Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Gaze-based intention estimation: principles, methodologies, and applications in HRI

Anna Belardinelli

2024

Preprint:

Copyright ACM 2024. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in ACM Transactions on Human-Robot Interaction, https://doi.org/https://doi.org/10.1145/3656376.

Check for updates

Gaze-based intention estimation: principles, methodologies, and applications in HRI

ANNA BELARDINELLI, Honda Research Institute Europe, Germany

Intention prediction has become a relevant field of research in Human-Machine and Human-Robot Interaction. Indeed, any artificial system (co)-operating with and along humans, designed to assist and coordinate its actions with a human partner, would benefit from first inferring the human's current intention. To spare the user the cognitive burden of explicitly uttering their goals, this inference relies mostly on behavioral cues deemed indicative of the current action. It has been long known that eye movements are highly anticipatory of the single steps unfolding during a task, hence they can serve as a very early and reliable behavioural cue for intention recognition. This conceptual review aims to draw a line between insights in the psychological literature on visuomotor control and relevant applications of gaze-based intention recognition in technical domains, with a focus on teleoperated and assistive robotic systems. Starting from the cognitive principles underlying the relationship between intentions, eye movements, and action, the use of eye tracking and gaze-based models for intent recognition in Human-Robot Interaction is considered, with prevalent methodologies and their diverse applications. Finally, special consideration is given to relevant human factors issues and current limitations to be factored in when designing such systems.

CCS Concepts: • **General and reference** → Surveys and overviews; • **Computing methodologies** → *Cooperation and coordination*; • **Human-centered computing** → **HCI theory, concepts and models**.

Additional Key Words and Phrases: Intention recognition, Human-Robot Interaction, Gaze-based interaction

1 INTRODUCTION

Intention recognition algorithms are currently at the heart of fast developing technologies involving the interaction of human and artificial systems, may these be teleoperated robotic arms, driving assistance systems, robotic wheelchairs, or social robots acting in an environment shared with humans [Vernon et al. 2016]. These systems usually rely on a number of machine learning techniques for early estimation of the ongoing or upcoming action, typically based on partially observed time series of behavioral cues. Further, as small and cheap wearable cameras are nowadays commonplace, a privileged ego-centric vantage point can be used to effectively observe how intentions evolve into action. In recent years, it has become clear that a conveniently natural and timely way to predict intention in many applications is to use gaze and eye movements. Considering intentions in physical actions, for example, in light of the need to cope with sensorimotor delays [Miall and Reckess 2002], gaze control itself in task-based scenarios can be considered as inherently predictive of a number of action-relevant aspects. Indeed, in moving our eyes we make use of knowledge- and sensorimotor-based experience and memory [Fiehler et al. 2019; Hayhoe 2017; Henderson 2017] to quickly retrieve the information needed to plan limb motion. We further use our eyes to initiate and supervise action during its whole execution. Such gaze cues (among others) are also used by other humans to estimate our intentions and possibly help with those or react accordingly

Author's address: Anna Belardinelli, anna.belardinelli@honda-ri.de, Honda Research Institute Europe, Carl-Legien-Str.30, Offenbach, Germany, 63073.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2024 Copyright held by the owner/author(s).
ACM 2573-9522/2024/4-ART
https://doi.org/10.1145/3656376

Just Accepted

[Emery 2000; Huang et al. 2015; Tomasello et al. 2005]. Gaze is indeed a strong cue in social interaction and it has been often used also in Human-Robot interaction (HRI) [Admoni and Scassellati 2017]. These considerations show that intention recognition is a highly cognitive task and it constitutes a critical feature for intelligent cooperative systems. In general, as pointed out by Bulling and Zander [2014], situation- or context-aware systems have been developed to proactively help users, but more attention has been paid to the environmental context, somehow neglecting, until recently, the cognitive context of the user. Taking this into account, for example by inferring intentions, would push cognition-aware computing forward. Indeed, as argued by Zhang et al. [2020], while robotic systems relying on human-centered AI are becoming pervasive in human social environments, such agents need to perceive and leverage the human gaze to better understand and interact with humans. As for other cognitive functionalities and models in HRI, one can approach gaze-based intention recognition in a purely data-driven way or in a more hypothesis- or model-driven way [Tsotsos et al. 2019], with the former often preferred lately in applicative fields. Gaze data indeed have currently become very easy to collect in large amounts for any specific application one might need, yet here we would like to make the point that a better comprehension of the human visuomotor system could go a long-way also in optimizing such data harnessing. In this respect, admittedly machine learning approaches can be used in an agnostic way to learn relevant associations between gaze and intentions. Nonetheless, knowing how intentions and actions emerge and coordinate in the human mind through the eyes interplay might help identify relevant features and design better systems depending on the context and activity of interest. Since this background knowledge is often missing or not sufficiently systematized in current technical systems, before surveying the literature on gaze-based intention recognition, we first make an excursus reviewing cognitive science studies on how intentions are formulated, what is their function in action production, and how the gaze and visual systems facilitate this connection in an integrated way, underlining the predictive nature of eye movements in natural eye-hand coordination. We further consider what gaze features are usually considered, how they are computed, and what kind of insight they can deliver while also looking at which predictive models have been mostly adopted. This also can depend on the technical fields in which intention recognition has found relevant application (see Figure 1 for a summary). In recent years the use of gaze tracking devices for inferring user intention has increased but a wide-spread, multi-purpose use is still hampered, that is, intentions have been considered both as overarching activities or at the level of primitive actions but deliberate manipulation intentions of higher semantic level seem more elusive and hard to capture. In this respect, we finally consider a number of limitations, challenges, and related open issues that need to be taken into account in order for intention recognition systems to be more effective and acceptable to human users.

Just Accepted

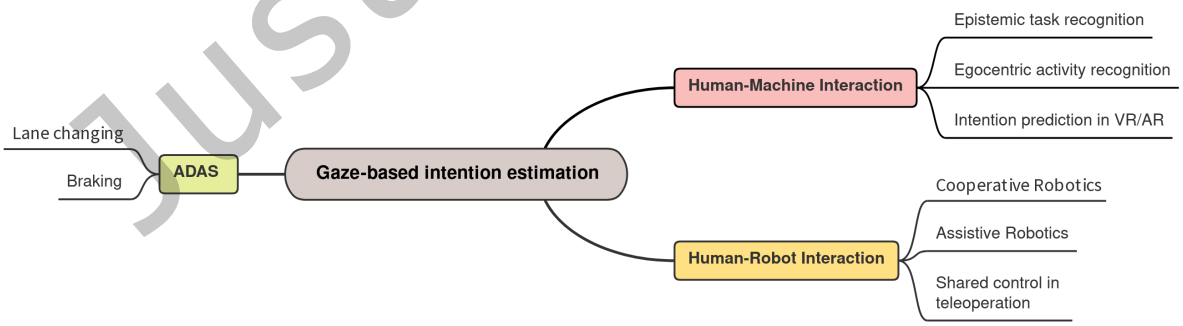

Fig. 1. Fields and sub-fields of application for gaze-based intention estimation, as surveyed in this review.

2 INTENTIONS, GAZE AND ACTION

By surveying the technical literature a number of terms occur that appear to be concerned with the same concept as intention recognition: intention estimation, intention prediction, intention inference, action prediction, activity recognition, task prediction, and, in some cases, plan recognition. All of them assume that there is a purposeful, deliberate behaviour displayed by a human agent that can be recognized/predicted by an artificial system on the basis of different observable cues. These cues range from body motion as externally tracked by cameras, to movements detected by body-worn inertial units or mouse/joystick controllers, to signals from myographic activity or brain-computer interfaces, to gaze control, again as observed from external cameras or by egocentric, head-mounted eyetracking devices. For the purpose of plan recognition, these cues do not allow to look too far ahead in the future and also a kind of Theory of Mind would be needed, along with a planner [Baker and Tenenbaum 2014], hence we will not review such systems. As pointed out by Kong and Fu [2018], even though at times the terms are still not so clearly differentiated, *action recognition* is usually done offline and when the action to recognize has been completed, while *action prediction* relies on preliminary observation of the above-mentioned cues to estimate (possibly online) the ongoing action. Still, this distinction is hardly applicable when speaking of intentions. As reviewed in the following section, intentions are hidden, unobservable states, which can be first inferred when translated into motor behavior. In this sense, intentions are always recognized or inferred *post-hoc*, since they have been already formed, while the corresponding actions might still be predicted before redinitiation or completion. In this sense, we will consider intention estimation, prediction, recognition, and inference as synonyms here, while using the former as the more generic term.

For the reasons stated above this review focuses on gaze as the paramount cue for intention estimation, which can help estimate the intention even before the manual action has started. In this perspective, what is an intention? What is its relation to action? How is this relation mediated by our visual and gaze control systems?

2.1 From intentions to actions

Intentions as mental states have long been considered as a key element in philosophical action theory. The philosopher Dennett proposed the *intentional stance* as the foundational perspective to predict others' behaviour and actions [Dennett 1989]. Ascribing to any rational agent beliefs, desires, and intentions, we are able to predict what they will do to achieve those intentions. The concept of intention is indeed strongly related to those of action generation and agency within the field of motor cognition and is at the core of action explanation [Pacherie 2012]. Early approaches [Davidson 1980; Goldman 1970] considered only beliefs and desires as causal antecedents, more specifically as mental causes of actions. However, action can be considered as a spectrum going from low-level motor primitives to carefully planned, deliberate actions. In this perspective, intentions were introduced, beside beliefs and desires, to account for future-directed intentions followed by no immediate action, as well as for instinctively, almost automatically performed actions that still denote a minimal purpose [Bratman et al. 1987]. Following this consideration, intentions have been assumed to have a number of functions and have been characterized according to their temporal relation to the consequent action. Specifically, before any actual action is initiated, intentions prompt plans and coordinate the plan activities in time [Bratman et al. 1987], while they further play the three critical functions of initiating, guiding and controlling actions up to their completion [Brand 1984; Mele 1992]. These distinctions have led to speaking of prior intentions vs. intentions-in-action [Searle 1983] or future-directed vs. present-directed intentions [Bratman et al. 1987] or distal vs. proximal intentions [Mele 1992].

In this line of thinking, Pacherie [2008] has proposed an even finer characterization: D-intentions, P-intentions, and M-intentions. This distinction relies on the consideration that distal intentions do not terminate their function once proximal intentions take their monitoring and guidance role. The three proposed intentions (distal, proximal,

Just Accepted

and motor) do instead work in parallel on different levels of representation content and at different time scales. In this framework, D-intentions define which goal to pursue and can be flexible regarding the time and context, leaving the planned action parameters and constraints rather unspecified. P-intentions receive an action plan from D-intentions and trigger off the start of the action, situating it in a specific time and context, relying on explicit perception, situation awareness, and thought. P-intentions, still, may act on higher-level action descriptions while M-intentions work with specific action primitives and motor representations (see Figure 2). The three levels of course operate on different time scales, being in the range of minutes and up for D-intentions, of few seconds for P-intentions, and of hundreds of milliseconds for M-intentions.

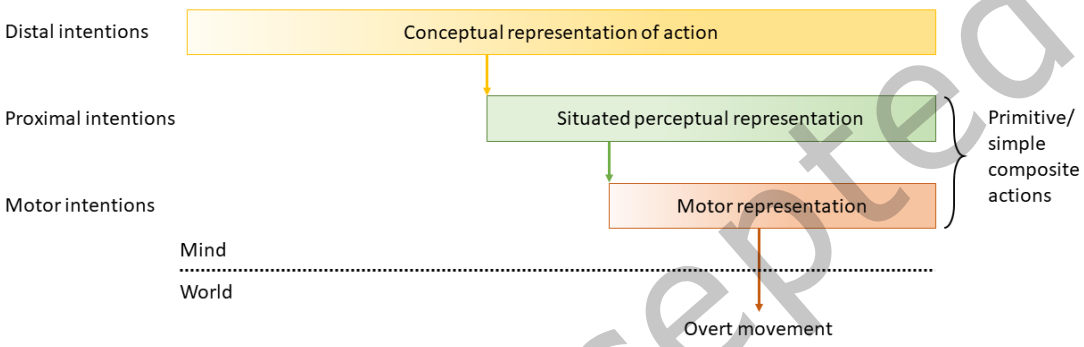

Fig. 2. A graphical representation of the cascade of intention levels, from abstract distal intentions to motor intentions directly translating into motor actions in a situated way (redrawn from [Pacherie 2008]).

These formalizations show that action emerges from plans and motor programs combined in a functional architecture relying on forward and inverse models. This means that any behavioral observation that can be related to the formulation of a motor plan entails the potential to reveal the action associated with the current proximal intention. In this sense, although a number of approaches have relied on motion cues to infer intentions [Aarno and Kragic 2008; Hauser 2013; Javdani et al. 2015; Tahboub 2006; Tanwani and Calinon 2017; Yu et al. 2005], an even earlier indicator, one which already displays during action preparation, is gaze control. Why is this the case? How do the visual and gaze systems mediate the transformation of an intention into motor action?

