Journal on Multimodal User Interfaces manuscript No.
(will be inserted by the editor)

A Dataset for Point of Gaze Detection using Head Poses and Eye Images

Christopher D. McMurrough · Vangelis Metsis · Dimitrios Kosmopoulos · Ilias Maglogiannis · Fillia Makedon

Received: date / Accepted: date

Abstract This paper presents a new, publicly available dataset[1], aimed to be used as a benchmark for Point of Gaze (PoG) detection algorithms. The dataset consists of two modalities that can be combined for PoG definition: (a) a set of videos recording the eye motion of human participants as they were looking at, or following, a set of predefined points of interest on a computer visual display unit (b) a sequence of 3D head poses synchronized with the video. The eye motion was recorded using a Mobile Eye-XG, head mounted, infrared monocular camera and the head position by using a set of Vicon motion capture cameras. The ground truth of the point of gaze and head location and direction in the three-dimensional space are provided together with the data. The ground truth regarding the point of gaze is known in advance since the participants are always looking at predefined targets on a monitor.

Keywords gaze tracking · head tracking · point of gaze · dataset

1 Introduction

Eye tracking is a research problem of great interest, due to its large array of applications ranging from medical research, to human-computer interaction and marketing research. In the era of ubiquitous and mobile computing, the utilization of eye tracking modules enables the development of easy to use communication interfaces with assistive devices and systems. In most applications, accurate detection or tracking of the Point of Gaze (PoG) is a prerequisite for the development of successful systems to utilize input via eye motion. Various researchers have investigated the problem of gaze tracking and PoG detection and have proposed methods to deal with it [2, 4, 6, 7, 9–12, 16].

Authors in [2] present one of the first computer vision based methodologies proposed for non-intrusive detection of the position of a user's gaze from the appearance of the user's eye. A three-layer feed forward-network, trained with standard error back propagation, is used for this purpose, while an accuracy of 1.5 degrees is reported. A real-time stereo-vision face tracking and gaze detection system is presented in [9]. In [12] authors describe the FreeGaze tracking system, which detects gaze position by the pupil and the Purkinje images. The gaze position is computed from these two images by using an eyeball model, specific for each user. Then, a personal calibration is proposed in order to achieve accurate gaze direction estimation. In [16] the cornea

This work is supported in part by the National Science Foundation under award numbers CNS 0923494 and CNS 1035913. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Christopher D. McMurrough (Corresponding author)
The University of Texas at Arlington
E-mail: mcmurrough@uta.edu

Vangelis Metsis
The University of Texas at Arlington
E-mail: vmetsis@uta.edu

Dimitrios Kosmopoulos
Rutgers, The State University of New Jersey
E-mail: dk598@cs.rutgers.edu

Ilias Maglogiannis
University of Central Greece
E-mail: imaglo@ucg.gr

Fillia Makedon
The University of Texas at Arlington
E-mail: makedon@uta.edu

[1] The dataset can be downloaded from: heracleia.uta.edu/eyetracking

of the eyeball is modeled as a convex mirror, while a method is proposed to estimate the 3D optic axis of the eye. The visual axis, which is the true 3D gaze direction of the user, can be determined subsequently after knowing the angle deviation between the visual axis and optic axis by a calibration procedure, which is required for a new individual. Another work based on a stereo-vision approach is presented in [6] that yields the position and orientation of the pupil in 3D space. This is achieved by analyzing the pupil images of two calibrated cameras and by a subsequent closed-form stereo reconstruction of the original pupil surface. Under the assumption that the gaze-vector is parallel to the pupil normal vector, the line of sight can be calculated without the need for the usual calibration that requires the user to fixate targets with known spatial locations. A methodology that does not use calibration procedures is presented in [4]. The specific methodology uses multiple infrared light sources for illumination and a stereo pair of video cameras to obtain images of the eyes. Each pair of images is analyzed and the centers of the pupils and the centers of curvature of the corneas are estimated. These points, which are estimated without a personal calibration procedure, define the optical axis of each eye. To estimate the point-of-gaze which lies along the visual axis, the angle between the optical and visual axes is estimated by a procedure that minimizes the distance between the intersections of the visual axes of the left and right eyes with the surface of a display while participants look naturally at the display (e.g., watching a video clip). Finally a new physical model of the eye for remote gaze tracking is proposed in [11]. This model is a surface of revolution about the optical axis of the eye. Authors in their work determine the mathematical expression for estimating the PoG on the basis of the specific model and report high accuracy.

Most of the approaches [2, 4, 6, 7, 12, 16] try to detect the PoG at each time point by detecting the pupil center or other features, while some other approaches track the eye motion over time [9, 11].

In all cases, the inherent difficulty of tracking the eye itself, plus the fact that the head position has to be taken into consideration for the estimation of the final PoG, prevent most systems from giving very accurate results. In addition, direct comparison of the effectiveness of each of the proposed methods is not possible, due to the different parameters of the datasets on which each method has been tested. Although many of the proposed methods manage to roughly estimate the PoG on a computer display unit, none of them provides enough accuracy to allow convenient input through eye motion in a standard visual computer interface, especially when head motion is involved. Such fine grained detection of the PoG would be of great benefit to people, for example, that cannot use their hands to interact with a computer, due to some type of disability.

