A general method for the Point of Regard estimation in 3D space

Fiora Pirri
pirri@dis.uniroma1.it

Matia Pizzoli
pizzoli@dis.uniroma1.it

Alessandro Rudi
alessandro.rudi@gmail.com

Abstract

A novel approach to 3D gaze estimation for wearable multi-camera devices is proposed and its effectiveness is demonstrated both theoretically and empirically. The proposed approach, firmly grounded on the geometry of the multiple views, introduces a calibration procedure that is efficient, accurate, highly innovative but also practical and easy. Thus, it can run online with little intervention from the user. The overall gaze estimation model is general, as no particular complex model of the human eye is assumed in this work. This is made possible by a novel approach, that can be sketched as follows: each eye is imaged by a camera; two conics are fitted to the imaged pupils and a calibration sequence, consisting in the subject gazing a known 3D point, while moving his/her head, provides information to 1) estimate the optical axis in 3D world; 2) compute the geometry of the multi-camera system; 3) estimate the Point of Regard in 3D world. The resultant model is being used effectively to study visual attention by means of gaze estimation experiments, involving people performing natural tasks in wide-field, unstructured scenarios.

1. Introduction

Gaze localisation in space is one of the most hot and challenging topic in cognitive vision and in general in vision research as it attacks important issues such as attention, feature detection, visual localisation, human-machine communication, and most of the factors concerning the human response to surrounding visual stimuli. Several issues related to these aspects of cognitive vision can be experimentally corroborated if data collected from human gaze are made available, under different environment conditions.

A recent survey of [6] and a whole issue of [10] witness the wealth of research that has been covered on this topic in the last thirty years and the progress accomplished, although many problems are still open.

In this paper we present an innovative model for 3D gaze localisation of a subject freely moving in a dynamic environment and under varying light conditions. The main contributions of our work are twofold. On the one hand we present a mathematical model underpinning the full calibration and localisation of both eyes and cameras. On the other hand we provide a full implementation of the system and a demonstration of its effectiveness in different contexts and environments. The model is firmly grounded in the geometry of the multiple views. The correctness of our approach, under specific hypotheses, is proved and also supported by several experiments, some of which we present in these pages.

The goal of gaze localisation in space is to predict the position of eye fixations in 3D space, on the basis of the eyes visual axes estimation, and the calibration of the cameras system, simultaneously acquiring eyes and scene data. Crucial issues are localisation correctness confined, by foveated vision, to approximately one degree around the fixation point, and completeness. These include the ability to accommodate diversity of observers eyes, head and body motion and light change. Likewise time-space completeness includes the compatibility of frequency of acquisition with the receptive fields response and the coverage of the observer field of view. In particular, the percentage of fixations captured should cope with a natural time delay for a fixation such as 100ms and change blindness [13, 16]. In this paper we report on the essential aspects of the gaze localisation model and, in the experiments section, we show that our system correctly estimates the observed point, up to an error that we quantify. Indeed, an efficient, accurate, highly innovative but also practical and easy calibration procedure is introduced.

Most 3D, model-based approaches [4, 9, 15, 19, 21, 23] rely on estimating the parameters of some model of the human eye. Due to the eye complexity and the great number of axes (see Figure 1) given by a detailed model, this often requires some simplification assumption, like spherical cornea or coincidence of the axes. Moreover, it is common practice to estimate the cornea centre as if it were the intersection of optical and visual axes. As a consequence the calibration point appears enough to estimate the visual axes. Despite most of the existing systems rely on these kind of approximations, it is well known that the visual and optical axes intersect at the nodal point [2](see Figure 1).

