robotics

MDPI

Article

GazeEMD: Detecting Visual Intention in Gaze-Based Human-Robot Interaction

Lei Shi †, Cosmin Copot * and Steve Vanlanduit

Department of Electromechanics, Faculty of Applied Engineering, Campus Groenenborger, University of Antwerp, 2020 Antwerp, Belgium; lei.shi@uantwerpen.be (L.S.); steve.vanlanduit@uantwerpen.be (S.V.)
* Correspondence: cosmin.copot@uantwerpen.be
† Current address: G.Z. 324, Campus Groenenborger, Groenenborgerlaan 171, 2020 Antwerp, Belgium.

Abstract: In gaze-based Human-Robot Interaction (HRI), it is important to determine human visual intention for interacting with robots. One typical HRI interaction scenario is that a human selects an object by gaze and a robotic manipulator will pick up the object. In this work, we propose an approach, GazeEMD, that can be used to detect whether a human is looking at an object for HRI application. We use Earth Mover's Distance (EMD) to measure the similarity between the hypothetical gazes at objects and the actual gazes. Then, the similarity score is used to determine if the human visual intention is on the object. We compare our approach with a fixation-based method and HitScan with a run length in the scenario of selecting daily objects by gaze. Our experimental results indicate that the GazeEMD approach has higher accuracy and is more robust to noises than the other approaches. Hence, the users can lessen cognitive load by using our approach in the real-world HRI scenario.

Keywords: Human-Robot Interaction; fixation; gaze; EMD

Citation: Shi, L.; Copot, C.; Vanlanduit, S. GazeEMD: Detecting Visual Intention in Gaze-Based Human-Robot Interaction. *Robotics* **2021**, *10*, 68. https://doi.org/10.3390/robotics10020068

Academic Editor: Marco Ceccarelli

Received: 19 January 2021
Accepted: 27 April 2021
Published: 30 April 2021

1. Introduction

Mobile eye-tracking devices, i.e., eye-tracking glasses usually comprise eye camera(s) for detecting pupils and a world camera for capturing the image of the scene. Gaze is calculated from the pupil images and projected to the image of the scene, which can reveal the information of a human being's visual intention. Gazes can be identified as different eye movements. Fixation and saccade are two of the most common types of eye movement event. Fixation can be viewed as gaze being stably kept in a small region and saccade can be viewed as rapid eye movement [1]. They can be computationally identified from eye tracking signals by different approaches, such as Identification by Dispersion Threshold (I-DT) [2], Identification by Velocity Threshold (I-VT) [2], Bayesian-method-based algorithm [3] and machine-learning-based algorithm [4].

In gaze-based HRI, fixation is often used as an indication of the visual intention of a human. In [5], when a fixation was detected, an image patch is cropped around the fixation point and fed to a neural network to detect a drone. In [6], fixation is used to determine if a human is looking at Areas of Interests (AOIs) in a mixed-initiative HRI study. In [7–9], fixations were used for selecting an object to grasp. They were also used in the selection of a grasping plane of an object in [7]. However, there are limitations to using fixation to select an object for further actions. Consider a scenario such as that displayed in Figure 1, where a human is wearing a mobile eye-tracking device and can select detected objects on the table by gaze and then let a robot pick the object up for him or her. The robot will receive the information of the selected object and plan the grasping task; it does not share the human gaze points.

One approach to determine if the human visual intention is focused on an object is to use fixation. When a fixation event is detected, the gaze points in the event are in a small region in the world image. If the fixation center is on the object, the visual intention is

considered to be on the object. Using fixation to select an object requires that the human looks at a very small region of the object and tries to not move the eyes. When observing an object, the human gaze is not fixating on a single point (or a small region). The authors in [10] demonstrated that human gazes, when observing an object, spread over different regions of the object. Figure 2a displays the gazes within a duration spent looking at scissors. The human gazes over time are plotted in yellow dots. This implies that saccade may occur during the process of observation. When using the fixation-based approach, the visual intention may cause information loss in the sense that, when a saccade has happened, the human is still looking at the object.

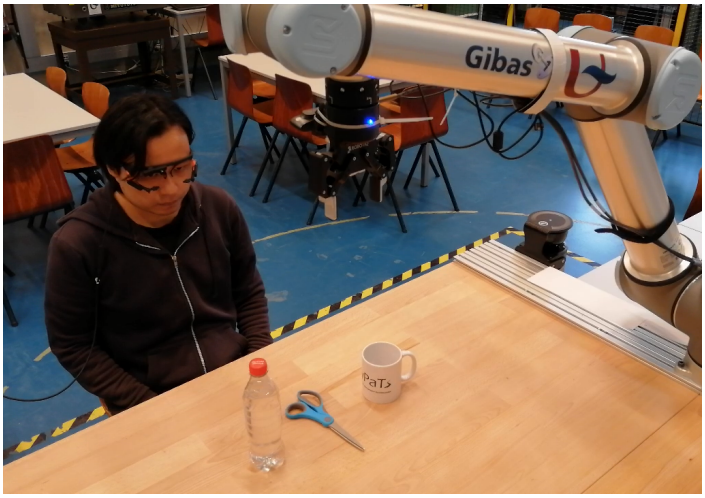
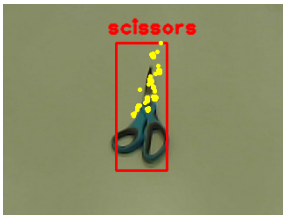
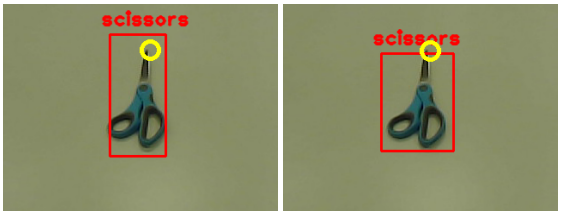


Figure 1. The scenario where a human can select one of the objects on the table using eye-tracking glasses and a robotic manipulator will pick up the selected object.



(**a**) Gaze patterns of observation over time. Yellow dots are the gaze points.



(**b**) Gaze outside of bounding box. Yellow circle is a single gaze point.

Figure 2. (**a**): The gaze patterns over time when a human is looking at the scissors. (**b**): Two consecutive frames with tracked gaze and detected bounding box. The variation in the size of the bounding box comes from the detection algorithm and can cause the gaze to fall outside of the bounding box while the human intention is still on scissors.

Intuitively, considering all gaze points can overcome the problem that the saccadic gazes on the object are lost. HitScan [11,12] uses all gaze points when determining if the visual intention is on the object. However, there exists one more kind of noise which we refer to as gaze drift error. Both fixation based method and HitScan have this problem. On some occasions, the center of a gaze point may fall out of the bounding box while the human is still looking at the object. This kind of noise has various sources. First is the fluctuation of the size of the bounding box caused by the object detection algorithm. Figure 2b gives one example. The figure shows two consecutive frames in one sequence. The gazes in two frames are located at the same position of the object, but the bounding box of the object changes. Second, a poor calibration would also result in this error. Moreover, the head-mounted mobile eye tracking device may accidentally be moved after calibration and the detected gaze will be shifted. Both fixation based method and HitScan based on

checking if the gaze/fixation center is inside the object bounding box. Both will suffer from the gaze drift error. In addition, fixation based method will have information loss due to the saccades inside the bounding box.

When using gaze to select an object to interact with a robot, one issue is that the robot does not know if the human has decided on an object for interaction, even if the human visual intention is on the object. This is the Midas problem [13]. One solution is to use a long dwell time to confirm the selection [14]. Using fixation or HitScan with a long dwell time will also be less efficient due to the saccades and gaze drift error.

We propose GazeEMD, an approach to detect human visual intention which can overcome the limitations mentioned above. We form the question of detecting visual intention from a different perspective than checking if the gaze points are inside a bounding box. We compare the hypothetic gaze distribution over an object and the real gaze distribution to determine the visual intention. For a detected object, we generate sample gaze points within the bounding box. They can be interpreted as a hypothesis of where a human being's visual focus is located. These sample points are formed as the hypothetic gaze distribution. The gaze signals from the mobile eye-tracking device provide information of the actual location a human is looking at. The actual gaze points are formed as the actual gaze distribution. The similarity between hypothetic gaze distribution and actual gaze distribution is calculated by Earth Mover's Distance (EMD) distance. The EMD similarity score is used to determine if the visual intention is on the object. We conduct three experiments and compare GazeEMD with a Fixation-based approach and HitScan. The results show that the proposed method can significantly increase accuracy in predicting human intention with the presence of saccades and gaze shift noise.