2.2 Gaze-based prediction and eye-hand coordination in natural tasks

Task strongly influences attention and consequently shapes eye movement patterns: in an early eyetracking experiment, Yarbus [1967] showed that presenting the same picture to subjects, asking them every time a different question (e.g., estimate the age of the pictured people, estimate their wealth), produced reliably different scanpaths. These days most of our visual activity might be concerned with passively processing information from texts, images, videos, and other media. Yet, our vision organs (and cognition altogether) actually developed and evolved in a moving organism that needed to act in the environment in order to adapt and survive. Even during a simple visual search, our performance is affected by action intentions and possibilities [Bekkering and Neggers 2002; Bub et al. 2013]. On a theoretical level, this was already postulated by the premotor theory of attention [Rizzolatti et al. 1987], arguing that spatial attention emerges as enhanced spatial processing at locations of possible actions, suggesting a tight relation and shared mechanisms for attention and action, specifically for

Just Accepted

selection-for-perception and selection-for-action [Deubel et al. 1998]. When we intend to act (or we need to react), that is the very moment our information-seeking perceptual behaviour must be most effectively tuned on the final goal. Action is indeed mostly sequential (we hardly coordinate movements with different goals at the same time) and it is relatively slowed down by inertia in execution. Thus, in order to maximize the probability of action success, motor planning needs to act on timely and certain information provided by perception. This corresponds to a visuomotor behaviour policy that strives to maximize the information gain and minimize uncertainty or surprise [Friston et al. 2015; Itti and Baldi 2009]. To do that, the visual system needs to rely on previous experience and knowledge, to anticipate the critical information to supply while preparing for and controlling a particular goal-directed action, in accordance with the current motor and subtask needs. Although this is apparent also in locomotion and navigation tasks [Einhäuser et al. 2007; Werneke and Vollrath 2012], it can be appreciated best in the interaction with objects. The advent of head-mounted eye-trackers in the last two decades has made possible the study of eye-hand coordination in an ecological way, with subjects performing daily tasks like making a sandwich [Hayhoe et al. 2003], preparing tea [Land et al. 1999], walking around ['t Hart and Einhäuser 2012] or practicing sports [for a thorough review see Land and Tatler 2009]. Specifically considering eye-hand coordination, [Johansson et al. 2001] looked at the timing correlations in the behaviour of gaze and hand when a simple wooden block had to be grasped, lifted, and carried to make contact with a target position, meanwhile avoiding a protruding obstacle.

A number of lessons have been learned from these studies, especially considering structured manual interaction with objects:

(1) for the most part only task-relevant objects are targeted by the gaze [Land 2006];
(2) objects are glanced at mostly right before they are needed for the task -"look-ahead" and "just-in-time" fixations [Hayhoe et al. 2003; Pelz and Canosa 2001]- and are further fixated until right before the hand grasps them or the task is completed. That is, the eyes might move ahead to the next sub-goal even if the motor task is not completely accomplished;
(3) the hand or the objects carried in the hand are rarely looked at [Johansson et al. 2001]. In general, an anticipation of the eye with respect to the hand (eye-hand span) of 0.5 up to 1 s is observed, which demonstrates that the gaze can be used to predict the next manipulation step.

This suggests that the way oculomotor and hand motor plans unfold during the execution of an eye-hand coordination (manipulation) task and with respect to the involved objects offers a substantial insight into the cognitive structure of action representation and planning, as orchestrated by intentions. Once we are set to achieve a specific goal, directions for a parallel and tightly interleaved execution of two behaviors (information seeking and object manipulation) are formed and delivered to the gaze and visual system, on the one hand, and to the motor system on the other.

Land [2009] proposes that four systems are involved in this process (see Fig. 3):

- The *schema control* sets the goal and the coarse plan to achieve it by defining the agenda of the necessary steps in the task sequence. Specifically, general directions about the objects of each subtask, where to find them, and the motor primitives necessary for the interaction need to be communicated to the other systems down the line.
- The *gaze system* encompasses control of each body part movement aimed at directing our eyes to the most likely locations where the objects to be acted upon can be found. Hence, body/trunk movements, head movements, and eye movements are all orchestrated by the gaze system[1].

[1]Even though one might argue that these movements are actually controlled by the motor system, in this case, such exploratory movements have an epistemic value (reducing the uncertainty) while the movements actually performing the action can be considered having a pragmatic value in reaching the goal state [Friston et al. 2015].

Just Accepted

- The *visual system* is responsible for the processing of the visual input and the recognition of the objects and context of the visual scene. It also monitors that the action unfolds successfully, achieving the goal of each task.
- The *motor system* physically executes the intended action, combining the available motor primitives (like reaching and grasping) with the contingent object configuration, as informed from the visual system.

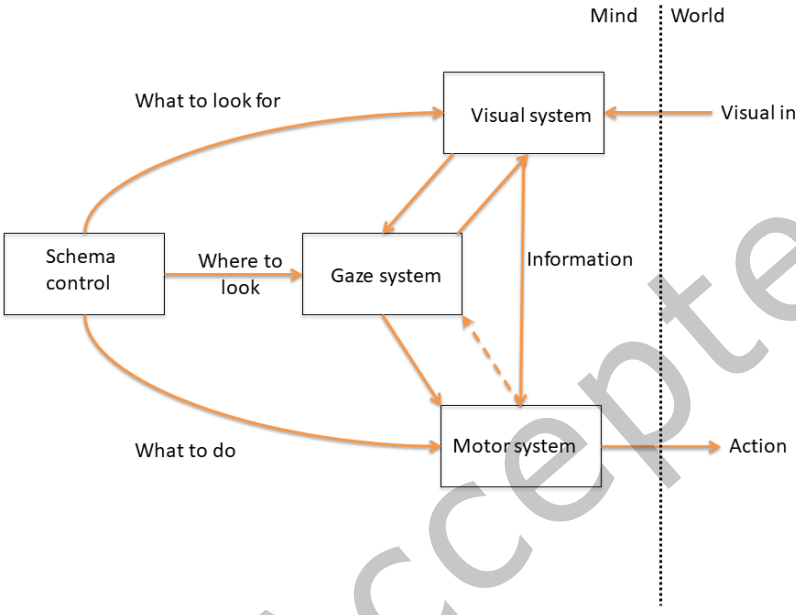

Fig. 3. Information exchange between the functional modules regulating perception and action: the relation between gaze control, visual processing, and action control is orchestrated by the schema system which sets the agenda for locations to look at, objects to look for, and movements to plan. The solid arrows show the information flow as sketched by [Land 2009], here redrawn; the dashed arrow represents a further information exchange suggested here and demonstrated in [Belardinelli et al. 2016a,b; Brouwer et al. 2009].

Clearly, these systems must work in tight concert with one another and information must be exchanged, especially between the gaze and visual system and between these two and the motor system. If, for instance, the object we are looking for is not detected by the visual system in the current field of view, the gaze system must know that further exploration is needed. Otherwise, if the target of the action was retrieved, the motor system must know where in the field of view and with respect to the rest of the body this is and how the object might be grasped. These information exchanges are represented in Fig. 3 by the solid arrows between the different systems.
The activity of the schema control system is behaviorally not observable, but its relation to the visual, gaze, and motor systems allows to draw inferences about its agenda by observing fixation locations, scan paths, and possibly hand movements and their temporal synchronization during a task. Not just what object is going to be grasped or used next can be inferred by fixations ahead of the hand reaching it. A number of studies have shown how the task, the intended grasp, and the user's experience with the object have an effect on fixations even before the hand reaches for the object [Belardinelli et al. 2016a, 2015, 2016b; Brouwer et al. 2009].
When planning a reaching-to-grasp movement, the eyes usually target the index finger contact location since this is usually the first finger to make contact with the object [Brouwer et al. 2009; Cavina-Pratesi and Hesse 2013].

Just Accepted

In general, the relevant affording points on the object are visually targeted, if possible also in parallel, hence either covertly or targeting a fixation point in the middle. In this way, a kind of attentional landscape is formed, enhancing perceptual processing in relation to the spatio-temporal action requirements [Baldauf and Deubel 2010]. That is, by observing the visual exploration of objects and the ensuing manipulation under different tasks and conditions, the diverse information requirements that are needed to plan and execute each task become apparent.

What these works additionally suggest is that a further arrow is missing from the schema above, namely information flowing from the motor system to the gaze system. That is, only once the motor plan for the appropriate grasp has been selected, the eyes can move to the most convenient location to guide the grasp. This exchange is denoted by the dashed arrow in Fig. 3.

The fact that the eyes precede the hand and satisfy its informational needs while the motor system communicates its kinematic needs to the gaze, all prior to or while initiating action execution, demonstrates how perception and action are substantially embodied and integrated processes, relying on learning and anticipation of the intended sensorimotor goal states.

What we have seen so far concerns the cognitive basis of visual attention and oculomotor control during grasping and manipulation. While still engaging eye-hand (and eye-foot) coordination, separate considerations must be done for a more complex activity such as driving. This is the domain of a number of applications that need to rely on timely intention estimation for human-machine cooperation. Gaze control, as seen above, works according to spatial predictions about where to find relevant information in the current scene [Henderson 2017], whereas these predictions are often made on the basis of episodic experience or knowledge. Coupled with top-down endogenous attention, this predictive ability makes vision particularly efficient and capable of focusing on relevant information within complex scenes in a matter of hundreds of milliseconds. Yet, at the same time we must be able to deal with unexpected stimuli, especially those not relevant to our current task but whose detection is critical to our and others' safety. This is particularly an issue, of course, when acting within dynamic environments, where an erroneous prediction requires a prompt reaction. Driving takes place in such environments and the scene to be scanned is usually very rich with information with different degrees of relevance. Driving is also inherently multitasking: it is not indeed just about steering the vehicle in the intended direction but it is made of different parallel subtasks (collision avoidance, navigation planning, lane keeping,...), each running with time-changing priorities, hence requiring continuous (re)scheduling of attentional resources allocation [Hayhoe and Ballard 2014; Rosner et al. 2019]. That is, if in tasks with a relative sequential structure, the gaze behaviour can be considered as a running commentary of the current cognitive agenda and can reveal to a certain degree the present subtask, fixation choices in a multi-tasking behaviour need to be made not according to a linear unfolding of the task, but depending on competing modules, which are to be periodically updated with fresh information, especially when the world state cannot be assumed to be static. To complicate the picture further, scanning behaviours might change depending on the context (urban vs. highway or country driving), on the current manoeuvre to be undertaken (overtaking or crossing an intersection), and on the current speed. These factors not only influence gaze location, but the whole visuo-motor control system. In urban traffic, the gaze might shift every 0.5 s, alternating between the car in the front, lane edges, pedestrians, road signs, traffic lights, or generic obstacles [Land and Tatler 2009]. Still, it is hard to make generalizations on gaze distribution independently of the situation at hand.

Particular attention has been given to gaze behavior during lane changing, a situation that requires tight coordination of situation awareness, decision making, and motor control because of the interplay with other moving vehicles. Salvucci and Liu [2002] looked at gaze dwell time in such case and found that if in the beginning (lane keeping) the gaze mostly lingers on the start lane, when changing from right to left lane, as the change time approaches, more time is spent on the mirror to check incoming vehicles and on the end lane. When changing

Just Accepted

from left to right instead, glances at the mirror increase shortly before lane change, but during and after lane change the gaze spends most of the time on the destination lane. The authors remark how the gaze, anticipating the manoeuvre, within 1-2 s of the onset of lane change is already concentrated on the destination lane, although the vehicle is actually still entirely in the start lane. All these results show that a number of factors can influence the way our gaze is controlled when driving [see Kotseruba and Tsotsos 2021, for a thorough review]. This could make the assessment of the driver's intentions particularly tricky, since the process that determined a series of fixations (intentional retrieval of information or reflexive capture of attention) might not be easily pinpointed. Still, a combination of behavioural cues could provide a much richer picture (see Section 4.3).

As a final remark, even if intentions, as we have seen, can have multiple temporal granularities, ranging from distal to proximal to primitive motor intentions, the studies above however show that 'just-in-time' fixations in physical actions denote an action planning policy according to which visual information about the immediate next action is retrieved right ahead (i.e., in the order of hundreds of milliseconds before starting the action). After that, the eyes stay anchored on the object of interest as long as visual guidance for manipulation is needed. In driving scenarios, although the time span between the eyes collecting information for a manoeuvre and the hand initiating the manoeuvre is usually longer, it still stays in the range of 1 to 5 seconds. Such time ranges hardly help retrieve long-term, distal intentions.
For these reasons, in the following we concentrate preeminently on models and systems handling proximal and motor intention recognition, driving single, often primitive, actions. We aim to show how indeed such cases represent thus far the most compelling use of gaze for intention recognition and how such intentions can be effectively retrieved by means of easily computed features, as presented in the next section.

