3D gaze cursor: continuous calibration and end-point grasp control of robotic actuators

Pablo M. Tostado1, William W. Abbott1 and A. Aldo Faisal1,2,3, Member IEEE

Abstract—Eye movements are closely related to motor actions, and hence can be used to infer motor intentions. Additionally, eye movements are in some cases the only means of communication and interaction with the environment for paralysed and impaired patients with severe motor deficiencies. Despite this, eye-tracking technology still has a very limited use as a human-robot control interface and its applicability is highly restricted to 2D simple tasks that operate on screen based interfaces and do not suffice for natural physical interaction with the environment. We propose that decoding the gaze position in 3D space rather than in 2D results into a much richer "spatial cursor" signal that allows users to perform everyday tasks such as grasping and moving objects via gaze-based robotic teleoperation. Eye tracking in 3D calibration is usually slow – we demonstrate here that by using a full 3D trajectory for system calibration generated by a robotic arm rather than a simple grid of discrete points, gaze calibration in the 3 dimensions can be successfully achieved in short time and with high accuracy. We perform the non-linear regression from eye-image to 3D-end point using Gaussian Process regressors, which allows us to handle uncertainty in end-point estimates gracefully. Our telerobotic system uses a multi-joint robot arm with a gripper and is integrated with our in-house "GT3D" binocular eye tracker. This prototype system has been evaluated and assessed in a test environment with 7 users, yielding gaze-estimation errors of less than 1cm in the horizontal, vertical and depth dimensions, and less than 2cm in the overall 3D Euclidean space. Users reported intuitive, low-cognitive load, control of the system right from their first trial and were straightaway able to simply look at an object and command through a wink to "grasp this" object with the robot gripper.

I. INTRODUCTION

The needs for assistive robotics and robotic rehabilitation has been driven by an ageing population with increasingly higher number of individuals that depend on others to perform daily-life activities due to movement disabilities. Moreover, paralysed patients and amputees have driven the development of brain-machine interfaces and neuromotor prosthetics. Accurate control of movements is a state-of-the-art problem of distinguished importance in the field of bioengineering. For years, Brain Machine Interfaces (BMIs) and electromyographic (EMG) signals have been the most commonly used approaches as interactive pathways from the user (to an external device such as a prosthesis [1], [2]. However, the implementation of EMG-technologies can be unfeasible in those cases in which the patient suffers from severe multiple disabilities, and brain computer interfaces are generally extremely invasive when fast, accurate control is sought and require considerable setup and training time – even for EEG base systems. In addition, the clinical and operational costs of such currently-emerging technologies often go beyond the means of potential users. Several different approaches have been attempted to circumvent these problems. In the recent years, eye-tracking systems have been suggested as a reliable non-invasive alternative to BMI and have been extensively developed as a cheaper and more compact interface between humans and robots[3]. The idea behind the use of eye movements as a control signal relies on the fact that our gaze is proactive and directly correlated to action intentions and cognitive control [4], [5]. Furthermore, function the ocular-motor system control is usually retained even in the most severe cases of paralysis related to muscular dystrophies or brain stroke, and movements of the eyes are preserved much longer then skeletal movement in neurodegenerative diseases affecting the motor system, such as Multiple Sclerosis, Parkinson’s, or ALS [6].

2D Eye-tracking is a long established technology that has been extensively used as a control interface in academic and industrial human-computer interface research. However, the use of 3D eye movements for control of gaze-based robotic interaction systems is not a straightforward process. One of the main issues limiting the direct implementation of eye-tracking systems is the difficulty to distinguish between intentional visual fixations and unintentional environment scrutiny, known as the Midas touch problem. Current approaches to this problem usually rely on differences in gaze-dwell times between natural fixations and saccades [7]. Yet, the use of this technology is significantly limited by the expensive cost of high performance commercial eye-tracking systems. Efforts have been made focusing on the enrichment of the signal interpretation and cheapening of eye-tracking devices. We have utilised an ultra low-cost eye-tracker "GT3D" custom built by the integration of two commercial Playstation3 cameras (10USD each) and two Infrared (IR) LEDs that yield pupil detection at high speed frame rates (up to 120 Hz) developed by us [3]. We have resolved the Midas touch problem here by using "Winks" for opening/closing of the gripper and "long fixations" to trigger movement to the gaze-end point – effectively proactively moving the arm to objects we examine for a long time.

