the (motor)cyclist.

then via an overhead check. D's monitoring attention and C's overhead check overlap in time and lead to a safe change of lane for C in front of the car. During the overhead check, C uses his peripheral vision to detect D and to confirm the auditory cue of car's engine, but C and D do not perform eye contact. Consequently, D and C are aware of each others presence and they achieve common knowledge about the events via recursive assumptions, inferences, and perspective-taking since there are no specific external behaviours (beyond monitoring attention). To examine the temporal coordination of actions we record head rotation angle, reaction times, the speed and acceleration changes of the car, and the duration of monitoring attention and overhead check. Two variations of explicit communication signals are developed to test how they may lead to higher levels of social attention.

4. Towards a Naturalistic Dataset of Human Interactions in Everyday Driving

Work is in progress to develop a multimodal interaction dataset (following the methodology discussed in Section 3) consisting of the original real-world scenes together with scenes representing variations (in VR). Specifically, we collect and analyse a set of 20 dynamic scenes, covering 15 scenarios (A-O, Table 3) recorded from the egocentric perspective of a driver, cyclist, or pedestrian. The scenes are chosen such that there exists diversity with respect to typically occurring events and hazardous situations published by the Accident Research report of the German Insurance Association ("Unfallforschung der Versicherer") [11].

Behavioural Analysis The overall analysis of the real-world scenarios suggests that the behaviour of roadside users varies significantly even for similar scenarios with the

SCENARIOS	ROADSIDE MULTIMODAL INTERACTIONS	MODALITIES
A	P with a kid's stroller is crossing a two-lanes road while D1-D2 are approaching.	HM, BP, GZ, FE
B	C changes lane / turns in front of a car.	HM, GZ, AU
C.	Inattentive group of P crossing the street, D approaches seeking attention and signalling	GE, GZ, BP
D.	P looks at the traffic light that turns red, signalling and taking to other P on the other side of the street and crossing inattentively, D approaches the crossing	SP, FE, GE, BP, GA
E.	P emerges between parked cars and enters a parked car. D approaches	HM, GZ, AU
F.	P on wheelchair approaches a zebra crossing, D and C approach from different sides and give priority to P	HM, FE, GE, GA
G.	D turns to the street and P who are walking on the street move to the side	AU, HM, BP
H.	P (or group of pedestrians) cross half way a double-way street, they do not check the second lane, D approaches	AU, BP
I.	Low traffic road, P on the side of the street negotiate crossing with D, while M and C are passing between stopped cars	GZ, AU, HM
J.	P exits a shop/parking slot and walks on the street, D approaches	HM, BP, GA
K.	P is close to a zebra crossing, talking on the phone or texting with no clear intention to cross, D approaches	BP, FE, SP
L.	C standing close to a bike, and get on the bike, with no clear intention to start driving	BP, HM
M.	P steps on the road because of an obstacle on the pavement, C avoids pedestrian and changes lane, while D approaches	HM, GE, BP, GA
N.	M overtakes a car, looking for occluded pedestrians, and gives priority to P who is crossing	GZ, HM
O.	Policemen regulates traffic, instruct D for the direction too follow	BP, AU, GE

Table 3. Select scenarios of multimodal interactions based on the real-world dynamic scenes. The modalities involved are represented by the acronyms as per Table 1: **SP** (Speech); **HM** (Head Movements); **FE** (Facial Expressions); **GE** (Gestures); **BP** (Body Postures); **GZ** (Gaze); **AU** (Auditory Cues). The stakeholders involved are: D (Driver), C (Cyclist), M (Motorcyclist), P (Pedestrian).

same user roles (Pedestrian-Driver, Cyclist-Driver), and the same goals (crossing, turning). This observation highlights the effect of external factors such as visuospatial complexity, traffic dynamics, culture. Moreover, in line with previous studies we observe some modalities to differentiate more between cases than others [20, 35]. For instance, the number of gestures seems to be the same on average across the users in similar interactions, while the number of head and body movement differ greatly. However, even though the use of multimodal cues differ a lot in manner and frequency, there are some underline commonalities rooted in human basic perception and cognition concerning visual attention, spatial cognition and decision-making on the use of communication signals. For instance, drivers are more likely to use a turn signal if they have to turn left instead of right, or if they gaze at the vehicle in front of them that approaches an intersection [32]. Additionally, analysis of interrelations between modalities suggests that different interaction modalities are closely related to different cognitive processes, e.g. gestures with thinking and motor control of speech, body movements and facial expression with emotions [8].