3 GAZE FEATURES AND MODELS FOR INTENTION PREDICTION

Before looking at how the gaze has been incorporated in different application domains, we give a brief overview of the gaze and eye movements features and measures that are mostly used and of the machine learning models that can make the most out of them, depending on the goal. Although there are a number of studies working with electro-oculography [see Bulling et al. 2010, for a review] or using just optical flow features caused by head movement [e.g., Kitani et al. 2011] or gaze direction extracted by external cameras [e.g., Dermy et al. 2019], we will focus here mostly on gaze- and eye tracking data collected with head-mounted devices, as this technique provides a more accurate position of the gaze and localizes it with respect to the scene, hence linking the eye movement patterns to the specific object or location of interest at any time.
Different abbreviations which will be used in the following are reported in Table 1.

3.1 Gaze features: position, movement, sequence, and semantic features

Without referring to any particular hardware, every eye tracking device delivers as raw data the eye(s) position as 2D coordinates on a calibrated surface or in the scene in terms of frame coordinates with respect to the streaming of a head-mounted scene camera. For many applications considering physical interaction with objects in the world, it might be necessary to localize the point of regard (POR) in 3D, and specific methods have been proposed for that, often relying on fiducial markers [e.g., Haji Fathaliyan et al. 2018; Huang and Mutlu 2016; Takahashi et al. 2018] but new solutions are trying to do without such aides [Liu et al. 2020; Weber et al. 2018].
In any case, once gaze coordinates are obtained, these can be used as such or to compute gaze events of interest, usually fixations and saccades, by means of specific detectors. Fixations are periods of time in which the gaze is relatively stationary on a certain point, while saccades denote fast displacements of the POR between fixations. Fixations are generally in the order of 200-300 ms while saccades can take between 30 and 80 ms. The reader is referred to Holmqvist et al. [2011] for a more thorough classification of eye movements. During saccades the

Just Accepted

Acronym	Meaning	Acronym	Meaning	Acronym	Meaning
ADAS	Advanced Driving Assisting Systems	**(G)HMM**	(Gaussian) Hidden Markov Models	**MAP**	Maximum A Posteriori
AOI	Areas Of Interest	**GMM**	Gaussian Mixture Model	**POMDP**	Partially Observable Markov Decision Process
AUC	Area-Under-The-Curve	**HMI**	Human-Machine Interaction	**POR**	Point Of Regard
CRF	Conditional Random Field	**HRI**	Human-Robot Interaction	**RBF**	Radial Basis Function
(D)BN	(Dynamic) Bayes Network	**k-NN**	k-Nearest Neighbors	**(R)NN**	(Recurrent) Neural Networks
DTW	Dynamic Time Warping	**LDA**	Linear Discriminant Analysis	**SVM**	Support Vector Machine
(GB)DT	(Gradient Boosted) Decision Tree	**LSTM**	Long Short-Term Memory	**VR**	Virtual Reality

Table 1. Abbreviations used in text.

visual input is blurred and hardly any visual processing can take place, hence often just fixations are considered when analyzing eye data. To this end, fixations can be detected via dispersion- or velocity-based algorithms [Salvucci and Goldberg 2000], which are yet based on predefined thresholds. In more recent years, some machine learning-based event detection schemes have been proposed, which achieve a higher accuracy [Tafaj et al. 2012; Van der Lans et al. 2011; Zemblys et al. 2018]. Still, when a system is working online, as needed for real-time prediction, it can make sense to use all gaze samples in order not to have an online fixation detector processing data before intention classification [Coutrot et al. 2018; Fuchs and Belardinelli 2021]. Simply using the PORs indeed accounts for both the gaze location and temporal distribution on the object of interest [as in Keshava et al. 2020].

For activity recognition or task prediction, in some cases, the pattern of gaze shifts can be used independently of the underlying visual input [as in, e.g., Bulling et al. 2010; Ogaki et al. 2012]. Such patterns still are usually not relatable to single motor intentions. Often, fixation and saccade statistics (such as fixation duration or saccade amplitude means and variances) along with the statistics of image features at fixation landing points (e.g., Histogram-Of-Gradients, saliency or gist features) are used to infer the current epistemic task [see Boisvert and Bruce 2016, for a review].

However, when considering the prediction of a physical interaction, higher-level - somehow more semantic - features need to be considered. Although gaze distribution on images or recurrent patterns of directional saccadic shifts can be revealing of specific repetitive activities, these features are hardly relatable to a single action intention (although, they can be revealing at least of the intent to interact [David-John et al. 2021]). The next proximal intention is indeed often revealed by the current object of interest and even by a specific affordance on that object [Belardinelli et al. 2016a, 2015], independently of their position in the camera image or of specific low-level features at fixation points. The relevant objects in the scene should hence be defined as Areas-Of-Interest (AOI) and any fixation within these areas gets considered for processing, with related statistics. In this case, AOI-specific measures can be derived, e.g. the number of fixations within each AOI, the time of entry in each AOI and the overall dwell time. More importantly, considering the temporal evolution of intentions and actions, sequences of glanced AOIs can be used as time series of discrete symbols, revealing the shifting of attention between different

Just Accepted

objects relevant for the current activity, as for example in Haji Fathaliyan et al. [2018]. Further, it has long been known in vision science, that human saccades often undershoot (and sometimes overshoot) their target by up to 10% depending on visual eccentricity [Becker and Fuchs 1969]. Hence, to account for such inaccuracy or for the strategic use of parafoveal vision, instead of assigning at any time the gaze in full to just one object, it can makes sense to represent the POR as a Gaussian distribution [e.g, as in Admoni and Srinivasa 2016; Fuchs and Belardinelli 2021] possibly intersecting different AOIs or time bins of gaze counts are considered in which a discrete distribution over AOIs is computed [Lengyal et al. 2021]. Alternatively, if objects of interest are well separated, one might consider larger AOIs to capture undershooting fixations and assign them to the intended object.

AOIs are often used as semantic inertial areas within vehicles. While estimating the current object of interest in a dynamic scene with ego-motion is technically quite complex [but see Jiang et al. 2018, for a recent solution], intentions like lane keeping or lane changing can be assessed considering which part of the field of view within the car the driver is currently attending: often the windscreen, the rear and lateral mirrors and the the dashboard are taken as AOIs.

In any context, sequences of AOIs can be considered as strings where typical, recurrent transitions lend themselves nicely to be learnt by Markov models, as discussed in the next subsection.

3.2 Prediction models

Typically, intention recognition systems feed gaze features considered indicative of the current intent into some classifier outputting the most likely intention. For this purpose often Support Vector Machines (SVM) [Huang and Mutlu 2016; Kanan et al. 2014], random forests [Boisvert and Bruce 2016], logistic regression [David-John et al. 2021], and neural networks [Lethaus et al. 2013] have been used. In some cases, like task prediction in viewing images [Henderson et al. 2013] or driving manoeuvres [Lethaus et al. 2013], a Naive Bayes classifier produced a satisfying intention recognition, yet most established approaches use more sophisticated generative models. Indeed, as mentioned earlier, for intention prediction we want to estimate a cognitive state which is *per se* not observable, that is, a hidden state. For this reason intention recognition has often been modeled via probabilistic graphical models, like Dynamic Bayesian Networks and Hidden Markov Models [Bader et al. 2009; Boccignone 2019; Fuchs and Belardinelli 2021; Haji-Abolhassani and Clark 2014; Tahboub 2006; Yi and Ballard 2009]. In general, if the intention I represents the hidden state and what can be observed (the eye features of interest) is denoted with E, intention recognition can be typically defined in Bayesian terms:

$$P(I|E) = \frac{P(E|I)P(I)}{P(E)} \tag{1}$$

If we have a model M_I for each intention and a sequence of features $E_{0..t}$, inference of the current intention I_t is given by the model which maximizes the posterior of the observed data:

$$I_t = \underset{I}{\operatorname{argmax}}\, P(M_I|E_{0..t}) = \frac{P(E_{0..t}|M_I)P(M_I)}{P(E_{0..t})} \tag{2}$$

A number of recent approaches have also used recurrent neural networks such as Long Short Term Memory (LSTM) [González-Díaz et al. 2019; Min et al. 2017; Schydlo et al. 2018; Wang et al. 2020] and Partially Observable Decision Markov Models (POMD) [Admoni and Srinivasa 2016], which are again algorithms that deal well with time series of different length.

In general, it makes sense to use models that account for the temporal evolution of the scanpath on the scene. To consider just the current fixation or POR, in fact, would produce a very erratic prediction whenever the gaze lands on anything which is not strictly related to the current intention. Even the approaches relying on discriminative classifiers mentioned above do not consider gaze features at a specific point in time, but often

Just Accepted

compute some summary statistics within a certain time window. By doing so, a larger time history is taken into consideration, still the temporal information is lost. Generative models also rely on data observed in a certain time window but they do incorporate an internal representation of the task evolution. This makes them also more robust against spurious fixations. Still, the time window to consider must be carefully chosen in order to capture the time scale of the intentions to be estimated. For reaching, grasping, and simple manipulations a time window between 0.5 and 2 s is usually delivering best results [Fuchs and Belardinelli 2021; Huang and Mutlu 2016], while in driving scenarios a longer time frame is usually considered, between 5 and 10 s [Lethaus et al. 2013].

Further, while Hidden Markov Models can handle sequences of different lengths in a relatively flexible way, sometimes time series need to be reduced to comparable lengths in order for a distance or similarity measure to be computed. This problem has been mostly tackled by dynamic time warping [Haji Fathaliyan et al. 2018; Wu et al. 2019] or by encoding sequences in a fixed length representation by means of an encoder-decoder LSTM architecture [Schydlo et al. 2018].
Considering both the oculomotor aspects of eye-hand coordination illustrated in section 2.2 and how those are tuned by the visuomotor system for the task at hand might give important indications about the choice of the most appropriate model. On the one hand, HMMs or BNs offer a model-based structure with states that usually have a concrete meaning, i.e., a current state of the system or phenomenon of interest. Such states might represent an action/subtask [Yi and Ballard 2009] or the real targeted object as opposed to distractors [Fuchs and Belardinelli 2021] and the model designer can choose such parameters by considering the visuomotor and semantic characteristics of the task. Yet, such models often come with strong assumptions, e.g., the Markov property according to which each state depends only on the previous state. On the other hand, longer time dependencies might occur in the data, which are then better captured by RNNs, especially in large datasets. These models are yet more difficult to tune and, indeed, might require large labeled datasets.
Finally, a further aspect to consider is that in many situations instead of a classification scheme returning just the winning intention, it is better to have a distribution over all possible intentions. The selection can then happen by means of an argmax function (as shown in Equation 2), still a measure of prediction confidence can be given, for example by taking the difference between the two most likely intentions [Jain and Argall 2018]. In this way the following assistance can be tuned more or less strongly, depending on how sure the system is about the estimated intention. In the following section, we illustrate which of the presented techniques and features have been prevalently used in multiple applicative fields, depending on the use case and nature of the intention to recognize.

4 GAZE-BASED INTENTION ESTIMATION IN TECHNICAL DOMAINS

As anticipated in Figure 1 there are three main fields where gaze features have found prominent use for intention recognition. In Computer Vision, often intent has been recognized *a posteriori*, offline, that is, inferring the epistemic task offline from collected data. We do not consider such applications and refer instead the reader to [Boisvert and Bruce 2016]. While some more interactive applications could be envisioned for video games, VR, or any further application of Extended Reality (XR), often in the domain of human-machine interaction we see intention recognition as activity/task classification or item selection, relying on features related to dwell time or overall fixation distributions. Conversely, when prompt reaction to or support for human real-world activities from robotic systems, such as teleoperated arms or driver-assistance systems, AOIs related to the object of interest and time series acquire prominence, with the earliness of recognition being an important parameter. Considering the main lessons learned from gaze shifts in real-world activities and in manual object interaction, we will show in the following how these have been often translated into technical recommendations, that is, referring to the same order as in subsection 2.2:

Just Accepted

(1) AOIs (or fixation distributions) should be defined only for objects relevant to the intentions of interest, excluding background surfaces or objects that can be incidentally fixated, but disrupt the typical fixation pattern.
(2) The time window for collecting the features of interest (either in aggregated or time-series form) should be tailored to the typical task dynamics.
(3) While the hand is rarely fixated, it might be reasonable to consider it as a gaze target to account for and measure deviations from natural eye-hand coordination, e.g, during teleoperation, in rehabilitation, or in HMI scenarios.

We survey here some paradigmatic examples of systems across the three domains. We do not aim at an exhaustive review, but at offering to the reader a wide picture of use cases, methodologies, and learned lessons as emerged in the literature. We considered the following inclusion criteria: 1) peer-reviewed papers in English; 2) prominent use of gaze-derived features for intention estimation, possibly in combination with other behavioral features; 3) prediction or estimation of upcoming or ongoing physical actions, that is, an action producing effects in a physical or virtual environment, as opposed to epistemic task recognition and to offline activity recognition; 4) the use of eye-tracking to extract the measure eye movements, as opposed to electrooculography or just head orientation. Google Scholar, ScienceDirect, IEEExplore, and ACM databases were searched with the following kewords: gaze (OR eye movements) + intention (OR intent) recognition (OR prediction OR estimation) + robot. In some cases, relevant related references were derived from bibliographies of the selected papers.