To our best knowledge, we are the first to deploy EMD similarity to detect if the visual intention is on an object. Until recently, the fixation-based method and the method checking all gazes are still widely used in HRI applications, although both have the problem of gaze drift error. Little research focusing on solving this problem has been reported. The novel contributions of our work compared to the state-of-the-art are:

1. We proposed GazeEMD, which can overcome the problem of gaze drift error which the current state-of-the-art methods do not solve and capture saccadic eye movements when referring objects to a robot;
2. We show that GazeEMD is more efficient than the fixation-based method and HitScan when confirming selection, using a long dwell time which has not been reported in the literature before. The eye gaze is not required to be held in a small region.

The rest of the paper is organized as follows, In Section 2, we review the related work. In Section 3, we explain our method in detail. We describe the experimental setup and evaluation in Sections 4 and 5. In Section 6, the experimental procedure and evaluation method are presented, and Section 7 is the discussion.

2. Related Work

The I-DT and I-VT [2] are two widely used algorithms to identify fixation events. I-DT detects fixation based on the location of gazes. If the gazes are located within a small region, i.e., the coordinates of gazes are under de ispersion threshold, the gazes are considered as a fixation event. I-VT detects fixation based on the velocities of gazes. Gazes with velocities are under the velocity threshold are considered as a fixation event.

When a fixation event is detected, the fixation center is compared with the bounding box of the object to determine the visual intention. Some works use all gaze points to detect visual intention on objects. In [14,15], the accumulated gazes on objects are used to determine the object at which the user is looking. HitScan [11] detects the visual intention by counting the number of gazes entering and the number exiting a bounding box. If the counts of gazes inside a bounding box is higher than the entering threshold, then an event is started. If the count of gazes consecutively located outside of a bounding box is higher, then the event is closed.

In gaze-based HRI, an important issue in referring to objects is the Midas Touch problem [13]. If the gaze dwell-time fixating on an object is too short, then a selection is activated even if that is not the human intention. To overcome this problem, an additional activation needs to be made to confirm the human intention. The activation can be additional input devices [16,17], hand gestures [18,19] and eye gestures [20,21]. A common solution to overcome the Midas Touch problem is using a long gaze dwell-time [14,22] to distinguish the involuntary fixating gazes, which are rarely higher than 300 ms. Several HRI works have adopted this solution. In [23,24], dwell time is set to 500 ms to activate the selections of AOIs. In [8,9], the 2D gazes obtained from eye-tracking glasses are projected into 3D with the help of an RGB-D camera. The fixation duration is set to two seconds to confirm the selection of an object in [8]. A total of 15 gaze points on the right side of the bounding box are used to determine the selection in [9]. In [21], gaze is used to control a drone. A remote eye-tracker is used to capture the gazes on a screen. Several zones are drawn on the screen with different commands to guide the drone. Dwell time from 300 ms to 500 ms is tested to select a command to control the drone. In [25], a mobile eye-tracker and a manipulator are used to assist surgery in the operating theatre. A user can select an object by looking at the object for four seconds. Extending gaze dwell time to overcome the Midas Touch problem in HRI has proven valid regardless of the type of eye-tracking device and application scenario. However, a long dwell time would increase the user's cognitive load [17]. The users deliberately increase the duration time, which means extra effort is needed to maintain the gaze fixation on the object or AOI. Furthermore, if a user fails to select an object, the long dwell time will make the process less efficient and also increase the user's effort. Such disadvantages are critical to users who need to use the device frequently, such as disabled people using gaze to control wheelchairs.

Dwell time has also been evaluated with other modalities of selection. In [26], dwell time is compared with clicker, on-device input, gesture and speech in VR/AR application. Dwell time is preferred as a hands-free modality. In [17], dwell time obtains a worse performance than the combination of dwell time and single-stroke gaze gesture in wheelchair application. When controlling drone [21], using gaze gestures is more accurate than using dwell time, although it takes more time to issue a selection. Depending on the different dwell times and applications, the results of these works differ. There is no rule of thumb to select the best specific modality; dwell time still has the potential to reduce the user discomfort and increase efficiency, provided that the problems mentioned in Section 1 are overcome.

Our proposed approach uses EMD as the metric to measure the similarity of two distributions. EMD was first introduced into the computer vision field in [27,28]. The EMD distance was also used in image retrieval [27,29]. The information of histograms of images were derived to construct the image signatures $P = \{(p_1, w_{p_1})...(p_m, w_{p_m})\}$ and $Q = \{(q_1, w_{q_1})...(q_n, w_{q_n})\}$ where p_i, w_{p_i}, m and q_j, w_{q_j}, n are the cluster mean, weighting factor and number of clusters of the respective signature. Distance matrix $\mathbf{D}$ is the ground distance between P and Q and flow matrix $\mathbf{F}$ describes the cost of moving "mass" from P to Q. EMD distance is the normalized optimal work for transferring the "mass". In [29], EMD is compared with other metrics, i.e., Histogram Intersection, Histogram Correlation, χ^2 statistics, Bhattacharyya distance and Kullback–Leibler (KL) divergence to measure image dissimilarity in color space. EMD had a better performance than the other metrics. It was also shown that EMD can avoid saturation and maintain good linearity when the mean of target distribution changes linearly.

A similar work [30] also uses EMD distance to compare the gaze scan path collected from the eye tracker and the gaze scan path generated from images. The main differences between their work and ours are: (i) In [30], the authors use a remote eye tracker to record gaze scan paths when the participants are watching images. We use First Person View (FPV) images and gazes record by a head-mounted eye tracking device. (ii) We want to detect if the human intention is on a certain object, and their work focuses on generating the gaze scan path based on the image.

3. Methodology

Our methodology will be applied in the scenario described in Section 1. There are three objects, namely, cup, scissors and bottle, in the scene (Figure 1). All objects are placed on a table. The user can select one of the objects and a robotic manipulator can pick up the desired object. Figure 3 shows the overview of the GazeEMD visual intention detection system. We use a head-mounted eye tracking device that provides the world image I_w and gaze point $g(x, y)$, where x and y are the coordinates in I_w. We first detect all the objects by feeding the world image I_w to the object detector. Then, we generate hypothetic gaze samples on the detected objects and compare them with actual gazes obtained from the head-mounted eye tracking device. Finally, the similarity score between the hypothetic gaze distribution and actual gaze distribution is used to determine if the human visual intention is on an object.

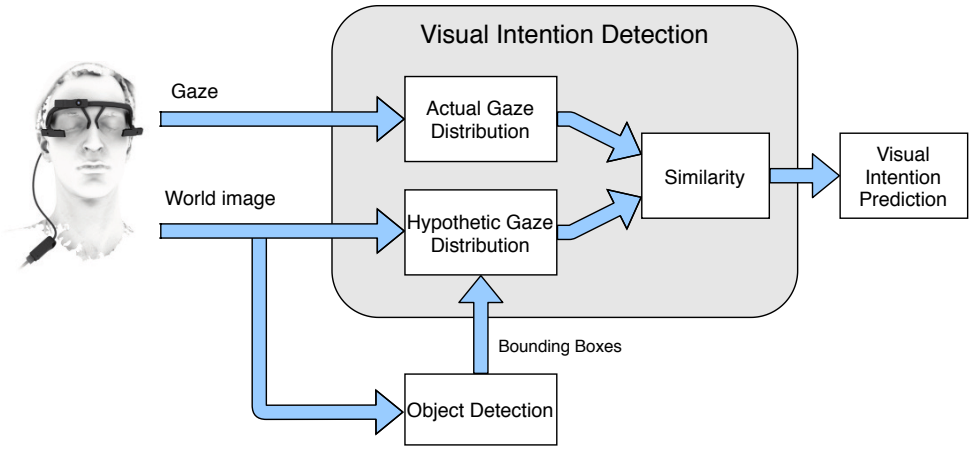


Figure 3. The block diagram of the GazeEMD system.