Additionally, we observe more implicit than explicit interactions, and many of the explicit ones in more hazardous situations. This is in line with studies suggested that there is a trade off between the complexity of the communication mode and the reaction time required to respond [36]. Explicit interaction requires more cognitive processing to be perceived and it occasionally leads to slower reactions and collisions. However, this does not mean that explicit communication is counter-productive, but it shows the need for the communication strategies to start well in advance. Moreover, a major weight of interpersonal communications in this dataset was held by gaze. Pedestrians employ direct gaze to indicate intention to cross, to make the drivers to yield and more. Eye contact many times supported by facial expressions, is used by pedestrians to gain attention, while lack

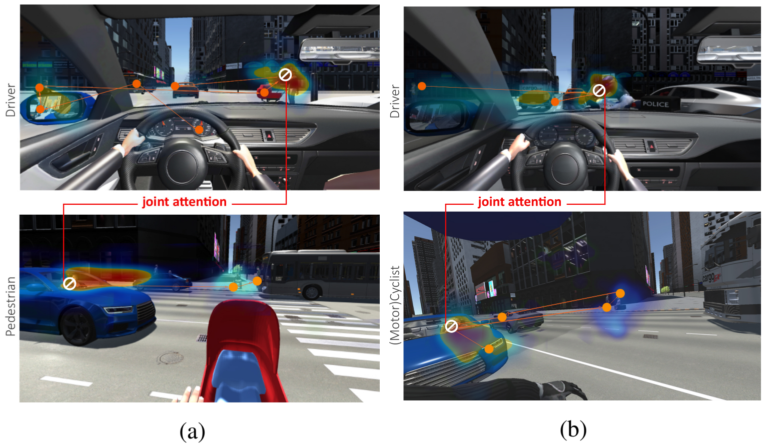

Figure 4. Sample eye-tracking data (presented as scanpath and heatmap) corresponding to the moment of joint attention (fixations during joint attention represented by ⊘): **(a)** Joint attention established between a Pedestrian with a stroller and a Driver (Situation corresponding to Scenario A in Fig. 2); **(b)** Joint attention established between a (Motor)Cyclist and a Driver (Situation corresponding to Scenario B in Fig. 3).

of gaze coordination, as a result of gaze deviation to distractors or aversion, indicates low level of social attention and it is related to hazardous scenes.

Empirical Evaluation in VR[3] Considering that the real-world scenes analysis shows a lot of variance between the behavioural patterns mostly because of external factors, the naturalistic VR scenes provide the controlled conditions for a behavioural study on fine-grained behavioural traits during interactions. By developing the scenarios in VR, we project the roles of pedestrians, drivers or (motor)cyclists to a pair of a user and a virtual agent. We manipulate variables related to the establishment of embodied interactions (mode of deliverance, formalisation and configuration of the signals), and the modalities used and we examine how they affect the establishment of different levels of social attention (Scenarios A-B). By manipulating aspects of the original events in VR we also explore how the complexity of the event may trigger the use of different modalities for interpersonal communication. In the ongoing behavioural study we collect physiological measurements (e.g. eye-tracking – Fig. 4), as well as observation on behavioural patterns and expressions (e.g. head rotation, steering wheel rotation, acceleration, intensity of gestures). This work adds empirical knowledge to the process of fine-grained modelling of interactions, concerning typical everyday scenarios in streetscape that may seem trivial and monotonic however they are complex problems for today's autonomous systems. It also contributes to evaluation of the human-centred interaction modelling and experimentation in respect to behavioural patterns and differences between people in the course of interaction.

[3] **Technical Setup (VR and Immersive Eye-Tracking)**. We implement full-body animated VR characters and several urban scenes built within the Unity Game Engine (v2019.2.2). The virtual scenarios are inherently multiperspective, e.g., from the POV of driver(s), pedestrian(s), cyclists(s). For the behavioural study, we use the HTC Vive Pro Eye system with embedded eye-tracking, accelerometer, gyroscope, and dual front-facing cameras with display resolution of 2880x1600 and 90Hz refresh rate. For the control of motion by the participants, we use a Logitech steering wheel with two pedals for the driver / cyclist, and a hand-held controller for the pedestrian.