4.1 Computer Vision and Human-Computer Interfaces

While a number of studies have looked at the classification of gaze features during information retrieval or screen-related activities using an egocentric perspective, as said, these approaches are more concerned with overarching activity recognition, hence for this we refer the reader to previous surveys [Boisvert and Bruce 2016; Nguyen et al. 2016]. Here we consider preeminently simple action or proximal/motor intention recognition, something particularly useful in digital environments and interactive games. In such cases, interaction possibilities might be limited or discretized according to the input control device, but in some cases even just the detection of the intention to interact offers valuable information [David-John et al. 2021]. Chen and Hou [2022] reasoned that often in VR the only actions available to users are item selection and teleportation by means of hand-held controllers. In the study, different classifiers were compared to identify, on the one hand, the generic intention to interact or not and, on the other hand, the two mentioned intentions or an idle intention. Features derived by descriptive statistics about gaze and pupil size yielded already good performance, while case-specific eye-hand coordination feature (the distance between the POR and the intersection point of the controller ray with the scene) improved classification only slightly. As said, given the goal-driven nature of our cognitive sensorimotor system, semantic features related to the scene content might help further in multinomial classification scenarios.
Considering more precise actions, such as block alignment, Keshava et al. [2020] have shown that the sheer proportion of PORs on different AOIs constitutes already a reliable signature, discriminating among different tasks. In this case, different AOIs on the same object were considered, as affording points for different interactions, similarly to the more semantic regions considered in natural object manipulation [**?**]. Previously, intention recognition in manipulation of virtual objects for pick-and-place on a large tabletop screen had already been proposed by Bader et al. [2009]. There a Bayes net was used to predict intentions on the basis of proactive fixations in order to improve the robustness of the input device or increase interaction possibilities.
On the other end of the spectrum, an approach aimed at longer-term plan recognition was presented by Singh et al. [2020]. In a digitized version of a board game considering routes in a graph form, the number and duration of fixations on any AOIs determined scores used as priors over possible trajectories leading to distal intentions. The authors show that by means of such gaze priors both proximal and distal intentions could be inferred earlier

Just Accepted

Study	Task	Features	Recognition method	What is recognized
[Chen and Hou 2022]	Object selection and teleporting in VR	Gaze/Pupil/eye-hand statistics	LR, RF, GBDT, SVM	Intent of interacting, interaction type
[Keshava et al. 2020]	Aligning cubes in VR	POR distribution in AOIs	Radial-basis SVM	Type of alignment
[Bader et al. 2009]	Virtual pick and place on screen	Fixation location and reactive/proactive behaviour	Bayes Net	Picking/placing target
[David-John et al. 2021]	Ingredient fetching in VR	Multiple features from gaze dynamics	Logistic regression	Onset of interaction
[Singh et al. 2020]	Board game *(Ticket to ride)*	Number and duration of fixations on AOIs	Normalized fixations weights	Route to target destination
[Min et al. 2017]	Open-world educational videogame	Fixations on virtual objects	LSTM + CRF	Goal/Task in the gameworld

Table 2. Approaches to task and intention recognition in computer vision and human-machine interaction.

Just Accepted

and more accurately by means of AI planning. In cases like this or for the approach by [Yi and Ballard 2009], gaze can be used for distal intention recognition if the structure or segmentation of the overall task (the plan) is known to the system.
In the field of digital games plan and goal recognition by means of gaze would help tailor the gameplay to the intentions of the player, showing personalized narratives or interaction with the non-player characters. [Min et al. 2017] showed how combining gaze and game events in an open-world game helped recognize the player's high-level goal (e.g. speak to a character) with an LSTM architecture better than only with game event features. The studies presented in this section are summed up in Table 2. All these studies are of interest for the approaches they present with respect to intent recognition, still most set ups do not involve cooperative, shared tasks, nor the user trying to control a physical system with effects in the real world, where timely intention recognition is critical for smooth HRI. Works related to this theme are reviewed next.

4.2 Human-Robot Interaction

Intention recognition finds probably its most appropriate domain of application in robotics since in this field a physical system can become aware of what goal the human is pursuing and proactively provide its support, by acting in the same environment (Section 4.2.1) or in a remote/not accessible one (Section 4.2.2).

4.2.1 Human-Robot Cooperation and Social Robotics. For human-robot collaboration often the robot partner is aware of the activity context and for effective cooperation it just needs to detect the current action intention of the human partner to help them with it. In one of the first attempts at using eye tracking in a human-robot cooperative scenario, Sakita et al. [2004] proposed a system helping a human in an assembly task. The robot would hand the next piece to assemble if it had confidently detected that as being the next step intended by the user or directly execute the step in parallel with the human. The gaze is apparently localized in the real world setup but very few details were given as to the implementation and accuracy.

Study	Task	Eye Features	Recognition method	What is recognized
[Haji Fathaliyan et al. 2018]	Powdered drink preparation	Gaze object sequence in sliding window	DTW Euclidean Distance	Subtask
[Huang and Mutlu 2016]	Smoothie preparation	Nr. of glances, duration of 1st fixation, total dwell time, if last glanced item	SVM	Intended ingredient
[Sakita et al. 2004]	LEGO assembly	Fixation time, fixations on functional points	Rule-based, voting scheme	Type of assembly
[Schydlo et al. 2018]	Placing, giving, eating, pouring	Gaze location	LSTM	Action and location

Table 3. Approaches to task and intention recognition in human-robot cooperation

In most cases, indeed, the user's gaze is mapped on the 2D scene acquired by the eye tracker frontal camera. This is rather inconvenient when investigating interaction with real 3D objects. To tackle this issue, Haji Fathaliyan et al. [2018] propose a method to localize gaze on 3D objects by projecting the gaze vector on point cloud representations of objects manipulated by a person preparing a powdered drink. 3D heat maps were produced, displaying the most gazed locations on each object depending on the performed subtask.
In some cases, the gaze is used to identify the object the user intends to select, given a predefined action. For example, Huang and Mutlu [2016] proposed a method for anticipatory control under which the robot predicts the intent of the human user and plans ahead of an explicit command. In the considered task, a robotic arm prepares a smoothie by picking the ingredients vocally selected by a human user looking at an illustrated list. By means of eye tracking the robot anticipates picking the intended ingredient, hence speeding up task completion. The intent recognition relies on the same features and training set as in a previous work [Huang et al. 2015], where an SVM was fed a feature vector for each ingredient, including the number of glances, duration of the first glance, total duration, and whether it was the most recently glanced item as predictors of the currently intended ingredient. Although such an approach is simple and effective, here the human customer is using their gaze purposely and solely for target selection, with no parallel visuomotor control task that could yield spurious fixations.
Considering action sequences of different lengths, Schydlo et al. [2018] proposed an architecture relying on gaze and body pose features fed to an encoder-decoder LSTM network, with the decoder layer expanding the K most probable action sequences. Training and testing were done on available datasets containing either placing and handing actions [Duarte et al. 2018] or eating and pouring action sequences. Gaze features improved the results obtained with only body pose, both in terms of accuracy and earliness of prediction.
The approaches reviewed for Human-robot cooperation are summed up in Table 3.

4.2.2 Shared autonomy grasping systems: assistive technologies and smart teleoperation. Robotic teleoperation in manipulation tasks aims at separating perception and action, by displacing the effects of motor action in a remote position or environment, still exploiting the precise eye-hand coordination of the operator. In recent years a number of intention estimation approaches have been developed for shared autonomy systems, specifically targeting assistive and teleoperation scenarios. Shared autonomy combines robotic and human control in the execution of remote tasks. Each partner is then tasked with the aspects for which they are better skilled: the lower

Just Accepted

kinematic aspects of action execution are usually left to the robot while higher-level cognitive skills, like task planning and handling unexpected events, are typically concurrently exercised by the human. This can happen with different degrees of autonomy for the robotic partner [Beer et al. 2014; Goodrich et al. 2013; Schilling et al. 2016]. Often there is a large asymmetry in terms of degrees of freedom or kinematic capabilities between the user input controller (such as a joystick) and the robotic effector, thus, shared autonomy can facilitate control, in so easing the operator's cognitive load and speeding up execution. The user is setting the goals and the ways to achieve them, hence this cooperative interaction cannot be brought about without the robot first inferring the current human intention (*intent prediction*) and only then deciding how much to assist with execution (*arbitration*).

Losey et al. [2018] review a large literature on shared control for physical HRI, specifically in the context of rehabilitation. The unifying view they propose in order to look at any system sharing task execution with a human counterpart states that the design of any system physically assisting a human should take care of three aspects:

- **Intention Recognition.** Intention recognition should happen as early and as naturally as possible in order for the user to be relieved of explicitly directing the robot and for the robot to timely initiate the assisting action. A number of approaches have been proposed that rely on *predict-then-act* policies [e.g., Dragan and Srinivasa 2013; Yu et al. 2005], predicting one goal and executing a corresponding plan, but a more flexible solution would compute a distribution over all possible goals and rely on continuous reassessment of the predicted intention as the action unfolds, in so providing on line correction and continuous adjustment [Hauser 2013; Javdani et al. 2015].
- **Arbitration.** Following intent detection, a policy must be designed to arbitrate to which degree the user input and the robot assistance are merged and executed. This can rely on a predict-then-blend policy, namely $Arbitration = (1-\alpha)\cdot UserInput - \alpha \cdot RobotPrediction,\ \ \alpha \in [0,1]$ [e.g., see Dragan and Srinivasa 2013], or a smoother assistance policy computing a probabilistic distribution over all possible user's goals and minimizing the cost of achieving the user's goal [Javdani et al. 2018, 2015].
- **Communication/feedback.** In most scenarios, a robotic system is substituting the motor functions of a human arm and in some cases, its sensors are also used to communicate back to the user properties of the environment.

In light of these premises, the possibility to infer manipulation intentions via gaze tracking has become an utmost appealing approach, especially in the context of assistive technologies. Here, the intended users are people with physical impairments or elderly for which full teleoperation with a joystick controller is a further challenge. Similar issues to some degree exist also in complex teleoperation scenarios, where the user needs to exert dexterous control in cluttered and possibly dangerous environments (e.g., handling nuclear waste or bomb defusing). Indeed, even in cases where the own hand is used to control a similarly anthropomorphic manipulator, there are low-level kinematic aspects that the robotic part can take care of, in stead of the user, if the right intention can be timely predicted. For example, the best grasp and grasping location can be automatically selected by the robot depending on the current hardware and the target object.

Some studies first looked at the theoretical formulation of gaze-based intention estimation. As a first attempt at integrating gaze input from the user, Admoni and Srinivasa [2016] presented a proposal relying on Javdani's framework [Javdani et al. 2015] where the probability distribution over the goals (hidden states) is updated by considering both user's eye movements and joystick commands as observations in a POMDP, using hindsight optimization to solve it in real time. Assuming the position of the manipulated objects to be known (by using fiducial markers), the probability of a certain goal given the current image is considered proportional to the distance between the current gaze position and the target object. Considering exclusively grasping actions,

Just Accepted

Study	Task	Features	Recognition method	What is recognized
[Admoni and Srinivasa 2016]	Pouring	Gaze distribution on AOIs	POMDP	Grasp target
[Razin and Feigh 2017]	Pick and place	Gaze location and velocity	LDA, decision trees, k-NN, HMM	Action ('move', 'ready', 'hold')
[Webb et al. 2016]	Grasping of a tennis ball	Gaze location and motion commands	Potential fields	Target location
[González-Díaz et al. 2019]	Object grasping	Aligned gaze points, object category probabilities	LSTM	Object to be grasped and when
[Li et al. 2017]	Grasping	Gaze 3D location	GMM	Grasp target and related contact point
[Li et al. 2020]	Grasping for manipulation	Gaze and motion	BN, SVM, NN	Grasp target and next manipulation (transfer, usage, handover)
[Shafti et al. 2019]	Pick and place, pouring	Gaze dwell time on object	Finite state machine	Primitive action (reach, grasp, pour) + object
[Jain and Argall 2018, 2019]	Pick and place	/	Bayesian filtering	Target object
[Wang et al. 2020]	Make instant coffee, make powdered drink, prepare cleaning sponge	Current AOI, AOI sequence, gaze-object angle/speed	RNN	Action primitive, target object
[Fuchs and Belardinelli 2021]	Pick and place	Time series of gaze distribution on AOIs	GHMMs	Action and target object
[Novak and Riener 2013]	Virtual recipe cooking	Gaze dwell time, hand velocity	Individual threshold on features	Target object

Table 4. Approaches to intention recognition in shared control systems

Just Accepted

[González-Díaz et al. 2019], used object recognition via Convolutional Neural Networks to extract the current active AOI, while LSTM predicted the onset of the intention to grasp.
In a more typical shared autonomy context, Jain and Argall [2018] put forward a mathematical formulation of intent recognition in a recursive Bayesian filtering framework. Assuming a finite number of available goals, the probability distribution over these is computed as the posterior conditioned on a number of observations (gaze, motion input commands,...) and depending on the previous goal estimation in a Markov model. The Maximum a posteriori (MAP) is then considered as the current goal, with confidence dependent on the difference between the first and the second most likely goal. Such measure then regulates the amount of assistance provided. Overall,

this study shows the importance of identifying the right trade-off between early assistance and high confidence in the predicted user's goal.
In a further work [Jain and Argall 2019], the same mathematical formulation is extended with a user-customized optimization of the rationality model, in so accounting for human action suboptimality. Still, in both works gaze is mentioned as a possible input to the system but not practically integrated and tested.