In cases were neurodegenerative diseases affect the capacity of communication, eye-tracking technology is suitable as communication tool via computers. In some cases, detection of fixations or blink patterns allow the user to operate interactive games or spelling programs [8]. Eye movement sequences including winks and blinks can be composed into...
“gaze gestures” are translated into commands. 2D eye tracking has therefore been used in human-computer interactive systems focused on the control of machines such as powered wheelchairs substituting current control mechanisms, for instance the joystick. Commands are displayed on a screen, and the direction of the eye gaze is decoded in real time in order to implement the desired instruction [9], [10]. The potential of a computer decoding natural eye movements (the same that occur during mouse-based computer use) to operate arcade video games has been shown in field trials with hundreds of users being able to play (and score) in under 15 seconds from first use of the system [11]. While able-bodied users can seamlessly interact with the physical world without thinking about it, movement impaired users have to learn to operate a gaze-based user interface to control a device – which is unnatural in their interaction, causes additional cognitive load, and divert (literally) their visual attention from the task at hand. The challenge lies in the user interface, when we manipulate with the physical world our gaze follows naturally our movement intentions and even our motor imagery (thinking about grasping a glass can trigger the same eye movements as actually grasping it). Thus, ideally we want to remove screen-based user interfaces and enable users to interact more naturally by directly looking at the objects they wish to manipulate with a robot. We previously demonstrated a gaze decoding solution deployed on an electrical wheelchair to freely drive the wheelchair (without having to look at a computer display or other user interface) by correlating naturally occurring eye movements during joystick-based driving into decoders for driving by eye-tracking [12]. Reaching, grasping and object manipulation are however more complex than driving, as they require 3D interactions. Thus, the main restriction that prevents binocular eye-tracking from being more widely used in fields such as teleoperation and prosthetics has been the difficulty to implement accurate 3D gaze estimation.

One of the fundamental challenges preventing this expected upgrading of eye-tracking systems is related to the problematic calibration of the system for individual users in the depth dimension. Some algorithms have approached the inference of 3D gaze position in virtual environments. Most of the studies aiming to acquire a depth directional gaze position examine the relationship between the inter-distance of the centre of both pupils and the depth of the gaze fixation [13], [14] and fit a mechanical model of the eye to the data. Nonetheless, the obtained accuracy in the depth direction is comparatively low. One of the only 3D gaze estimation systems implemented in a real-world environment uses the combined geometry of the eyeball and the eye-tracking cameras setup to estimate the gaze vectors that link the fovea and the center of the tracked pupil of both eyes. The 3D gaze fixating point is inferred from the calculation of the closest intersection point of the aforementioned vectors in the 3D space. The overall determined accuracy of the system was 3.93 cm in Euclidean space [15]. Alternative solutions to solve the insufficient accuracy in the depth dimension have attempted to use 2D gaze control while performing plane selection independently. A recent approach combined 2D eye-tracking control with an imaginary-based BMIs for depth plane selection, yielding a pseudo-3D interface. Assessment of the interface for 3D interaction was performed by combining the eye-tracker system with an arm reaching mode as two independent navigation methods, which were independently selected by imaginary-based BMI electroencephalography [16].

Here, we present a novel calibration approach that aims to exploit the huge potential of 3D gaze estimation in the field of robotic teleoperation and prosthetics control. A continuous trajectory in 3D space generated by the robot rather than a discrete set of points in a given 2D plane is used for calibration. This richer and continuous data set can be actually obtained faster and more naturally than screen-based interactions, by simply asking the user to track the robot end-point marker. Our integrated system can reliably estimate gaze fixations with high and near equal accuracy in all three spatial coordinates, and can be used to operate our robot arm including grasping of small objects with the gripper.

