MIDAS: Deep learning human action intention prediction from natural eye movement patterns

Paul Festor[1,2,3,4,+], **Ali Shafti**[1,2,3,+], **Alex Harston**[1,3], **Mickey Li**[1,2], **Pavel Orlov**[1,2,3], **and A. Aldo Faisal**[1,2,3,4,5,*]

[1]Brain and Behaviour Lab: Dept. of Bioengineering, Imperial College London, SW7 2AZ, London, UK
[2]Dept. of Computing, Imperial College London, SW7 2AZ, London, UK
[3]Behaviour Analytics Lab, Data Science Institute, SW7 2AZ, London, UK
[4]UKRI CDT in AI for Healthcare, Imperial College London, SW7 2AZ, London, UK
[5]MRC London Institute of Medical Sciences, W12 0NN, London, UK
[*]corresponding author: A. Aldo Faisal (aldo.faisal@imperial.ac.uk)
[+]these authors contributed equally to this work

ABSTRACT

Eye movements have long been studied as a window into the attentional mechanisms of the human brain and made accessible as novelty style human-machine interfaces. However, not everything that we gaze upon, is something we want to interact with; this is known as the Midas Touch problem for gaze interfaces. To overcome the Midas Touch problem, present interfaces tend not to rely on natural gaze cues, but rather use dwell time or gaze gestures. Here we present an entirely data-driven approach to decode human intention for object manipulation tasks based solely on natural gaze cues. We run data collection experiments where 16 participants are given manipulation and inspection tasks to be performed on various objects on a table in front of them. The subjects' eye movements are recorded using wearable eye-trackers allowing the participants to freely move their head and gaze upon the scene. We use our Semantic Fovea, a convolutional neural network model to obtain the objects in the scene and their relation to gaze traces at every frame. We then evaluate the data and examine several ways to model the classification task for intention prediction. Our evaluation shows that intention prediction is not a naive result of the data, but rather relies on non-linear temporal processing of gaze cues. We model the task as a time series classification problem and design a bidirectional Long-Short-Term-Memory (LSTM) network architecture to decode intentions. Our results show that we can decode human intention of motion purely from natural gaze cues and object relative position, with 91.9% accuracy. Our work demonstrates the feasibility of natural gaze as a Zero-UI interface for human-machine interaction, i.e., users will only need to act naturally, and do not need to interact with the interface itself or deviate from their natural eye movement patterns.

Introduction

The way we interface humans with machines can define the success of their interaction, and is in many cases the bottleneck for overall performance[1]. Human eye movements have long been studied as a window to human cognition, with the resulting gaze-based interfaces seen as a promising medium for low-cognitive-load human-machine interaction[2]. Interfaces of this kind aim to infer intention from the way in which a user looks at the world[3,4]. A limiting factor to their development however is that they suffer from the 'Midas Touch' problem – the problem of distinguishing whether a user is simply viewing an object to intake visual information about it, or actively viewing because they intend to interact with it[5]. The Midas Touch problem remains unsolved in gaze-based interfaces, due to the challenges within accurate eye-tracking, in variable environmental conditions, resulting in imperfect data[6] as well as the complexity of interpreting gaze dynamics in an ever-shifting visual environment and mapping these to a range of possible intentions[7].

Researchers have studied gaze as a proxy for human intentions for over a century[8–11], predominantly with respect to the major constituent parts of gaze, i.e. saccades and fixations[12,13]. Measures of intention in this regard have generally focused either on differences in the distribution of gaze fixations (defined as periods of comparatively little movement, for time periods of around 350ms[14]), or on differences in saccadic velocity profiles[15–17]. Since the days of Yarbus[9] the field has shown spatial differences in gaze scanpaths in different static task contexts. Such paradigms of showing a stimulus and observing a response lie at the heart of vision science[18]; but here we aim to invert this traditional approach using machine learning, to investigate how we can decode the intentions of a person by observing *how* they look at a scene, based on purely natural gaze traces. Such an approach would unlock not only the use of natural gaze for human interfacing by enabling us to overcome the Midas Touch problem but would also unlock our scientific understanding of what controls visual saliency and task goals.

Many prior works have gone beyond static fixation distributions and investigated gaze behaviour in natural settings[19–23],

to try and capture the implicit information held within natural gaze. For 'reverse engineering' gaze behaviour to understand intention, there have been attempts at building Inverse Yarbus models[24], though most eye movement research is based on participants looking at a computer monitor with a static image[25–28]. Approaches like these may be suitable for modelling simplistic static gaze behaviour where multiple saccades and fixations are made over a constant image, but are not applicable for more complex free-viewing behaviours. The efficacy and applicability of such Inverse Yarbus models outside of very limited static scenarios has been contested[25,26,29]. With the exception of Bulling et. al's work[30], which focuses on a simplified binary classification problem, no Inverse Yarbus models detailed in the literature have been developed to work in real-world settings – that is, no models have been trained using data gathered from real-world natural experiments, where participants freely move their heads and interact with real physical objects in 3D[31]. One of the reasons for this is the increased difficulty in overcoming the Midas Touch problem in natural settings – we do not necessarily want to interact with every object we look at, and in object-dense real-world environments this problem is magnified significantly.

Past efforts to overcome the Midas Touch problem have involved techniques such as analysis of dwell-time[32], detection of focal fixation[7], or co-actuation with other modalities, such as voice[33], winking[34], or keyboard input[35] but such attempts are only successful in limited and simplified settings. These methods have the advantage of correctly interpreting the user's intentions, but ignore all natural gaze cues and force end-users to direct their gaze in an artificial manner, which can prove both difficult for users, and unreliable in all but the most narrow of scenarios.

Co-actuation of gaze with a voluntary movement is also not suitable for those users who would be the most direct beneficiaries of natural gaze interfaces, i.e. users with severe motor impairments. For these users, many of the typical human-machine interfaces which require actuation by limbs or digits, are not usable. The human oculomotor system typically remains intact even in severe cases of tetraplegia and other motor impairments[2], and eye movements are preserved for a longer period than skeletal movement with respect to neurodegenerative diseases[36], making them a viable interface for disabled users. Ideally, a gaze interface would rely solely on natural gaze cues, such that the user would direct their gaze as they naturally do, with the system having the capability to detect action intention from subtle differences in the natural gaze signal, thereby minimising cognitive load. In addition, natural gaze contains temporal characteristics that models can leverage, such as a sequential structure[20,23] and just-in-time order[21,22] when actively viewing freely in natural tasks. Hidden Markov Models have already been applied to scanpath modelling[37] for predicting future actions based on gaze information, however these models suffer considerably when data is noisy and incomplete, as is often the case in real-world recordings, and are unreliable in complex situations with more flexible behaviours.

Here we detail a new approach by capturing multimodal data sources from real-world human behaviour. By combining spatiotemporal eye movement dynamics with contextual information extracted on-the-fly through the implementation of a 'semantic fovea' (an object recognition system that can recognise and categorise objects in the field of view in real-time[38]), we obtain rich natural gaze behaviour, allowing our models to learn and distinguish subtle differences in the user's gaze dynamics. We designed an experimental process to capture task-specific natural gaze behaviour data in a dining table scenario (see Fig. 1). Subjects sat in front of a dining table where different target objects were placed. Each subject was asked, through computer-generated voice commands, to perform actions of three types: manipulate one of the objects, imagine manipulating one of them, or inspect objects to answer a question. Experiments were conducted with 16 subjects, for about 3 hours of interactions each, resulting in 13,679 total trials.

This multimodal real-world experimental approach allows us to gather sufficient quantities of high resolution contextual gaze data, such that we can build deep recurrent neural network models that can capture and distinguish subtle differences in the temporal and spatial evolution of scanpaths over salient objects. Our models can thereby accurately identify task context, based solely on a participant's stream of eye movements in a given real-world task, providing a promising step forward for data-driven natural gaze control interfaces.