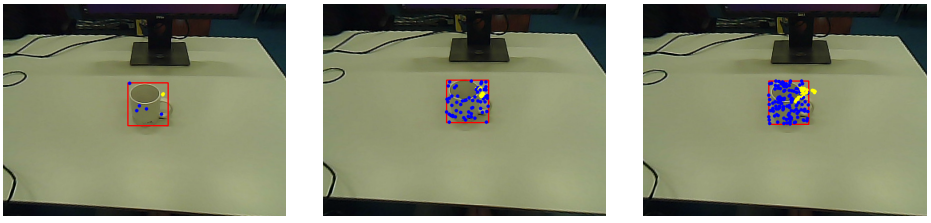
3.1. Object Detection

We use deep-learning-based object detector YOLOv2 [31] to detect the objects in our scene. The network of YOLOv2 is trained on a COCO dataset [32]. YOLOv2 detector $Y = [\mathbf{B}, \mathbf{C}]$ takes world image I_w as the input of network and predicts the bounding boxes $\mathbf{B}$ and class labels $\mathbf{C}$.

3.2. GazeEMD

3.2.1. Hypothetic Gaze Distribution and Actual Gaze Distribution

Figure 4 shows an example of the hypothetic gaze distribution and the actual gaze distribution. The hypothetic gaze instances are plotted in blue and the actual gaze instances are plotted in yellow.



(**a**) Gaze length k = 5. (**b**) Gaze length k = 60. (**c**) Gaze length k = 120.

Figure 4. The hypothetic gaze distributions and the actual gaze distribution. The hypothetic gaze instances are plotted in blue and the actual gaze instances are plotted in yellow.

For each detected object, we crop an image patch I_{obj} from I_w by the size of the object bounding box. From all pixels of I_{obj}, k pixels are sampled, following a Gaussian distribution. The sampled pixels are interpreted as hypothetic gaze points. Next, we

calculate the Euclidean distance between each of the k pixels and the center of the bounding box. This distance distribution is denoted as hypothetic gaze distribution π_s. To form the actual gaze distribution, we collect k gaze points from the eye tracking device and calculate their Euclidean distances to the center of the bounding box. The resulting distance distribution is denoted as actual gaze distribution π_g.

3.2.2. Similarity between Distributions

EMD is used as the measure of the similarity between distributions π_s and π_g. In order to use EMD, the distributions need to be transformed into signatures. We first calculate the geometric distance histograms $H_s = \sum_{i=1}^{m} b_s^i$ for π_s and $H_g = \sum_{j=1}^{n} b_g^j$ for π_g, where m and n are the number of bins. The range of histogram value depends on the size of the world image. If the image size is 640×480, the maximum value will be 800. The signatures $\mathbf{s}_s$ and $\mathbf{s}_g$ are calculated similarly to [29]

$$\mathbf{s}_s = \sum_{i=1}^{m} b_s^i w_s^i, \quad \mathbf{s}_g = \sum_{j=1}^{n} b_g^j w_g^j \tag{1}$$

where b_s and b_g are the bin values from H_s and H_g; the weighting factors w_s and w_g are the middle values of the respective bin intervals [29]. The distance matrix $\mathbf{D}_{sg} = [d_{ij}]$ is the ground distance between π_s^i and π_g^j. The flow matrix $\mathbf{F}_{sg} = [f_{ij}]$ is the cost of moving the "mass" from π_s and π_g. The work function is

$$Work(\mathbf{D}_{sg}, \mathbf{F}_{sg}) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}. \tag{2}$$

The EMD distance is calculated as

$$EMD(\pi_s, \pi_g) = \frac{min(Work(\mathbf{D}_{sg}, \mathbf{F}_{sg}))}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}. \tag{3}$$

Then, the similarity score, i.e., EMD distance, is used as a metric to determine whether the visual intention is on an object, given the detected object bounding boxes $\mathbf{B}$ and a set of consecutive gaze $\mathbf{g}$. The EMD visual intention $V_{EMD}(\mathbf{B}, \mathbf{g})$ is calculated as

$$V_{EMD}(\mathbf{B}, \mathbf{g}) = \begin{cases} C_i & \text{if } \underset{i}{\text{argmin}} \left(EMD(\pi_{s_i}, \pi_g)\right) < T_i \\ 0 & \text{else} \end{cases}. \tag{4}$$

where C_i and T_i is the ith label of object and the threshold for ith object. The EMD visual intention $V_{EMD}(\mathbf{B}, \mathbf{g})$ is the object a human is looking at. $V_{EMD}(\mathbf{B}, \mathbf{g}) = 0$ represents that the human is not looking at any object. The threshold T_i is required for binary classification. We use Receiver Operating Characteristic (ROC) curve to select appropriate threshold values for GazeEMD. The threshold T_i for ith object is calculated by

$$T_i = \underset{j}{\text{argmax}}(TPR_j - FPR_j) \tag{5}$$

where TPR_j and FPR_j are the jth True Positive Rate and the jth False Positive Rate in the ROC curve.

3.3. Fixation-Based Method

The GazeEMD is compared with the fixation-based method and the HitScan. We describe the fixation-based method here, and the HitScan in Section 3.4. The I-DT algorithm detects fixation events and calculates the fixation center $f_p = IDT(t_d)$ for each detected

fixation event. The parameter t_d is the fixation duration. The rest of the gaze instances are all considered as saccadic events. The fixation visual intention V_f is calculated as

$$V_f = \begin{cases} C_i & \forall i \in [0, 1, ...n], \exists f_p \in B_i, \\ 0 & \text{else} \end{cases} \tag{6}$$

where n is the number of detected objects. If the f_p is within the ith bounding box B_i, the intention is assigned to the object.

3.4. HitScan with Run Length Filtering

We implement the HitScan with run length filtering approach proposed in [11]. For a given gaze sequence g and the bounding box of one object B, HitScan H_i checks if a gaze point g_i is located inside a bound box

$$H_i = \begin{cases} 1 & \text{if} \quad g_i \in B, \\ 0 & \text{else} \end{cases} \tag{7}$$

Run length filter defines two constraints, T_1 and T_2. T_1 is the minimal number of consecutive gaze points, which are located inside a bound box. Similarly, T_2 is the minimal number of consecutive gaze points located outside a bound box. Run length filter uses T_1 and T_2 to define if a "look" L is on an object. L is equivalent to our visual intention for single objects and it meets the condition that, for a set of gazes, $\mathbf{g}_L$,

$$\prod_{i=0}^{T_1} H_i = 1, \sum_{j=n-T_2}^{T_2} H_j = 0 \tag{8}$$

where n is the length of $\mathbf{g}_L$. A HitScan event consists of the gaze points in the look L. In the case of multiple objects, the visual intention V_{hr} is calculated by iterating the HitScan and run length filter over all objects.

4. Experiment

We conduct three experiments to evaluate the performance of our proposed algorithm. First, we see the performance on single objects. Second, we evaluate the case with multiple objects and last is free viewing. In all experiments, three daily objects, i.e., bottle, cup and scissors, are used. Although we only test with three objects, GazeEMD can generalize well on different objects, since the hypothetic gaze distribution is generated based on the size of the bounding box. The objects are placed on the table. The participant wears the eye-tracking glasses and sits next to the table and conducts the experiments.

4.1. Experiment Procedure

4.1.1. Single Objects

In this experiment, each participant performs three experiment sessions. In each session, a different object is used. At the beginning of one session, the participants look at the object first and then look away from the object. During the "look away" period, the participants can freely look at any place in the scene except the object.

4.1.2. Multiple Objects

In this experiment, each participant performs one experiment session. In the session, all objects are placed on the table at the same time. Participants look at objects one by one, in order, and repeat the process several times. For instance, a participant looks at the scissors, bottle and cup sequentially, and then looks back at the scissors and performs the same sequence.

4.1.3. Free Viewing

The scene setup of the experiment is the same as Multiple Objects: instead of looking at objects with a sequence, the participants can freely look at anything, anywhere in the scene.

