

# A3D: A Device for Studying Gaze in 3D

Mahmoud Qodseya<sup>(✉)</sup>, Marta Sanzari, Valsamis Ntouskos, and Fiora Pirri

ALCOR Lab, DIAG, Sapienza University of Rome, Rome, Italy  
{qodseya,sanzari,ntouskos,pirri}@diag.uniroma1.it

**Abstract.** A wearable device for capturing 3D gaze information in indoor and outdoor environments is proposed. The hardware and software architecture of the device provides an estimate in quasi-real-time of 2.5D points of regard (POR) and then lift their estimations to 3D, by projecting them into the 3D reconstructed scene. The estimation procedure does not need any external device, and can be used both indoor and outdoor and with the subject wearing it moving, though some smooth constraint in the motion are required. To ensure a great flexibility with respect to depth a novel calibration method is proposed, which provides eye-scene calibration that explicitly takes into account depth information, in so ensuring a quite accurate estimation of the PORs. The experimental evaluation demonstrates that both 2.5D and 3D POR are accurately estimated.

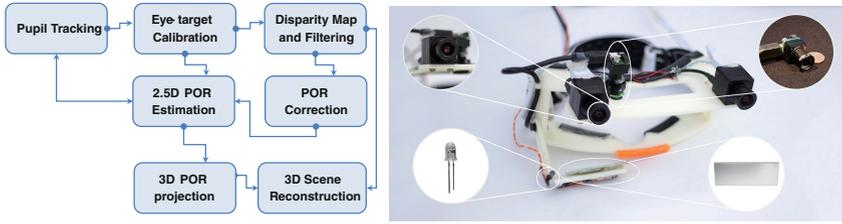
**Keywords:** Wearable device · 3D gaze estimation · Point of regard in 3D scene

## 1 Introduction

Eye tracking has developed in the context of studying human visual selection mechanisms and attention (see [1–3] for a review on eye detection and gaze tracking in video-oculography). It is indeed well known that the points toward which humans direct the gaze are crucial for studying human perception and his ability to select the regions of interest out of a massive amount of visual information [4].

In the last few years the use of head-mounted eye tracking has spread in several research areas such as driving [5,6], learning [7], marketing [8], training [9], cultural heritage [10] and prevalently in human computer interfaces [11,12]; just to cite few of an increasing number of applications where gaze direction is studied. All these applications spotlight the need to move beyond prior models of computational attention and saliency [13–16] and move toward a deeper experimental analysis of gaze direction and eye-head motion, by collecting data to better understand the relation between *point of regard* (POR) and visual behavior [17], likewise strategies of search [18] and detection [19] in natural scenes.

However only quite recently models for head-mounted eye tracking have been extended first to include head motion tracking [20,21] and further to 3D, so as to be employed in real life experiments in unstructured settings [22–25]. In [26]



**Fig. 1.** On the left: a schema of the methods involved for projecting the PORs in the 3D reconstructed scene. On the right: the head-mounted eye-tracker with all its components.

Paletta and colleagues propose an interesting solution based on POR projection in an already reconstructed environment, exploiting an RGB-D sensor and the approach of [27]. To localize the PORs in the reconstructed scene they use key point detection and matching against those key points that are already in the reconstructed map. This said, 3D projection of the gaze in natural scenes is still an open research problem due to the difficulties of both designing lightweight wearable devices supporting complex computations for solving an ill-posed problem.

There is a considerable advantage in extending head-mounted eye-tracking to 3D. The advantages are formidable not only for the purpose of providing the depth of the field of view, not only for effectively understanding eye motion in natural scene, the shift and the inhibition of return [28], having the possibility of collecting PORs in the scene, but also to understand the relation between saliency and combined motion of eyes, head and body. All these factors can induce quite relevant advances in the comprehension of visual perception, and also would provide the possibility of collecting a huge amount of data in several contexts and for several useful applications.

Despite the remarkable advances in capturing head motion and real world scenes, in many of the above cited papers extension to 3D does not necessarily imply full 3D reconstruction of the scene nor the subject localization in the visually explored scene. For example in [24] the 3D gaze projection is obtained by remote recording, and localization and reconstruction is not considered. Similarly in [25] the head position and orientation are determined either by a remote camera or by fiducial markers, which have to be placed in the environment in such a way that at least one marker is visible in the scene camera image of the eye-tracker. It follows that no localization nor reconstruction is provided in so implying that the subject is not free to move in the environment.

However in order to be fully usable as a 3D device, both indoor and outdoor, and to provide a deep insight of visual perception, while the subject is performing simple activities, the point of regard needs to be localized in space likewise the head-mounted tracker. Therefore a 3D device requires not only eye and head tracking but also localization and depth perceptivity within the reconstructed scene, also to grant free motion to the subject.

In this paper, we propose a head-mounted device A3D for gaze estimation in both the 2D scene images (actually in 2.5D as images are RGB-D, due to the eye-scene calibration procedure) and in the 3D reconstructed scene, ensuring mobility of the subject wearing the device, both indoor and outdoor (see Fig. 1). This contribution extends significantly the work of [23, 29] with respect to pupil estimation, eye-scene calibration and 3D reconstruction.