Fig. 1. High-resolution eye tracking enables tracking of a user 3D gaze target location (red arrow). Two type of gaze-gesture commands (“wink” and “fixations”) enable intuitive and seamless operation of a 5DOF robotic gripper arm. Inset: IR camera images of user’s left and right eye cameras and tracking ellipse of the pupils (blue circles).
II. METHODS

Our integrated system (see Fig. 1) consists of two eye-tracking cameras, an OptiTrack motion tracking system designed and a WidowX robot arm. The streaming of data from and to each individual piece of software is synchronized by an integrated C++ interface.

a) WidowX robotic arm: This low-end robotic arm operated in this system is the WidowX Reactor Robot Arm (TrossenRobotics, Downers Grove IL). The robot arm possesses 5 DOF granted by 5 different servos, including the ability to open and close a parallel gripper with 500g holding strength. Communication to and from the servos of the WidowX is enabled by an ArbotiX micro-controller that also performs the inverse kinematic calculations for a given target position. This communication is integrated in our C++ interface, from which the endpoint position is computed. The current specifications of this robot arm describe a maximum radius reach of 40 cm.

b) GT3D binocular eye-tracker: The GT3D binocular eye-tracker utilised in our system was developed by Abbott & Faisal, [3]. The eye-tracker is built from two commercial PlayStation 3 cameras yielding a resolution of 320 x 240 pixels at a maximum frame-rate of 120Hz. Pupil recognition is achieved by the inclusion of two infrared (IR) light LED sources and modifying the cameras’ filter to allow IR light flow. This adjustment optimises the imaging conditions easing pupil detection by increasing the contrast between the pupil and the iris. Common image-processing methods (OpenCV) are implemented to generate a binary image of the eye from the greyscale recorded image by threshold image segmentation in each video frame. Subsequently, erosion and dilation morphological operations are performed over the binary image preceding a connecting component labelling and a shape-based filter leading to pupil classification. An ellipse is fitted to the classified pupil using a least-squares regression algorithm that allows extraction of the center-of-pupil coordinates. Further zoom and threshold settings are accessible during the eye-tracker setup. The hardware costs for the system are at 20 USD.

c) Motion tracking and synchronisation: The novelty of our system relies on the form of calibration of the eye-tracker. This process is performed using the robot arm itself to define a continuous set of calibration points in 3D space given by a pre-defined trajectory. During robot movement across this trajectory, the user must look at the endpoint position of the gripper in a non-stopping fashion. Normally the inverse kinematics of a robot arms provides a high resolution end-point position. These 3D positions need to be synchronised with the eye movement information from the eye tracker. This proved challenging with the WidowX arm as feedback signals were transmitted a serial-USB communication back to our computer – which caused unreliable time synchronisation (>300 ms variability). Therefore, we monitored during calibration the endpoint position of the gripper by using an OptiTrak optical motion capture system (Natural Point Inc., Corvallis, OR) and recorded continuously the position of the system. A synchronisation algorithm was run to match the simultaneous acquisition of data from both systems at every time step. Both the center-of-pupils data and the continuous spatial position of the robot are recorded and saved respectively for posterior calibration. The motion tracking system requires a single-time self-calibration after being setup. OptiTrack technology can resolve the location of markers with millimetre accuracy. A restricted working volume of 40cm x 30cm x 20cm (width x height x depth) to operate the robot arm was defined. Fig.2 shows a hypothetical trajectory followed by the robot for 3D user calibration.

A. Gaussian Processes Regression for calibration

The non-linear regression algorithm used to infer the correlation between the pupil-centers and the gaze point of the user is the core part of our continuous eye-tracker calibration. We used Gaussian Processes (GP) Regression, which is a non-parametric Bayesian supervised learning algorithm [17]. GPs are a family of statistical distributions characterised by the association of every data point in the input space with a normally distributed random variable. Moreover, any collection of \( n \) input random variables is treated as a \( n \)-dimensional normal distribution, and the resulting distribution of a GP is the joint distribution of all input random variables. Henceforth, GP may be understood as an infinite-dimensional generalization of multivariate normal distributions.