4.2. Data Collection and Annotation

We asked seven people to participate in the experiments. All participants are aged between 20 and 40, and all of them are researchers with backgrounds in engineering. All the people voluntarily participated in the experiments. One of the participants has experience in eye tracking. The rest had no prior experience in eye tracking.

A researcher with eye-tracking knowledge and experience labelled the dataset. The annotator used the world image to label the data. The object-bounding boxes and gaze points are drawn in the world images. All datapoints are annotated sample by sample. For the Single Objects experiment, an algorithm can be viewed as a binary classifier, i.e., whether the visual intention is on an object or not. The annotations are clear since the visual intention on and off the object is distinguishable. In the phase of looking at the object, even if gazes are outside of the bounding box, they can be labeled as "intention on object". Conversely, in the phase of looking away from the object, all data can be labeled as "intention not on object". In the Multiple Objects and Free Viewing experiment, an algorithm acts as a multi-class classifier. If a participant looks at none of the objects, it is treated as a null class. The labeling in the Multiple Objects experiment is also clear, since the sequence of shifting visual intention between objects is distinguishable. During the free viewing period, the annotator subjectively labels the data by experience.

4.3. Implementation

We used Pupil Labs eye-tracking glasses [33] for eye tracking. The frame rate of the world image and the eye-tracking rate are both set to 60 fps. The YOLOv2 object detector is implemented by [34].

For the GazeEMD, we calculated the optimal thresholds for each object by Equation (5). The calculation used the data from the Single Objects experiments. The thresholds were applied in the Multiple Objects and the Free Viewing cases too. The number of bins and the histogram range in GazeEMD is 10 and 715. They were used in all experiments. The I-DT for fixation detection we used is also implemented by [33]. The dispersion value for the I-DT is three degrees. It was used in all experiments. The selection of parameter T_1 and T_2 are described in Section 5.1.

5. Evaluation

For all experiments, we carry out sample-to-sample analysis and event analysis. We compared our algorithm with the fixation-based approach and HitScan With Run Length Filtering proposed in [11].

The Single Objects experiment serves three purposes. First, the optimal thresholds are determined. As described in Section 3.2.2, our algorithm needs a threshold value for the binary classification in GazeEMD. We first evaluated the performances with different thresholds with the data from the Single Objects experiment. For each object in the experiment, we selected one threshold for further evaluation and the threshold value is also used in the Multiple Objects experiment and the Free Viewing experiment. Second, we evaluated the sample-to-sample accuracy with different gaze lengths to show that GazeEMD can deal with the gaze drift error better than fixation and HitScan. Third, the event analysis with a long gaze length is equivalent to using a long dwell time to confirm the object selection. We evaluated the performance on the event level to show that GazeEMD is more efficient in confirming the selection.

The Single Objects experiment setup is constrained. The participants are asked to look at the object and look away. There is only one object in the scene. We conduct the same evaluation to Multiple Objects experiment and the Free Viewing experiment to see whether GazeEMD can overcome gaze drift error when the constraints are fewer.

5.1. Selection of Event Length

All algorithms are evaluated with three different event lengths: 90 ms, 1000 ms and 2000 ms. They correspond to a short, medium and long dwell time. The short dwell time is within the normal fixation duration [35], and medium and long dwell times can be used to distinguish the involuntary fixations. Since the algorithms detect events differently, it is not possible to set the event lengths exactly the same. We set the parameter of each algorithm so that they have approximately the same event length. The event length of our algorithm depends on the size of distribution k. k is set to 5, 60 and 120, respectively. The fixation duration t_d determines the event lengths of the fixation-based approach, which iareset to 90 ms, 1000 ms and 2000 ms. For detection by HitScan with run length filtering, the parameter T_1 decides the event length. T_1 is sample related; thus, it is also set to 5, 60 and 120, respectively. T_2 is set to 17 according to the experiment in [11].

5.2. Metrics

5.2.1. Sample-to-Sample Analysis

The prediction of algorithm contains a set of gaze points. For labeling the ground truth for sample-to-sample analysis, each gaze in the data is assigned a label. To analyze the results sample-wise, the algorithm predictions are compared with the ground truth sample by sample, i.e., the predicted label of each gaze point in one prediction is compared with the ground truth label of each gaze point. We use Cohen's Kappa to evaluate sample-to-sample accuracy instead of other commonly used metrics such as Precision-Recall and F1 score. Cohen's Kappa measures the agreement between two sets of data. The value zero means no agreement and a value of one means perfect agreement. Cohen's Kappa is commonly used to evaluate the agreement between different annotators. However, it can also be used to compare the predictions by algorithms with ground truth [4]. Then, the Kappa score can be interpreted as a measure of accuracy. As pointed out in [4] Cohen's Kappa is a better option than Precision-Recall and F1 score when evaluating imbalanced data. In our experiment, we will compare the results of algorithms with annotated data. When determining visual intention using the fixation-based method, the data is imbalanced due to the high number of fixations. Thus using Cohen's Kappa will give a better understanding of the results.

5.2.2. Event Analysis

We first define the events for the event analysis. An event consists of a set of consecutive gaze points whose label is on the object. For the data in a session, two sets of events—the events in the predictions and the events in the ground truth—are obtained. The calculations of predictions of GazeEMD, fixation-based approach and HitScan events are described in Sections 3.2–3.4. The event analysis metrics are similar to the ones in [36,37]. We derive the segments from the events in the predictions and the events in the ground truth and score the segments for evaluation. Segments are the partitions of a sequence that have one-to-one relations between the events in predictions and ground truth. A segment is either partitioned by the boundary of the events in prediction or the boundary of the events in ground truth. Figure 5 shows how segments are partitioned from the events.

The segments are scored with True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). In the single-object case, the segment scoring is a binary case. A TP segment means that, in this segment, the algorithm has an intention event detection and there is also an intention event in the ground truth. An FP segment has algorithm detection but no detection in ground truth. TN means neither algorithm nor ground truth has a detection in segment, and FN means that the algorithm fails to detect the intention event in ground truth. In the multiple-object case, all scoring is the same as in the single-object case, except for FN. An FN in multiple objects still holds the definition in the single-object case. In addition, if an algorithm detects an intention event but its label is not the same as in the ground truth, the segment is also an FN segment.

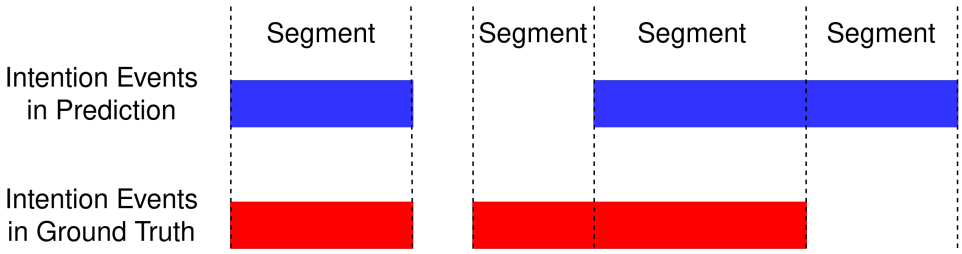


Figure 5. The segments derived from intention events.

Then, the scores of all segments are summed and we calculate the F1 score for the full sequence. We use F1 instead of Cohen's Kappa for the following reason. For long gaze length (2000 ms), the predictions made by Fixation and Hitscan are extremely low. There are cases where it is not possible to calculate the Kappa score.

6. Results

6.1. Single Objects

As indicated in Equations (4) and (5), the GazeEMD acts as a binary classifier. We use Area Under the Curve (AUC) scores under the ROC curve to see the binary classifier performance. Table 1 shows the means and standard deviations of AUC scores of all participants. As shown in the table, our algorithm performs almost perfectly on all objects with different gaze lengths. This implies that there are clear boundaries in the EMD scores.

Table 1. AUC Scores of All Participants.