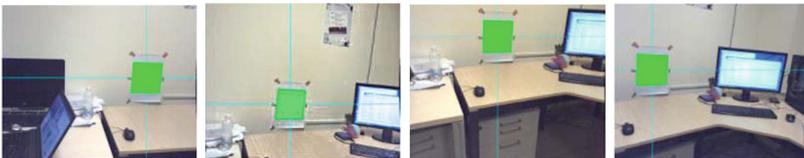
An overview of the methods involved, to ensure PORs projection in a 3D dense scene, is illustrated in Fig. 1. The proposed system works in quasi-real-time on a GPU, NVIDIA Corporation GK106GLM [Quadro K2100M]. However we are still far from a true gaze measurement device. In fact, the system has still severe limitations on freedom of movements. Indeed, approximatively correct localization and dense reconstruction is ensured under the proviso of smooth motion. In other words, erratic motion of the head, likewise sudden motions of the body must be avoided.

The remainder of this paper is organized as follow. In Sect. 2 and its subsections, we present the pipeline of the proposed system and the model for pupil detection, disparity map and filtering, localization and dense reconstruction, and finally the projection of PORs, or the visual axes in the reconstructed scene. In Sect. 3 and its subsections we present the details of the A3D device, the software, the performance and an analysis of errors computation for both the 2.5D and 3D PORs projections. We conclude the work with some considerations on present limits and future work.

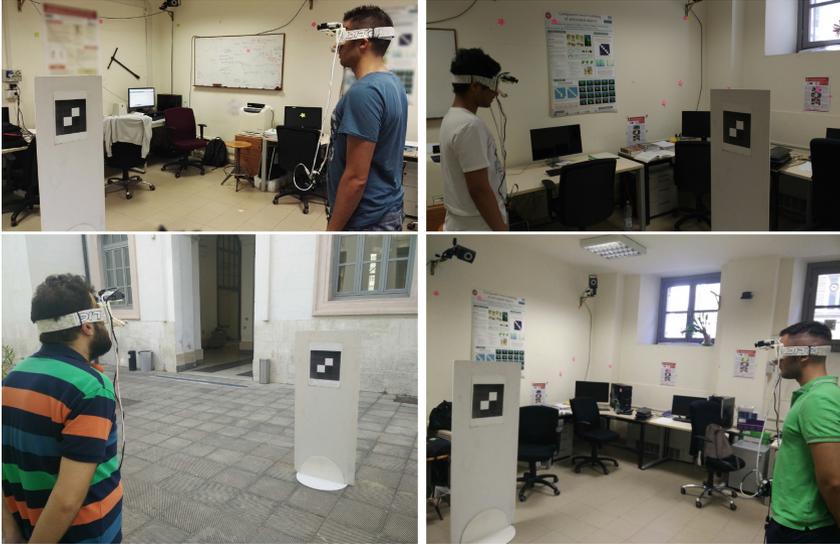
## 2 Proposed Model

In this section we introduce the methods that are necessary to project PORs in the dense reconstructed scene. Note that we assume that the scene is static though the subject wearing the A3D device can move, however we shall show in Sect. 3 an example in slow motion. The A3D device is described in Sect. 3.

The methods we introduce take care of pupil detection, pupil and scene images calibration, disparity map construction together with filtering, the device localization, 3D scene reconstruction and, finally, the projection of the PORs in the scene. Though the solution proposed improves the Gaze machine of [23, 29], we use mainly state of the art methods to solve each of the above mentioned problems. The whole pipeline extends the approaches where is needed, to make the whole system work properly.



**Fig. 2.** The calibration sequence: back, forward, head left, right, up and down

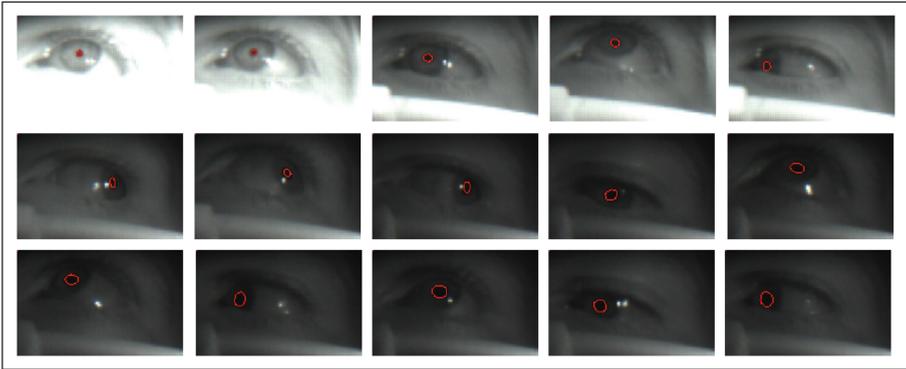


**Fig. 3.** Subjects performing the calibration procedure of the A3D device under different conditions.

## 2.1 Pupil Detection

A wealth of literature has faced the problem of pupil detection, as reported in [1–3]. Most approaches relies heavily on laboratory conditions, histogram thresholding and shape fitting, such as [30], which is also implemented in *Pupil* of [31]. Most 3D model-based approaches [21, 32, 33] rely on estimating the parameters of some model of the human eye. Due to the high complexity of the eye [34], shape based models are often hard to fit and usually require sophisticated settings like multiple light sources and cameras. On the other hand, those simpler models that are not complex enough to take into account eye shape and kinematics have to rely on simplifying assumptions, resulting in less accurate estimations. To overcome the limitations related to special assumptions on the eyeball or cornea shape and especially to cope with the eye dynamics, we introduce a model for estimating the POR on the basis of a set of learned parameters.