Results

Dataset description

We ran data collection experiments with 16 subjects, for approximately 3 hours each. During each experiment trial, subjects interacted with 6 different target objects (a banana, a bottle, a bowl, a cup, a doughnut, and an orange) on a table in front of them, see Fig. 1. Interactions were prompted by instructions communicated through computer-generated voice commands. Each instruction was composed of what to do and which target object to do this on e.g., *"Pick and place in the center: banana"*. See supplementary materials for a comprehensive list of all instructions. The instructions covered two classes of behaviour: *inspection*, where users would visually inspect objects' attributes to answer a two-choice question, and *manipulation* where users would manipulate an object, either in physical or imaginary interaction. The structure of each instruction sentence was designed to reveal the target of the task to the subjects at the very end of the command, to minimise task-object related gaze cues appearing earlier than the end of an instruction communication.
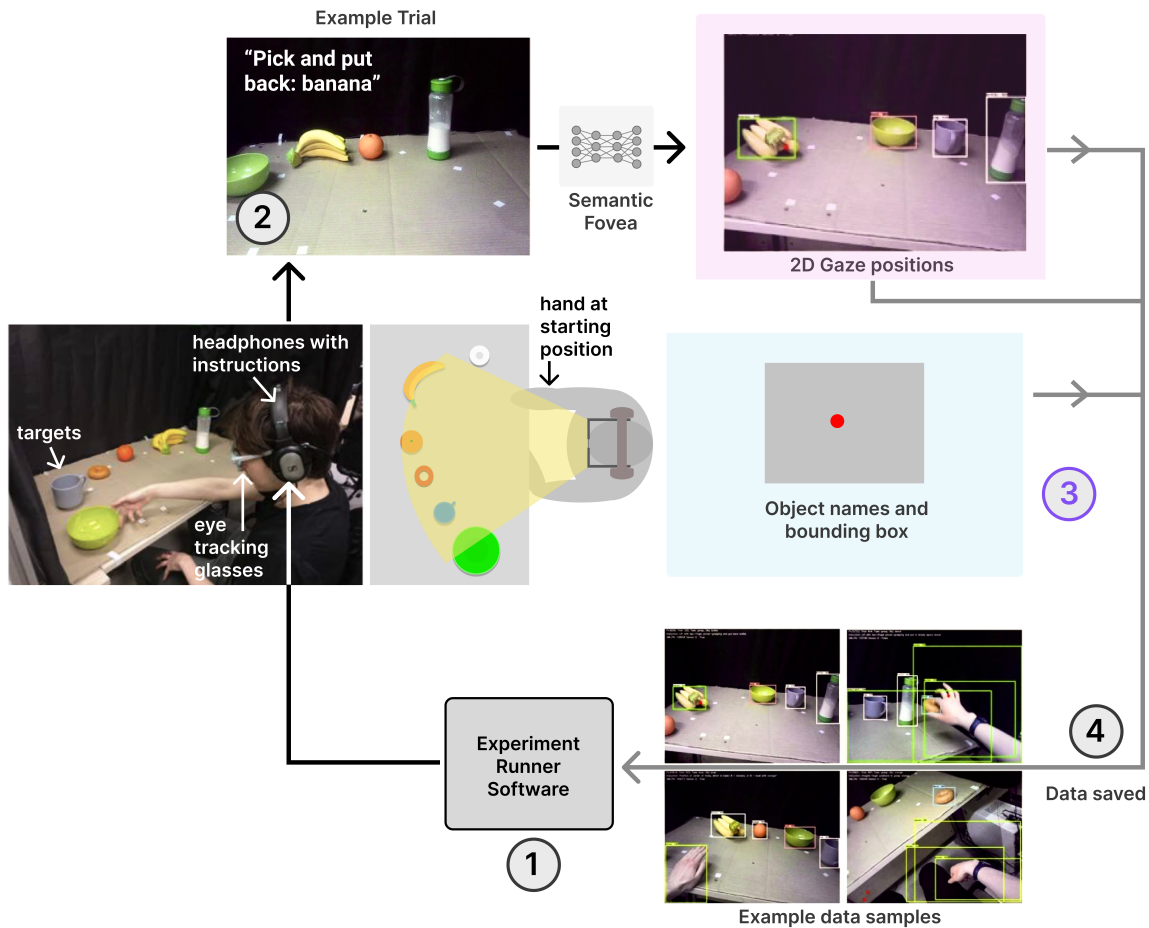
Figure 1. Data collection experiments. (1) Our experiment runner software manages the experiment, randomising the tasks and keeping track of progress. The software produces voice commands, asking the user to perform tasks requiring physical/imaginary manipulation of objects or inspection of attributes of the objects on the table. These commands are communicated to human participants through headphones (left). The participants are seated at the table with the objects placed on it, and are wearing eye-tracking glasses that record their eye-movements, and an ego-centric video stream. The subjects will then proceed to fulfil the task given to them by the experiment runner. An example of this can be seen in an ego-centric view in (2). The ego-centric video frames are passed through our Semantic Fovea[38] which detects objects in the scene and their position with respect to gaze. The generated data, i.e. 2D gaze positions, object names and bounding box positions are then all saved by the experiment runner software (4). All frames shown in the example trial, 2D gaze positions and example data samples subfigures are egocentric views recorded from our dataset.

A behavioural dataset was collected through these experiments. Participants wore a pair of eye-tracking glasses (SMI ETG 2W, SensoMotoric Instruments, Germany), which recorded both ego-centric video and gaze position in pixels within the video frame, at average frequencies of 30Hz and 120Hz, respectively. Video frames were post-processed, going through our Semantic Fovea[38], a convolutional neural network trained for object detection within ego-centric videos, relative to gaze points, producing object names and bounding boxes, see Fig. 1. Our experimental setup was equipped with optical sensors, to monitor which object slots were occupied, and for the subjects to answer the two choice question on object attributes presented to them in each inspection task by placing their hand on one of two sensors. See Methods for details. The sensor values are part of the dataset but were not used for our classification study. The intention decoding models we present here take as features the gaze and object positions within a video frame. These signals were not recorded at the same frequency; gaze points were recorded at 120Hz while object positions were extracted from video frames recorded at 30Hz. Moreover, the object detection algorithm can occasionally miss objects in a given frame. We used interpolation to make up for the gaps resulting from recording frequency mismatches and object detection errors. See Methods for details.

The experiment was designed to be conducted in 1-hour batches of trials, with 10 minutes of break every hour. In total, 16 subjects took part in the experiments, all right-handed, completing two or three 1-hour long batches each, resulting in a total of 13, 679 trials, evenly split between manipulation (itself split into physical or imaginary, 53.4% and 46.6% respectively) and inspection tasks. See Fig. 1 for an overview of the experimental procedure. Gaze data recorded from all 16 subjects follows normal Gaussian distributions both in the horizontal and vertical axes of the egocentric image frame, with the emergence of a central bias[39] of gaze behaviour in head-centric coordinates, see Fig. 2A. There is slight but noticeable difference when comparing the centrally-biased distributions of the two classes of tasks, with the manipulation task showing a consistent ≈100 pixel offset in the horizontal axis across all subjects, and a negligible difference in the vertical axis (Fig. 2A).

Gaze speeds across all subjects follow similar patterns, showing a chi-squared distribution (Fig. 2B). When comparing mean gaze distance covered over all subjects' respective trials, we can see that the variance on a per-subject basis is considerably higher than the difference between the two tasks (Fig. 2C). We note however that there is a small but systematic difference in the mean gaze distance covered between inspection and manipulation trials, with the distance covered across inspection trials systematically higher than that of manipulation trials (Fig. 2C). We investigate these systematic differences in the analysis and validation of our classifier.