This dataset aspires to provide a resource for the development of new, more accurate eye tracking methods with focus on PoG detection, as well as a benchmark for the comparison of such methods. The dataset includes a set of videos recording the eye motion of 20 human participants as they looked at predefined positions of a computer display, or followed a target while it moved inside the display dimensions.

The ground truth of where the participant was looking at, at every time point, is known in advance, as the participants were advised to keep their eyes on the target at all times. The participant's head position and orientation relative to the display is tracked in three dimensions using a Vicon Motion Capture System[2], which provides sub-millimeter and sub-degree accuracy for the translation and rotation of the tracked object respectively, in the 3D space. This guarantees a virtually perfect ground truth regarding the participant's head position and orientation, which allows researchers to only worry about dealing with the eye tracking accuracy when trying to determine the exact point of gaze. In the collected data, the motion of the right eye of each participant as they were staring at, or following, predefined targets on the display has been recorded using an Applied Science Laboratories Mobile Eye-XG[3], head mounted, infrared monocular camera. The synchronized and timestamped eye recordings and head tracking data have been made publicly available to be used for educational and research purposes.

The main contribution of this work is that this is the first time that a public dataset provides synchronized modalities of eye images, head poses, and known gaze points. One of the fundamental problems of eye tracking is the necessity of a calibration process in order to create a mapping between pupil positions and gaze vectors in order to compute the PoG [8]. In most 2D approaches (such as interacting with a computer display), the calibration mapping directly correlates pupil position with a PoG and does not explicitly compute a gaze vector. This results in a mapping that is not robust to head movement, as any rotation or translation of the head will introduce error in the calibration (even a head resting on a pillow will drift over time, which degrades the accuracy of the calibration). A solution to this problem is to estimate the PoG by computing the intersection of the gaze vector originating from the eye with a point in the environment. This, however, requires knowledge of the head position and orientation

[2] http://www.vicon.com/
[3] http://www.asleyetracking.com/

in 3D. We anticipate that our dataset will be of great benefit to research in this area given that the included high-fidelity motion capture data will make it possible to evaluate the effect of sensor noise on the resulting PoG estimate. We also expect that the dataset will facilitate the development of advanced calibration and human eye modeling techniques, since we provide several video calibration sequences synchronized with the head tracking data. Finally, the large corpus of video data from several different subjects provides a valuable benchmarking tool for developers of computer vision based eye tracking algorithms, which tend to perform differently for various eye shapes, eye colors, etc. Our dataset makes it possible to evaluate the performance of new and existing algorithmic approaches while removing the constraints imposed by hardware and participant availability. We demonstrate such an application by evaluating the performance of some mainstream eye-tracking methods using the dataset.

The rest of this paper is organized as follows. Section 2 gives an overview of existing related datasets created for similar purposes and explains how our dataset differentiates from them. In section 3 we describe the system setup. Section 4 describes the head pose and the video data streams, while section 5 explains how to combine them. Section 6 describes the data collection and data format, while section 7 gives our experimental results. Finally, in section 8 we give our concluding remarks and discuss our plans for future work.

2 Related Work

Since the problem of eye tracking in general has been studied by previous researchers, there have already been efforts toward the creation of datasets to facilitate experiments with different aspects of the problem. However, to the best of our knowledge, all the previous publicly available datasets are either not well suited to the problem of PoG detection, or other limitations, such as insufficient head tracking accuracy or lack of ground truth, do not allow a reliable evaluation of methods developed for PoG detection via eye tracking. In this section we give an overview of existing eye tracking datasets and we explain their limitations regarding the problem of PoG detection.

In [1], the authors have collected a head pose and eye gaze dataset of ten participants using a web camera. The participants were instructed to perform a set of head and eye motions and then specific head pose and eye gaze estimation methods were tested on the collected data. The ground truth about the head's pose was extracted using 3 LEDs mounted at the participant's head. Using this method, one can determine the location and rotation of the head relative to the camera, but only in the 2D space, i.e., the participants cannot move towards or away from from the camera. As for the ground truth regarding the gaze position, only three discrete classes of gaze directions were used: looking straight forward, looking to the extreme left and to the extreme right. This is insufficient for applications where the exact point of gaze needs to be detected in continuous space.

A similar dataset is provided in [14]. In this dataset the ground truth regarding the head pose was obtained by asking the participants to point a laser beam mounted on their head to a specific location. An extra limitation of this dataset is that it only provides a set of images instead of video, which makes unsuitable for tracking problems.

The eye tracking dataset found in [5] has been generated for the purposes of examining visual attention models on a set of video sequences. The eye point of gaze on the display was determined using a commercially available Locarna "Pt-Mini" head mounted eye tracker, therefore the tracking accuracy depends on the tracker and no ground truth is provided.