Dark pupil images are acquired with active IR illumination and a light filter is attached to the camera pointing the eye. The goal here is to predict whether a pixel  $\mathbf{x} = (u, v, w)$ , at location  $(u, v)$  with intensity  $w$ , in a test eye image belongs to the pupil or not, given a number of vectorized images and a vector of responses in  $\{-1, +1\}$ . The problem can be seen in terms of binary classification, by modelling a latent function  $g$ . Namely, given training data  $X$ , for a single pixel the probability of success  $p(y = 1|X)$  is related to a latent function  $g(X)$  which is mapped to the unit interval by a probit function  $\varphi$ ,  $p(y = 1|X) = \varphi(g(X))$ , with  $\varphi(g(X)) = 0.5(1 + \text{erf}(g(X)/\sqrt{2}))$ . Due to the symmetry of  $\varphi$ , the likelihood of a single pixel, given the probit, is  $p(y_i|g_i(X)) = \varphi(y_i g_i(X))$ . Hence we rely on



**Fig. 4.** Pupil detection with different eye orientations and sunlight intensities.

a Gaussian process as a prior on  $g$ , whose mean and variance gives a value for the prior  $p(y = 1|X)$ , somehow encoding the min and max variation of the pupil in the eye image. Using Gaussian process approximation  $q(g|X, y)$  for binary classification [35, 36], the posterior expectation and variance for  $g_*$ , given a test pixel  $\mathbf{x}_*$  of a new test image can be computed analytically, namely, the predictive probability of  $\mathbf{x}_*$  to be of class 1 is given by

$$\int \varphi(g_*)q(g_*|X, \mathbf{y}, \mathbf{x}_*)dg_* \quad (1)$$

Both GP and visible light filter enable the method to work well for both indoor and outdoor environments, where the GP provides a dynamic threshold and the filter eliminates the effect of the visible light on the intensity of the eye image. An ellipse is finally fitted to bound the pupil. See Fig. 4.

## 2.2 Depth Map and Pupil-Target Calibration

Differently from other approaches (e.g. [37–39]) a real scene, as opposed to a screen, is here considered, hence we cannot exploit a calibration-free method based on a saliency map. Once the pupils are estimated, the optical axis is computed as the normal to the plane tangent to the cornea, approximated by a sphere, at the pupil center. The normal however does not correspond to the visual axis. This fact, together with noise, induces an error that increases with distance. To overcome this difficulty a calibration procedure is provided, minimizing this error. The important contribution here is that error minimization is parametric with respect to the distance measured along the optical axis.

In few words, the calibration requires the subject wearing the device to do the following motions in front of a calibration device as illustrated in Fig. 2: go forward toward the device, go backward, turn head up, down, left and right, keeping the eyes fixed to the target. Sequences of calibration performed by four subjects are illustrated in Fig. 3. By approximating the cornea with a sphere we

obtain a function  $f(\mathbf{x})$ , with center  $\mathbf{c}$ . This last is computed exploiting anatomic ratios between the cornea center and the plica and between the plica and the camera centre. The camera centre is given in fact by visual odometry as detailed below. Given  $f(\mathbf{x})$  then the normal  $\mathbf{n} = \nabla f(\mathbf{p})$ , with  $\mathbf{p}$  the pupil center, where the  $z$ -coordinate of  $\mathbf{p}$  is computed exploiting the above reasoning and recalling that  $\mathbf{n}$  pass through  $\mathbf{c}$ . Hence, during the calibration the data

$$U = [(X_1, Y_1, x_1, y_1, d_1, \mathbf{n}_1^\top), \dots, (X_N, Y_N, x_N, y_N, d_N, \mathbf{n}_N^\top)]^\top \in \mathbb{R}^{N \times 8} \quad (2)$$

are collected, with  $X_i, Y_i$  the target location in the scene image,  $x_i, y_i$ , the pupil center in the eye image,  $d_i$  the distance between the camera center and the target plane,  $\mathbf{n}_i$  the normal,  $i = 1, \dots, N$ , along the optical axis.

To obtain the distance  $d_i$  along the normal the target depth needs to be computed. Having the camera already calibrated, an early disparity map is computed by semiglobal matching [40]. Semiglobal matching considers pixelwise matching based on mutual information, and it approximates a global smoothness constraint. With semiglobal matching input images are not required to be rectified. We have used the approach of [41, 42] for GPU implementation of semiglobal stereo matching. To obtain a dense disparity map we apply finally a global image smoothing based on weighted least square [43], following the real time implementation of [44].

Now we have to consider that at training time, each sample from  $U$  has size  $1 \times 8$ , while at test time we are given a measure  $1 \times 6$ , returning the pupil centre and the normal, and that the values  $X, Y$  of the real target in the image, corresponding to a specific eye orientation, we resort to multi task Gaussian process prediction, or to multi-output prediction (for a review see [45]), computing a Gaussian process for a vector valued function.