The advantage of this approach is that we do not make assumptions about the eye geometry. Thus, we can perform model-free gaze end-point calibration, making the system automatically also suitable for users with strabism or other eye alignment defects – as long as the eyes show repeatability in how they are looking at a target. This is possible because we collect more data from the user (continuous trajectories instead of discrete points) that are more accurate (using smooth pursuit eye movements tracking an object instead of discrete points) that are more accurate (using smooth pursuit eye movements tracking an object instead of discrete points) that are more accurate.
of saccades between jumping calibration targets), but also because GP based approaches allows us to perform non-linear regression in a sdata efficient regime as they operate with outliers and can track associated uncertainty when reconstructing gaze, which becomes a crucial aspect in the safety aspects of robotic operation The key distinction between this regression estimator and other non-linear function approximators is that the resulting output from given set of arguments is considered to be a stochastic process. This means that the output is randomly drawn from a joint normal distribution of potential calibration functions consistent with the data. Additionally, it is important to realise that GP do not explicitly estimate functions, but rather evaluates correlations between locations where data was collected.

We have a stochastic process \( t : \mathbb{R} \rightarrow \mathbb{R} \) where all \((t(x_1), \ldots, t(x_n))\) are normally distributed. Like any normal distribution a GP has a mean and a covariance defined for its potentially infinite domain. The correlation between the outputs of a GP is given by the corresponding covariance function, which is at the same time determined by a certain covariance kernel. For a certain unknown model, a set of output samples obtained from given chosen inputs is assumed to be the result of the normal distribution of the expected mean and covariance:

\[
\begin{bmatrix}
  f_1 \\
  \vdots \\
  f_n
\end{bmatrix} \sim N \left( \begin{bmatrix} 
  \mu_1 \\
  \vdots \\
  \mu_n
\end{bmatrix}, \begin{bmatrix} 
  K_{11} & \cdots & K_{1n} \\
  \vdots & \ddots & \vdots \\
  K_{n1} & \cdots & K_{nn}
\end{bmatrix} \right)
\]

The covariance function is determined by the covariance kernel, which evaluates a presumed relation between observed variables. Generally, any kernel assumes higher correlation between variables closer to each other in space, and lower correlation otherwise. Furthermore, for a small variation in the input arguments, a small variation in the estimation is expected. We used one of the most simple kernel functions, the Squared Exponential Kernel:

\[
K_{ij} = e^{-\|x_i - x_j\|^2}, K_{ij} = \begin{cases} 
  0, & \text{if } \|x_i - x_j\| \to \infty \\
  1, & \text{if } x_i \equiv x_j
\end{cases}
\]

Then in a GP, for a given random input argument \(x_\star\), the estimated output is similarly assumed to be drawn from a normal distribution as follows:

\[
\begin{bmatrix}
  f_1 \\
  \vdots \\
  f_n
\end{bmatrix} \sim N \left( \begin{bmatrix} 
  \mu_1 \\
  \vdots \\
  \mu_n
\end{bmatrix}, \begin{bmatrix} 
  K_{11} & \cdots & K_{1n} \\
  \vdots & \ddots & \vdots \\
  K_{n1} & \cdots & K_{nn}
\end{bmatrix} \right)
\]

\[
\begin{bmatrix}
  K_1 \\
  \vdots \\
  K_n
\end{bmatrix} \sim N \left( \begin{bmatrix} 
  \mu_1 \\
  \vdots \\
  \mu_n
\end{bmatrix}, \begin{bmatrix} 
  K_{11} & \cdots & K_{1n} \\
  \vdots & \ddots & \vdots \\
  K_{n1} & \cdots & K_{nn}
\end{bmatrix} \right)
\]

From this assumption, the mean and standard deviation of the estimation is:

\[
\mathbb{E}(f_\star) = \mu(x_\star) + K_\star^T K^{-1} (f - \mu(x))
\]

\[
\sigma^2 = -K_\star^T K^{-1} K_\star + K_{\star\star}
\]

GP Regressors were trained with eye-parameters for each of the spatial coordinates of the robot’s working environment at different positions and used to estimate the gaze fixation point for new eye-data. The input was 4 dimensional (pupilX-left, pupilY-left, pupilX-right, pupilY-right) and the mapping was (constrained by the GP formalism) to 1-dimension, so we trained 3 independent regressors, one for each 3D coordinate. This could be further refined, e.g. by implementing a joint prior distribution of joint locations.