On the other hand taking care of a detailed model of
the eye requires a sophisticated setting with multiple light sources or cameras. Both these methods are not suitable for operating in large, outdoor, unconstrained scenarios. Indeed, to overcome the limitations related to special assumptions on the eyeball or cornea shape and especially to cope with the eye dynamics, we decoupled the problem of Gaze estimation into two separated steps. In the first step we estimate the 3D pupil axes given the left and right pupils, in the second step we estimate the 3D point of regard (POR) given the 3D pupil axes. The problem of the pupil axis estimation is clearly specified by the information from the pupils imaged on the image plane, in such a way so as to avoid an explicit model of the eye morphology and dynamics, and it is solved geometrically. The key ingredient in the formalisation of the pupil axis estimation is the locus of the pupil centres. The 3D POR is estimated using the optical axes, namely the axes orthogonal to the corneal edge [2], which is a point on the locus of pupil centres, and the relative positions of the cameras and the loci, in space. We shall not illustrate here how the visual axes are obtained from the optical ones, as the process deserves a description by its own. We address an overview of the general geometric settings and of the paper organisation in Section 2.

2. Overview of the Geometric Model

The gaze estimation model is formed by four projective cameras, \( P_{el}, P_{er} \) looking at the eyes, \( P_{wl}, P_{wr} \) looking at the world, and two differentiable manifolds \( \Gamma_l \) and \( \Gamma_r \) which are the pupils envelopes for the left and right eyes (as shown in Figure 2).

The manifold \( \Gamma \) is the geometrical locus of the pupil centres, obtained by the dynamics of the eye. This manifold is a surface on which a pupil centre lies and for each pupil there exists a plane tangent to \( \Gamma \) precisely at the pupil centre. This is the pupil plane \( \pi \). We represent the pupil as a circle of free ray, whose centre is indeed the pupil centre, denoted by \( \psi \). The circle lies on the pupil plane \( \pi \) hence, under this assumption, the pupil observed by the camera \( P_{ex}, x = l, r \), is imaged as a conic \( C \) on the image plane. A pupil is, thus, represented in space by the pupil representation \( \Theta(\lambda) = (\psi(\lambda), \pi(\lambda)) \). The problem of pupil axis estimation, given the conic \( C \) and the manifold \( \Gamma \), is to find the pupil representation \( \Theta \) compatible with \( C \), having \( \pi \) tangent to \( \Gamma \) in \( \psi \). Indeed, according to [2] the optical axis is precisely the normal to \( \pi \) in \( \psi \) (see Figure 1).

In the following section we show how to estimate the pupil representations \( \Theta \), in space, compatible with the observed imaged pupil \( C \). These will be used to estimate the manifold \( \Gamma \). The estimation of the manifold, whose algebraic characterisation (specified in Section 4) allows to identify an initial quadric approximation, is performed by an error estimation and minimisation on the set of observed imaged pupils \( C \). The error estimation is illustrated in Section 5.

The estimation of the relative positions of the cameras and the loci in space is addressed in Section 7, once the optical axes and their relation with the visual axes is briefly addressed in Section 6. Finally, in Section 8 some experiments are presented to prove the correctness of our implemented model, under challenging conditions, called the Gaze Machine. The contribution is resumed in the conclusions.

3. Pupil representation in space

Let \( C \) be the conic of the imaged pupil, in this section we show how to obtain from \( C \) the set of compatible pupil representations \( \Theta(\lambda) = (\psi(\lambda), \pi(\lambda)) \) in space, specified by a line \( \psi(\lambda) \) and a family of planes \( \pi(\lambda) \), with \( \lambda \) a scale factor. This representation brings an ambiguity that we further solve in Section 6.

Let \( P_{ex} = K[I, 0], x = l, r \), be the known camera looking at the eye, that in this section we abbreviate with \( P \). The problem of identifying both the circle, its centre, and its supporting plane which, projected by \( P \), forms the given conic \( C \), can be stated as follows: let \( P' \) be the virtual projective
camera, centred in the centre of the camera $P$, having image axes orthogonal and isometric, and projecting the pupil as a circle centred on the $x$-axis. Under these conditions the pupil plane would be parallel to the principal plane of $P'$. Consider the cone formed by the back-projection of $C$ through $P$, this cone is the same cone formed by the back-projection of the pupil circle through $P'$. Let $Z$ be the pupil circle, the problem can be stated as follows:

$$P^T C P = P'^T Z P'. \tag{1}$$

Here $P'$ is of the form $P' = [R|0]$, to ensure orthogonal and isometric image axis, with $R$ a rotation matrix; while $Z$ is a circle whose centre is on the $x$-axis:

$$Z = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & -y \\ 0 & -y & \beta \end{pmatrix}. \tag{2}$$