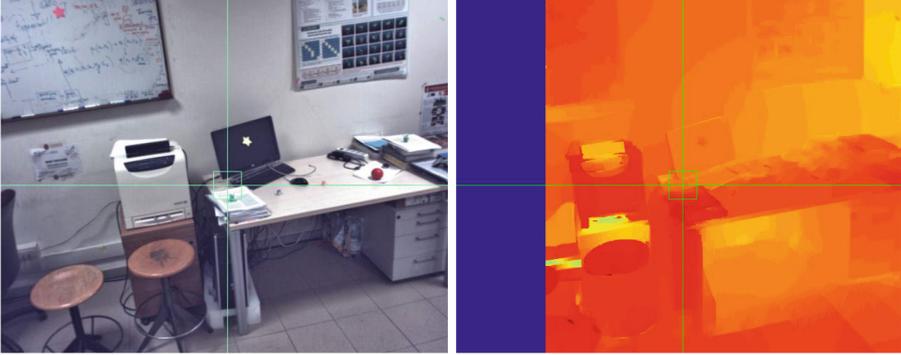
Given the depth map, we set the  $N$  distinct input vectors as  $\mathbf{u}_i = (x_i, y_i, d_i, \mathbf{n}_i^\top)^\top$ ,  $i = 1, \dots, N$  and we set the  $2N$  responses as  $\mathbf{v} = (v_{x1}, \dots, v_{xN}, v_{y1}, \dots, v_{yN})^\top$  where  $v_{ij}$  is the response for task  $i$  to the input  $v_j$ ,  $i \in \{X, Y\}$ . We note that the goal is to compute a latent function for predicting a good estimate for the  $X$  coordinate and a latent function for predicting a good estimate of the  $Y$  coordinate in the image, corresponding to the true reprojection of the observed target in the scene image. Placing a Gaussian process prior over the latent functions  $\mathbf{f} = \{f_x, f_y\}$  a correlation is induced between the two tasks. Assuming zero means we have:

$$\langle f_X(\mathbf{u}) f_Y(\mathbf{u}') \rangle = K_{xy}^{\mathbf{f}} \kappa^{\mathbf{u}}(\mathbf{u}, \mathbf{u}') \text{ with } v_{ix} \sim \mathcal{N}(f_X(\mathbf{u}_i), \sigma_i^2) \quad (3)$$

where  $K^{\mathbf{f}}$  is a positive semidefinite matrix that specifies the intertask correlation,  $\kappa^{\mathbf{u}}$  is a covariance function on the input and  $\sigma_i^2$  is the noise variance of the  $i$ -th task,  $i \in \{X, Y\}$ .

Inference is done using standard Gaussian process model. The mean prediction of a new point  $\mathbf{u}_*$  for task  $X$  is:

$$\begin{aligned} \hat{f}_X(\mathbf{u}_*) &= (\kappa_X^f \otimes \kappa_*^u)^\top \Sigma^{-1} \mathbf{v} \\ \Sigma &= K^f \otimes K^u + D \otimes I \end{aligned} \quad (4)$$



**Fig. 5.** On the left the 2.5POR in the RGB image of the scene, on the right the 2.5POR on the depth map.

Here  $\otimes$  is the Kronecker product,  $\kappa_x^f$  selects the first column of  $K^f$ ,  $\kappa_\star^u$  is the vector of covariances between the point  $u_\star$  and the training points,  $K^u$  is the matrix of covariances between the training data points, and  $D$  is a  $2 \times 2$  matrix in which the element  $(i, i)$ ,  $i \in \{X, Y\}$  is  $\sigma_i$ , and  $\Sigma$  is an  $2N \times 2N$  matrix, we invite the reader to consult [46] for more details, also concerning hyperparameters estimation.

Once both hyperparameters and the matrix  $K^f$  are estimated, then the estimated values return the predicted position of the 2.5D POR (since being located in the image it is located in the depth map).

### 2.3 3D Reconstruction and POR Projection

Given the computed depth maps, the corresponding RGB image, and the location of the POR in the scene image, the 2.5D PORs are mapped from the left to the right image so that by epipolar geometry there is no more ambiguity between the visual axes. However to both localize the scene camera center and the 3D PORs in the images it is necessary to localize the scene camera with respect to the reference frame, hence we need to resort to visual SLAM. Recently [47] has proposed ElasticFusion as dense visual SLAM. We approach the problem of 3D reconstruction of the scene using the ElasticFusion of [47] that is based on GPU programming. The model uses an unordered list of surfels which are descriptors encoding position, normal, color, weight, time stamp update and initial time stamp. Surfels are divided into active and inactive, where the inactive parts are those no more observed. The concept is to verify loop closure when some active part of the model successfully registers with some inactive part. Successful loop closure induces non-rigid deformations of all the surfels, according to a deformation graph. Local loop closure are further incorporated in global loop closures.

The ElasticFusion model is defined over the space of RGB-D images, taking as input a depth and a color map. Therefore having them available we have been

able to integrate the ElasticFusion model into our model, and obtain the dense visual slam, limited by the depth map estimation provided by the stereo pair. The advantage is that for outdoor experiments the stereo rig still provides interesting results for PORs projection, while most RGB-D sensors have some flaws. Furthermore RGB-D sensors at the current time are not enough lightweight for a wearable device designed to study perception.

Having the PORs in the depth maps the dense visual SLAM finally allows to locate the PORs in space, and then projected as spheres with radius inverse proportional to the depth.

### 3 Experiments

The hardware components of the A3D consist of a stereo camera rig, an Arduino unit, an eye camera, a mirror, and an IR LED. These components are supported on an ergonomically designed 3D printed frame (see Fig. 1). The stereo rig is composed by 2 XIMEA cameras (model: MQ013CG-E2) with a baseline equal to 14.4 cm, and it is configured to acquire synchronized image pairs at a frame rate of 25 fps with a resolution of  $640 \times 512$  pixels. To guarantee this frame rate the upper limit for the exposure time is set to 30 ms. This has as a results less bright images at indoor scenes. The Arduino unit (model: Uno R3) is used to provide hardware synchronization to the stereo rig. The eye camera is a XIMEA camera (model: MU9PC-MH) and it is used to capture the motion of the eye. It captures images at a rate of 105 fps. The pupil detection software processes these images at the same frequency, as it needs 9.5 ms to detect the pupil in the eye images. The IR LED is required to facilitate pupil detection since it increases the overall contrast of the eye images, and in particular, the contrast between the pupil and the iris. The mirror is used in order to place the eye camera at the top of the A3D device, avoiding the scene occlusion caused by placing the camera in front of the eyes of the subject.