III. RESULTS

The Matlab Gaussian Processes Regression for Machine Learning (GPML) toolbox by Rasmussen & Nickisch was used for implementation of our GP model [18]. The number of calibration points and hence the calibration accuracy is dependent on the calibration process explained above. The number of acquired points usually ranges between 500 and 1500 points, depending on the trajectory selected and the performance of the user during the procedure. The GP model selected for the system was run with settings (1) Linear Mean, (2) Matérn covariance, (3) Gaussian Likelihood and (4) Exact Inference method. The training method retrieves the optimal intrinsic constant hyperparameters that define the chosen functions to accurately describe the model’s behaviour. Once these hyperparameters are defined, simple regression may be applied to infer the gaze fixating point from new eye-data.

After calibration, the user is in free control of the robot arm in 3D space within the working volume. The system developed allows for two deterministic, non-overlapping
by projection on each individual dimension. The overall correspondence at least to 200 data points for each subject. were performed over 30% of the data (testing data), which train the Gaussian Process Regression. All the analyses different subjects. 70% of the recorded data was used to
of the robot from the corresponding pupils' data for seven
is performed by evaluating the prediction of the 3D position
for the same subject is shown in Fig. 4.
A histogram of the Euclidean Error in the predicted position the confident intervals of the prediction for a sample subject.
position of the robot versus the predicted position from eye-
on average for all subjects tested. Fig.3 shows the actual
dimensions) in the predicted gaze position lower than 2cm regression accuracy shows Euclidean errors (across x,y,z
preferred eye for the user is used to send the command to open/close the gripper successively. The analysis of the regression accuracy shows Euclidean errors (across x,y,z dimensions) in the predicted gaze position lower than 2cm on average for all subjects tested. Fig.3 shows the actual position of the robot versus the predicted position from eye-data for the X, Y and Z coordinates respectively, as well as the confident intervals of the prediction for a sample subject. A histogram of the Euclidean Error in the predicted position for the same subject is shown in Fig. 4.
Assessment of the accuracy of the calibration algorithm is performed by evaluating the prediction of the 3D position of the robot from the corresponding pupils' data for seven different subjects. 70% of the recorded data was used to train the Gaussian Process Regression. All the analyses were performed over 30% of the data (testing data), which corresponded at least to 200 data points for each subject. Performance of the estimation was independently analysed by projection on each individual dimension. The overall prediction accuracy is assessed as the 3D Euclidean distance between the actual and predicted point. The mean absolute error and standard deviation of this analysis across subjects is shown in Table I.

IV. DISCUSSION

We have developed and demonstrated continuous control of a robotic arm and gripper operation in 3D space using 3D eye-tracking technology. We have taken gaze interaction from 2D to 3D by using a continuous trajectory for calibration rather than a discrete grid of points in a 2D plane. This novel procedure allows for a very fast high calibration data throughput (up to 600 points per minute), while keeping the calibration time very similar to 2D calibration processes that often use a 3 x 3 set of points. This high data throughput enables Gaussian Processes Regression to be applied to estimate in a model-free manner 3D gaze end-point. The use of GP regression yielded the lowest estimation Euclidean error reported by any 3D gaze-controlled system up to date, operating below 2cm. Moreover, outliers of larger than 4cm are very rare (see Fig. 4). Crucially, these outliers coincided with uncertainty reported by the GP (e.g. outlier vs GP 95% confidence interval, see Fig. 3). Thus, when high reliability end-point precision is necessary, e.g. for safety considerations during robot operation, come into play, we can use the explicit representation of uncertainty by the GP regressor to detect and trigger an appropriate action.

One of the most important limitations preventing a wider use of eye-tracking technology as a robot control interface is its deficient performance in the interaction with the real 3D world. Our system shows promising improvement towards fluent eye-tracking control in 3 dimensions. It allows for free direct interaction with the environment incorporating the possibility of grasping and moving objects around. An accurate gaze-estimation in the 3D domain supposes a direct application to the field of prosthetics. The setup and calibration of our system can be performed by the user and takes only a few minutes from the user sitting down at the robot, calibration to direct control of reaching and grasping. This is a rather short time compared to technology like EEG used for neuroprosthetics control, which needs protracted calibration procedures (usually longer than 30 minutes) and expert user assistance [19] or invasive brain-machine interfaces that require (after surgery and rehabilitation on the time scale of months) hour-long calibration before being able to control robotic actuators.