Gaze-based intention classification

As a first model for predicting gaze context from scanpaths, we investigate the horizontal shift of gaze points observed in the data between the two types of tasks (see Fig. 2A). We focus on static 2D gaze positions on the target object, attempting to classify intentions based on the average horizontal gaze position within the target object bounding box. Fig. 3A shows the average normalised image of an orange obtained by normalising all the orange object frames given by the Semantic Fovea[38].

The naive hypothesis here is that as all our participants are right-handed, and would reach to manipulate objects from the right-hand side, the distribution of gaze positions within the object bounding box would show a higher concentration on the right-hand portion, where they would place their hand – we refer to this as handedness bias. To check for this hypothesis, we implemented our handedness bias classifier which takes as input the average of the horizontal positions of the gaze points on the object, with gaze positions normalized to the object bounding box, and classifies the user's intention as manipulation if this average is above a given threshold. The threshold could vary from 0 (left border of the object) to 1 (right border of the object). If the average normalized horizontal gaze position is above the threshold, then the trial is considered as manipulation, otherwise it is identified as inspection.

Fig. 3B shows data from a sample trial where the target object was the orange and Fig. 3C shows a sketch of the handedness bias model. This model was evaluated with thresholds ranging from 0 to 1 by steps of 0.01 and with best accuracy reached on our dataset being 51%, i.e. almost chance level. Such low accuracy on a balanced binary classification task indicates that the average horizontal gaze position on the object does not carry distinguishable information with regards to the context of gaze, i.e. the intended task, and is therefore not a good feature to infer the user's intention from.

For a more general evaluation of gaze position on target object as a feature for intention decoding, we set out to determine whether the distribution of gaze positions on the target object differs from one intention to the other. We used Gaussian Mixture Models (GMMs)[40] to learn the static distribution of gaze positions on the target object for each intention class. A GMM is a flexible model for probability distributions, which fits a weighted set (or mixture) of Gaussian distributions to a target distribution. Here we are trying to model two target distributions, one per intention, so we fit two different GMMs to our data. To classify a trial, the set of gaze points on the target object is extracted, and the likelihood of this set under each of the models is computed. The intention with the highest likelihood is then associated to the observed gaze data. The number of Gaussians in each GMM was determined by optimizing the Bayesian Information Criterion (BIC)[41].

Fig 3E and Fig 3F show the learnt distributions for inspection and manipulation trials respectively, on an orange. To help in reading the difference, Fig 3F shows the difference between the two learnt models. The accuracy of the GMM approach in
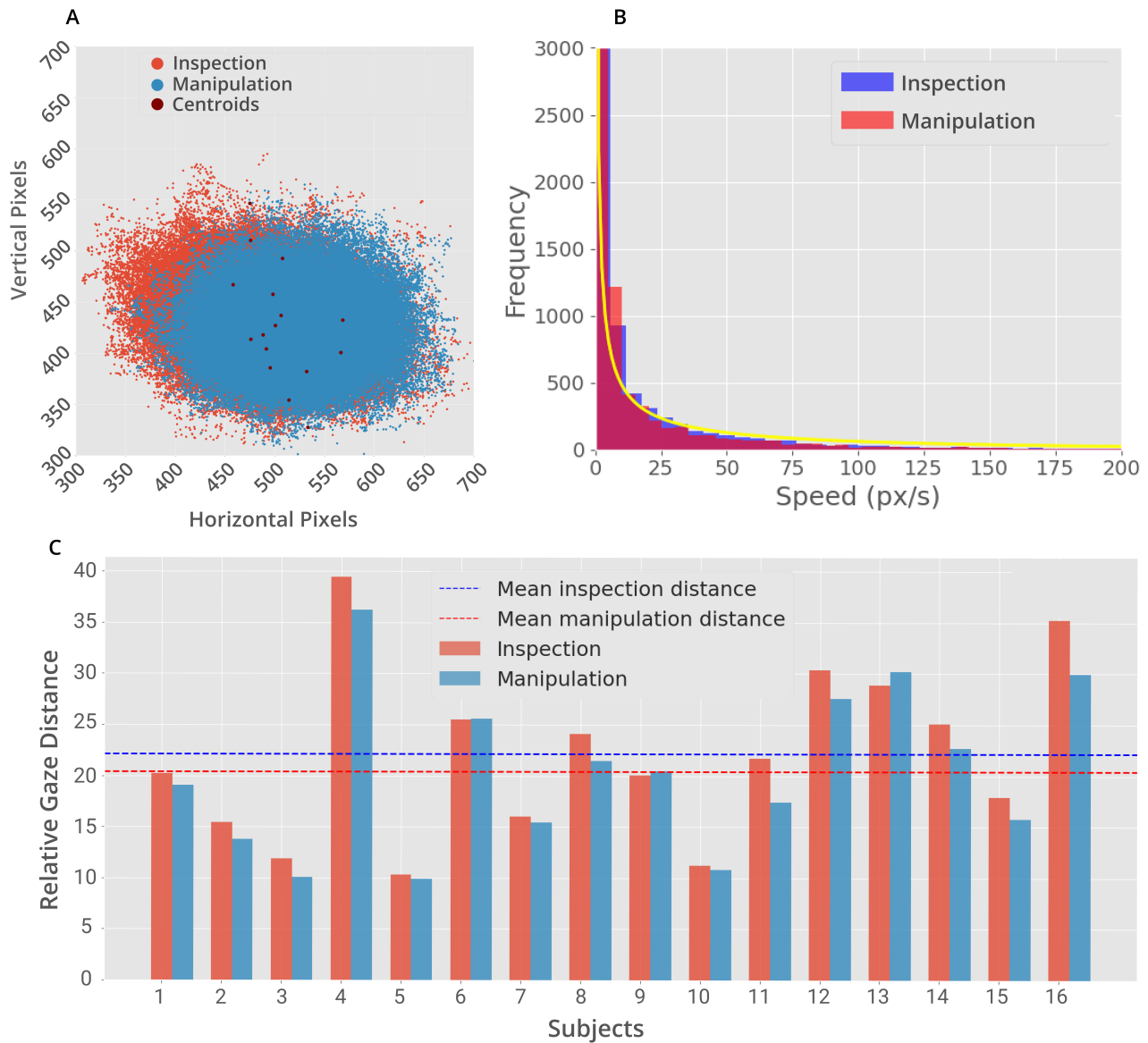
Figure 2. Dataset overview. A) shows a slight difference in the mean 2-D distributions of gazepoints across all subjects in the two different tasks, with the manipulation task right-shifted across all subjects. We also overlay in purple the centroids of the subjects' individual gaze distributions on the grand average per-task distributions. B) shows a histogram of the speed distribution difference between the two different modalities, both fitting a chi squared distribution. C) Mean gaze distance covered over all subjects' respective trials, showing variance on a per-subject basis as considerably higher than the difference between the two tasks