The main software modules of the A3D device consists of the data acquisition software, the pupil detection software, the depth computation software, and a



**Fig. 6.** Examples of the estimated 2.5D PORs for the first experiment

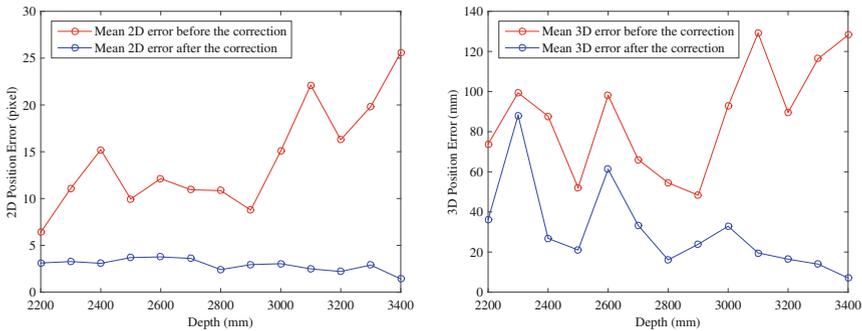
**Table 1.** Depth map average processing processing time for different image sizes

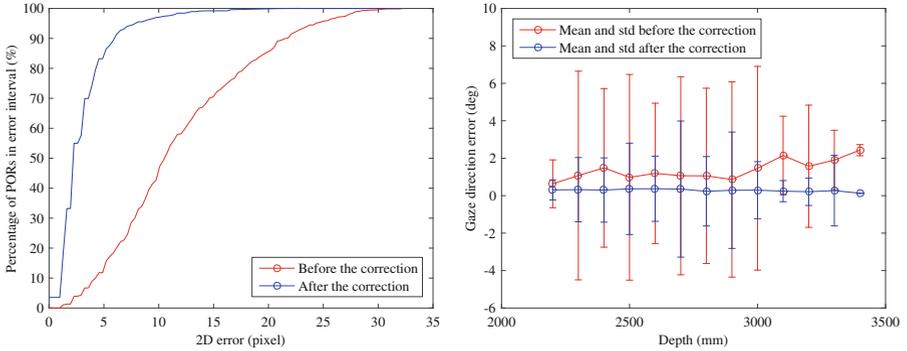
Image size	$2828 \times 1924$	$2262 \times 1539$	$1696 \times 1154$	$1131 \times 769$	$565 \times 384$
Processing time (ms)	366	284	118	54	9

modified version of ElasticFusion [47]. The acquisition software, implemented in C++, manages the acquisition and rectification of the stereo images, the acquisition of the eye images, and the 2.5D POR estimation based on the calibration procedure. The depth estimation software consists of 2 parts: one based on the SGBM algorithm [40] and the other on the WLS filter [43]. Both are implemented in GPU in order to allow for real-time computation of the depth maps from the stereo image pairs. Finally, the modified version of ElasticFusion [47], also implemented in GPU, exploits here the computation of the depth map in place of the RGB-D sensor, which cannot be mounted on the device, and it is used to compute a 3D reconstruction of the scene and the projection of 3D PORs from the depth maps to the reconstructed scene.

### 3.1 Time Measurements

We report now the time performance for each software part of the A3D device. For each part we process 1000 frames and compute the average processing time. Pupil detection requires 9.5 ms on a region of interest with dimensions  $180 \times 120$  pixels, while the depth estimation needs 25 ms for images of resolution  $640 \times 512$  pixels. ElasticFusion and 3D POR estimation need 75 ms on the image resolution  $640 \times 512$  pixels. Table 1 shows the average time for the depth map computation for different image resolutions. The computation time is computed using images taken from the Middlebury dataset [48], resized according to the different sizes required by the above measurements.

**Fig. 7.** 2.5D (left) and 3D (right) POR error before and after correction



**Fig. 8.** On the left: cumulative error distribution for 2.5D PORs. On the right: angular error in gaze direction before and after correction.

### 3.2 Error Evaluation in Indoor Environments

In this section we describe a number of experiments that have been performed by 4 subjects. Note that PORs that are computed on the depth image are illustrated also on the 2D images.

The first experiment involves searching interest points in an indoor environment. Interest points taken in consideration are star-shaped stickers located within a workroom on walls, desks, devices or on the floor. The scene is considered static, so in the room there are no moving objects/persons.

The experiment is performed by a person wearing the A3D device and consists in performing first the calibration procedure described in Sect. 2, and then in a search task, which amounts to find all the stickers also moving inside the room. Some examples of scene images with overlaid POR visualizations are shown in Fig. 6.