Most shared autonomy approaches relying on eye tracking for intention recognition thus far are rather recent and often do not entail all three components presented above. Such systems usually focus on the first component or subcomponents of it, like gaze 3D localization or investigation of gaze patterns in shared autonomy contexts. The following studies considered the characterization of gaze behavior during teleoperation and how it could be learned for intention estimation.
For an assistive robot arm spearing food bits from a plate to feed an impaired user, Aronson et al. [2018] presented a preliminary eye-tracking experiment comparing user behavior within-subjects in different operation modalities: full teleoperation, complete autonomy, shared autonomy according to Javdani et al. [2015], blend-policy autonomy as in Dragan and Srinivasa [2013]. Two patterns of fixations were observed: monitoring glances, meant to check the translational behavior of the arm approaching the intended food morsel, and planning glances, which select the target morsel before starting the arm actuation, as in natural eye-hand coordination. Yet, an effect of operation condition on the frequency of this kind of fixations was not found but some patterns of repeated monitoring glances were observed, as the gaze was switching between the effector and the target morsel. Still, eye data were further not really localized with respect to the scene content, and analyses were focused just on gaze dynamics, hence limiting the conclusiveness of some results. The lack of localization of the gaze on the object of interest and the possible drop in data quality was further addressed in a subsequent study [Aronson and Admoni 2019], with semantic labeling, while the same group also proposed to use gaze tracking for error detection during manipulation [Aronson and Admoni 2018].
3D gaze localization and its use in intention prediction was investigated by Li et al. [2017]. In this case, a visuomotor grasping model also predicts how a user intends to grasp the chosen object. This latter feature still is formalized for just one cuboid object, resulting in the inference of one contact point and direction of approach of the hand, depending on where the cuboid was fixated, as learnt from a subject physically grasping the object in a training phase. Recently, relying on the approach first put forward by Haji Fathaliyan et al. [2018], Wang et al. [2020] proposed a framework for action primitive recognition by means of 4 features: the currently gazed object, a gaze object sequence complemented with the gaze object angle (angular distance between the gaze vector and the vector to each object) and the gaze object angular speed. Four action primitives (reach, move, set down, manipulate) were envisioned in the context of 3 activities (make instant coffee, prepare a powdered drink, prepare a sponge for cleaning). The proposed system based on a recurrent neural network achieves good recognition accuracy and observational latency, that is a correct prediction was available on average after 10% of the action was observed. Although aimed at shared autonomy in assistive contexts, the system was trained on natural eye-hand coordination, which in the actual scenario might be strongly mediated and influenced by the visual input and by the robotic actuator. Fuchs and Belardinelli [2021] partially addressed this issue by presenting the scene in a VR headset and letting the user teleoperate a rendered robotic hand by means of natural arm movements (tracked via Vive controllers). The sequences of gazed AOIs during a pick-and-place task were used to train two action-specific HMMs able to generalize their accuracy to new users, number of objects, and object configurations. Still the system was not yet tested in a shared control scenario.

As in the previous section, a number of studies combined gaze and motion features, in the effort to improve prediction and increase robustness. Razin and Feigh [2017] tested a number of modalities, features, and classification techniques to predict intent during robotic eye-hand coordination. Hand motion, eye, and gaze features are used

Just Accepted

to predict 3 actions ("move", "ready to move", and "hold") in a scenario where a robotic claw is used for pick and place tasks. In such simplified conditions, however, the authors find that motion features suffice to predict the three tasks, with eye and gaze features achieving good levels of prediction but yet not improving performance when considered together with motion. Decision trees, k-Nearest Neighbor, and Hidden Markov Models achieved the best performance in terms of average accuracy and Area-Under-the-Curve (AUC). Still, when a cluttered scene is presented to the user, the gaze provides more precise and earlier information with respect to motion trajectory features, which in the beginning might be compatible with multiple targets, as shown by Aronson et al. [2021] for motor control with a joystick and by Belardinelli et al. [2022] for natural movement control in VR.

In the works reviewed thus far, only the immediate action and object of interest are predicted.
In a more recent study, Li et al. [2020] considered whether besides the target object also the manipulation intent can be recognized before the reaching motion is completed, in order for the robot to select the best grasp for the following task. In this study, hand tracking is used to control a robotic arm with a three-finger gripper. Since the same grasp could lead to different tasks (and sometimes the same task is still executed with different grasps) the whole problem is treated in a multi-label classification framework. The intent inference is modelled via Bayesian Network (and for comparison SVM and NN) computing the MAP likelihood of an intent given an object, on which the gaze is lingering, and a grasping configuration. Experiments were conducted with 4 objects and 3 tasks, with the training set collected in a human grasping experiment and testing conducted on both real grasping sequences and teleoperated sequences. Again, unlike in Razin and Feigh [2017], gaze information seems to increase performance in the teleoperated case, less so in the real grasping case.

Other systems present some form of intent detection already integrated in a shared autonomy framework. Webb et al. [2016] consider a remote teleoperation scenario and a hybrid joystick-gaze control strategy to overcome the hampered depth perception when manipulating objects via screen feedback, by restoring a kind of more natural eye-hand coordination. "Look-ahead" fixations are used to predict the target of the movement and a Gaussian potential field is used to combine this input with the direct joystick commands, in a predict-then-blend framework. This strategy resulted in fewer grasp attempts and shorter execution times and was more liked by the participants compared to a joystick-only control. Yet, in this system, the gaze was actively used to indicate the target. To discriminate between voluntary and involuntary gaze shifts a button press on the joystick was required when looking at the target object, while the grasping was also triggered via button press, reducing the claim of natural eye-hand coordination.
Gaze-based intention estimation has found large application also in rehabilitation robotics. Usually, a robotic arm is attached to the patient's arm to support reaching and manipulation exercises during physical therapy, e.g. after a stroke. For example, Novak and Riener [2013] used dwell time and intentional movement initiation as features in a recipe cooking task in a virtual environment. The robotic arm there would support the trajectory to the intended ingredient. While enhancing the patients' freedom of movement and helping to restore natural eye-hand coordination, the system uses basic features and no trained recognition system, hence in cluttered conditions the wrong object was at times selected. A more elaborate assistive framework for rehabilitation is presented in Shafti et al. [2019]. The system here relies on 3D gaze localization, convolutional neural networks for object recognition, head tracking, and a finite state machine (FSM) implementing a basic action grammar according to the predefined 'graspability' and 'pourability' of the objects in the scene. Indeed, besides pick and place on a table, also pick and place in a container and pour into a container were chosen as tasks. Here too, the user is required to actively fixate the object on the right side to signal an interaction intention, while the robotic arm carries out the action.

Just Accepted

The studies presented here and summarised in Table 4 show the fast increasing interest in using gaze in assistive HRI and teleoperation, but, more so than in other fields, also show a number of critical limitations and challenges that will be further addressed in Section 4.3.

4.3 Inferring driver's intentions in ADAS

Advanced Driving Assistance Systems (ADAS) have been developed in the last decades in order to improve driving safety, not just by protecting passengers in case of an accident, like passive systems (seat belts, airbags, ...) do, but by preemptively decreasing accident risks [Koesdwiady et al. 2017]. Until autonomous driving becomes commonplace in multiple contexts, it is important to also invest in systems that support the driver at any automation level less than level 5. As we have seen, driving is a rather long and complex behavior that assumes a number of high-level cognitive capabilities (sustained attention, multi-tasking, situation awareness, decision making, eye-hand-foot coordination,...) to constantly work at full capacity. In this context, a number of solutions have been explored: on the one hand, many research efforts consider gaze tracking to assess the driver's cognitive state and possibly warn him if they get distracted [e.g., Ahlstrom et al. 2013]; on the other hand, especially in the context of semi-autonomous driving (up to level 3), efforts are being directed at assisting in the execution of difficult maneuvers. In this latter case, of course, the earliness of the intention estimation output could be critical, since the time range for the ADAS to act on this estimation is even shorter than in teleoperation, especially considering vehicle speed compared to human motion speed. This also depends on what the ADAS needs to do with information about the driver intention: it can just warn (by accounting for human reaction time), assist (by planning and executing a supporting plan) or simply avoid taking decisions conflicting with the driver's intent, like triggering a lane departure warning when a lane change is indeed intended. [Xing et al. 2019].

Although the idea of using gaze to infer driving intentions dates back to a couple of decades ago [e.g, in Oliver and Pentland 2000, where up to 7 maneuvers were recognized via HMMs], in the following years mostly other behavioral cues have been used for intention estimation in this context, due to the technical difficulties of performing eye tracking in a car, with changing light conditions and a moving scene.
In particular, Trivedi and colleagues have long worked on driver's intent estimation, yet they mostly used head dynamics data, showing that in lane changing this provides a more significantly discriminative cue than gaze data [Doshi and Trivedi 2009; McCall and Trivedi 2007].
Focusing on gaze, Lethaus et al. [2013] compared a neural network, a Bayesian network, and a Naive Bayes classifier in discriminating a lane change versus lane keeping (or lane change right versus left). By collecting gaze data in a driving simulator, the authors looked both at the amount of data necessary for the model to make a prediction and how soon a prediction could be made before the maneuver started. Gaze time on 5 AOIs (rear mirror, left/right mirrors, windscreen, and speedometer), in time windows of 10 and 5 s before the lane change, was considered as input for the models. As to be expected, the closer the prediction to the maneuver start, the more accurate the estimate for all models. A 5s window turned out to produce better predictive models since less noise from other activities was included, with neural networks delivering slightly better results than Bayesian Networks. Peng et al. [2015] found a similar window of 5 seconds necessary when using head rotation standard deviation along with other features concerning the vehicle motion, the state of the indicator, and lane changing conditions.
[Wu et al. 2019] considered intention estimation within a semi-autonomous vehicle. In this context, if the system is uncertain about the best maneuver when following a lead vehicle that starts decelerating, it should alert the driver to take control. To estimate the driver's intention to brake or change lane just by means of gaze/fixation data, the authors designed an HMM coupled with probabilistic Dynamic Time Warping to model the probability of each new observation given last observations and the last estimated intention in a recursive Bayesian estimation

Just Accepted

scheme. An accuracy of over 90% for the 3 intentions could be demonstrated with an anticipation time over 3s before the maneuver actually started.
While most intention estimation systems in ADAS deal with the prediction of lane change and braking, [Jiang et al. 2018] consider a different application: here the user is supposed to be a passenger in an autonomous vehicle and the system is trying to retrieve their point of interest, such as shop signs or a location the user intends to navigate to, as the vehicle moves, to possibly offer related information. By using Markov Random Fields, the proposed solution estimates dynamic interest points and shows robustness to noise sources due to eye tracking inaccuracy as well as to the moving scene and the vehicle bumpy motion.
As the studies presented in this section show, although intention estimation is critical for ADAS, gaze-based intent prediction is still in a rather early phase. In such scenarios where prediction errors might have fatal consequences the use, even for warning systems, is still very moderate.

5 CURRENT LIMITATIONS, ISSUES AND OPEN CHALLENGES

The literature surveyed in the previous sections shows that scientific research and technical progress on gaze-based intention estimation have reached a consistent level of maturity to let foresee forthcoming application in technological products, especially in cognitive robots. Still, even though eyetracking capabilities are increasingly embedded in multiple interfaces, the same speed has not been witnessed for robust algorithms leveraging eye movements. As seen, applications and solutions have usually a limited scope and rely on carefully designed supervised training. While we hope that a better understanding of the cognitive underpinnings might help with that, it is also the case that certain limitations derive from open issues and challenges both in the current understanding of how the visuomotor system serves the schema control and how technical systems harness eye movement data. Further, although perhaps not central to the mere functioning of an intention prediction framework, there are still other issues and themes that can be critical for the design of a user-friendly and effective system, one which the user is also able to trust and hence willing to interact often with. These issues can indeed also influence the user's gaze behavior and should be taken into account. We present here some of the most relevant ones, along with positions and approaches that have been put forward in the literature.