Finally, in [13], the authors provide a database of visual eye movements from 29 observers as they look at 101 calibrated natural images. The eye movements of the participants as they look at the images are recorded using a Forward Technologies Generation V dual-Purkinje eye tracker, while holding a fixed head position. Again, the final PoG provided relies on the capabilities of the eye tracking system used.

What differentiates our dataset from the above ones, is that the ground truth of both the PoG on the display and head rotation/translation in the 3D space are known in advance and with a high degree of accuracy, which makes it ideal for testing the accuracy of methods estimating the point of gaze based on the eye motion, assuming that the head pose is known.

3 System setup

3.1 Overview

The experimental system comprises mainly:

- a Mobile Eye-XG device that monitors the right eye (see Figure 1).
- A network of 16 Vicon MX motion capture cameras, which can be used for extracting the ground truth position of reflective markers at 100Hz with sub-millimeter accuracy (see Figure 3).

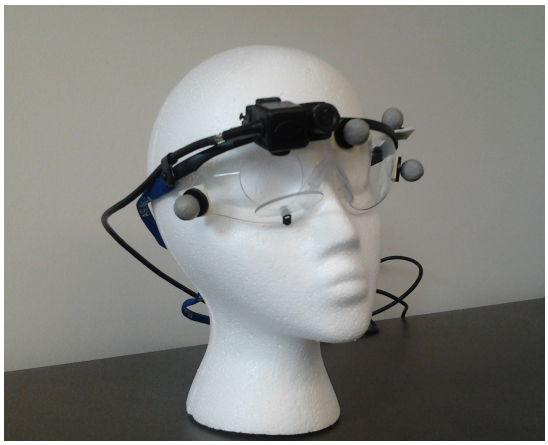


Fig. 1 Picture of Mobile Eye-XG, head mounted, infrared monocular camera. Reflective markers used for head tracking have been attached to the glasses.

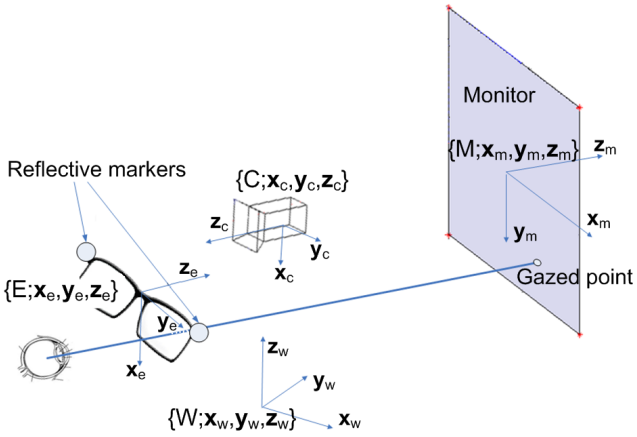


Fig. 2 The coordinate systems for the world $\{W; \mathbf{x}_w, \mathbf{y}_w, \mathbf{z}_w\}$, the eye camera $\{C; \mathbf{x}_c, \mathbf{y}_c, \mathbf{z}_c\}$ Mobile Eye-XG $\{E; \mathbf{x}_{ce}, \mathbf{y}_{ce}, \mathbf{z}_{ce}\}$ and monitor $\{M; \mathbf{x}_m, \mathbf{y}_m, \mathbf{z}_m\}$.

– An LCD *Samsung LN32C350 (32 Inch, 1360x768 pixels)* display unit, on which we display several patterns (see Figure 4).
– A PC in which the images from the camera are stored, as well as the head pose and PoG data.

3.2 Coordinate Systems

In this section we describe the coordinate systems (CSs) of the proposed setup. We have the following CSs:

1. $\{W; \mathbf{x}_w, \mathbf{y}_w, \mathbf{z}_w\}$, the world coordinate system located on the ground behind the test participant;
2. $\{M; \mathbf{x}_m, \mathbf{y}_m, \mathbf{z}_m\}$, attached to the upper left corner of the monitor;
3. $\{C; \mathbf{x}_c, \mathbf{y}_c, \mathbf{z}_c\}$, attached to the eye camera;
4. $\{E; \mathbf{x}_{ce}, \mathbf{y}_{ce}, \mathbf{z}_{ce}\}$, attached to the center of the Mobile Eye-XG device.

Each CS is defined using the right-hand notation. For the CS {W} , $\mathbf{z}_w$ points up and $\mathbf{x}_w$ points to the right. The Vicon Motion Capture System returns position and orientation data relative to this CS. CS {M} is attached to the center of the monitor, with $\mathbf{z}_w$ pointing up and $\mathbf{x}_w$ pointing to the right. CS {E} is attached to the center of the Mobile Eye-XG device, with $\mathbf{z}_w$ pointing up and $\mathbf{x}_w$ pointing to the right. This CS also corresponds to head position, as the Mobile Eye-XG is worn on, and does not move relative to the head. Figure 2 shows the relative assignment of CSs used in the dataset.

4 Video and Head pose Streams

This section describes the video and head pose data streams captured by the system. In each recording session, the user is instructed to visually track target points on a fixed video display while wearing an eye recording device. This device collects video data of the participant's eye while being tracked in 3D within a motion capture system. The data streams from the motion capture system are synchronized with the video frames provided by the eye recording device, as well as the pixel coordinates of the video display target points.