	Bottle			**Cup**			**Scissors**		
	90 ms	**1000 ms**	**2000 ms**	**90 ms**	**1000 ms**	**2000 ms**	**90 ms**	**1000 ms**	**2000 ms**
mean	0.983	0.979	0.974	0.989	0.989	0.99	0.985	0.982	0.987
std	0.023	0.021	0.028	0.018	0.017	0.018	0.017	0.028	0.016

Figure 6 displays the EMD distances and Euclidean distances of the object bottle from one participant with different gaze lengths. The blue dots are the EMD distance between the gaze distribution and the hypothetic gaze distribution. The Euclidean distance is represented by red dots, showing the geometric distance between actual gaze points and the center of the bounding box. Figure A1a–c shows the distances with all tested objects in Appendix A. The phase of "looking at the object" and "looking away" can be observed from the figures. For instance, in Figure 6a, the participant switches from "looking at the object" to "looking away" at the 1570th sample. The EMD distance values can be easily separated between the two phases. With a longer gaze length, it is easier to distinguish if a participant is looking at an object. Moreover, the EMD distances demonstrate a good correlation with the Euclidean distances, which means that, in addition to its use as similarity score, the EMD value also indicates the information of geometric distance between gaze and bounding box, i.e., a higher EMD value means that the gazes in a gaze distribution are farther from the center of the bounding box.

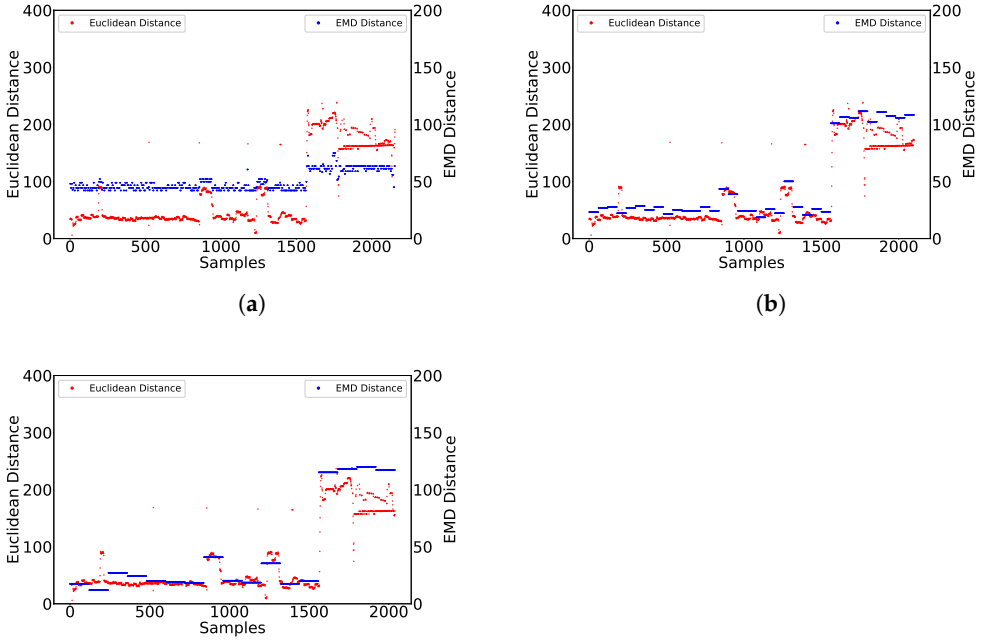


Figure 6. EMD and Euclidean distances of the bottle with different gaze lengths. The red dots represent the Euclidean distances and the blue dots represent the EMD distance. (**a**) EMD distance and Euclidean distance with gaze length of 5. (**b**) EMD distance and Euclidean distance with a gaze length of 60. (**c**) EMD distance and Euclidean distance with a gaze length of 120.

Table 2 shows the means and standard deviations of the Kappa scores in sample-to-sample analysis. The Kappa scores suggest that our algorithm generally performs better than the other two algorithms. For all three objects with all gaze lengths, the GazeEMD has the highest mean Kappa scores, which are all above 0.9 and lowest standard deviations. GazeEMD is less affected by the gaze length. As shown in Table 2, the mean Kappa scores for all objects in all gaze lengths are comparable. The fixation-based method is severely affected by the gaze length. For the object bottle, when the gaze length is increased from 90 ms to 1000 ms and 2000 ms, the mean Kappa score is dropped to 0.398 and 0.143. For the object cup and scissors, it is dropped to 0.389, 0.13 and 0.539, 0.283, respectively. The mean Kappa of HitScan is also decreased to 0.657 and 0.559 for bottle, 0.75 and 0.67 for cup, 0.794 and 0.733 for scissors. For each object, it contains a certain percentage of gaze points which locate outside of the bounding box while the human intention is still on the object, 19.8% of the gazes when looking at the bottle is located outside of the bounding box. For cup and scissors, the percentages are 11.2% and 8%. The analysis in 90 ms shows that GazeEMD can better deal with the gaze drift error. The predictions with 90 ms assign sample labels on the basis of five samples, which is more precise than 1000 ms and 2000 ms. Although the mean Kappa of all algorithms decreases with more gaze drift error, GazeEMD has the highest mean Kappa and lowest standard deviation. This shows that GazeEMD is more accurate with the presence of the gaze drift error. In this experiment, the participants are asked to look at the object and then look away. The majority of gazes either belong to the period of looking at the object or the period of looking away, and the gazes are easily distinguishable. The factor affecting the performance is the gaze drift error. The higher Kappa of GazeEMD indicates that it can better overcome the drift error compared to Fixation and the HitScan.

Table 2. The Mean And Standard Deviation of Cohen's Kappa In Single-Object Experiment.

		90 ms mean (std)	1000 ms mean (std)	2000 ms mean (std)
Bottle	GazeEMD	**0.901 (0.132)**	**0.93 (0.079)**	**0.924 (0.056)**
	Fixation	0.717 (0.257)	0.398 (0.197)	0.143 (0.075)
	HitScan	0.745 (0.272)	0.657 (0.309)	0.559 (0.273)
Cup	GazeEMD	**0.968 (0.043)**	**0.925 (0.072)**	**0.928 (0.037)**
	Fixation	0.809 (0.067)	0.389 (0.105)	0.13 (0.124)
	HitScan	0.828 (0.07)	0.75 (0.113)	0.67 (0.113)
Scissors	GazeEMD	**0.958 (0.041)**	**0.936 (0.045)**	**0.928 (0.054)**
	Fixation	0.833 (0.162)	0.539 (0.165)	0.283 (0.259)
	HitScan	0.874 (0.176)	0.794 (0.219)	0.733 (0.255)

Table 3 shows the means and standard deviations of the event F1 scores. The GazeEMD has the best mean and the best standard deviation for all objects with all gaze lengths, and the effect of a longer gaze length is trivial. Similarly to the sample-to-sample analysis, the mean F1 scores of the fixation and the HitScan decrease with the increase in gaze length.

Table 3. The Mean And Standard Deviation of Event F1 In Single-Object Experiment.

		90 ms mean (std)	1000 ms mean (std)	2000 ms mean (std)
Bottle	GazeEMD	**0.965 (0.054)**	**0.971 (0.022)**	**0.962 (0.026)**
	Fixation	0.831 (0.126)	0.661 (0.131)	0.41 (0.105)
	HitScan	0.872 (0.167)	0.765 (0.226)	0.692 (0.233)
Cup	GazeEMD	**0.989 (0.012)**	**0.97 (0.032)**	**0.952 (0.035)**
	Fixation	0.878 (0.024)	0.664 (0.084)	0.374 (0.173)
	HitScan	0.93 (0.033)	0.843 (0.056)	0.773 (0.071)
Scissors	GazeEMD	**0.984 (0.017)**	**0.975 (0.014)**	**0.965 (0.022)**
	Fixation	0.892 (0.065)	0.75 (0.083)	0.506 (0.272)
	HitScan	0.951 (0.073)	0.876 (0.117)	0.794 (0.176)

Table 4 shows the numbers of detected events by all algorithms. The number of detected events of all participants for each object and each gaze length are summed together. HitScan is not affected by the gaze length. The number of detected events is close in all three gaze lengths. HitScan decides when an event is closed by checking if the gaze remains located outside of the bounding box for T_2 frames. This merges the gazes inside the bounding box into a long event, and thus the total number of events is lower. Furthermore, our algorithm detects more events than the fixation algorithm in all cases (Table 4). The fixation algorithm cannot capture the saccadic gaze movements while the GazeEMD does. The additional events are mostly detected during saccades. In addition, when using a long dwell time to confirm the selection, more events and higher event accuracy means that GazeEMD is more efficient than fixation and HitScan. Fewer events in Fixation means that the remaining gazes are detected as saccade events. When a saccade event occurs inside the object bounding box, the fixation is interrupted and the participant needs to start a new fixation event in order to confirm the selection. For the HitScan, fewer events are caused by the algorithm itself. A HitScan event contains more gazes than the GazeEMD, which means that it will take a longer time to detect the event. Fewer detected events and a lower F1 in Fixation and HitScan means that a participant may attempt to confirm the selection.