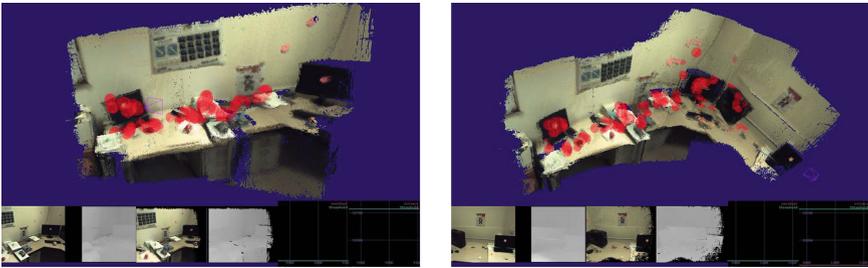
In Fig. 7 the corresponding 2.5D and 3D errors are shown with respect to the depth of the interest points. The errors are considered to be the mean distances between estimated and ground truth PORs. In each plot the mean errors before and after the 2.5D and 3D corrections are shown, as described in Sect. 2. The error graphs point out the great improvement of the correction phase refining both 2.5D and 3D POR position estimation. The smaller improvement caused by the correction in the 3D case is due to occlusions and the great variance of 3D position estimation near edges.

In Fig. 8 the cumulative error distribution of the 2.5D PORs is shown. The area under the curve corresponding to the 2.5D PORs after the correction is much larger with respect to the curve obtained from the 2.5D PORs before the correction. This is representative of the large improvement achieved by the correction process.

Finally, for further evaluation we computed the error of the gaze direction corresponding to the estimated PORs before and after the correction. Denoting as  $\mathbf{K}$  the camera calibration matrix and  $\mathbf{p}$  the 2.5D POR position in homogeneous



**Fig. 9.** Examples of the estimated PORs for the second experiment.



**Fig. 10.** Examples of the reconstructed 3D scene during the second experiment showing projected PORs as red spheres. (Color figure online)

coordinates, the direction of the corresponding viewing ray is computed according to (see [49] for details):

$$\mathbf{d} = \mathbf{K}^{-1}\mathbf{p}. \tag{5}$$

Based on this, the angular error between the estimated and the ground truth PORs is given by

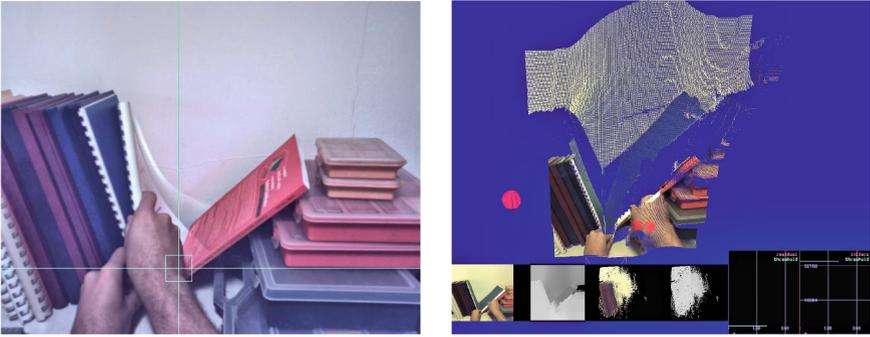
$$\theta = \arccos \frac{\mathbf{d}_{est}\mathbf{d}_{gt}}{\|\mathbf{d}_{est}\|\|\mathbf{d}_{gt}\|} \tag{6}$$

Here too the experiments show the improvements caused by the correction phase, which has as a result the decrease of both the mean and the standard deviation of the gaze direction error.

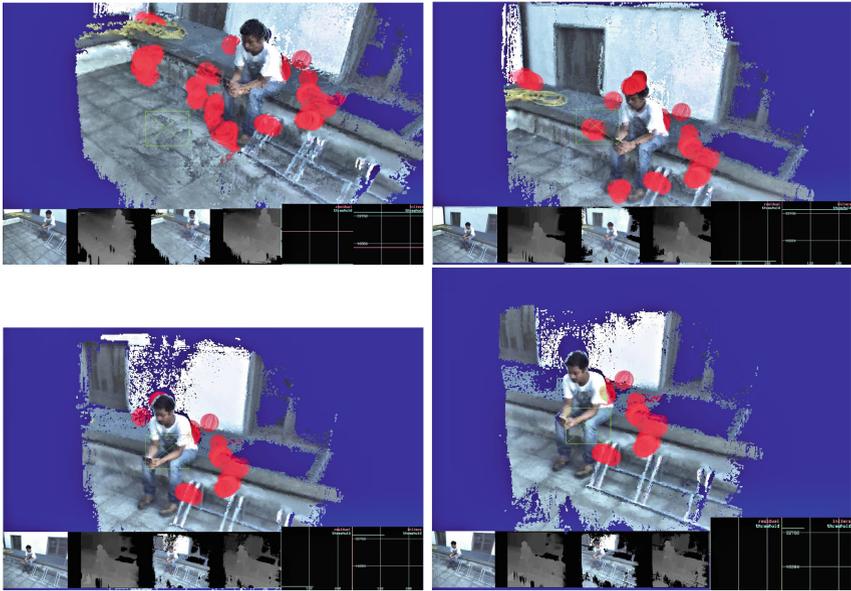
### 3.3 Qualitative Analysis in Indoor Environments

The second experiment is a qualitative study involving predefined objects search in the same indoor environment. The scene is considered to be static and the predefined objects are small plastic toys.

During the experiment subjects wearing the A3D have to look for several unknown toys inside the room. Some examples of the 2D images of the scene



**Fig. 11.** Examples of the estimated PORs (left) and reconstructed 3D scene (right) during the third experiment.



**Fig. 12.** 3D PORs in an outdoor sequence, with reconstruction and 3D PORs projection.

with PORs visualization are shown in Fig. 9. In Fig. 10 some examples of the 3D reconstructed scene are shown with the relative 3D PORs.