Therefore we can rewrite (1) as

$$R^T W R = Z \tag{3}$$

where $W = K^T C K$. It follows that $W$ and $Z$ are similar matrices on the same field, and they have the same characteristic polynomial, namely $\text{det}(\lambda I - W) = \text{det}(\lambda I - Z)$. Hence we can use the fact that $W$ is known, together with the above specified similitude, to find first the unknown parameters of $Z$ and further the rotation $R$. Let $W = U \Lambda U^T$, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$ the eigenvalue matrix with $\lambda_1 \geq \lambda_2 > 0$ and $\lambda_3 < 0$. The equation of the two characteristic polynomials has six solutions. Only the following two solutions are admissible, under the constraint that $Z$ is a proper circle: $\{\alpha_i = \lambda_2, \beta_i = \lambda_1 - \lambda_2 + \lambda_3, y_i = \pm s\}$ where $i = 1, 2$ and $s = \sqrt{\lambda_1 \lambda_2 - \lambda_1 \lambda_3 + \lambda_2 \lambda_3 - \lambda_2^2}$.

It follows that there are two possible circles which, once projected, generate the given conic. Now, the two circles $Z_i$ are given by the obtained parameters $\alpha, \beta$ and $y_i, i = 1, 2$, then, by the eigen decomposition $H_i \Lambda H_i^T$ of $Z_i$, equation (3), becomes:

$$R_i U \Lambda U^T R_i^T = H_i \Lambda H_i^T \tag{4}$$

whence $R_i = H_i U^T$.

The pupil planes compatible with the given conic $C$ are two families of planes parallel to the principal planes of the two camera matrices $P_i' = [R_i|0]$ with $i = 1, 2$. The two families are

$$\pi_i(\lambda) = (H_i^3 U^T \lambda)^T \tag{5}$$

here $H_3^3$ is the third row of $H_i$, $i = 1, 2$ and $\lambda$ a scalar.

Finally, the centres $c_i$ of the circles $Z_i, i = 1, 2$, are those point satisfying $Z_i c_i = [0 \ 0 \ 1]^T$, as the polar line related to the centre of a circle is the line at infinity. Now, given a scalar $\lambda$, an $i$ and a pupil plane $\pi_i(\lambda)$, the 3D point $\psi_i(\lambda)$ is the back-projection of $c_i$ lying on $\pi_i(\lambda)$. Whence:

$$\begin{cases} \psi_i(\lambda) = P_i^T c_i + [0 \ 0 \ 0 \ \gamma]^T \\ \psi_i(\lambda)^T \pi_i(\lambda) = 0 \end{cases} \tag{6}$$

Here $P_i^T = P_i^T (P_i^T P_i)^{-1}$ is the pseudo inverse of $P_i$. Now, eliminating the scale factor $\gamma$ in (6) we are left with:

$$\psi_i(\lambda) = (-\lambda U H_i^T c_i \ 1) \tag{7}$$

Here $c_i$ is the vector $c_i$ normalised by its third element, namely $c_i = c_i/\lambda c_i^3$. Therefore two families of pupil representation in space are associated to each conic $C$ specifying a pupil: $\Theta_i(\lambda) = (\psi_i(\lambda), \pi_i(\lambda)), i = 1, 2$, with $\psi_i(\lambda), \pi_i(\lambda)$ as defined by equations (5) and (7), (see Figure 3).

4. Estimation of the eye manifold: the initial quadric

Given the pupil representation $\Theta_i(\lambda) = (\psi_i(\lambda), \pi_i(\lambda))$, with $\lambda$ a free scale factor, we show how to estimate the eye manifold $\Gamma$ into two steps. In the first step, described in this section, a set of candidate quadrics is obtained algebraically from a suitable set of pupils. This candidate set of quadrics is designated to be an initial approximation of the true envelope $\Gamma$.