decoding intention was 53%, barely above chance level. This result indicates that static gaze distributions on the target object do not carry enough information to discriminate between user intentions. Two factors limit the approaches we have shown here: ignoring the time dimension, and the dismissal of all gaze positions which did not land in the target object's bounding box.
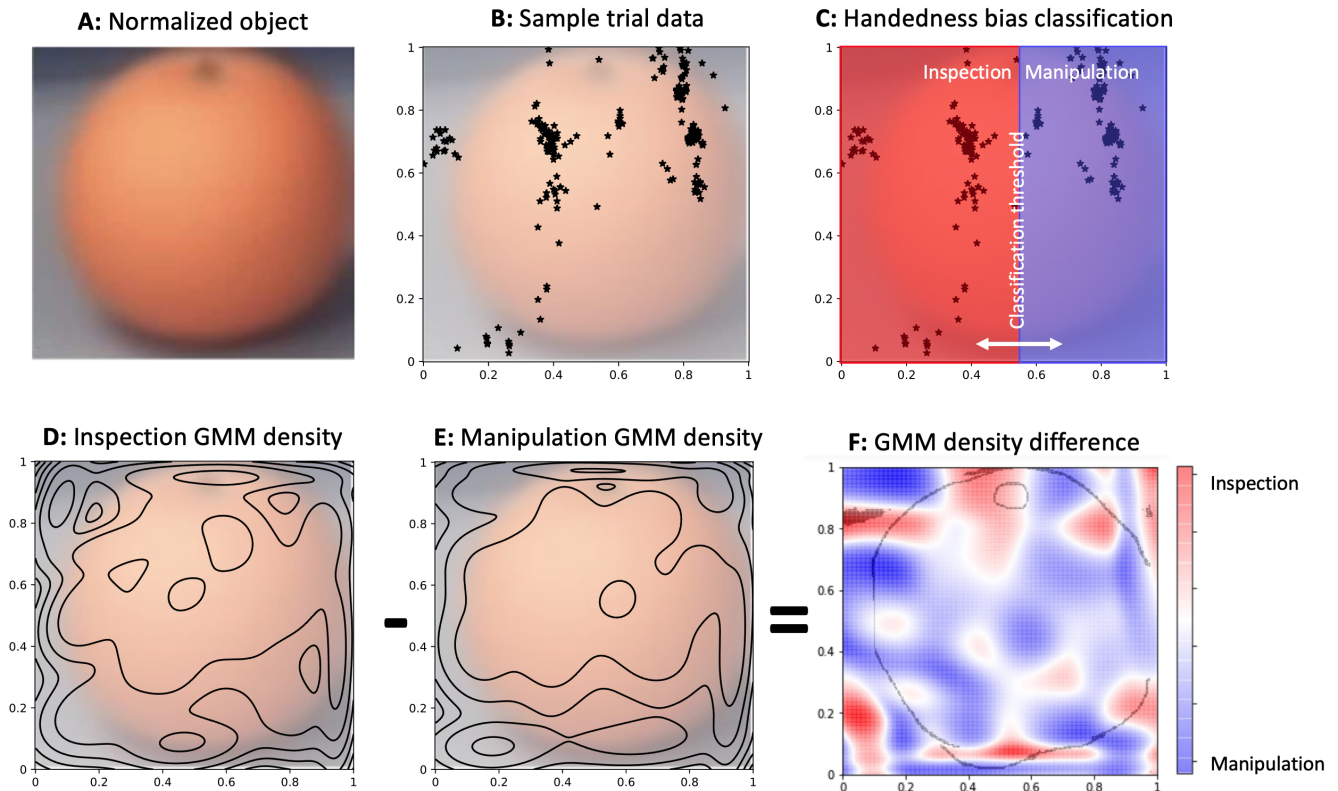

Figure 3. Illustration of the handedness bias and GMM intent classifiers with the orange as an example target object. **(A)** Normalised picture of the orange. **(B)** Gaze points landing on the orange on a sample trial where it is the target object. **(C)** Illustration of the handedness bias classifier, which looks at the relative amount of eye movements to the left or right of the vertical decision boundary between the red section (inspection) and the blue section (manipulation) - for a right handed user using the right hand for interaction. **(D)** Contour plot of the learnt GMM density for the inspection trials. **(E)** Contour plot of the learnt GMM density for the manipulation trials. **(F)** Heatmap of the difference between the learnt GMM densities for inspection and manipulation trials.

The observations from the results above confirm findings described in the literature regarding the importance of gaze dynamics and the order of motion in gaze scanpaths over time[42]. To incorporate this dimension into our classifiers, we implemented models which use time series data as input. We considered different sets of features: gaze position, gaze and target object positions, gaze-target distance and gaze speed. We also trained models to classify intention from gaze-target distances distribution and gaze speeds distribution. See Fig. 5D for a summary of the different features considered. The purpose is now to find a model which does well at classifying the different input sequences to recognize the user's intent.

Treating intention decoding as a sequence classification problem is well studied in Natural Language Processing (NLP), particularly in the task of sentiment analysis, where algorithms are designed to decode the main sentiment of a sentence[43]. Here, we took inspiration from the work of Baziotis et. al.[44] who have used an LSTM-based sentiment decoder running on tweets. Fig. 4 presents the architecture of our gaze intention classifier, MIDAS. First, input time series (2D gaze position alone or with target object position, gaze speed or gaze-target distance) are fed to a bidirectional LSTM network[45], which captures the time dynamics of the input features. The output of the LSTM block is then fed to an attention layer[46] which focuses the model on the most meaningful chunks of the time series. The output of the attention layer is fed into a fully connected neural network which outputs the user's intention. To benchmark MIDAS's performance, we also trained standard machine learning models to decode user intention from our behavioral gaze dataset, namely Gradient Boosting, Support Vector Machines (SVMs), and Logistic Regression.

Different feature sets are explored to classify intent. Given its architecture, MIDAS only allows time series data as input (2D gaze positions, gaze-target distance or gaze speed, see Fig. 5D). All other machine learning models used in this study only
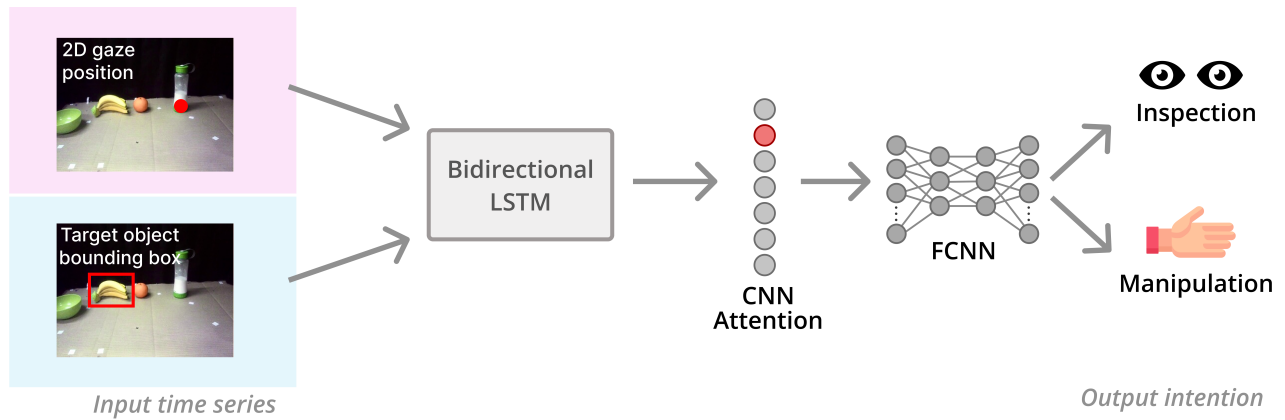
Figure 4. Sketch of MIDAS's architecture. Inputs are 2D gaze and target position time series (or any other time series feature such as gaze speed or gaze-target distance), color coding refers to Fig. 1. The input time series are fed to a bidirectional LSTM layer, an attention layer and finally a fully connected network which outputs the intention class.

require fixed-size inputs, meaning that they accept both fixed-length time series and their histograms. The models were trained and evaluated under cross-validation schemes, in which the available data is split into several chunks, and the model is trained successively on all except one left out for validation. We used two cross validation schemes: 5-fold where the full dataset is randomly split into 5 chunks of data, and leave one subject out where each chunk is the data from one subject. The latter allows to evaluate how well the model would generalize when being fed data from new, unseen users.

Fig. 5A shows the mean and standard deviations of MIDAS's accuracy from leave one subject out cross-validation. The patterns on the bars in Fig. 5A correspond to those shown in the Venn diagram in Fig. 5D. The highest mean accuracy reached by MIDAS on leave one subject out cross-validation is 91.1%, for the case where an attention layer is used in the architecture, and no target object position information is provided. However, the MIDAS accuracy does not vary much without attention, or with inclusion of target object position data, as can be seen in Fig. 5A, reaching a minimum of 89.13% mean accuracy in the case with no attention layer and no target object position data. This indicates that the result is robust to the inclusion/exclusion of an attention layer in the architecture, and also that knowing the target object a priori is not a requirement to maintain the model's performance level.