The third experiment is a qualitative study to visualize the gaze projection in slow motion in close range and dynamic indoor environments. During the experiment the subjects wearing the A3D have to pick up a book from the books on the shelf, which will cause changes in the environment and make it dynamic. Some examples of the 2D image of the scene with PORs visualization and the 3D reconstructed scene with relative reprojected PORs are shown in Fig. 11.



Fig. 13. 3D PORs in an outdoor sequence, observing two people talking.

### 3.4 Qualitative Analysis in Outdoor Environments

Finally a number of experiments have been performed for a qualitative study about focus on subjects in outdoor environments. The experiments consist first in a calibration procedure described in Sect. 2 and then the subjects wearing the A3D observe the people in the scene, moving around them. Some of the experiments are illustrated in Figs. 12 and 13. It is interesting to observe the projected PORs, in Fig. 13 as the subject is focusing on the person on the left, who is talking.

## 4 Conclusions

In this paper, we propose A3D, a head-mounted device for gaze estimation both in indoor and outdoor environments. By a suitable calibration procedure we can estimate the PORs on the depth images, hence in 2.5D, using the depth maps we exploited [47] ElasticFusion to both obtain the 3D reconstructed scene and 3D position of PORs in the scene.

Experiments on error analysis of the taken measures, together with cumulative error distribution and gaze direction errors studies, demonstrate the reliability of the introduced correction step.

The proposed system works in quasi-real-time, but limitations such as smooth head motion restrict its applications.

Future work include a much more reliable 3D reconstruction procedure to allow for fast head motions, and the possibility to extend the work to cover dynamic scenarios.

**Acknowledgment.** Supported by EU FP7 TRADR (609763) and EU H2020 SecondHands (643950) projects. The authors thank the anonymous reviewers for their comments.

## References

1. Ji, Q., Wechsler, H., Duchowski, A.T., Flickner, M.: Special issue: eye detection and tracking. *Comput. Vis. Image Underst.* **98**(1), 1–3 (2005)
2. Hansen, D.W., Ji, Q.: In the eye of the beholder: a survey of models for eyes and gaze. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(3), 478–500 (2010)
3. Fuhl, W., Tonsen, M., Bulling, A., Kasneci, E.: Pupil detection for head-mounted eye tracking in the wild: an evaluation of the state of the art. *Mach. Vis. Appl.* 1–14 (2016). doi:[10.1007/s00138-016-0776-4](https://doi.org/10.1007/s00138-016-0776-4)
4. Tsotsos, J., Culhane, S., Wai, W., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. *Artif. Intell.* **78**, 507–547 (1995)
5. Ahlstrom, C., Victor, T., Wege, C., Steinmetz, E.: Processing of eye/head-tracking data in large-scale naturalistic driving data sets. *IEEE Trans. Intell. Transp. Syst.* **13**(2), 553–564 (2012)
6. Kasneci, E., Sippel, K., Aehling, K., Heister, M., Rosenstiel, W., Schiefer, U., Papageorgiou, E.: Driving with binocular visual field loss? A study on a supervised on-road parcours with simultaneous eye and head tracking. *PloS One* **9**(2), e87470 (2014)

7. Lai, M.L., Tsai, M.J., Yang, F.Y., Hsu, C.Y., Liu, T.C., Lee, S.W.Y., Lee, M.H., Chiou, G.L., Liang, J.C., Tsai, C.C.: A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educ. Res. Rev.* **10**, 90–115 (2013)
8. Wedel, M.: Attention research in marketing: a review of eye tracking studies. Robert H. Smith School Research Paper No. RHS 2460289 (2013)
9. Rosch, J.L., Vogel-Walcutt, J.J.: A review of eye-tracking applications as tools for training. *Cogn. Technol. Work* **15**(3), 313–327 (2013)
10. Alletto, S., Abati, D., Serra, G., Cucchiara, R.: Wearable vision for retrieving architectural details in augmented tourist experiences. In: 2015 7th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN), pp. 134–139 (2015)
11. Jacob, R., Karn, K.S.: Eye tracking in human-computer interaction and usability research: ready to deliver the promises. *Mind* **2**(3), 4 (2003)
12. Jaimes, A., Sebe, N.: Multimodal human-computer interaction: a survey. *Comput. Vis. Image Underst.* **108**(1), 116–134 (2007)
13. Treisman, A., Gelade, G.: A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980)
14. Koch, C., Ullman, S.: Shifts in selective visual-attention: towards the underlying neural circuitry. *Hum. Neurobiol.* **4**(4), 219–227 (1985)
15. Treisman, A.: Preattentive processing in vision. *Comput. Vis. Graph. Image Process.* **31**(2), 156–177 (1985)
16. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI* **20**(11), 1254–1259 (1998)
17. Carmi, R., Itti, L.: Visual causes versus correlates of attentional selection in dynamic scenes. *Vision. Res.* **46**(26), 4333–4345 (2006)
18. Duncan, J., Humphreys, G.W.: Visual search and stimulus similarity. *Psychol. Rev.* **96**(3), 433–458 (1989)
19. Pessoa, L., Exel, S.: Attentional strategies for object recognition. In: Mira, J., Sánchez-Andrés, J.V. (eds.) *IWANN 1999*. LNCS, vol. 1606, pp. 850–859. Springer, Heidelberg (1999). doi:[10.1007/BFb0098243](https://doi.org/10.1007/BFb0098243)
20. Beymer, D., Flickner, M.: Eye gaze tracking using an active stereo head. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings, vol. 2, pp. 451–458 (2003)
21. Zhu, Z., Ji, Q.: Novel eye gaze tracking techniques under natural head movement. *IEEE Trans. Biomed. Eng.* **54**(12), 2246–2260 (2007)
22. Munn, S.M., Pelz, J.B.: 3D point-of-regard, position and head orientation from a portable monocular video-based eye tracker. In: Proceedings of the 2008 Symposium on Eye Tracking Research and Applications, pp. 181–188. ACM (2008)
23. Pirri, F., Pizzoli, M., Rudi, A.: A general method for the point of regard estimation in 3D space. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 921–928. IEEE (2011)
24. Bulling, A.: Pervasive attentive user interfaces. *IEEE Computer* **49**(1), 94–98 (2016)
25. Pfeiffer, T., Renner, P., Pfeiffer-Leßmann, N.: EyeSee3D 2.0: model-based real-time analysis of mobile eye-tracking in static and dynamic three-dimensional scenes. In: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research and Applications, pp. 189–196. ACM (2016)
26. Paletta, L., Santner, K., Fritz, G., Mayer, H., Schrammel, J.: 3D attention: measurement of visual saliency using eye tracking glasses. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp. 199–204. ACM (2013)