Let $\lambda$, possibly with subscript, be a scale parameter and let $C^h$ be any conic of an imaged pupil. As shown in the previous section there are two representations in space for each imaged pupil, namely, $\Theta_i^1(\lambda) = (\psi^1_i(\lambda), \pi^1_i(\lambda))$ and $\Theta_i^2(\lambda) = (\psi^2_i(\lambda), \pi^2_i(\lambda))$. Although only four $\Theta^h$ are needed to determine a quadric, taking any possible combination of four $\Theta^h$ ensures that precisely one is specified by the correct representation set.

Let us given four conics $\{C^1, \ldots, C^4\}$ and let $S^j$ be the $j$-th set of the form $\{\Theta^1_i, \ldots, \Theta^k_i\}$, with $i_u \in \{1, 2\}, u = 1, \ldots, 4$ and $j = 1, \ldots, 16$. Each $S^j$ enumerates the four pupil representations in space, induced by the chosen conics, as specified in the previous section, and for each $j$-th of them we obtain a quadric $Q^0_i$. All the quadrics $Q^0_i$ are candidates to be a good initial approximation of the real manifold, although only one is induced be the true representations of
the pupils. In the following we show how to obtain each candidate \( Q_0 \) from its set of four pupil representations \( S^3 \).

Given for some \( j \) a set \( S \), first observe that any pupil plane must be tangent to the associated \( Q_0 \), precisely at the pupil centre. Therefore, given a family of free scale parameters \( \mu_u \) and \( \lambda_u \), with \( |\lambda_u|, |\mu_u| \in (0, +\infty) \), the following equations tell us that the quadric \(Q_0\) has to satisfy the condition that one plane from the \( u \)-th family must be tangent to it precisely at a point of the family:

\[
\begin{align*}
\psi_u(\lambda_u)\mathbf{Q}_0\psi_u(\lambda_u) &= 0 \\
\pi_u(\lambda_u) &= \mu_u Q_0 \psi_u(\lambda_u) \\
\psi_u(\lambda_u)\pi_u(\lambda_u) &= 0 \\
\mu_u \pi_u(\lambda_u) &= Q_0 \psi_u(\lambda_u) \\
\end{align*}
\]

Note that the first equation in the system (8) is tautological, hence the above mentioned conditions can be simplified into the equation

\[
Q_0 H = K \quad \text{where} \\
H &= [\psi_1(\lambda_1) \psi_2(\lambda_2) \psi_3(\lambda_3) \psi_4(\lambda_4)] \quad \text{and} \\
K &= [\mu_1 \pi_1(\lambda_1) \mu_2 \pi_2(\lambda_2) \mu_3 \pi_3(\lambda_3) \mu_4 \pi_4(\lambda_4)]
\]

Furthermore, \( Q_0 \) must be a symmetric matrix of rank 4, in order to be a non-degenerate quadric. This means that \( H \) and \( K \) have to be, in turn, full rank matrices. So from the rank condition \( Q_0 = KH^{-1} \) and from the symmetry condition \( KH^{-1} = H^{-1} K^\top \) it follows that:

\[
H^\top K - K^\top H = 0
\]

(9)

Noting that \( \pi_u(\lambda_u) = [\Pi_u \lambda_u] \) and \( \psi_u(\lambda_u) = [\lambda_u x_u] \) we introduce \( \mathbf{P} = [\Pi_1 \ldots \Pi_4] \) and \( \mathbf{X} = [x_1 \ldots x_4] \) the related points of tangency. Let \( L = \operatorname{diag}(\lambda_1 \ldots \lambda_4) \), \( M = \operatorname{diag}(\mu_1 \ldots \mu_4) \), and let \( I = (1, 1, 1, 1) \). Then equation (10) can be specified as

\[
L \mathbf{X}^\top \mathbf{P} \mathbf{M} - \mathbf{M} \mathbf{P}^\top \mathbf{X} L = LM \mathbf{1}^\top 11^\top - 11^\top ML
\]

(11)

which corresponds to the system

\[
\{ \lambda_j \mu_i a_{ij} - \lambda_i \mu_j a_{ji} + \lambda_i \mu_i - \lambda_j \mu_j = 0 \quad \forall 0 < i < j \leq 4 \}
\]

(12)

Here the \( a_{ij} \) are the \( ij \)-th elements of the matrix \( A \), with \( A = \mathbf{X}^\top \mathbf{P} \). Finally, in order to make the system inhomogeneous we set \( \lambda_1, \mu_1 = 1 \). The polynomial system (12) of six equations in six unknowns is solved with the linear-algebra method of the multivariate resultant, we refer the reader to [20] for the details.