Fig. 5C shows means and standard deviations for MIDAS and standard machine learning approaches, on 5-fold cross validation, with different feature set inputs; here as well, the bar chart patterns are indicative of feature type and correspond to those shown in the Venn diagram in Fig. 5D. Given raw gaze traces and the target object's position as input MIDAS outperforms standard machine learning approaches on the intention decoding task, rising up to 91.9% mean accuracy, compared to highest mean accuracy of other models being 84.5% in the case of gradient boosting. The same observation can be made with gaze speed and gaze-target distance time series as input, where MIDAS reaches respectively 82.6% and 73.1% accuracy, each at about 10% above the next best performing model. Gradient boosting outperforms logistic regression and SVM on this task, no matter the feature set. It can also be noticed that the histogrammed feature sets carry some information about the user's intent as classifiers can reach an accuracy significantly above chance with them. Table 1 shows detailed results of the classification accuracy for all models on 5-fold cross validation.

Input Feature	MIDAS	Gradient boosting	SVM	Logistic regression	Handedness bias
Raw gaze	**89.8±0.5**	83.1±0.6	56.9±0.9	49.7±0.2	N/A
Raw gaze and target position	**91.9±0.5**	84.5±0.6	56.8±0.9	49.7±0.2	50.5±0.1
Gaze - target distances	**73.1±0.9**	63.7±0.7	56.1±1.2	50.5±1.0	N/A
Gaze speeds	**82.6±0.5**	70.8±1.0	56.4±0.8	52.9±0.7	N/A
All time series	**91.9±0.7**	84.3±0.8	57.8±0.2	50.6±0.6	N/A
Histogrammed distances	N/A	**76.8±0.9**	75.1±0.7	62.5±0.7	N/A
Histogrammed speeds	N/A	**72.0±0.8**	70.9±1.0	67.5±1.3	N/A

Table 1. Accuracy and its standard deviation (5-fold cross validation) of all trained models and input data. N/A indicates impossible combinations of input and classifier. Models within two standard deviations of chance 50% are greyed out. The top-performing model for each feature set is highlighted in bold, the top performing models have a shaded table cell.
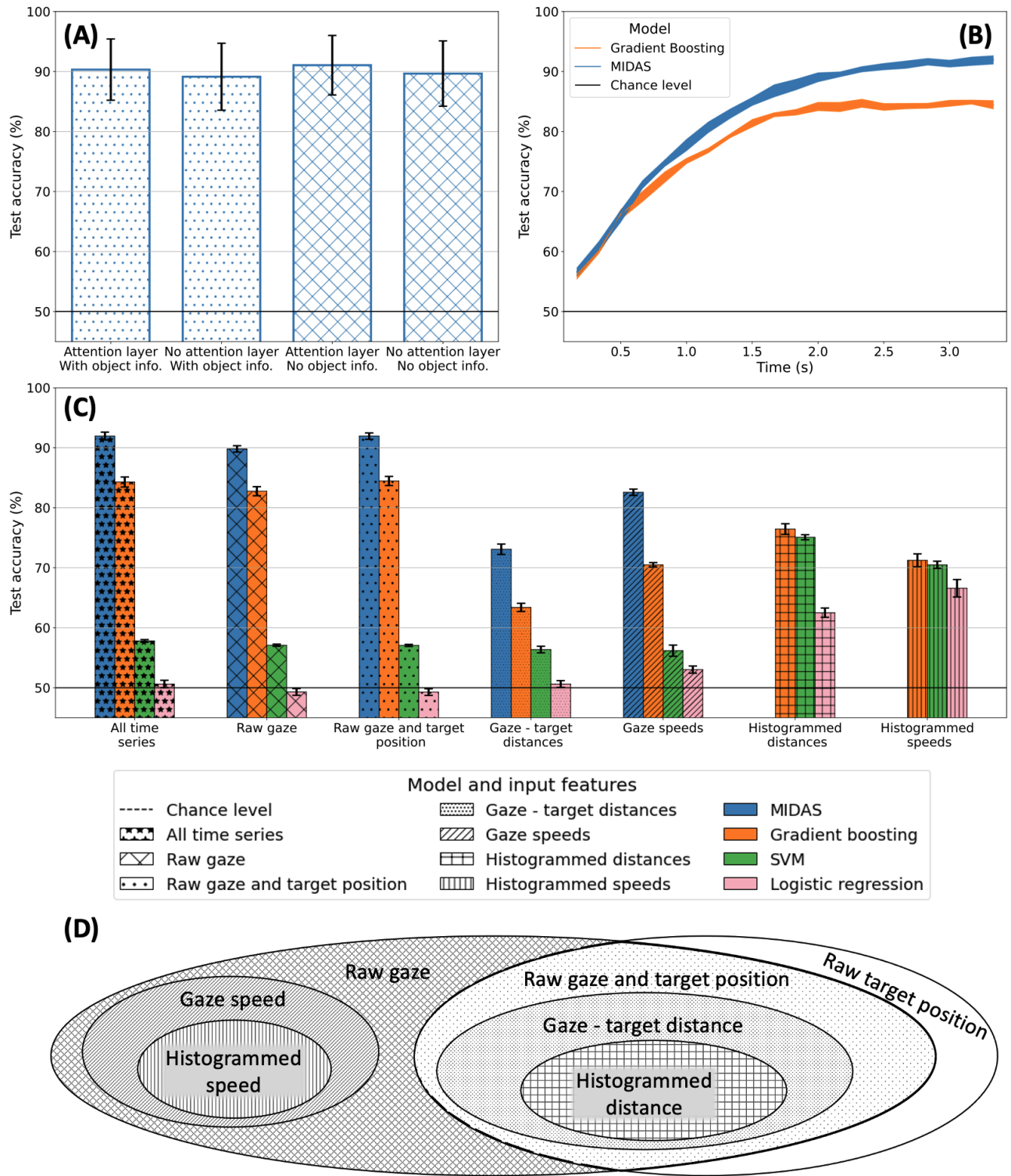
Figure 5. The performance of our model, MIDAS, in ablation studies and in comparison with standard models. (a) Mean and standard deviation of MIDAS's accuracy on leave-one-subject-out cross-validation, with ablations on inclusion of the attention layer in the deep network, and the target object position. (b) Accuracy of MIDAS, gradient boosting and random forest over time from the moment when the subject has heard the full task description. (c) The performance of MIDAS on 5-fold cross-validation compared to more standard machine learning classification models. In all plots, the dashed black line indicates chance level.

The results presented in Fig. 5A show that the presence or lack of an attention layer has very little impact on MIDAS's performance, and so does having the target object's position. Moreover, it can be seen that MIDAS generalizes well to unseen users, even if only the raw gaze data is available (Fig. 5A). To better understand the performance of MIDAS over different lengths of input data streams, in Fig. 5B we present the evolution of MIDAS's accuracy with respect to the time series duration it is being given to infer an intention. After only 600 milliseconds, MIDAS can determine the user's intent with 65% accuracy and this only goes up with time, reaching a peak performance of around 92% accuracy with 3.3 seconds of input time series. This result shows that there is a latency versus accuracy trade-off when using MIDAS, which could potentially be set by users to suit their convenience. We see a similar trend for the best performing standard machine learning model, i.e. gradient boosting (Fig. 5B), showing an initial tie with MIDAS in shorter sequence lengths, with MIDAS taking over as sequence lengths increase. It must be noted that MIDAS has the advantage that a single model is trained to cope with time series of all lengths, whereas gradient boosting models need to retrain for different sequence lengths.