5.1 Multitasking nature of gaze, gaze as a proxy of attention, and bimanual manipulations

As already mentioned, at any time multiple cognitive processes are ongoing in our brains. This is especially true when considering more complex activities than grasping or selecting items, like during driving or carrying out a manipulation task made of multiple steps in a loosely constrained sequence. Already Land et al. [1999] recognized how fixations on objects could be categorized according to four different functions (locating, guiding, directing, and checking), yet these could also be interleaved with fixations from different subtasks: so even though the eyes are monitoring the current execution some fixations could proactively target already the next object to check its location and then go back to the current one. This kind of scheduling is rather inscrutable currently and has high variability across contexts, tasks, and people.
On the one hand, if eye fixations can reveal a lot about the information accrual process and are usually considered as a behavioural proxy for the attentional focus, it is at times an unwarranted assumption that everything in a fixated region is consciously detected. It is, for example, easy to experience time gaps when driving, when our internal focus is directed to unrelated thoughts and still the driving and visual scanning behaviour can proceed normally, according to automatized patterns. On the other hand, especially when controlling multiple effectors, some monitoring could be done covertly by peripheral vision or by strategically fixating some middle point, hence not hitting relevant AOIs. In natural bimanual eye-hand coordination, we usually visually monitor more closely the dominant effector, executing the most difficult action and rely on covert attention, proprioception and haptic feedback to control the other limb [Srinivasan and Martin 2010]. Alternatively, we shift our eyes

Just Accepted

between the two targets in an interleaved fashion [Riek et al. 2003]. However this kind of scheduling is not well researched, especially across people and handedness.
Most current gaze-based systems in robotics, to our knowledge, deal with the control of just one arm, while recent bimanual frameworks in shared control rely on learning directly both hand motion patterns or the second robotic arm is autonomously coordinated with the motion of the operator's right arm [Laghi et al. 2018; Rakita et al. 2019]. That is, gaze-based intention estimation for multi-tasking and multiple effectors remains still an open problem, especially when it is difficult to discern what has been covertly perceived or altogether overlooked.

5.2 Bidirectional effects on users' gaze behaviour

In Human-Computer Interaction, gaze-based control of an input device has been limitedly used and mostly in assistive contexts [e.g. Zeng et al. 2020], that is, in the case of patients with impaired limb control. A gaze-based controller indeed can be at times problematic because of the "Midas touch" effect [Jacob 1991]. We are in fact not used to controlling our eyes as effectors, but to using them for perception in a mostly automatized way, without producing intentional effects in the physical world. Thus, gaze-based selection in such interfaces can lead to a number of false selections, as discussed for instance in [Novak and Riener 2013]. To ameliorate such issue, often the gaze information is used in combination with an intentional hint by the user, such as intentionally blinking, fixating for a certain time, clicking, or movement initiation. Still, distinguishing glances to perceive and glances to act becomes difficult for the system, while it can lead the user to very controlled, unnatural eye movements, in turn increasing the cognitive load. This might be less the case in a gaze-based intention estimation system, still: if the user becomes aware that the system is using their gaze to produce an assisting effect, they might modify their natural scanpaths to actively obtain the intended effect. As a result, scan patterns would diverge from the behavior seen during training without assistance, consequently throwing off the system. The general assumption, as in Javdani et al. [2015], is in fact that the user is not minding the system while pursuing their own plan. If instead, the user learns that their gaze has an effect on the robot motion, they might actively use it as a further control input to guide the robot: this might be on purpose so or just a side effect that needs to be considered. For example, in a driving context, if users learned that failing to look at the mirror when changing lanes would trigger a warning, they might actively plan a gaze check to the mirror to avoid the nuisance, triggering instead the lane change assistance.

In teleoperated manipulation settings, the visuomotor task needs first to be learned, by acquiring the novel mapping between own movements and the distant visual sensory consequences. This implies that until a ceiling level of training is reached, gaze patterns change along with training. Sailer et al. [2005] investigated how the gaze behavior evolved when learning to control a cursor on a screen for an intercept task by means of an unfamiliar device applying isometric forces and torques. They observed a first exploratory stage where the gaze was mostly pursuing the controlled effector (the cursor) and the hit rate was rather low. In the second stage, the eyes became more predictive of the striven cursor position, while the hit rate improved. Finally, once the visuomotor task was mastered, the eyes would jump directly to the target, without the need to monitor the cursor movement. While this was a rather unfamiliar task, it could be possibly comparable to controlling a robotic limb by means of some input device utterly different than the own hand motion.
Effects on scanpaths while controlling a robotic arm have recently been discussed by Aronson et al. [2018]; Fuchs and Belardinelli [2021]; Wang et al. [2020], but further investigation will be needed to shed more light on the way robot-mediated eye-hand coordination might influence how healthy and disabled users manage to orchestrate perception and action processes. Such an issue is in general also related to user adaptation and the the degree of trust in the system, as discussed next.

Just Accepted

5.3 Adaptation, co-adaptation and trusting the assistance system

Considering predict-then-blend approaches in shared autonomy, there is usually an arbitration parameter α determining how strongly the robot assistance will be implemented with respect to the user's input. Small values can determine too timid assistance, hence not really supporting, while too large values can lead to aggressive assistance, which can be annoying [Dragan and Srinivasa 2013]. The amount of assistance, of course, should depend also on the confidence in the inferred intention, to avoid aggressively assisting with the wrong goal. To adapt arbitration to the user and hence find the optimal amount of assistance, Oh et al. [2019] propose to learn the best arbitration parameter online from the user interaction with the system. Considering in general human-machine interfaces, Gallina et al. [2015] proposed an approach in between goal-oriented design (which maximizes performance) and human-centered design (which maximizes usability): *progressive co-adaptation*. The idea here is that as the user through interaction with the system adapts and improves their skills , concurrently the system should monitor these improvements and adjust its parameters accordingly. As shown by Sailer et al. [2005], this learning has also an effect on gaze, which should be taken into account. Yet, in most HRI approaches a model is learned by collecting data during the targeted task and usually not retrained in light of long-term use and adaptation by the user. Again this is an open problem and a possible solution might entail online learning to adapt the parameters of the time series prediction model as the data distribution changes [e.g., Guo et al. 2016].
While Javdani et al. [2015] make the assumption that the user is agnostic as to the robot assistance and does not consider it in its control strategy, recently Amirshirzad et al. [2019] investigated more systematically instead exactly how humans adapt in teleoperation versus shared control. Performance sharply improves in the latter condition compared to the former, after an initial exploration phase, where users implicitly learn the robot behavior and adapt to it.
Finally, a further aspect to consider is that, as suggested by Haji Fathaliyan et al. [2018] and Aronson and Admoni [2018], by means of gaze also the degree of trust in the system, anomalies due to malfunctioning, or other issues can be assessed. Indeed, recalling the results of Sailer et al. [2005], in teleoperation scenarios, for example, once training is completed and the user is familiarized with the system, less fixations will be placed on the robot and the gaze will become more predictive of the targeted objects/locations. Indeed, brief visual assessment of the success of the current subtask will be followed by a gaze shift to acquire information for the following subtask [see e.g., Fuchs and Belardinelli 2021]. This will probably not be the case if the user is constantly monitoring the robot to make sure that it does its part. On the other hand, if after some time of trusted use of the system the gaze deviates from the usual predictability, it can be a reliable sign that something either in the system or in the mental state of the user has changed. Such hypotheses of course will need suitably designed user studies and the collection of both objective and subjective measures to test performance, user satisfaction, and experience.

5.4 Trade-off sense of agency/speed of execution: the case for personalization

Letting the user retain their sense of agency is a crucial aspect for most human-machine interaction systems, even when control is shared transparently and the machine can correctly read the user's intention. Also in semi-autonomous vehicles, indeed, if the human driver takes a decision and executes a maneuver, an override from the system might substantially damage the trust in the system.
In [Pacherie 2008], the sense of agency is defined as the sense of intentional causation and it manifests in the subjective perception of a match between the prior intention and the following action [Wegner 2005]. This is strongly related to the ability to predict the consequences of our actions. If an artificial system interferes with this prediction and produces an action execution that does not match our expectations, we might feel we are not the agent causing the action. This might be sometimes at odds with the assistance the system is trying to deliver. Further, it will cause eye movement patterns to change, becoming again less anticipative. Still, behavioral measures for such sense, and in particular oculomotor measures, have been hard to identify [Schwarz et al. 2018].

Just Accepted

Most works on shared autonomy systems, for example, report user studies with objective and subjective measures. If task completion times usually benefit from human-robot cooperation, subjective opinions give a less clear-cut picture. Sometimes for complex tasks more autonomy is preferred, while for other simple tasks more control to the human is preferred. Dragan and Srinivasa [2013] found that these preferences nonetheless interact with other factors, like the arbitration style (timid vs. aggressive) and the intention prediction accuracy (right vs. wrong), and in general there are also subjective differences in these preferences.
In this context, contrasting results which found a preference for execution speed [Dragan and Srinivasa 2013; Hauser 2013] suggest that the different ways participants interact with the system call for some personalization of the cost function modeling the user.
Gopinath et al. [2016] picked up on this suggestion, noting that in the assistive context, users differ in their motor capabilities (and these can even both deteriorate or improve) and might need different degrees of assistance. An approach grounded in optimal control theory is proposed, with the tuning of the arbitration parameters formulated as an optimal control problem, where the user's and robot's control signals are given. The optimization is carried out by the users themselves, increasing or decreasing assistance by verbal commands. While tasking the user with a further parameter to control in another input modality, the study still paves the way for a more thorough investigation of personalization within shared autonomy systems, with the aim of optimizing the balance between user satisfaction and system effectiveness.
This is still not just a matter of arbitration as individual differences in scanpath behavior might call for personalized training also in intention estimation [Wiebel-Herboth et al. 2020]. In simple motor tasks like pick and place scanpath patterns might look very similar across users, while in less constrained, more complex activities the oculomotor behavior might be more directed according to the way each person schedules the subtasks [Yi and Ballard 2009].

5.5 Extracting more information from 3D gaze data

As seen, collecting quality gaze data in 3D and localizing them on real-word objects, is still a matter of research (see Section 4.2). As argued in Section 2.2, also the type of grasp and further manipulation intent in some cases can be inferred from fixations. Still, these aspects have been hardly tackled in technical systems, because it is often computationally hard to recognize 3D objects in any pose and appearance, while also localizing online the gaze with head motion on their functional parts without fiducial markers. This would also need the definition of multiple *ad-hoc* AOIs for each object and infusing in the system task, grasp, and functional knowledge about predefined objects. In this sense, gaze-based intention estimation can achieve better results when embedded in more advanced cognitive system, not just relying on precise eye-tracking capabilities but also on a sophisticated scene and world understanding, as currently developed social and assistive robots.

6 CONCLUSIONS

As stated by [Vernon et al. 2016], intentions are the cognitive bridge integrating goals and actions in a prospective framework. Inferring human intentions is thus crucial for assistive and cooperative systems. In this survey, methods and applications of gaze-based intention estimation have been reviewed and summarized, along with the underlying cognitive science underpinnings concerning perception and action. Eye tracking capabilities have made large progress in the last decades, allowing moving from constrained and controlled laboratory settings to real-world activities. Concurrently, machine learning, robotics, and ADAS research endeavors have reached a point where integrated cooperation of humans and machines can finally find application in common tasks such as daily living activities, driving, or human-robot cooperation. This interaction can be made more effective by inferring the human intention based on gazed objects, endowing the robotic partner with an awareness of what the user wants to achieve, and letting it assist in a seamless way. In this respect, surveyed studies and

Just Accepted

frameworks show promising results and perspectives, at least for relatively constrained tasks and applications. They compellingly show how gaze features are crucial for early and accurate inference of intentions and for assessing the cognitive state of the user. Yet, on the one hand, a more principled consideration of the targeted cognitive processes and subsequent oculomotor manifestations, as reviewed here, could improve the design of such systems, especially selecting the most adequate features and prediction models. On the other hand, some open issues and limitations need still to be investigated and improve for intention estimation to achieve wider spread in cognitive systems.

Important challenges and opportunities are connected to these technologies, especially regarding physical systems. Robots that could be operated remotely in a natural way and support the operator by taking care of low-level motor aspects would find wide applications in crises like the pandemic recently afflicting the world. Similarly, assistive systems able to infer and execute actions in peripersonal space would enable people with certain disabilities to act in their environment and regain a certain amount of independence. Further, gaze-based driving assistance systems could support the driver in a smarter and more reliable way, also possibly providing a basis for learning from their perceptual decision making and paving the way for a more human-like full autonomy.

Just Accepted

REFERENCES

Daniel Aarno and Danica Kragic. 2008. Motion intention recognition in robot assisted applications. *Robotics and Autonomous Systems* 56, 8 (2008), 692–705.

Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63.

Henny Admoni and Siddhartha Srinivasa. 2016. Predicting user intent through eye gaze for shared autonomy. In *2016 AAAI Fall Symposium Series*.

C. Ahlstrom, K. Kircher, and A. Kircher. 2013. A Gaze-Based Driver Distraction Warning System and Its Effect on Visual Behavior. *IEEE Transactions on Intelligent Transportation Systems* 14, 2 (2013), 965–973. https://doi.org/10.1109/TITS.2013.2247759

Negin Amirshirzad, Asiye Kumru, and Erhan Oztop. 2019. Human adaptation to human–robot shared control. *IEEE Transactions on Human-Machine Systems* 49, 2 (2019), 126–136.