Table 4. Number of Detected Events In Single-Object Experiment.

		90 ms	1000 ms	2000 ms
Bottle	GazeEMD	1899	155	74
	Fixation	1259	73	16
	HitScan	25	18	13
Cup	GazeEMD	1882	146	70
	Fixation	1336	72	13
	HitScan	24	15	14
Scissors	GazeEMD	1967	159	74
	Fixation	1477	101	30
	HitScan	19	13	11

6.2. Multiple Objects

In this experiment, we extend the evaluation from Single Objects to Multiple Objects. The means and standard deviations of the sample-to-sample Kappa are shown in Table 5. The GazeEMD has the best mean Kappa scores in all gaze lengths. It also has the lowest standard deviations in gaze lengths of 90 ms and 1000 ms. Although the lowest standard deviation in gaze length 2000 ms is the fixation-based method, the mean Kappa of GazeEMD is still 0.307 higher than the fixation-based method. In the Single Objects experiment, we showed that GazeEMD outperformed fixation and HitScan with the presence of gaze drift error. When there are multiple objects in the scene, the GazeEMD also has the best mean Kappa, showing that it can deal with gaze drift error in a multi-object scene.

Table 5. Means and Standard Deviations of Sample-to-Sample Kappa Scores and Event F1 Scores In Multiple-Object Experiment.

		90 ms mean (std)	1000 ms mean (std)	2000 ms mean (std)
Sample-to-sample (Kappa)	GazeEMD	**0.871 (0.082)**	**0.658 (0.124)**	**0.469** (0.155)
	Fixation	0.693 (0.153)	0.326 (0.174)	0.162 **(0.123)**
	HitScan	0.235 (0.168)	0.185 (0.143)	0.136 (0.138)
Event (F1)	GazeEMD	**0.961 (0.014)**	**0.755 (0.053)**	**0.591 (0.117)**
	Fixation	0.944 (0.024)	0.571 (0.134)	0.327 (0.168)
	HitScan	0.847 (0.06)	0.466 (0.32)	0.365 (0.365)

The means and standard deviations of the event F1 are shown in Table 5. GazeEMD has the best means and standard deviation in all gaze lengths. The numbers of detected events are displayed in Table 6. HitScan can detect considerably more events than in the case of a single object when the gaze length is 90ms. The detected events are 1559 in this experiment. In the Single Objects experiment, the detected events of bottle, cup and scissors are 25, 24 and 19, respectively. The reason for this is that the participants keep switching the visual intention onto different objects in this experiment. The exit parameter T_2 can, therefore close an event accordingly. When the gaze lengths are 1000 ms and 2000 ms, the enter parameter T_1 for entering an event is set to 1000 ms and 2000 ms, which means a participant needs to look at an object for 1000 ms and 2000 ms to start an event. Thus, the number of detected events is 92 and 29, which is significantly lower than the events when the gaze length is 90 ms. Overall, GazeEMD still has the highest number of detected events and event F1, which indicates that the confirmation of object selection is more efficient in the multiple-objects scenario.

Table 6. Number of Detected Events In Multiple-Object And Free-Viewing Experiment.

		90 ms	1000 ms	2000 ms
Multiple Objects	GazeEMD	2821	196	80
	Fixation	2262	96	28
	HitScan	1559	92	29
Free Viewing	GazeEMD	1511	95	35
	Fixation	1151	42	10
	HitScan	1204	74	33

6.3. Free Viewing

Table 7 shows the means and deviations of the Kappa and F1 scores. In the sample-to-sample analysis, GazeEMD has the highest mean Kappa for all gaze lengths. HitScan has the fewest standard deviations on 90 ms and 1000 ms. Fixation has the fewest standard deviations when the gaze length is 2000 ms. The sample-to-sample results show that GazeEMD is more accurate in dealing with gaze drift error when the participants are not instructed. On the event level, GazeEMD has the highest mean F1 on 1000 ms. On 90 ms and 2000 ms, HitScan has the highest mean F1 scores. However, on 2000 ms gaze length, the mean F1 of GazeEMD is 0.334, close to the 0.365 of HitsScan. The event analysis does not represent the scenario of confirming selection by a long dwell time, since the participants are freely looking at anything, without instruction to look at a particular object for a long time.

Table 7. Means and Standard Deviations of Sample-to-Sample Kappa Scores and Event F1 Scores In Free-Viewing Experiment.

		90 ms mean (std)	1000 ms mean (std)	2000 ms mean (std)
Sample-to-sample (Kappa)	GazeEMD	**0.736** (0.216)	**0.521** (0.168)	**0.301** (0.264)
	Fixation	0.727 (0.169)	0.221 (0.161)	0.091 **(0.096)**
	HitScan	0.197 **(0.107)**	0.134 **(0.129)**	0.12 (0.134)
Event (F1)	GazeEMD	0.864 (0.143)	**0.544 (0.198)**	0.334 (0.268)
	Fixation	0.896 (0.07)	0.377 (0.25)	0.193 **(0.171)**
	HitScan	**0.919 (0.06)**	0.466 (0.32)	**0.365** (0.29)

On both sample-to-sample level and event level, the Kappa scores and F1 scores in Free Viewing (Table 7) are lower than the ones in Multiple Objects (Table 5). One of the causes of this is confusion in annotation. The annotations in Single Objects and Multiple Objects are clear, since whether the gaze is on an object is distinguishable. However, in Free Viewing, the gaze intention is not as clear as in the other two experiments. Especially when the gazes are close to the bounding boxes, whether a participant is looking at the edge of an object or deliberately looking at the area around the object cannot be determined. Although the annotations of the gazes contain uncertainties, they create a scenario with noisier data. All algorithms will have the same uncertainties and we can see how they perform on the noisier data. The higher Kappa of GazeEMD on sample level shows it has better performance not only in the constrained experiments (clean data), but also in the experiment without constraints (noisier data). This demonstrates that GazeEMD can generalize well.

The numbers of detected events are displayed in Table 6. Similar to the Multiple Objects, the number of events detected by the HitScan when the gaze length is 90 ms is significantly higher than 1000 ms and 2000 ms. The reason for this is described in Section 6.2.

7. Discussion and Conclusions

In this work, we propose a new approach to determine visual intention for gaze-based HRI applications. More specifically, our algorithm GazeEMD determines which object a human is looking at by calculating the similarity between the hypothetical gaze points on the object and the actual gaze points acquired by mobile eye tracking glasses. We evaluate our algorithm in different scenarios by conducting three experiments: Single Objects, Multiple Objects and Free Viewing. There are two constraints in the Single Objects. The scene is rather clean—only one object exists at a time—and the participants are asked to look at the object first and then look away. We use this constrained setting for three reasons. First, it is easier to evaluate the performance of GazeEMD as a binary classifier and select appropriate threshold values for different gaze lengths. Second, we can remove the noise from annotation to better evaluate the noises caused by gaze drift error and variations in bounding boxes. Finally, we can create sequences with medium and long gaze lengths (1000 ms and 2000 ms), which are essential for confirming the selection of an object by long dwell time. Evaluating the long gaze lengths is equivalent to solving the Midas problem with a long dwell time in HRI applications.