These results lead to two essential conclusions. First, we show the possibility of learning patterns in natural gaze behavior data which allows us to infer a users' intent. Second, the longer the gaze sequence, the more accurate the decoding of the intention behind it. These two observations, confirmed by MIDAS and standard machine learning models as well, imply that there are characteristic patterns to specific intentions in our dataset of natural gaze behaviour, and that these cues are spread across time.

Discussion

We set out to investigate whether we could utilise deep recurrent neural networks to overcome the Midas touch problem of intention decoding in hand-based manipulations in a real-world, desk-based scenario. Our deep classifier, MIDAS, learns to distinguish between different gaze dynamics in real-world behaviour, allowing us to overcome the Midas Touch Problem[5] with a high degree of accuracy to distinguish between mere visual inspection and hand manipulation intentions. MIDAS can infer user intent with relatively little data, with the accuracy rising quickly from the first few hundred millisecond (including processing delays of verbal instruction) at 65% to plateauing with 92% at 3.3 seconds (see Fig. 5B). This performance is in contrast to standard machine learning methods including methods developed previously to help resolve the Midas Touch problem.

We investigated whether the model was simply learning large spatial or temporal differences in the data. The gaze distance covered during the task varies considerably more across individuals than across the two classes, although there is a slight but systematic increase in gaze distance covered in the inspection task versus the manipulation task, see Fig. 2. This is perhaps to be expected, given the more active information-gathering for object comparisons that needs to take place in our inspection tasks, compared with a more focused approach in the manipulation tasks. There is considerable overlap between the two classes with respect to their spatial distributions, with the natural central-biased 2D Gaussians of both fitting with gaze behaviour described in the literature[39]. Gaze speed is consistent between the two classes, with both classes' speeds fitting a chi-squared distribution closely, see Fig. 2. We investigated the effect of these potential biases within gaze-target distance, and gaze speed by training conventional machine learning algorithms on these features. The results (see Fig. 5C) show that MIDAS outperforms all of these models, suggesting that MIDAS is capturing dynamics beyond gaze-target distance and gaze speed.

A body of literature show how gaze temporal dynamics implicitly hold information both about intentional state (e.g.[27,49]) and task structure (e.g.[50]. The weak performance of the GMM model indicates that gaze locations alone do not carry sufficient information to build reliable inverse Yarbus models. It is important to investigate the evolution of temporal structure across these natural tasks, if we are to properly build predictive models of natural behaviour that are consistent with the dynamic and ever-changing nature of our top-down goals. There is an unresolved question in the field as to what extent eye movements are driven by bottom-up processes[51,52], in response to visual features in a scene, versus by top-down processes, by our intentions and task goals. Our results with MIDAS lend weight to the combination of the two, but highlight the importance of evolving top-down subgoals in determining gaze location and future scanpaths.

Our approach uses unconstrained natural gaze behaviour as a action intention signal from free head and body movement based eyetrscking in a real-world setting involving physical manipulation of ovjects. We can now distnguish the subtle differences within eye movement trajectories across two major modes of intention that triggers eye movements. Taking our approach into practice means,e.g. as a user interface in VR or as an asstivie technology for motor impaired, we can implement a Zero User interface (Zero UI) system[53], that detects the user's intentions while they act naturally, and will therefore not require any interactions with a user interface nor users having to learn custom gaze "gestures" (such as blink timings, dwell times, or complex eye movement patterns) to resolve the Midas Touch problem. MIDAS operates and is trained on natural eye movements occurring during real-world interactions. Our approach is entirely data-driven, and based on a new machine learning model that only requires free movement eye-tracking data combined with ego-centric video to operate. This means that our approach is customisable to many real-world, natural tasks, simply by including further scenarios in the training phase. We have previously shown how navigation intentions can be successfully decoded from natural eye movements with a Zero UI

approach and deployed it successfully in wheel chair driving in real-world settings[54]. However, while the system there opeated on natural eye movement behaviour, the temporal component was less important in the navigational setting of a wheelchair. Here, we demonstrate the ability to resolve the Midas Touch problem for actual touch and physical manipulation, where we show that the temporal component of eye movements are essential for intention decoding.

We have shown a data-driven route to resolve the Midas Touch problem and decode natural eye movements associated with action intentions versus visual inspection, previously in navigation tasks[53,54] and now here in manual ones. Our findings prompt the question whether a general-purpose, multi-task, gaze-based intention decoder robust to the Midas Touch problem can be created simply through collecting sufficiently rich and diverse task data. The tools for ubiquitous data collection using wearable eye-tracking and mobile computing are there. The scale of the required labelling of the action intention could be a challenge if done by hand, but automatic labelling for action intention either through automatic action recognition[55] or latent embeddings of movements of the body and the eyes[56,57] may be timely solutions. It has now been 60 years since yarbus, the rcognittion that eye movement interacxes have benefits with limitations due to the Midas Touch problem[15]. Resolving the Midas Touch problem using our approach may help further boost the deployment of eyetracking technologies from controlled lab spaces into e.g. inside vehicles to support consumers[58], to non-invasive easy to setup-and-go assistive technology for motor impaired users[59], and to create more intuitive human-computer and human-robot interaction settings.

Methods

Experimental setup

Experimental design - For our data collection experiments, we designed an experimental setup to collect natural gaze data for two specific intentions: Inspection and Manipulation. Figure 1 gives an overview of the experimental pipeline. Participants sit in front of a dining table with six different objects on it. The experiment was split into trials. A trial would start with the subject hearing a computer generated audio description of the task through an over-ear headset they were wearing. The subject then proceeds to fulfil the given task, after which the trial is complete and a new one begins. Trials were designed to capture the two different intentions when looking at an object, i.e. Inspection or Manipulation.

The tasks in the inspection class relate to comparing objects on a given feature, e.g. *"Weight, which is higher: A - Bottle, or B - Bowl?"*. The manipulation class of tasks involved different types of object manipulation, either requiring for the task to be physically performed, or imagined, e.g. *"Pick and put back: Banana."* or *"Imagine finger positions to eat from: Bowl."*. A full list of given tasks are provided in supplementary materials. As observed in above examples, instructions are given in such a way so as to reveal the target object of interaction as the last word. The participants would therefore have to wait for the instructions to be finished before proceeding to complete the task. This enables us to control which part of the trial carries task-relevant gaze cues for our classification model to be trained on. Trials were taken by participants in sessions of one hour. In total, 16 subjects, all right-handed, took 2 to 3 sessions each, leading to a grand total of 13,679 trials and about 45 hours of recorded data.

Setup - An experiment runner software was designed and coded for these experiments, which would handle the progression of experiments, including randomisation of tasks, and generation of voice commands, as well as recording all the streams of data and producing the final dataset for our classifier. The software was run on a laptop using an NVIDIA GTX 1080 with 64GB RAM. The participants wore a pair of over-ear headphones and eye-tracking glasses (SMI ETG 2W, SensoMotoric Instruments GmbH, Teltow, Germany). Optical proximity sensors were integrated into the dining table setup with an Arduino Uno which was then interfaced with by the experiment runner software via USB, allowing participants to indicate an answer in the comparison tasks by holding their hand over the respective sensor, to select A or B, and also keeping track of object places, i.e. which object space on the table is empty. Egocentric video and gaze position on that video were recorded from the eye-tracking glasses. Gaze position was recorded at an average frequency of 120 Hz while video frame recordings are at 30 Hz. As the different recording devices had different recording rates, the time code of each of their recordings were also saved to ensure proper reconstruction of the events' timeline when working with the dataset later on. The recorded images were then fed into Semantic Fovea[38] for object recognition and bounding box definition. The sensor values, as well as the given task and objects involved are also recorded. In summary, 3 main classification-relevant assets were recorded: ego-centric video, ego-centric gaze position and object names and bounding boxes.