Our subjects reported natural control of the system through the use of detecting long natural fixations to encode movement commands, avoiding the cumbersome procedure of learning gaze gestures. Moreover, user visual attention was always on the scene of their interactions, and not distracted by a screen-based user interface. Binocular eye tracking allowed us not only to perform 3D gaze measurement, but also to add a further dimension. We used voluntary 'winks' rather than natural occurring blinks and to include extra action selection features to our system: here, opening and closing of the robot gripper. New commands could be easily

<p>| Table I |</p>
<table>
<thead>
<tr>
<th>Mean absolute error(cm)</th>
<th>Standard deviation(cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>1.0</td>
</tr>
<tr>
<td>0.6</td>
<td>1.1</td>
</tr>
<tr>
<td>0.8</td>
<td>1.3</td>
</tr>
<tr>
<td>1.6</td>
<td>1.7</td>
</tr>
</tbody>
</table>
implemented including the detection of left vs right ‘winks’ independently or sequences of them, that may potentially add complexity and applicability towards prosthetic control.

In addition, the effective integration of the different software and algorithms into a single common interface facilitates the portability of the system, that could be potentially utilised to actuate other operational devices such as an eye-driven vehicle or drone.

A potential issue affecting the calibration procedure and posterior control of the actuator is the user. Although strict constraints (such as simultaneous detection of both pupils) are built-in during the data acquisition process in order to consider a certain data point for calibration, the active participation of the subject during the process is of major importance. This may preclude safe use by patients with more advanced motor deficiencies in executive control or autistic disorders. Additionally, the system as it does currently not perform head-tracking and thus assumes a fixed position of the user’s head, which may introduce calibration errors if movement occurs (such as smooth movements occurring during smooth pursuit). This could be solved by 1. monitoring the user’s head position and accounting for displacements to achieve increased precision and user comfort and 2. by using a head-mounted eye-tracker (eyetracking glasses).

In the system herein described, we used optical motion trackers in order to achieve high-temporal accuracy in the synchronisation gripper’s position monitoring during calibration and reduce the overall position-prediction error during control. An alternative approach intended to estimate the robot’s trajectory from the end-effector’s position commands (robot’s forward kinematics) was attempted. However, the signal feedback from our amateur-grade ArbotiX microcontroller timing was not accurate enough to allow good synchronisation with the eye-tracker (>300 ms synchronisation mismatch). Furthermore, the by design asynchronous commands sent to the servos in sequence compounded the a noisy movement estimation. Nonetheless, we are confident that time synchronisation with more professional robotic and prosthetic actuator will enable removing the need of additional motion tracking hardware.

In conclusion, we are exploiting that eye movements are highly correlated to our motor actions. They are a key factor in human motor planning and help us to direct and coordinate limb movements. Consequently, gaze-tracking becomes a rich source of information of user intention that may be taken advantage of. We have proven that accurate 3D gaze tracking is possible in a fast and accurate manner. This technology bears a tremendous potential as a signal for robotic teleoperation. Eye-movements are retained in the majority of motor disabilities enhances the applicability of eye-tracking technology in the field of brain-robot interfaces for the paralysed and in prosthetics. In contrast with the long training periods for BMI-based control, our system requires only a short calibration time and virtually no training from the user. We used our low-cost GT3D eyetracker, and with the rise of more low-cost, high-precision behaviour-based neuro technology [20] we expect that at least the sensor side can be further augmented at the same low equipment cost and even take the technology outdoors. Our system allows for straightforward control of a robot arm that permits the user to grab and move objects and manipulate the environment with the mere use of their eyes. Moreover, the high spatial accuracy and low latency (<100 ms) of our closed-loop system enables natural feeling interactions. This makes our 3D gaze-controlled robot system a vivid competitor to more conventional Brain-Machine Interfaces and screen-based assistive devices. The ability to freely look at the scene of our interaction and control the robot en passant enables novel ways of achieving embodiment in robotic solution that restore or augment human function.

REFERENCES