The device used to obtain the eye video data is an Applied Science Laboratories Mobile Eye-XG. The Mobile Eye-XG is worn on each participant's head during data collection, and is positioned such that the user's right eye is centered in the video frame. The recording of the right eye over the left is advantageous in that a higher percentage of the population exhibits right eye versus left eye dominance. The video data is recorded with a resolution of 768 x 480 pixels at a frame rate of 29.97 Hz. The provided resolution is considered adequate for most tracking applications due to the close proximity (less than 5 cm) of the camera to the participants eye and the use of a wide-angle lens, which results in the corneal area occupying the majority of the image space. Each video is provided as an individual AVI file encoded with the Motion JPEG Video (MJPG) codec. The Mobile Eye projects a triangular infrared glint pattern on the user's eye during recording, which can be used as an additional tracking feature.

The user head position and pose was measured during the data collection process using a Vicon motion capture environment. The motion capture system consists of 16 tracking cameras surrounding an area measuring roughly 10 x 10 meters. The system is able to track the position and orientation of multiple rigid structures equipped with reflective markers at a rate of 100 Hz with sub-millimeter accuracy. The tracked struc-

Fig. 3 Picture of the Vicon Motion Capture System setup at the Heracleia Laboratory.

tures of interest in our case were the display unit and the participant's head. The reflective markers used to track the head are actually attached to the skeleton of the glasses (see Figure 1). Therefore, different head shapes or hair styles should not affect head tracking parameters. After the glasses have been worn by each participant, every head movement will correspond to the same movement (direction and rotation) for the skeleton of the glasses. Camera position with respect to the eye stays fixed throughout the recording process for each participant. Figure 3 shows the Vicon setup that was used during the data collection process. The pose stream is composed of homogeneous transformation matrices from CS W to CS M ($^{M}\mathbf{H}_{W}$) and from CS W to CS E ($^{E}\mathbf{H}_{W}$). Generally, the transformation $\mathbf{H}$ is a 4x4 matrix, which relates a source point $\mathbf{P}_s$ and a destination point $\mathbf{P}_d$ through a rotation $\mathbf{R}$ and translation $\mathbf{t}$ as follows:

$$\mathbf{P}_d = \mathbf{H} \cdot \mathbf{P}_s \Leftrightarrow (x_d, y_d, z_d, 1)^T = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} (x_s, y_s, z_s, 1)^T \quad (1)$$

The $\mathbf{H}$ matrix is calculated by a least squares method, if no outliers are assumed. For more general cases a review of related methods can be found in [3].

5 Integration of gaze estimation and head position

To map the pupil coordinates to monitor coordinates, the homogeneous transformation matrices can be employed. This can be calculated from

$$^{M}\mathbf{x}_p = {}^{M}\mathbf{H}_{W} \cdot {}^{W}\mathbf{H}_{E} \cdot {}^{E}\mathbf{H}_{C} \cdot {}^{C}\mathbf{x}_p \quad (2)$$

where in general $^{A}\mathbf{H}_{B}$ is the homogeneous transformation matrix from the CS A to the CS B. $^{M}\mathbf{H}_{W}$ is generally constant but we measure it anyway through the motion capture system, since there might be some motion of the display unit due to vibrations. $^{W}\mathbf{H}_{E}$ is also measured online using motion capture and the reflective markers on the Mobile Eye-XG. $^{E}\mathbf{H}_{C}$ is calibrated during the first session experiments.

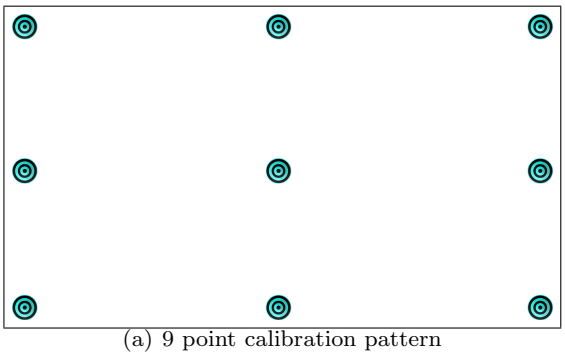
6 Data Collection and Dataset Structure

A total of 20 participants (2 women and 18 men) participated in our data collection sessions. The participants were graduate and undergraduate students of the University of Texas at Arlington. Some of them had normal vision and some of them wore contact lenses or spectacles. The participants that normally wore spectacles did not use them during the data collection process. However, their vision level was still good enough to locate the displayed target on the monitor. The special characteristics of each participant are given as metadata together with the dataset.