Reuben M Aronson and Henny Admoni. 2018. Gaze for Error Detection During Human-Robot Shared Manipulation. In *Fundamentals of Joint Action workshop, Robotics: Science and Systems*.

Reuben M Aronson and Henny Admoni. 2019. Semantic gaze labeling for human-robot shared manipulation. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. 1–9.

Reuben M Aronson, Nadia Almutlak, and Henny Admoni. 2021. Inferring Goals with Gaze during Teleoperated Manipulation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7307–7314.

Reuben M Aronson, Thiago Santini, Thomas C Kübler, Enkelejda Kasneci, Siddhartha Srinivasa, and Henny Admoni. 2018. Eye-hand behavior in human-robot shared manipulation. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 4–13.

Thomas Bader, Matthias Vogelgesang, and Edmund Klaus. 2009. Multimodal Integration of Natural Gaze Behavior for Intention Recognition during Object Manipulation. In *Proceedings of the 2009 International Conference on Multimodal Interfaces* (Cambridge, Massachusetts, USA) *(ICMI-MLMI '09)*. Association for Computing Machinery, New York, NY, USA, 199–206.

Chris L Baker and Joshua B Tenenbaum. 2014. Modeling human plan recognition using Bayesian theory of mind. *Plan, activity, and intent recognition: Theory and practice* (2014), 177–204.

Daniel Baldauf and Heiner Deubel. 2010. Attentional landscapes in reaching and grasping. *Vision research* 50, 11 (2010), 999–1013.

Wolfgang Becker and Albert F Fuchs. 1969. Further properties of the human saccadic system: eye movements and correction saccades with and without visual fixation points. *Vision research* 9, 10 (1969), 1247–1258.

Jenay M Beer, Arthur D Fisk, and Wendy A Rogers. 2014. Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of human-robot interaction* 3, 2 (2014), 74–99.

Harold Bekkering and Sebastiaan FW Neggers. 2002. Visual search is modulated by action intentions. *Psychological science* 13, 4 (2002), 370–374.

Anna Belardinelli, Marissa Barabas, Marc Himmelbach, and Martin V Butz. 2016a. Anticipatory eye fixations reveal tool knowledge for tool interaction. *Experimental brain research* 234, 8 (2016), 2415–2431.

Anna Belardinelli, Oliver Herbort, and Martin V Butz. 2015. Goal-oriented gaze strategies afforded by object interaction. *Vision Research* 106 (2015), 47–57.

Anna Belardinelli, Anirudh Reddy Kondapally, Dirk Ruiken, Daniel Tanneberg, and Tomoki Watabe. 2022. Intention estimation from gaze and motion features for human-robot shared-control object manipulation. https://doi.org/10.48550/ARXIV.2208.08688

Anna Belardinelli, Madeleine Y Stepper, and Martin V Butz. 2016b. It's in the eyes: Planning precise manual actions before execution. *Journal of vision* 16, 1 (2016), 18–18.

Giuseppe Boccignone. 2019. Advanced statistical methods for eye movement analysis and modelling: a gentle introduction. In *Eye Movement Research*. Springer, 309–405.

Jonathan FG Boisvert and Neil DB Bruce. 2016. Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing* 207 (2016), 653–668.

Myles Brand. 1984. *Intending and acting: Toward a naturalized action theory*. MIT Press (MA).

Michael Bratman et al. 1987. *Intention, plans, and practical reason*. Vol. 10. Harvard University Press Cambridge, MA.

Anne-Marie Brouwer, Volker H Franz, and Karl R Gegenfurtner. 2009. Differences in fixations between grasping and viewing objects. *Journal of Vision* 9, 1 (2009), 18–18.

Daniel N Bub, Michael EJ Masson, and Terry Lin. 2013. Features of planned hand actions influence identification of graspable objects. *Psychological Science* (2013), 0956797612472909.

Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Troster. 2010. Eye movement analysis for activity recognition using electrooculography. *IEEE transactions on pattern analysis and machine intelligence* 33, 4 (2010), 741–753.

Andreas Bulling and Thorsten O Zander. 2014. Cognition-aware computing. *IEEE Pervasive Computing* 13, 3 (2014), 80–83.

Cristiana Cavina-Pratesi and Constanze Hesse. 2013. Why do the eyes prefer the index finger? Simultaneous recording of eye and hand movements during precision grasping. *Journal of vision* 13, 5 (2013), 15–15.

Xiao-Lin Chen and Wen-Jun Hou. 2022. Gaze-Based Interaction Intention Recognition in Virtual Reality. *Electronics* 11, 10 (2022). https://doi.org/10.3390/electronics11101647

Antoine Coutrot, Janet H Hsiao, and Antoni B Chan. 2018. Scanpath modeling and classification with hidden Markov models. *Behavior research methods* 50, 1 (2018), 362–379.

Brendan David-John, Candace Peacock, Ting Zhang, T Scott Murdison, Hrvoje Benko, and Tanya R Jonker. 2021. Towards gaze-based prediction of the intent to interact in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications*. 1–7.

Donald Davidson. 1980. *Essays on Actions and Events: Philosophical Essays Volume 1*. Vol. 1. Oxford University.

Daniel Clement Dennett. 1989. *The intentional stance*. MIT press.

Oriane Dermy, François Charpillet, and Serena Ivaldi. 2019. Multi-modal intention prediction with probabilistic movement primitives. In *Human Friendly Robotics*. Springer, 181–196.

Heiner Deubel, Werner X Schneider, and Ingo Paprotta. 1998. Selective dorsal and ventral processing: Evidence for a common attentional mechanism in reaching and perception. *Visual Cognition* 5, 1-2 (1998), 81–107.

A. Doshi and M. M. Trivedi. 2009. On the Roles of Eye Gaze and Head Dynamics in Predicting Driver's Intent to Change Lanes. *IEEE Transactions on Intelligent Transportation Systems* 10, 3 (2009), 453–462. https://doi.org/10.1109/TITS.2009.2026675

Anca D Dragan and Siddhartha S Srinivasa. 2013. A policy-blending formalism for shared control. *The International Journal of Robotics Research* 32, 7 (2013), 790–805.

Nuno Ferreira Duarte, Mirko Raković, Jovica Tasevski, Moreno Ignazio Coco, Aude Billard, and José Santos-Victor. 2018. Action anticipation: Reading the intentions of humans and robots. *IEEE Robotics and Automation Letters* 3, 4 (2018), 4132–4139.

Wolfgang Einhäuser, Frank Schumann, Stanislavs Bardins, Klaus Bartl, Guido Böning, Erich Schneider, and Peter König. 2007. Human eye-head co-ordination in natural exploration. *Network: Computation in Neural Systems* 18, 3 (2007), 267–297. https://doi.org/10.1080/09548980701671094 PMID: 17926195.

Nathan J Emery. 2000. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews* 24, 6 (2000), 581–604.

Katja Fiehler, Eli Brenner, and Miriam Spering. 2019. Prediction in goal-directed action. *Journal of vision* 19, 9 (2019), 10–10.

Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. 2015. Active inference and epistemic value. *Cognitive neuroscience* 6, 4 (2015), 187–214.

Stefan Fuchs and Anna Belardinelli. 2021. Gaze-Based Intention Estimation for Shared Autonomy in Pick-and-Place Tasks. *Frontiers in Neurorobotics* 15 (2021), 33. https://doi.org/10.3389/fnbot.2021.647930

Paolo Gallina, Nicola Bellotto, and Massimiliano Di Luca. 2015. Progressive co-adaptation in human-machine interaction. In *2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, Vol. 2. IEEE, 362–368.

Alvin I. Goldman. 1970. *A Theory of Human Action*. Princeton University Press.

Iván González-Díaz, Jenny Benois-Pineau, Jean-Philippe Domenger, Daniel Cattaert, and Aymar de Rugy. 2019. Perceptually-guided deep neural networks for ego-action prediction: Object grasping. *Pattern Recognition* 88 (2019), 223–235.

Michael A Goodrich, Jacob W Crandall, and Emilia Barakova. 2013. Teleoperation and beyond for assistive humanoid robots. *Reviews of Human factors and ergonomics* 9, 1 (2013), 175–226.

Just Accepted

Deepak Gopinath, Siddarth Jain, and Brenna D Argall. 2016. Human-in-the-loop optimization of shared autonomy in assistive robotics. *IEEE Robotics and Automation Letters* 2, 1 (2016), 247–254.

Tian Guo, Zhao Xu, Xin Yao, Haifeng Chen, Karl Aberer, and Koichi Funaya. 2016. Robust online time series prediction with recurrent neural networks. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Ieee, 816–825.

Amin Haji-Abolhassani and James J Clark. 2014. An inverse Yarbus process: Predicting observers' task from eye movement patterns. *Vision research* 103 (2014), 127–142.

Alireza Haji Fathaliyan, Xiaoyu Wang, and Veronica J Santos. 2018. Exploiting three-dimensional gaze tracking for action recognition during bimanual manipulation to enhance human–robot collaboration. *Frontiers in Robotics and AI* 5 (2018), 25.

Kris Hauser. 2013. Recognition, prediction, and planning for assisted teleoperation of freeform tasks. *Autonomous Robots* 35, 4 (2013), 241–254.

Mary Hayhoe and Dana Ballard. 2014. Modeling task control of eye movements. *Current Biology* 24, 13 (2014), R622–R628.

Mary M Hayhoe. 2017. Vision and action. *Annual review of vision science* 3 (2017), 389–413.

Mary M Hayhoe, Anurag Shrivastava, Ryan Mruczek, and Jeff B Pelz. 2003. Visual memory and motor planning in a natural task. *Journal of vision* 3, 1 (2003), 6–6.

John M Henderson. 2017. Gaze control as prediction. *Trends in cognitive sciences* 21, 1 (2017), 15–23.

John M Henderson, Svetlana V Shinkareva, Jing Wang, Steven G Luke, and Jenn Olejarczyk. 2013. Predicting cognitive state from eye movements. *PloS one* 8, 5 (2013), e64937.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology* 6 (2015), 1049.

Chien-Ming Huang and Bilge Mutlu. 2016. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 83–90.

Laurent Itti and Pierre Baldi. 2009. Bayesian surprise attracts human attention. *Vision research* 49, 10 (2009), 1295–1306.

Robert JK Jacob. 1991. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems (TOIS)* 9, 2 (1991), 152–169.

Siddarth Jain and Brenna Argall. 2018. Recursive Bayesian human intent recognition in shared-control robotics. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3905–3912.

Siddarth Jain and Brenna Argall. 2019. Probabilistic Human Intent Recognition for Shared Autonomy in Assistive Robotics. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 1 (2019), 1–23.

Shervin Javdani, Henny Admoni, Stefania Pellegrinelli, Siddhartha S Srinivasa, and J Andrew Bagnell. 2018. Shared autonomy via hindsight optimization for teleoperation and teaming. *The International Journal of Robotics Research* 37, 7 (2018), 717–742.

Shervin Javdani, Siddhartha S Srinivasa, and J Andrew Bagnell. 2015. Shared autonomy via hindsight optimization. In *Robotics science and systems: online proceedings*, Vol. 2015. NIH Public Access.

Yu-Sian Jiang, Garrett Warnell, and Peter Stone. 2018. Inferring user intention using gaze in vehicles. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 298–306.

Roland S Johansson, Göran Westling, Anders Bäckström, and J Randall Flanagan. 2001. Eye–hand coordination in object manipulation. *Journal of Neuroscience* 21, 17 (2001), 6917–6932.

Christopher Kanan, Nicholas A Ray, Dina NF Bseiso, Janet H Hsiao, and Garrison W Cottrell. 2014. Predicting an observer's task using multi-fixation pattern analysis. In *Proceedings of the symposium on eye tracking research and applications*. 287–290.

Ashima Keshava, Anete Aumeistere, Krzysztof Izdebski, and Peter Konig. 2020. Decoding Task From Oculomotor Behavior In Virtual Reality. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Short Papers)*. Association for Computing Machinery, New York, NY, USA, Article 30, 5 pages.

Kris M Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. 2011. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR 2011*. IEEE, 3241–3248.

A. Koesdwiady, R. Soua, F. Karray, and M. S. Kamel. 2017. Recent Trends in Driver Safety Monitoring Systems: State of the Art and Challenges. *IEEE Transactions on Vehicular Technology* 66, 6 (2017), 4550–4563. https://doi.org/10.1109/TVT.2016.2631604

Yu Kong and Yun Fu. 2018. Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230* (2018).

Iuliia Kotseruba and John K Tsotsos. 2021. Behavioral Research and Practical Models of Drivers' Attention. *arXiv preprint arXiv:2104.05677* (2021).

Marco Laghi, Michele Maimeri, Mathieu Marchand, Clara Leparoux, Manuel Catalano, Arash Ajoudani, and Antonio Bicchi. 2018. Shared-autonomy control for intuitive bimanual tele-manipulation. In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*. IEEE, 1–9.