The results demonstrate that the GazeEMD has excellent performance, as well as the ability to reject the gaze drift error. Tables 2 and 3 show that GazeEMD has the highest mean Kappa and F1 scores on both the sample-to-sample level and event level. When the gaze length is 1000 ms and 2000 ms, the mean Kappa and F1 scores are significantly higher than the fixation-based method and HitScan. For the bottle, when the participants look at the object, 19.8% of the gazes are outside of the bounding box. This indicates that GazeEMD has a better performance than the fixation-based method and HitScan, when gaze drift errors occur. The same conclusion can be drawn for the cup and scissors. We extend the evaluation from the single-object case to the scene with multiple objects and free viewing. GazeEMD still has higher Kappa and F1 scores (Tables 5 and 7) than Fixation and HitScan, except the cases of 90 ms and 2000 ms in the Multiple Objects experiment, where the F1 scores are 0.055 and 0.031 lower than the HitScan. Nevertheless, the results are still comparable in these two cases.

In a lot of gaze-based HRI applications, a human needs to interact with objects. A common case is the selection of an object to be picked up by a robotic manipulator. One key issue in this kind of interaction is confirming the selection of an object. The robot knows which object a human is looking at, but does not know that he or she has confirmed the selection of a certain object without additional information, i.e., the Midas problem. One approach to this is looking at the intended object for a longer time, i.e., a long dwell time. This scenario is equivalent to the 2000 ms in the Single Objects experiment. If the gaze dwells on the object for 2000 ms, the object is considered to be selected and confirmed. The robot can perform further steps. For the fixation-based approach, voluntarily increasing the fixation duration will be helpful to increase the number of successful confirmations, but it will also increase the cognitive load [38]. Another downside of fixation is that it cannot capture the saccadic gazes located within the bounding box. This means that using the fixation-based approach will miss the information from gained from the human gaze moving to different parts of the object. This will interrupt a long fixation; hence, the human needs to try harder to select an object for interaction, which will increase the cognitive load. By using an algorithm such as HitScan, which considers all gazes within the bounding box, this problem could be eliminated. However, both the fixation-based approach and HitScan still cannot deal with the gaze drift problem. The trials would potentially be increased to confirm the selection. However, GazeEMD can overcome these problems. GazeEMD detects more events and has higher accuracy than fixation and HitScan (Tables 3 and 4). This indicates that GazeEMD has more successful confirmations of the selection than the other two methods. This is important in the real application, where the users constantly use gaze for interaction, such as disabled people who will need wheelchairs and manipulators to help with daily life. The GazeEMD also has an excellent performance with a short dwell time (90 ms). It can be applied to the cases in which the gaze and object

need to be evaluated but interaction with the object is not required, such as analyzing gaze behavior during assembly tasks [39], or during the time in which an object is handed between humans or human and robot [40].

We propose using GazeEMD to detect whether the human intention is on an object or not. We compared GazeEMD with the fixation-based method and HitScan in three experiments. The results show that GazeEMD has a higher sample-to-sample accuracy. Since the experimental data contain gaze drift error, i.e., the intention is on the object while the gaze points are outside of the object bounding box, the higher accuracy of GazeEMD indicates that it can overcome the gaze drift error. In HRI applications, a human often needs to confirm the selection of an object so that the robot can perform further actions. The event analysis with long gaze lengths in Single Objects experiments shows the effect of using a long dwell time for confirmation. The results show that GazeEMD has higher accuracy on the event level and more detected events, which indicates that GazeEMD is more efficient than the fixation. The proposed method now can detect the human intention in the scenario where the detected bounding boxes of the objects are not overlapped. One future research direction could be further developing the algorithm so that the gaze intention can be detected correctly when two bounding boxes are overlapped.

Author Contributions: Conceptualization, L.S. and C.C.; methodology, L.S.; software, L.S.; resources, L.S., C.C. and S.V.; writing—original draft preparation, L.S.; writing—review and editing, C.C. and S.V.; supervision, C.C. and S.V. All authors have read and agreed to the published version of the manuscript.

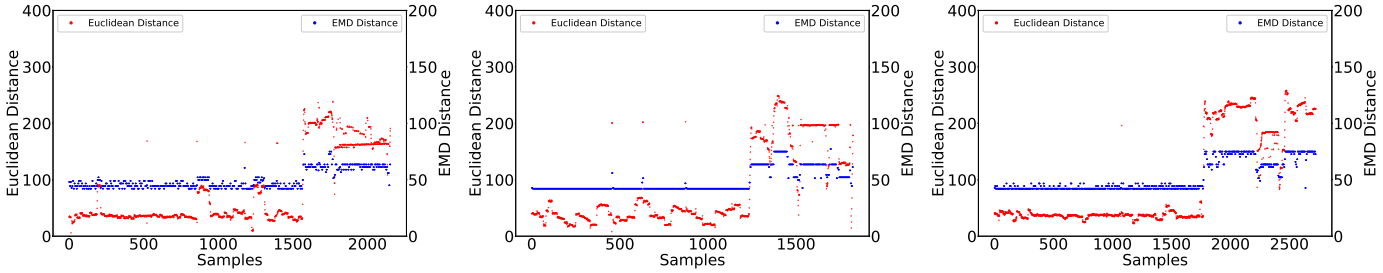
Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

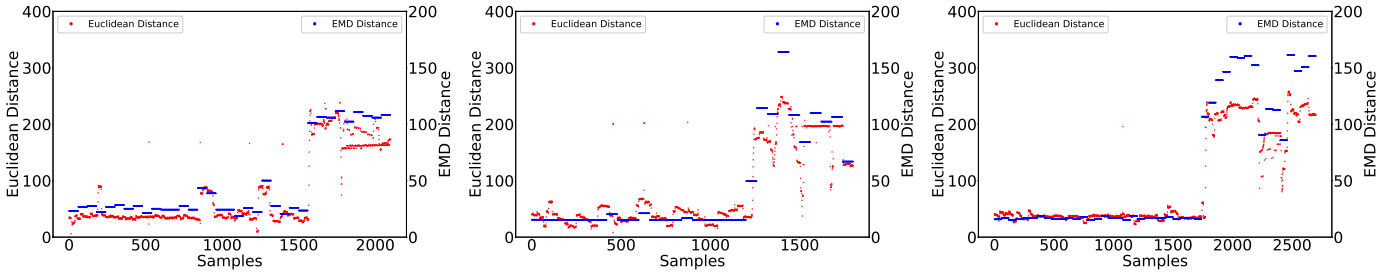
Data Availability Statement: Data sharing not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

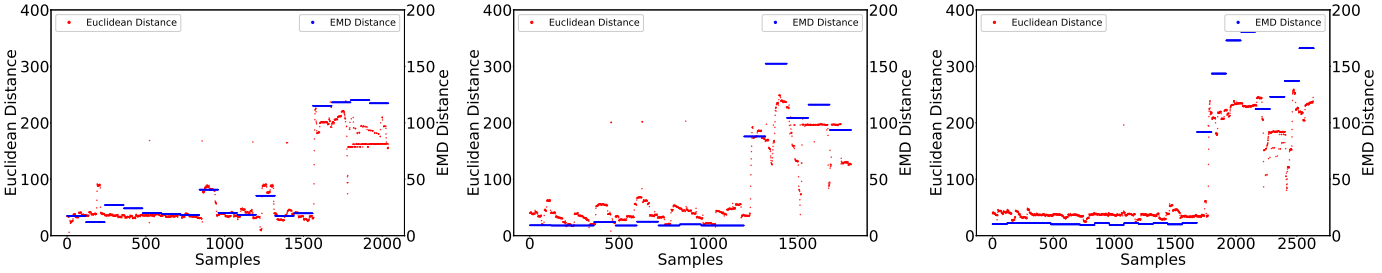


(**a**) EMD distance and Euclidean distance with gaze length of 5. Left is bottle, middle is cup and right is scissors.



(**b**) EMD distance and Euclidean distance with gaze length of 60. Left is bottle, middle is cup and right is scissors.

Figure A1. *Cont.*



(**c**) EMD distance and Euclidean distance with gaze length of 120. Left is bottle, middle is cup and right is scissors.

Figure A1. EMD and Euclidean distances of all objects with different gaze length. The red dots represnet the Euclidean distances and the blue dots represent the EMD distance.