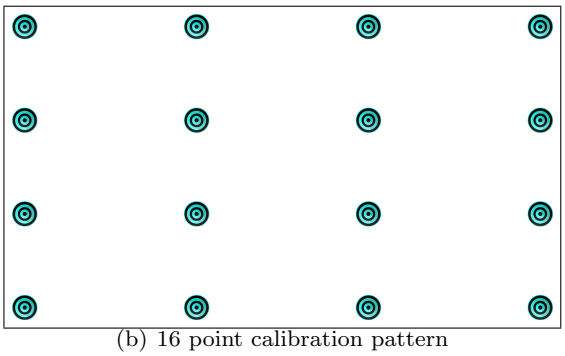
Each participant performed two video recording sessions in which they looked at (or followed) a target on the display unit. Both sessions recorded data from three different video target patterns, resulting in six videos total per participant. In the first session, the participants were allowed to move only their eyes while keeping their head still, whereas in the second session, they were allowed to freely move their heads and eyes at their convenience. The videos provided from both recording sessions capture both smooth pursuit and gaze fixation data in a continuous form (i.e., the ground-truth PoG locations are not discretized into a small number of labels, as they are in other datasets). Figure 4 shows a photo taken during the data collection process. In the photo, the reader can see the experimental setup used for the data collection.

In the first session, the participants were asked to keep their head still and their eye motion, while looking at different patterns of targets appearing on the computer display, was recorded. In the first pattern, a target appeared at nine different positions of the display for a few seconds and the participants were instructed to look at the target as soon as it appears and until it disappears. Note that since the human eye may require a few milliseconds from the moment the target appears on display until the moment the eye point of gaze moves to fall onto it, the target location and the point of gaze may not align for a few milliseconds after the appearance of each target. However, they should be aligned soon after.

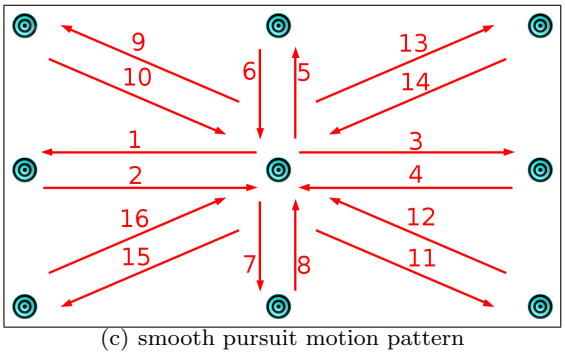
Fig. 4 Photo taken during the data collection process. The photo shows a participant wearing the eye video recording device, as well as the rest of the experimental setup including the display unit, the data collection hardware and some of the Vicon system cameras.



(a) 9 point calibration pattern



(b) 16 point calibration pattern



(c) smooth pursuit motion pattern

Fig. 5 An example of the target locations on which the participants have to focus their point of gaze during the data collection process. For convenience, the figures show all of the target locations at once. During the data collection process only one target at a time is displayed.

Similarly, in the second pattern, 16 targets appeared on the display and the participant had to repeat the same procedure. The number of targets and their location on the display was chosen so as to resemble common calibration patterns used by eye trackers. The third pattern, instead of static targets, involved a target moving inside the display and the participants had to follow the target with their eyes at all times. This pattern is particularly useful for eye tracking methods which do not statically determine the point of gaze but follow the center of the pupil or other eye features over time. Figure 5 shows an example of the three target patterns shown on the display during the recording sessions.

Since in real life people do not move only their eyes but also their heads to look at different targets (or follow a target as it is moving), in the second session the participants repeated the same process as in the first, but this time they did not have any constraints regarding their head motion. The exact position of the head in the 3D space was tracked by the Vicon system using a set of markers attached to the head mounted eye video recorder. For convenience, four markers were also attached at the corners of the computer display. This allows us to determine the exact location of the head and the display monitor in the same coordinate system. Figure 6 visualizes the locations of the head and the computer display in the 3D space as captured by the Vicon System.

Requiring from the users to keep their head still while moving only their eyes, in the first session, is a common practice used by many eye tracking systems which incorporate some kind of calibration before using the eye input. Such systems are used in cases of patients with full body paralysis (e.g. Amyotrophic lateral sclerosis (ALS)). In those situations, the patients are not able to voluntarily move their heads, however, a slight involuntary drifting may occur even if the head is supported by a pillow or headrest. The accuracy of such systems deteriorates over time even with subtle head movement, and thus, re-calibration is required for the system to become usable again. One of the reasons that we included the session of trying to hold the head still in our collected data, was to consider such cases and provide the option for experiments that evaluate the accuracy loss due the head position drifting. The purpose here is not to make sure that the head remains completely still throughout the session, but to simulate

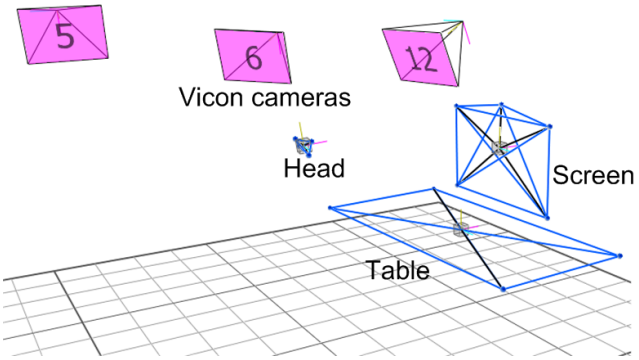


Fig. 6 A representation of the 3D space and the tracked objects as perceived by the Vicon Motion Capture System using world coordinates.