27. Pirker, K., R  ther, M., Schweighofer, G., Bischof, H.: GPSlam: marrying sparse geometric and dense probabilistic visual mapping. In: BMVC, pp. 1–12 (2011)
28. Posner, M.I., Rafal, R.D., Choate, L.S., Vaughan, J.: Inhibition of return: neural basis and function review. *Cogn. Neuropsychol.* **2**, 211–228 (1985)
29. Ntouskos, V., Pirri, F., Pizzoli, M., Sinha, A., Cafaro, B.: Saliency prediction in the coherence theory of attention. *Biol. Inspir. Cogn. Archit.* **5**, 10–28 (2013)
30. Świrski, L., Bulling, A., Dodgson, N.: Robust real-time pupil tracking in highly off-axis images. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 173–176. ACM (2012)
31. Kassner, M., Patera, W., Bulling, A.: Pupil: an open source platform for pervasive eye tracking and mobile, gaze-based interaction. CoRR abs/1405.0006 (2014)
32. Hennessey, C., Nouredin, B., Lawrence, P.: A single camera eye-gaze tracking system with free head motion. In: Proceedings of the 2006 Symposium on Eye Tracking Research and Applications, pp. 87–94. ACM (2006)
33. Ariz, M., Bengoechea, J.J., Villanueva, A., Cabeza, R.: A novel 2D/3D database with automatic face annotation for head tracking and pose estimation. *Comput. Vis. Image Underst.* **148**, 201–210 (2016)
34. Carpenter, R.H.: *Movements of the Eyes*, 2nd edn. Pion Limited, London (1988)
35. Williams, C.K., Rasmussen, C.E.: *Gaussian processes for regression* (1996)
36. Nickisch, H., Rasmussen, C.E.: Approximations for binary gaussian process classification. *J. Mach. Learn. Res.* **9**, 2035–2078 (2008)
37. Sugano, Y., Matsushita, Y., Sato, Y.: Appearance-based gaze estimation using visual saliency. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(2), 329–341 (2013)
38. Sugano, Y., Matsushita, Y., Sato, Y.: Calibration-free gaze sensing using saliency maps. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2667–2674. IEEE (2010)
39. Alnajar, F., Gevers, T., Valenti, R., Ghebreab, S.: Calibration-free gaze estimation using human gaze patterns. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 137–144 (2013)
40. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(2), 328–341 (2008)
41. Zhu, K., Butenuth, M., d’Angelo, P.: Comparison of dense stereo using CUDA. In: Kutulakos, K.N. (ed.) ECCV 2010. LNCS, vol. 6554, pp. 398–410. Springer, Heidelberg (2012). doi:[10.1007/978-3-642-35740-4\\_31](https://doi.org/10.1007/978-3-642-35740-4_31)
42. Zhu, K., Butenuth, M., d’Angelo, P.: Efficient dense stereo matching using CUDA. Technical report, TUM (2013)
43. Farbman, Z., Fattal, R., Lischinski, D., Szeliski, R.: Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Trans. Graph. (TOG)* **27**, 1–67 (2008). ACM
44. Li, Q., Zhao, H.: Real-time implementation for weighted-least-squares-based edge-preserving decomposition and its applications. In: Pan, Z., Cheok, A.D., M  ller, W. (eds.) Transactions on Edutainment VI. LNCS, vol. 6758, pp. 256–263. Springer, Heidelberg (2011). doi:[10.1007/978-3-642-22639-7\\_25](https://doi.org/10.1007/978-3-642-22639-7_25)
45. Alvarez, M.A., Rosasco, L., Lawrence, N.D.: Kernels for vector-valued functions: a review. *Mach. Learn.* **4**(3), 195–266 (2011)
46. Bonilla, E.V., Chai, K.M., Williams, C.: Multi-task Gaussian process prediction. In: Advances in Neural Information Processing Systems, NIPS, pp. 153–160 (2007)
47. Whelan, T., Leutenegger, S., Salas-Moreno, R.F., Glocker, B., Davison, A.J.: ElasticFusion: dense SLAM without a pose graph. In: Robotics: Science and Systems (RSS), Rome, Italy, July 2015

48. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: Jiang, X., Hornegger, J., Koch, R. (eds.) GCPR 2014. LNCS, vol. 8753, pp. 31–42. Springer, Heidelberg (2014). doi:[10.1007/978-3-319-11752-2\\_3](https://doi.org/10.1007/978-3-319-11752-2_3)
49. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2003)