Data preprocessing, bias mitigation and model verification

Gaze position data was used raw, except for the very few missing data points which were set to 0. Object position was recorded at around $30Hz$ while gaze positions are at around $120Hz$. It was therefore necessary to interpolate the object positions to fit the recording rate of the gaze traces. There are three major reasons for gaps in object positions' time series: the difference in recording frequencies, a failure of the object detection algorithm and the object actually not being in the frame. An interpolation strategy was put in place to decide which gaps should be filled in and which ones should be left empty. The decision criterion was the length of the gap, consider that if data is missing due to a failure of the detection algorithm, the gap is likely to

be smaller than the case if the object is simply not in the frame. From the CDF of object position gap sizes presented in Supplementary Fig. S1, it was decided to interpolate all gaps of less than 100 data points ($\simeq 0.8$ seconds).

With the experimental design above, we anticipate two potential sources of bias: (1) Inspection task descriptions are longer than Manipulation task descriptions. (2) Two or three objects are named in Inspection tasks against a single object in Manipulation ones. We studied these potential biases and mitigated where necessary.

The first major difference between two trials with a different intention is their length. Indeed, instruction texts for inspection trials are longer on average (therefore longer to pronounce) than those with manipulation intention. As such, it could be possible to get a trial's corresponding intention solely from its length. To mitigate this bias, we only considered data which was recorded after the audio task description ended. However, even when considering only the post-instruction section of a trial, not all trials were the same length, still representing a bias. Fig. S2C presents the inverse trial lengths CDF for each intention class. The difference in trial lengths start to appear after 400 points of recording so the rest of the trials was discarded to present from any bias. We thus cropped each sample to keep only a central part from the end of the audio description to the point where most recordings stop. This left us with around 3.33 seconds of recording for each trial to train our classifier on. See the effect of cropping the trials in Fig. S2.

The second bias which was studied comes from the fact that the number of objects named in the task descriptions was different with the associated intention. More precisely, two objects were named in inspection tasks against only one in manipulation tasks. This difference could impact the gaze trace by making the eyes move further from the target object during inspection trials. To check that the impact of this difference was light, we trained classifiers only on distance features. Fig. 5.b. shows that these models couldn't perform well at the task of intention decoding, therefore the bias was considered as not significant.

MIDAS and other classifiers

All classification models mentioned in this paper were trained on the gaze and target object traces or a function of these (e.g. target-object distance, histograms, ...). Models were trained on different variants of cross-validation. Cross-validation consists in splitting a dataset into a given number N of folds, setting a specific fold as test set, training the model on all the $N - 1$ others and do this for each fold. This allows to get N samples of the model accuracy, giving an idea of the distribution of accuracies for a specific model. In this paper, two types of cross-validation are used. In k-fold cross-validation, the original dataset is split randomly into k equally sized smaller sets: the folds. In leave one subject out, each fold is the data of a specific subject.

All the standard models were trained with Python's `sklearn` and MIDAS was implemented using `Tensorflow 2`. All instances of MIDAS waas trained during 200 epochs. In details, MIDAS is composed of 2 bidirectional LSTM layers with 256 neurons each, one attention layer leading to 3 dense layers with respectively 32, 16 and 2 neurons, with ReLU activation for the first two and sigmoid activation on the output layer. MIDAS was trained to minimise a binary cross-entropy loss using the Adam optimiser with a learning rate of 2.5×10^{-4}.

References

1. (ed.) Laurel, B. The art of human-computer interface design. *New York: Addison-Wesley.* (1990).

2. Abbott, W. & Faisal, A. Ultra-low-cost 3d gaze estimation: an intuitive high information throughput compliment to direct brain–machine interfaces. *J. neural engineering* **9**, 046016 (2012).

3. Adjouadi, M., Sesin, A., Ayala, M. & Cabrerizo, M. Remote eye gaze tracking system as a computer interface for persons with severe motor disability. In *International conference on computers for handicapped persons*, 761–769 (Springer, 2004).

4. Chin, C. A., Barreto, A., Cremades, J. G. & Adjouadi, M. Integrated electromyogram and eye-gaze tracking cursor control system for computer users with motor disabilities. *J. Rehabil. Res. & Dev.* (2008).

5. Velichkovsky, B., Sprenger, A. & Unema, P. Towards gaze-mediated interaction: Collecting solutions of the "midas touch problem". In *Human-Computer Interaction INTERACT'97*, 509–516 (Springer, 1997).

6. Niehorster, D. C. *et al.* The impact of slippage on the data quality of head-worn eye trackers. *Behav. Res. Methods 2019 52:3* **52**, 1140–1160, 10.3758/S13428-019-01307-0 (2020).

7. Velichkovsky, B. B., Rumyantsev, M. A. & Morozov, M. A. New solution to the midas touch problem: Identification of visual commands via extraction of focal fixations. *Procedia computer science* **39**, 75–82 (2014).

8. Buswell, G. T. *How people look at pictures: a study of the psychology and perception in art* (University of Chicago Press, 1935).

9. Yarbus, A. L. Eye movements during perception of complex objects. In *Eye movements and vision*, 171–211 (Springer, 1967).

10. Grasso, R., Prévost, P., Ivanenko, Y. P. & Berthoz, A. Eye-head coordination for the steering of locomotion in humans: an anticipatory synergy. *Neurosci. Lett.* **253**, 115–118 (1998).

11. Hollands, M. A., Patla, A. E. & Vickers, J. N. "look where you're going!": gaze behaviour associated with maintaining and changing the direction of locomotion. *Exp. brain research* **143**, 221–230 (2002).

12. Becker, W. & Fuchs, A. F. Further properties of the human saccadic system: eye movements and correction saccades with and without visual fixation points. *Vis. research* **9**, 1247–1258 (1969).

13. Fuchs, A. F. The neurophysiology of saccades. *Eye movements psychological processes* 39–53 (1976).

14. Rayner, K. Eye movements and visual cognition: Introduction. In Rayner, K. (ed.) *Eye Movements and Visual Cognition: Scene Perception and Reading*, 1–7 (Springer New York, New York, NY, 1992).

15. Jacob, R. J. What you look at is what you get: eye movement-based interaction techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 11–18 (1990).

16. Salvucci, D. D. & Goldberg, J. H. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, 71–78 (2000).

17. Duchowski, A. T. A breadth-first survey of eye-tracking applications. *Behav. Res. Methods, Instruments, & Comput.* **34**, 455–470 (2002). Publisher: Springer.

18. Lorenz, R. *et al.* The automatic neuroscientist: A framework for optimizing experimental design with closed-loop real-time fMRI. *NeuroImage* **129**, 320–334 (2016).

19. Rothkopf, C. A. & Pelz, J. B. Head movement estimation for wearable eye tracker. *Proc. 2004 symposium on eye* (2004).

20. Land, M., Mennie, N. & Rusted, J. The roles of vision and eye movements in the control of activities of daily living. *Perception* **28**, 1311–1328 (1999). Publisher: SAGE Publications Sage UK: London, England.

21. Hayhoe, M. M., Shrivastava, A., Mruczek, R. & Pelz, J. B. Visual memory and motor planning in a natural task. *J. vision* **3**, 6 (2003).

22. Keshava, A. *et al.* Just-in-time: gaze guidance behavior while action planning and execution in vr. *bioRxiv* (2021).

23. Johansson, R. S., Westling, G., Bäckström, A. & Flanagan, J. R. Eye–Hand coordination in object manipulation. *The J. neuroscience: official journal Soc. for Neurosci.* **21**, 6917–6932 (2001).

24. Haji-Abolhassani, A. & Clark, J. J. An inverse yarbus process: predicting observers' task from eye movement patterns. *Vis. research* **103**, 127–142 (2014).

25. Greene, M. R., Liu, T. & Wolfe, J. M. Reconsidering yarbus: A failure to predict observers' task from eye movement patterns. *Vis. research* **62**, 1–8 (2012).