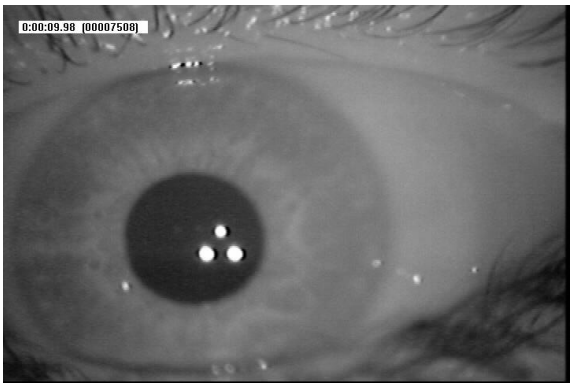


Fig. 7 An example video frame contained in the dataset

metadata	value(s)
gender	male
age	24
glasses	no
contacts	no
${}^{E}\mathbf{H}_{W}$	$\begin{bmatrix} 0.99628 & -0.04734 & 0.07197 & -367.86231 \\ 0.04230 & 0.99666 & 0.06990 & 704.13371 \\ -0.07504 & -0.06660 & 0.99495 & 1219.87140 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
${}^{M}\mathbf{H}_{W}$	$\begin{bmatrix} 1.00000 & -0.00124 & -0.00015 & -350.59819 \\ 0.00124 & 1.00000 & 0.00063 & 1677.67326 \\ 0.00015 & -0.00063 & 1.00000 & 1081.51075 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
${}^{M}\mathbf{x}_{p}$	$\begin{bmatrix} 683 \\ 384 \end{bmatrix}$

Table 1 Example metadata for the video frame shown in Figure 7. Metadata for each frame is provided in both MATLAB and CSV files for convenience.

real life situations where minor movements may occur when the head is not fixed using some external support. The exact head position and orientation is still tracked by the Vicon system and provided as ground truth. This allows researchers to detect if motion has occurred, and if so, it can be taken into consideration in order not to throw off the calibration.

6.1 Dataset Structure

In this section we describe the structure of the dataset and the details of the provided format. For each one of the twenty participants, the dataset includes six videos in avi format and corresponding metadata. File sets 00001, 00002, 00003, come from the first session in which the participant was to remain as still as possible, and file sets 00004, 00005, 00006, come from the second session in which the participant was freely allowed to move their head.

File sets 00001 and 00004 contain the videos of the first pattern of each session (a target appearing in 9 different locations of the display unit), file sets 00002 and 00005 contain the second pattern (16 target locations), and file sets 00003 and 00006 contain the videos coming from third pattern (target moving within the display).

The metadata includes two homogeneous transformation matrices, H from {W} to {M} and H from {W} to {E}, and it also includes the (i, j) pixel location of the target location for each video frame. This data is packaged as a MATLAB (.mat) data file and as a CSV file. The .mat file contains a structure which includes the two homogeneous transformation matrices and the pixel locations. The data is written into the CSV file with each column corresponding to a frame in the video, rows 1:16 represent the flattened homogeneous transformation matrix from {W} to {E}, rows 17:32 represent the flattened homogeneous transformation matrix from W to M, and rows 33:34 represent the flattened pixel locations. The transformation matrices are unflattened by their (i, j) value being the $(((i-1)\times 4)+j)$ of their respective columnar data. The pixel locations are unflattened by their (i, j) value being the first and second values respectively. Pixel locations can be easily converted to metric units using H from {W} to {M} and the known display resolution.

Figure 7 shows a single frame from the video set, while Table 1 shows the metadata corresponding to the video frame and participant.

7 Experimental results

To demonstrate the utility of the proposed dataset, we used it to evaluate the performance of the popular starburst algorithm [15] compared to some other conventional approaches.

The starburst algorithm works by roughly estimating the pupil center (simple thresholding), fitting an

ellipse to the pupil 'blob', and then refining the ellipse by considering pupil edge points, which lie on rays projected outward from the center of the first ellipse. It then finds a corneal reflection (a small dot made from an IR LED), and computes the vector between the pupil ellipse center and corneal reflection center.

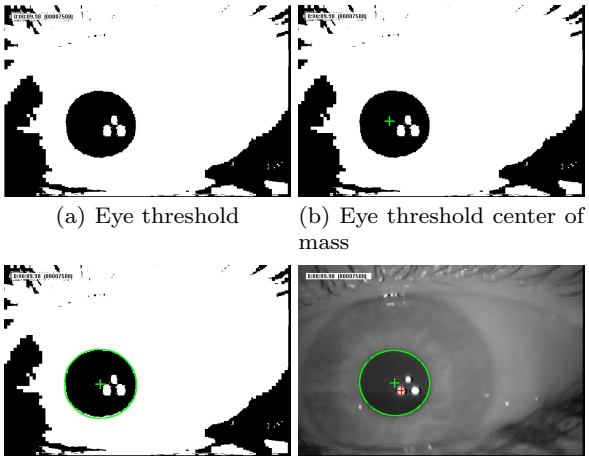
In the experiments, we evaluated the effectiveness of the starburst algorithm versus simple pupil blob center tracking and rough ellipse fitting. The blob center method tracks only the center of mass of the thresholded pupil image, while the fitted ellipse method tracks the center of the best-fit ellipse around the pupil blob with no other refinements. These methods are each performed using the same 5 test video sets which were selected from the overall dataset of 20 participants.