References

1. Holmqvist, K.; Nyström, M.; Andersson, R.; Dewhurst, R.; Jarodzka, H.; Van de Weijer, J. *Eye Tracking: A Comprehensive Guide to Methods and Measures*; OUP: Oxford, UK, 2011.
2. 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, Palm Beach Gardens, FL, USA, 6–8 November 2000; pp. 71–78.
3. Santini, T.; Fuhl, W.; Kübler, T.; Kasneci, E. Bayesian identification of fixations, saccades, and smooth pursuits. In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, Charleston, SC, USA, 14–17 March 2016; pp. 163–170.
4. Zemblys, R.; Niehorster, D.C.; Komogortsev, O.; Holmqvist, K. Using machine learning to detect events in eye-tracking data. *Behav. Res. Methods* **2018**, *50*, 160–181. [CrossRef] [PubMed]
5. Yuan, L.; Reardon, C.; Warnell, G.; Loianno, G. Human gaze-driven spatial tasking of an autonomous MAV. *IEEE Robot. Autom. Lett.* **2019**, *4*, 1343–1350. [CrossRef]
6. Chanel, C.P.; Roy, R.N.; Dehais, F.; Drougard, N. Towards Mixed-Initiative Human-Robot Interaction: Assessment of Discriminative Physiological and Behavioral Features for Performance Prediction. *Sensors* **2020**, *20*, 296. [CrossRef] [PubMed]
7. Li, S.; Zhang, X.; Webb, J.D. 3-D-gaze-based robotic grasping through mimicking human visuomotor function for people with motion impairments. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 2824–2835. [CrossRef] [PubMed]
8. Wang, M.Y.; Kogkas, A.A.; Darzi, A.; Mylonas, G.P. Free-View, 3D Gaze-Guided, Assistive Robotic System for Activities of Daily Living. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 2355–2361.
9. Shafti, A.; Orlov, P.; Faisal, A.A. Gaze-based, context-aware robotic system for assisted reaching and grasping. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 863–869.
10. Takahashi, R.; Suzuki, H.; Chew, J.Y.; Ohtake, Y.; Nagai, Y.; Ohtomi, K. A system for three-dimensional gaze fixation analysis using eye tracking glasses. *J. Comput. Des. Eng.* **2018**, *5*, 449–457. [CrossRef]
11. Chukoskie, L.; Guo, S.; Ho, E.; Zheng, Y.; Chen, Q.; Meng, V.; Cao, J.; Devgan, N.; Wu, S.; Cosman, P.C. Quantifying gaze behavior during real-world interactions using automated object, face, and fixation detection. *IEEE Trans. Cogn. Dev. Syst.* **2018**, *10*, 1143–1152. [CrossRef]
12. Venuprasad, P.; Dobhal, T.; Paul, A.; Nguyen, T.N.; Gilman, A.; Cosman, P.; Chukoskie, L. Characterizing joint attention behavior during real world interactions using automated object and gaze detection. In Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, Denver, CO, USA, 25–28 June 2019; pp. 1–8.
13. 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, Seattle, WA, USA, April 1990; pp. 11–18.
14. Blattgerste, J.; Renner, P.; Pfeiffer, T. Advantages of eye-gaze over head-gaze-based selection in virtual and augmented reality under varying field of views. In Proceedings of the Workshop on Communication by Gaze Interaction, Warsaw, Poland, 14–17 June 2018; pp. 1–9.
15. Tanriverdi, V.; Jacob, R.J. Interacting with eye movements in virtual environments. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, The Hague, The Netherlands, 1–6 April 2000; pp. 265–272.
16. Stellmach, S.; Dachselt, R. Still looking: Investigating seamless gaze-supported selection, positioning, and manipulation of distant targets. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 285–294.
17. Meena, Y.K.; Cecotti, H.; Wong-Lin, K.; Prasad, G. A multimodal interface to resolve the Midas-Touch problem in gaze controlled wheelchair. In Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea, 11–15 July 2017; pp. 905–908.
18. Chatterjee, I.; Xiao, R.; Harrison, C. Gaze+ gesture: Expressive, precise and targeted free-space interactions. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 131–138.

19. Pfeuffer, K.; Mayer, B.; Mardanbegi, D.; Gellersen, H. Gaze+ pinch interaction in virtual reality. In Proceedings of the 5th Symposium on Spatial User Interaction, Brighton, UK, 16–17 October 2017; pp. 99–108.
20. Istance, H.; Bates, R.; Hyrskykari, A.; Vickers, S. Snap clutch, a moded approach to solving the Midas touch problem. In Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, Savannah, GA, USA, 26–28 March 2008; pp. 221–228.
21. Yu, M.; Lin, Y.; Schmidt, D.; Wang, X.; Wang, Y. Human-robot interaction based on gaze gestures for the drone teleoperation. *J. Eye Mov. Res.* **2014**, *7*, 1–14.
22. 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 Comput. Sci.* **2014**, *39*, 75–82. [CrossRef]
23. Krishna Sharma, V.; Saluja, K.; Mollyn, V.; Biswas, P. Eye gaze controlled robotic arm for persons with severe speech and motor impairment. In Proceedings of the ACM Symposium on Eye Tracking Research and Applications, Stuttgart, Germany, 2–5 June 2020; pp. 1–9.
24. Araujo, J.M.; Zhang, G.; Hansen, J.P.P.; Puthusserypady, S. Exploring Eye-Gaze Wheelchair Control. In Proceedings of the ACM Symposium on Eye Tracking Research and Applications, Stuttgart, Germany, 2–5 June 2020; pp 1–8.
25. Kogkas, A.A.; Darzi, A.; Mylonas, G.P. Gaze-contingent perceptually enabled interactions in the operating theatre. *Int. J. Comput. Assist. Radiol. Surg.* **2017**, *12*, 1131–1140. [CrossRef] [PubMed]
26. Esteves, A.; Shin, Y.; Oakley, I. Comparing selection mechanisms for gaze input techniques in head-mounted displays. *Int. J. Hum. Comput. Stud.* **2020**, *139*, 102414. [CrossRef]
27. Rubner, Y.; Tomasi, C.; Guibas, L.J. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121. [CrossRef]
28. Peleg, S.; Werman, M.; Rom, H. A unified approach to the change of resolution: Space and gray-level. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 739–742. [CrossRef]
29. Bazan, E.; Dokládal, P.; Dokladalova, E. Quantitative Analysis of Similarity Measures of Distributions In Proceedings of the British Machine Vision Conferences, Cardiff, UK, 9–12 September 2019; pp 187.
30. Yoo, B.S.; Kim, J.H. Evolutionary fuzzy integral-based gaze control with preference of human gaze. *IEEE Trans. Cogn. Dev. Syst.* **2016**, *8*, 186–200. [CrossRef]
31. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. *arXiv* **2016**, arXiv:1612.08242.
32. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 740–755.
33. Kassner, M.; Patera, W.; Bulling, A. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Seattle, WA, USA, 13–17 September 2014; ACM: New York, NY, USA, 2014; pp. 1151–1160. [CrossRef]
34. Bjelonic, M. YOLO ROS: Real-Time Object Detection for ROS. Available online: https://github.com/leggedrobotics/darknet_ros (accessed on 6 July 2019).
35. Rayner, K. The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Q. J. Exp. Psychol.* **2009**, *62*, 1457–1506. [CrossRef]
36. Ward, J.A.; Lukowicz, P.; Tröster, G. Evaluating performance in continuous context recognition using event-driven error characterisation. In Proceedings of the International Symposium on Location-and Context-Awareness, Dublin, Ireland, 10–11 May 2006; pp. 239–255.
37. Ward, J.A.; Lukowicz, P.; Gellersen, H.W. Performance Metrics for Activity Recognition. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*. [CrossRef]
38. Shojaeizadeh, M.; Djamasbi, S.; Trapp, A.C. Density of gaze points within a fixation and information processing behavior. In Proceedings of the International Conference on Universal Access in Human-Computer Interaction, Toronto, ON, Canada, 17–22 July 2016; pp. 465–471.
39. Wang, H.; Shi, B.E. Gaze awareness improves collaboration efficiency in a collaborative assembly task. In Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, Denver, CO, USA, 25–28 June 2019; pp. 1–5.
40. Moon, A.; Troniak, D.M.; Gleeson, B.; Pan, M.K.; Zheng, M.; Blumer, B.A.; MacLean, K.; Croft, E.A. Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, Bielefeld, Germany, 3–6 March 2014; pp. 334–341.