26. Borji, A. & Itti, L. Defending yarbus: Eye movements reveal observers' task. *J. vision* **14**, 29–29 (2014).

27. Iqbal, S. T. & Bailey, B. P. Using eye gaze patterns to identify user tasks. https://www.interruptions.net/literature/Iqbal-GHC04.pdf (2004). Accessed: 2021-6-17.

28. Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G. & Olejarczyk, J. Predicting cognitive state from eye movements. *PloS one* **8**, e64937 (2013).

29. Castelhano, M. S., Mack, M. L. & Henderson, J. M. Viewing task influences eye movement control during active scene perception. *J. vision* **9**, 6.1–15 (2009).

30. Bulling, A., Weichel, C. & Gellersen, H. EyeContext: recognition of high-level contextual cues from human visual behaviour. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 305–308 (Association for Computing Machinery, New York, NY, USA, 2013).

31. Boisvert, J. F. G. & Bruce, N. D. B. Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing* **207**, 653–668 (2016).

32. Hansen, J. P., Johansen, A. S., Hansen, D. W., Ito, K. & Mashino, S. Command without a click: Dwell time typing by mouse and gaze selections. In *INTERACT*, vol. 3, 121–128 (Citeseer, 2003).

33. Parisay, M., Poullis, C. & Kersten, M. EyeTAP: A novel technique using voice inputs to address the midas touch problem for gaze-based interactions. *arxiv* (2020). 2002.08455.

34. Dziemian, S., Abbott, W. W. & Faisal, A. A. Gaze-based teleprosthetic enables intuitive continuous control of complex robot arm use: Writing & drawing. In *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 1277–1282 (IEEE, 2016).

35. Hansen, J. P., Alapetite, A., MacKenzie, I. S. & Møllenbach, E. The use of gaze to control drones. In *Proceedings of the symposium on eye tracking research and applications*, 27–34 (2014).

36. Tostado, P. M., Abbott, W. W. & Faisal, A. A. 3d gaze cursor: Continuous calibration and end-point grasp control of robotic actuators. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 3295–3300 (IEEE, 2016).

37. Coutrot, A., Hsiao, J. H. & Chan, A. B. Scanpath modeling and classification with hidden markov models. *Behav. Res. Methods* **50**, 362–379 (2018).

38. Auepanwiriyakul, C., Harston, A., Orlov, P., Shafti, A. & Faisal, A. A. Semantic fovea: real-time annotation of ego-centric videos with gaze context. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 1–3 (2018).

39. Tatler, B. W. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *J. vision* **7**, 4 (2007).

40. Reynolds, D. A. Gaussian mixture models. *Encycl. biometrics* **741** (2009). Publisher: Springer City.

41. Neath, A. A. & Cavanaugh, J. E. The Bayesian information criterion: background, derivation, and applications. *Wiley Interdiscip. Rev. Comput. Stat.* **4**, 199–203 (2012). Publisher: Wiley Online Library.

42. Land, M., Mennie, N. & Rusted, J. The roles of vision and eye movements in the control of activities of daily living. *Perception* **28**, 1311–1328 (1999). Publisher: SAGE Publications Sage UK: London, England.

43. Rosenthal, S., Farra, N. & Nakov, P. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 502–518 (2017).

44. Baziotis, C., Pelekis, N. & Doulkeridis, C. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 747–754 (2017).

45. Graves, A. & Schmidhuber, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks* **18**, 602–610 (2005).

46. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

47. Lin, T.-Y. *et al.* Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755 (Springer, 2014).

48. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255 (Ieee, 2009).

49. Pelz, J., Hayhoe, M. & Loeber, R. The coordination of eye, head, and hand movements in a natural task. *Exp. brain research. Exp. Hirnforschung. Exp. cerebrale* **139**, 266–277 (2001).

50. Lengyel, G., Carlberg, K., Samad, M. & Jonker, T. Predicting visual attention using the hidden structure in eye-gaze dynamics. *CHI EMICS 2021* (2021).

51. Tatler, B. W. & Vincent, B. T. Systematic tendencies in scene viewing. *J. eye movement research* **2** (2008).

52. Einhäuser, W., Spain, M. & Perona, P. Recognition and attention: Relation of eye-position and object recall to bottom-up models of saliency. *J. Vis.* (2008).

53. Subramanian, M., Park, S., Orlov, P., Shafti, A. & Faisal, A. A. Gaze-contingent decoding of human navigation intention on an autonomous wheelchair platform. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, 335–338 (IEEE, 2021).

54. Subramanian, M., Songur, N., Adjei, D., Orlov, P. & Faisal, A. A. A. eye drive: Gaze-based semi-autonomous wheelchair interface. In *IEEE Engineering in Medicine and Biology (EMBC)*, vol. 41, 5967–5970 (IEEE, 2019).

55. Thomik, A. A., Haber, D. & Faisal, A. A. Real-time movement prediction for improved control of neuroprosthetic devices. In *IEEE Neural Engineering (NER)*, vol. 6, 625–628 (IEEE, 2013).

56. Xiloyannis, M., Gavriel, C., Thomik, A. A. & Faisal, A. A. Gaussian process autoregression for simultaneous proportional multi-modal prosthetic control with natural hand kinematics. *IEEE Transactions on Neural Syst. Rehabil. Eng.* **25**, 1785–1801 (2017).

57. Harston, J. A., Auepanwiriyakul, C. & Faisal, A. Prediction of visual attention in embodied real-world tasks. *J. Vis.* **21**, 2741–2741 (2021).

58. Faisal, A. Predicting visual attention of human drivers boosts the training speed and performance of autonomous vehicles. *J. Vis.* **21**, 2819–2819 (2021).

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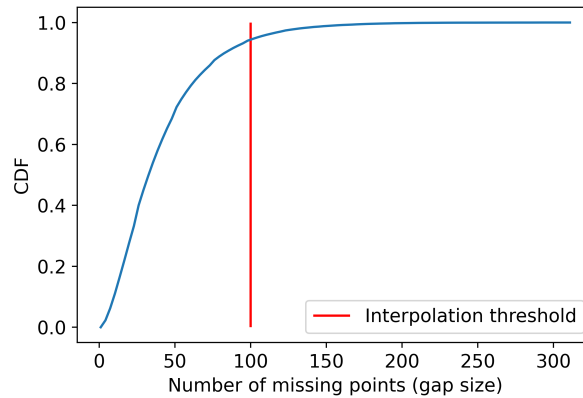
Supplementary Material

Figure S1. Cumulative Density Function (CDF) of the distribution of the length of gaps in the object position time series.
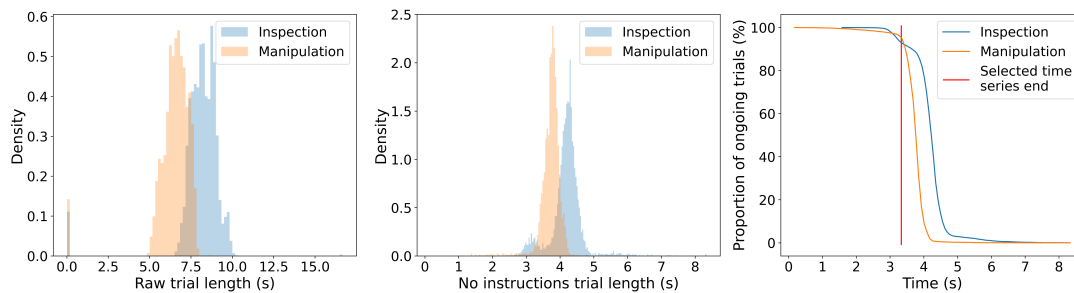

Figure S2. Illustration of the effect of our preprocessing to mitigate the trial length bias. (a) shows the distribution of raw trial lengths (task description + task execution) for both intention classes.. (b) shows the distribution of task execution length for each trial (trial length minus instruction length) for both intention classes.. (c) shows the inverse CDF of the task execution lengths for both intention classes as well as the cutting point we used as end of input sequences.