Video based eye trackers are generally sensitive to factors such as eye color, eye shape, user age, previous corrective surgery, camera position, etc. This can often be mitigated by adjusting tracking parameters for each user, but occasionally a tracker may fail completely under certain conditions. The test video sets were chosen such that they gave meaningful, easy to compare results for each of the methods that we evaluate, and also in order to perform the necessary processing in a reasonable amount of time.

In each case, the tracker was calibrated using the 9 point calibration pattern video corresponding to the test set. A linear mapping between the pupil/CR vectors and known screen coordinates was created, one for each of the evaluated methods for each participant. Each method is run on the remaining two head motion-free videos acquired during the first participant recording sessions. Each estimated point-of-gaze (PoG) was measured against the known screen target locations, provided by the dataset metadata files for each video. The videos containing head motion from the second participant recording sessions were not used, since the trackers we evaluated were not designed with this in mind. Figure 8 shows the application of each tracking method to a single video frame from the test set.

The RMSE error of each method is presented in Table 2. Simple outlier rejection was performed on the tracking results (an error over 100 pixels was removed from consideration). This was generally caused by blinking and/or delay as the user fixates on a new target location. No other filtering of the data was performed. Using simple outlier rejection, 8.49% of the tracking estimates were removed (i.e. 91.51% of the frames where kept to compute the results in Table 2).

In Figure 9 we showcase the results of the gaze estimation experiments using simple blob center tracking, ellipse center tracking, and full starburst algorithm implementation. The starburst algorithm shows an overall higher degree of accuracy than the other methods. The sequential application of various refinement methods (threshold centering, rough ellipse fitting, then ellipse refinement via starburst) results in a clear gain in accuracy at each step. This fact can be valuable to designers of flexible tracking systems that may have accuracy requirements and available processing resources that change over time. It is also shown that the number of outliers decreases as each refinement step is applied, especially once an ellipse is fit to the threshold blob area. This is likely due to an increased robustness to motion blur and corneal reflection location as the eye changes position, since the blob tracking is based on dark pixel center of mass and does not assume any knowledge of the pupil geometry.



(a) Eye threshold (b) Eye threshold center of mass

(c) Eye threshold ellipse (d) starburst algorithm results

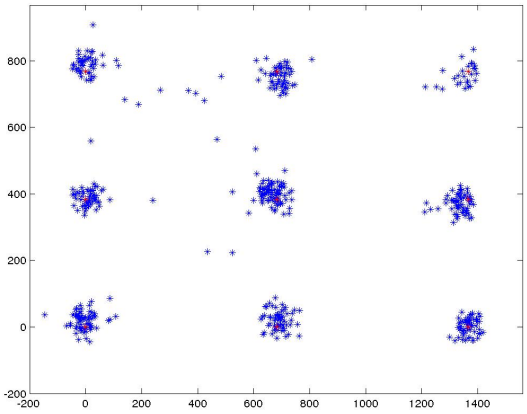
Fig. 8 Application of the various pupil tracking methods to a test video frame. Each of the three methods were performed for every frame in the test set

Method	blob center	ellipse center	starburst
Outliers	93	51	42
Average error	38.22	26.49	17.75
MSE	1869.18	1017.11	513.22
Root MSE	43.23	31.89	22.65

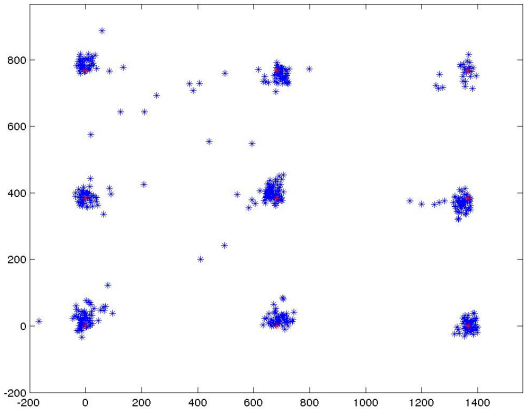
Table 2 Experimental results of the 3 tracking methods. The error is given in pixels.