5. Outlook

Within autonomous driving, the need for ethical regulation has most recently garnered interest [4, 16, 31]; therefore, qualitatively specified human-centred behavioural and normative benchmarks and evaluation for machine intelligence are imminent. Embedding evidence-based modelling in the process of designing agent-user multimodal interactions provides an ecologically rooted naturalistic basis for the development of human-centred technologies such as autonomous vehicles and social robotics.

As an example application of the proposed methodology, we have reported work-in-progress concerning the development of a dataset (including real-world and VR scenes) emphasising a cognitive characterisation of roadside multimodal interactions. In synergy with recent work on evaluating the visuospatial complexity of dynamic scenes [16], such a dataset provides an empirical foundation for the human-centred design of cognitive (computational) vision components within cognitive interaction technologies in general, and autonomous vehicles in particular [3, 30, 31]. That said, even from the singular viewpoint of behavioural research alone, we believe that ecologically valid naturalistic datasets such that ones resulting from this research (and [16]) can provide a shared foundation for conducting naturalistic studies in perception and interaction, e.g., in the context of established paradigms such as event perception [34], ensemble perception [38], visual search and foraging [17], change blindness [25]. We posit that such a confluence of computational and behavioural studies combining cognitive psychology, AI, digital media, HCI, and design science [2] are needed to better appreciate the complexity and spectrum of varied human-centred challenges in the design of cognitive (assistive) technologies and other artefacts in everyday life and work.

Publication Note.

This paper is a pre-publication draft of the contribution to appear as part of Proceedings of the 6th International Digital Human Modeling Symposium (**DHM 2020**), August 31 - September 2, 2020, Skövde, Sweden. `http://www.dhm2020.org/`.

References

[1] R. Beckers, O. Holland, and J. Denebourg. From local actions to global tasks: Stigmergy and collective robotics. pages 181–189, Cambridge, 1994. Fourth International Workshop on the Synthesis and Simulation of Living Systems (Artificial Life IV), MIT Press.

[2] M. Bhatt. Minds. movement. moving image. *Cognitive Processing*, 19(Suppl. 1):S5–S5. doi: 10.1007/s10339-018-0884-3. URL `www.codesign-lab.org/www/ICSC2018.pdf`.

[3] M. Bhatt. and J. Suchan. Cognitive vision and perception: Deep semantics integrating AI and vision for (declarative) reasoning about space, action, and motion. In *24th European Conference on Artificial Intelligence (ECAI)*, Santiago de Compostela, Spain, 2020.

[4] BMVI. Report by the ethics commission on automated and connected driving. *BMVI*, 2018.

[5] M. Botero. Tactless scientists: Ignoring touch in the study of joint attention. *Philosophical Psychology*, 29(8):1200–1214, 2016.

[6] R. Buck and C. VanLear. Verbal and nonverbal communication: Distinguishing symbolic, spontaneous, and pseudo-spontaneous nonverbal behavior. *Journal of Communication*, 52 (3):522–541, 2002.

[7] G. Butterworth and E. Cochran. Towards a mechanism of joint visual attention in human infancy. *International Journal of Behavioral Development*, 3(3):253–272, 1980.

[8] E. De Stefani and D. De Marco. Language, gesture and emotional communication: An embodied view of social interaction. *Frontiers in Psychology*, 10, 2019.

[9] W. V. Dube, R. P. MacDonald, R. C. Mansfield, W. L. Holcomb, and W. H. Ahearn. Toward a behavioral analysis of joint attention. *The Behavior Analyst*, 27(2):197, 2004.

[10] R. S. Feldman and B. Rim. *Fundamentals of Nonverbal Behavior*. Cambridge University Press, 1991.

[11] GDV. *Compact Accident Research by the German Insurance Association (Unfallforschung der Versicherer)*. 2017.

[12] P. Healey, M. Colman, and M. Thirlwell. *Analysing Multimodal communication. Repair-Based Measures of Human Communicative Coordination*. Springer, 2005.

[13] E. Heymann. The neglected sense-olfaction in primate behavior, ecology, and evolution. *Am J Primatol.*, 68(6):519–524, 2006.

[14] J. Higham and E. Hebets. An introduction to multimodal communication. *Behavioral Ecology and Sociobiology*, 67(9), 2013.

[15] P. Holth. An operant analysis of joint attention skills. *Journal of Early and Intensive Behavior Intervention*, 2(3):160, 2005.