Michael Land. 2009. Vision, eye movements, and natural behavior. *Visual neuroscience* 26, 01 (2009), 51–62.

Michael Land, Neil Mennie, and Jennifer Rusted. 1999. The roles of vision and eye movements in the control of activities of daily living. *Perception* 28, 11 (1999), 1311–1328.

Just Accepted

Michael Land and Benjamin Tatler. 2009. *Looking and acting: vision and eye movements in natural behaviour.* Oxford University Press.
Michael F Land. 2006. Eye movements and the control of actions in everyday life. *Progress in retinal and eye research* 25, 3 (2006), 296–324.
Gabor Lengyal, Kevin Carlberg, Majed Samad, and Tanya R Jonker. 2021. Predicting visual attention using the hidden structure in eye-gaze dynamics. In *CHI2021 Eye Movements as an Interface to Cognitive State (EMICS) Workshop Proceedings. ACM.*
Firas Lethaus, Martin RK Baumann, Frank Köster, and Karsten Lemmer. 2013. A comparison of selected simple supervised learning algorithms to predict driver intent based on gaze data. *Neurocomputing* 121 (2013), 108–130.
Songpo Li, Michael Bowman, Hamed Nobarani, and Xiaoli Zhang. 2020. Inference of Manipulation Intent in Teleoperation for Robotic Assistance. *Journal of Intelligent & Robotic Systems* (2020), 1–15.
Songpo Li, Xiaoli Zhang, and Jeremy D Webb. 2017. 3-D-gaze-based robotic grasping through mimicking human visuomotor function for people with motion impairments. *IEEE Transactions on Biomedical Engineering* 64, 12 (2017), 2824–2835.
Meng Liu, Youfu Li, and Hai Liu. 2020. 3D Gaze Estimation for Head-Mounted Eye Tracking System With Auto-Calibration Method. *IEEE Access* 8 (2020), 104207–104215.
Dylan P Losey, Craig G McDonald, Edoardo Battaglia, and Marcia K O'Malley. 2018. A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction. *Applied Mechanics Reviews* 70, 1 (2018).
Joel C McCall and Mohan M Trivedi. 2007. Driver behavior and situation aware brake assistance for intelligent vehicles. *Proc. IEEE* 95, 2 (2007), 374–387.
Alfred R. Mele. 1992. *Springs of Action: Understanding Intentional Behavior.* Oxford University Press.
RC Miall and GZ Reckess. 2002. The cerebellum and the timing of coordinated eye and hand tracking. *Brain and cognition* 48, 1 (2002), 212–226.
Wookhee Min, Bradford Mott, Jonathan Rowe, Robert Taylor, Eric Wiebe, Kristy Elizabeth Boyer, and James Lester. 2017. Multimodal goal recognition in open-world digital games. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference.*
Thi-Hoa-Cuc Nguyen, Jean-Christophe Nebel, Francisco Florez-Revuelta, et al. 2016. Recognition of activities of daily living with egocentric vision: A review. *Sensors* 16, 1 (2016), 72.
Domen Novak and Robert Riener. 2013. Enhancing patient freedom in rehabilitation robotics using gaze-based intention detection. In *2013 IEEE 13th international conference on rehabilitation robotics (ICORR).* IEEE, 1–6.
Keisuke Ogaki, Kris M Kitani, Yusuke Sugano, and Yoichi Sato. 2012. Coupling eye-motion and ego-motion features for first-person activity recognition. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.* IEEE, 1–7.
Yoojin Oh, Marc Toussaint, and Jim Mainprice. 2019. Learning Arbitration for Shared Autonomy by Hindsight Data Aggregation. *arXiv preprint arXiv:1906.12280* (2019).
N. Oliver and A. P. Pentland. 2000. Graphical models for driver behavior recognition in a SmartCar. In *Proceedings of the IEEE Intelligent Vehicles Symposium 2000 (Cat. No.00TH8511).* 7–12. https://doi.org/10.1109/IVS.2000.898310
Elisabeth Pacherie. 2008. The phenomenology of action: A conceptual framework. *Cognition* 107, 1 (2008), 179–217.
Elisabeth Pacherie. 2012. Action. In *The Cambridge Handbook of Cognitive Science*, Keith Frankish and William Editors Ramsey (Eds.). Cambridge University Press.
Jeff B Pelz and Roxanne Canosa. 2001. Oculomotor behavior and perceptual strategies in complex tasks. *Vision research* 41, 25-26 (2001), 3587–3596.
Jinshuan Peng, Yingshi Guo, Rui Fu, Wei Yuan, and Chang Wang. 2015. Multi-parameter prediction of drivers' lane-changing behaviour with neural network model. *Applied ergonomics* 50 (2015), 207–217.
Daniel Rakita, Bilge Mutlu, Michael Gleicher, and Laura M Hiatt. 2019. Shared control–based bimanual robot manipulation. *Science Robotics* 4, 30 (2019).
Yosef Razin and Karen Feigh. 2017. Learning to predict intent from gaze during robotic hand-eye coordination. In *Thirty-First AAAI Conference on Artificial Intelligence.*
Stephan Riek, James R Tresilian, M Mon-Williams, Vanessa L Coppard, and Richard G Carson. 2003. Bimanual aiming and overt attention: one law for two hands. *Experimental brain research* 153, 1 (2003), 59–75.
Giacomo Rizzolatti, Lucia Riggio, Isabella Dascola, and Carlo Umiltá. 1987. Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia* 25, 1 (1987), 31–40.
Agnes Rosner, Thomas Franke, Frederik Platten, and Christiane Attig. 2019. Eye Movements in Vehicle Control. In *Eye Movement Research.* Springer, 929–969.
Uta Sailer, J Randall Flanagan, and Roland S Johansson. 2005. Eye–hand coordination during learning of a novel visuomotor task. *Journal of Neuroscience* 25, 39 (2005), 8833–8842.
Kenji Sakita, Koichi Ogawara, Shinji Murakami, Kentaro Kawamura, and Katsushi Ikeuchi. 2004. Flexible cooperation between human and robot by interpreting human intention from gaze information. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai, Japan, September 28 - October 2, 2004.* IEEE, 846–851.
Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications.* 71–78.

Just Accepted

Dario D Salvucci and Andrew Liu. 2002. The time course of a lane change: Driver control and eye-movement behavior. *Transportation research part F: traffic psychology and behaviour* 5, 2 (2002), 123–132.

Malte Schilling, Stefan Kopp, Sven Wachsmuth, Britta Wrede, Helge Ritter, Thomas Brox, Bernhard Nebel, and Wolfram Burgard. 2016. Towards a multidimensional perspective on shared autonomy. In *2016 AAAI Fall Symposium Series*.

Katharina A Schwarz, Sebastian Burger, David Dignath, Wilfried Kunde, and Roland Pfister. 2018. Action-effect binding and agency. *Consciousness and Cognition* 65 (2018), 304–309.

Paul Schydlo, Mirko Rakovic, Lorenzo Jamone, and José Santos-Victor. 2018. Anticipation in human-robot cooperation: A recurrent neural network approach for multiple action sequences prediction. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1–6.

John R. Searle. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Ali Shafti, Pavel Orlov, and A Aldo Faisal. 2019. Gaze-based, context-aware robotic system for assisted reaching and grasping. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 863–869.

Ronal Singh, Tim Miller, Joshua Newn, Eduardo Velloso, Frank Vetere, and Liz Sonenberg. 2020. Combining gaze and AI planning for online human intention recognition. *Artificial Intelligence* (2020), 103275.

Divya Srinivasan and Bernard J Martin. 2010. Eye–hand coordination of symmetric bimanual reaching tasks: Temporal aspects. *Experimental Brain Research* 203, 2 (2010), 391–405.

Bernard Marius 't Hart and Wolfgang Einhäuser. 2012. Mind the step: complementary effects of an implicit task on eye and head movements in real-life gaze allocation. *Experimental Brain Research* 223 (2012), 233–249.

Enkelejda Tafaj, Gjergji Kasneci, Wolfgang Rosenstiel, and Martin Bogdan. 2012. Bayesian online clustering of eye movement data. In *Proceedings of the symposium on eye tracking research and applications*. 285–288.

Karim A Tahboub. 2006. Intelligent human-machine interaction based on dynamic Bayesian networks probabilistic intention recognition. *Journal of Intelligent and Robotic Systems* 45, 1 (2006), 31–52.

Ryo Takahashi, Hiromasa Suzuki, Jouh Yeong Chew, Yutaka Ohtake, Yukie Nagai, and Koichi Ohtomi. 2018. A system for three-dimensional gaze fixation analysis using eye tracking glasses. *Journal of Computational Design and Engineering* 5, 4 (2018), 449–457.

Ajay Kumar Tanwani and Sylvain Calinon. 2017. A generative model for intention recognition and manipulation assistance in teleoperation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 43–50.

Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences* 28, 5 (2005), 675–691.

John K Tsotsos, Iuliia Kotseruba, Alexander Andreopoulos, and Yulong Wu. 2019. A possible reason for why data-driven beats theory-driven computer vision. *arXiv preprint arXiv:1908.10933* (2019).

Ralf Van der Lans, Michel Wedel, and Rik Pieters. 2011. Defining eye-fixation sequences across individuals and tasks: the Binocular-Individual Threshold (BIT) algorithm. *Behavior research methods* 43, 1 (2011), 239–257.

David Vernon, Serge Thill, and Tom Ziemke. 2016. The role of intention in cognitive robotics. In *Toward Robotic Socially Believable Behaving Systems-Volume I*. Springer, 15–27.

Xiaoyu Wang, Alireza Haji Fathaliyan, and Veronica J Santos. 2020. Toward shared autonomy control schemes for human-robot systems: Action primitive recognition using eye gaze features. *Frontiers in Neurorobotics* 14 (2020), 66.

Jeremy D Webb, Songpo Li, and Xiaoli Zhang. 2016. Using visuomotor tendencies to increase control performance in teleoperation. In *2016 American Control Conference (ACC)*. IEEE, 7110–7116.

Sascha Weber, Rebekka S Schubert, Stefan Vogt, Boris M Velichkovsky, and Sebastian Pannasch. 2018. Gaze3DFix: Detecting 3D fixations with an ellipsoidal bounding volume. *Behavior research methods* 50, 5 (2018), 2004–2015.

Daniel M. Wegner. 2005. Who is the Controller of Controlled Processes? In *The New Unconscious. Oxford Series in Social Cognition and Social Neuroscience*, Ran R. Hassin, James S. Uleman, and John A. Bargh (Eds.). Oxford University Press, 19–36.

Julia Werneke and Mark Vollrath. 2012. What does the driver look at? The influence of intersection characteristics on attention allocation and driving behavior. *Accident Analysis & Prevention* 45 (2012), 610–619.

Christiane B. Wiebel-Herboth, Matti Krüger, and Martina Hasenjäger. 2020. Interactions between Inter- and Intra-Individual Effects on Gaze Behavior. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 35–40.

M. Wu, T. Louw, M. Lahijanian, W. Ruan, X. Huang, N. Merat, and M. Kwiatkowska. 2019. Gaze-based Intention Anticipation over Driving Manoeuvres in Semi-Autonomous Vehicles. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 6210–6216. https://doi.org/10.1109/IROS40897.2019.8967779

Y. Xing, C. Lv, H. Wang, H. Wang, Y. Ai, D. Cao, E. Velenis, and F. Wang. 2019. Driver Lane Change Intention Inference for Intelligent Vehicles: Framework, Survey, and Challenges. *IEEE Transactions on Vehicular Technology* 68, 5 (2019), 4377–4390. https://doi.org/10.1109/TVT.2019.2903299

Alfred L Yarbus. 1967. Eye movements during perception of complex objects. In *Eye movements and vision*. Springer, Boston, MA, 171–211.

Just Accepted

Weilie Yi and Dana Ballard. 2009. Recognizing behavior in hand-eye coordination patterns. *International Journal of Humanoid Robotics* 6, 03 (2009), 337–359.

Wentao Yu, Redwan Alqasemi, Rajiv Dubey, and Norali Pernalete. 2005. Telemanipulation assistance based on motion intention recognition. In *Proceedings of the 2005 IEEE international conference on robotics and automation.* IEEE, 1121–1126.

Raimondas Zemblys, Diederick C Niehorster, Oleg Komogortsev, and Kenneth Holmqvist. 2018. Using machine learning to detect events in eye-tracking data. *Behavior research methods* 50, 1 (2018), 160–181.

Hong Zeng, Yitao Shen, Xuhui Hu, Aiguo Song, Baoguo Xu, Huijun Li, Yanxin Wang, and Pengcheng Wen. 2020. Semi-Autonomous Robotic Arm Reaching With Hybrid Gaze–Brain Machine Interface. *Frontiers in Neurorobotics* 13 (2020), 111.

Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. 2020. Human gaze assisted artificial intelligence: A review. In *IJCAI: Proceedings of the Conference*, Vol. 2020. NIH Public Access, 4951.

Just Accepted