8 Conclusion and Future Work

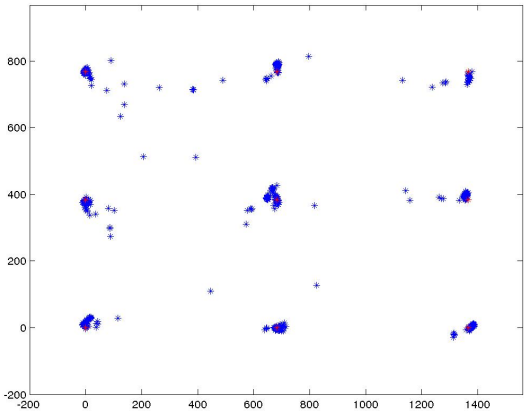
In this paper we have described a new publicly available eye tracking video dataset. The dataset has been collected and published with the purpose of facilitating the research in point of gaze detection or any other related eye tracking applications. The hardware used, the



(a) blob center



(b) ellipse center



(c) starburst algorithm

Fig. 9 Gaze extraction for the experiments using the starburst algorithm. The 9 red points are the known target location and the blue points are the point of gaze (PoG) computed using the linear calibration.

collection methodology and related metadata accompanying the dataset, such as the ground truth about the point of gaze and head pose, are presented and explained. We expect that this dataset will be a useful resource to the eye tracking community. We have already evaluated the dataset using three popular approaches and namely the starburst algorithm, simple pupil blob center tracking, and simple fitted ellipse tracking.

A resource still missing from the eye tracking community is a dataset where objects located or moving in the 3D space, instead of a 2D computer display, are followed by the human eye. Such a resource would be of great interest to robotics applications where the communication of humans with robots involves interaction in the 3D space. In the near future we are planning to publish another dataset that covers this gap.

Acknowledgements This work is supported in part by the National Science Foundation under award numbers CNS 0923494 and CNS 1035913. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Asteriadis S, Soufleros D, Karpouzis K, Kollias S (2009) A natural head pose and eye gaze dataset. In: Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots - AFFINE '09, ACM Press, New York, New York, USA, pp 1–4, DOI 10.1145/1655260.1655261
2. Baluja S, Pomerleau D (1993) Non-Intrusive Gaze Tracking Using Artificial Neural Networks. In: Working Notes: AAAI Fall Symposium Series, Machine Learning in Computer Vision: What, Why and How?
3. Eggert D, Lorusso A, Fisher R (1997) Estimating 3-D rigid body transformations: a comparison of four major algorithms. Machine Vision and Applications 9(5-6):272–290, DOI 10.1007/s001380050048
4. Guestrin ED, Eizenman M (2008) Remote point-of-gaze estimation requiring a single-point calibration for applications with infants. In: Proceedings of the 2008 symposium on Eye tracking research & applications - ETRA '08, ACM Press, New York, New York, USA, p 267, DOI 10.1145/1344471.1344531
5. Hadizadeh H, Enriquez MJ, Bajić IV (2012) Eye-tracking database for a set of standard video sequences. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 21(2):898–903, DOI 10.1109/TIP.2011.2165292

6. Kohlbecher S, Bardinst S, Bartl K, Schneider E, Poitschke T, Ablassmeier M (2008) Calibration-free eye tracking by reconstruction of the pupil ellipse in 3D space. In: Proceedings of the 2008 symposium on Eye tracking research & applications - ETRA '08, ACM Press, New York, New York, USA, p 135, DOI 10.1145/1344471.1344506
7. Lee EC, Park KR (2008) A robust eye gaze tracking method based on a virtual eyeball model. Machine Vision and Applications 20(5):319–337, DOI 10.1007/s00138-008-0129-z
8. Majaranta P, Aoki H, Donegan M, Hansen DW, Hansen JP, Hyrskykari A, Räihä KJ (2011) Gaze Interaction and Applications of Eye Tracking. IGI Global, DOI 10.4018/978-1-61350-098-9
9. Matsumoto Y, Zelinsky A (2000) An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), IEEE Comput. Soc, pp 499–504, DOI 10.1109/AFGR.2000.840680
10. Morimoto CH, Mimica MR (2005) Eye gaze tracking techniques for interactive applications. Computer Vision and Image Understanding 98(1):4–24, DOI 10.1016/j.cviu.2004.07.010
11. Nagamatsu T, Iwamoto Y, Kamahara J, Tanaka N, Yamamoto M (2010) Gaze estimation method based on an aspherical model of the cornea. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10, ACM Press, New York, New York, USA, p 255, DOI 10.1145/1743666.1743726
12. Ohno T, Mukawa N, Yoshikawa A (2002) FreeGaze: a gaze tracking system for everyday gaze interaction. In: Proceedings of the symposium on Eye tracking research & applications - ETRA '02, ACM Press, New York, New York, USA, p 125, DOI 10.1145/507072.507098
13. Van Der Linde I, Rajashekar U, Bovik AC, Cormack LK (2009) DOVES: a database of visual eye movements. Spatial vision 22(2):161–77, DOI 10.1163/156856809787465636
14. Weidenbacher U, Layher G, Strauss PM, Neumann H (2007) A comprehensive head pose and gaze database. 3rd IET International Conference on Intelligent Environments (IE 07) 2007(CP531):455–458, DOI 10.1049/cp:20070407
15. Winfield D, Parkhurst D (2005) Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops, IEEE, vol 3, pp 79–79, DOI 10.1109/CVPR.2005.531
16. Zhu Z, Ji Q (2007) Novel Eye Gaze Tracking Techniques Under Natural Head Movement. IEEE Transactions on Biomedical Engineering 54(12):2246–2260, DOI 10.1109/TBME.2007.895750