[16] V. Kondyli, M. Bhatt, and J. Suchan. Multimodal interaction in autonomous driving. towards human visual perception driven standardisation and benchmarking. In *STAIRS @ ECAI 2020: 9th European Starting AI Researchers Symposium (STAIRS)., at ECAI 2020, the 24th European Conference on Artificial Intelligence (ECAI)*, 2020.

[17] T. Kristjánsson, I. Thornton, A. Chetverikov, and A. Kristjánsson. Dynamics of visual attention revealed in foraging tasks. *Cognition*, 194, 2020.

[18] D. McNeill. Hand and mind: What gestures reveal about thought. *Leonardo*, 27(4), 1992.

[19] C. Moore, M. Angelopoulos, and P. Bennett. The role of movement in the development of joint visual attention. *Infant Behavior and Development*, 20(1):83–92, 1997.

[20] S. Oviatt, R. Coulston, and R. Lunstord. When do we interact multimodally? cognitive load and multimodal communication patterns. Pennsylvania, USA, 2004. ICMI.

[21] S. Partan and P. Marler. Issues in the classification of multisensory communication signals. *The American Naturalist*, 166(2):231–45, 2005.

[22] A. Rasouli, I. Kotseruba, and K. J. Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *IEEE Inter Confon Computer Vision Workshops*, pages 206–213, 2017.

[23] K. Renge, G. Weller, B. Schlag, M. Peraaho, and E. Keskinen. Comprehension & evaluation of road users' signaling - an international comparison between finland, germany, & japan. pages 91–100. International Conference on Traffic and Transport Psychology- ICTTP, 2000.

[24] R. Risser. Behavior in traffic conflict situations. *Accident Analysis and Prevention*, 17(2): 179–197, 1985.

[25] D. J. Simons and D. T. Levin. Change blindness. *Trends in Cognitive Sciences*, 1(7):261 – 267, 1997.

[26] B. Siposova and M. Carpenter. A new look at joint attention and common knowledge. *Cognition*, 189:260–274, 2019.

[27] C. Smith and C. Evans. A new heuristic for capturing the complexity of multimodal signals. *Behavioral Ecology and Sociobiology*, 67(9), 2013.

[28] S. Stanciu, D. Eby, L. Molnar, R. Louiss, N. Zanier, and L. Kostyniuk. Pedestrians/ bicyclists and autonomous vehicles: How will they communicate? *Transportation Research Record*, 2672(22):58–66, 2018.

[29] M. Sucha, D. Dostal, and R. Risser. Pedestrian-driver communication and decision strategies at marked crossings. *Accident Analysis and Prevention*, 102:41–50, 2017.

[30] J. Suchan and M. Bhatt. Driven by commonsense: On the role of human-centred visual explainability for autonomous vehicles. In *24th European Conference on Artificial Intelligence (ECAI)*, Santiago de Compostela, Spain., 2020.

[31] J. Suchan, M. Bhatt, and S. Varadarajan. Out of sight but not out of mind: An answer set programming based online abduction framework for visual sensemaking in autonomous driving. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1879–1885, 2019.

[32] J. Sullivan, S. Bao, R. Goudy, and H. Konet. Characteristics of turn signal use at intersections in baseline nauralistic driving. *Accident Analysis and Prevention*, 74:1–7, 2015.

[33] M. Tomasello and M. Carpenter. Shared intentionality. *Developmental science*, 10(1):121–125, 2007.

[34] B. Tversky and J. M. Zacks. Event perception. In D. Reisberg, editor, *The Oxford Handbook of Cognitive Psychology*, Oxford Library of Psychology. ISBN 9780195376746. doi: 10.1093/oxfordhb/9780195376746.013.0006.

[35] P. Wagner, Z. Malisz, and S. Kopp. Gesture & speech in interaction: An overview [editorial]. *Speech Communication*, 57:209–232, 2014.

[36] I. Walker. Signals are informative but slow down responsess when drivers meet bicyclists at road junctions. *Accident Analysis and Prevention*, 37(6):1074–1085, 2005.

[37] I. Walker and M. Brosnan. Drivers gaze fixations during judgements about a bicyclists intentions. *Transportation research part F: Traffic psychology and behaviour*, 10(2):90–98, 2007.

[38] D. Whitney and A. Yamanashi-Leib. Ensemble perception. *Annual review of Psychology*, 2017.

[39] G. Wilde. Immediate and delayed social interaction in road user behaviour. *Applied Psychology*, 29(4):439 – 460, 1980.