5. Error measures, estimation and minimisation

Up to this point we have shown how to compute the pupil representation in space, namely \( \Theta_i(\lambda) = (\psi_i(\lambda), \pi_i(\lambda)) \), \( i = 1, 2 \), and an initial estimation set of \( j = 1, \ldots, 16 \) candidates \( Q_0 \) of the manifold \( \Gamma \) for the eye envelope. We have, indeed, accumulated the ambiguity in the pupil representation in space affecting also the initial estimation of the pupil envelope. Given the observations, namely the conics \( C \) imaging the pupils on the image plane, these ambiguities can be resolved using the candidate set \( \{Q_0^n\} \) provided in the previous section. In this section we introduce two error measures that are strictly interlaced and illustrate the estimation and minimisation steps with two algorithms. The first measure concerns the pupil representation and the second the initial quadric estimation, given the pupil representation. More precisely, let \( x \) be a point on \( \Gamma \) specifying the true pupil location in space, corresponding to \( C \), and let \( \Theta_i(\lambda) = (\psi(\lambda), \pi(\lambda)) \) be the induced pupil representation (as specified in Section 3). Now, let \( l \) be a line joining the camera centre and the point \( x \) and let \( \Pi \) be the plane tangent to \( \Gamma \) at \( x \), namely \( \Pi = \nabla_x \Gamma \). There is an error between the correct representation given by \( (x, \Pi) \) and the computed representation \( (\psi(\lambda), \pi(\lambda)) \), for some \( \lambda \). This error can be defined in terms of two angles, \( \theta = \angle(\psi(\lambda), l) \) and \( \varphi = \angle(\pi, \Pi) \) as follows:

\[
E(x, \Gamma, \Theta) = 1 - \cos^2 \theta \cos^2 \varphi
\]

(13)

The angles interpret how distant is the estimation from the true representation. We can note that the above error function is invariant to similarity transformations. The angle between two intersecting lines is the same angle of the related normal planes and, furthermore, as the camera centre is the world origin, a plane normal to \( l \) is \( \hat{I} x \) and a plane normal to \( \psi \) is \( \hat{I} \psi \), where \( \hat{I} \) is the canonic absolute dual quadric, namely \( \hat{I} = \operatorname{diag}(1, 1, 1, 0) \). Thus, from 3D projective geometry, given that \( \Pi \) is the plane tangent to \( \Gamma \) at \( x \), it follows that:

\[
\begin{align*}
\cos^2 \theta &= (\psi_1 \hat{I} x)^2 (\psi_1 \hat{I} \psi x^\top \hat{I} x)^{-1} \\
\cos^2 \varphi &= (\pi \hat{I} \Pi)^2 (\pi \hat{I} \pi \Pi ^\top \Pi)^{-1}
\end{align*}
\]

(14)

We define the point \( \hat{I} \ell \) at the intersection between the envelope \( \Gamma \) and the line \( l \) specified above. If there are several intersection points on \( \Gamma \) we can choose the one closest to the image plane and in front of the camera. Given the above definitions we introduce the following error measure, independent of the ambiguity on \( \Theta \):

\[
\text{err}(\hat{Q}_0|C^h) = \min\{E(Q_0, \psi_1^h, Q_0, \Theta_1^h), E(Q_0, \psi_2^h, Q_0, \Theta_2^h)\}
\]

(15)

here \( E(\cdot) \) is the error function defined in equation (13). Indeed, for any candidate \( \hat{Q}_0 \) the mean of the \text{err} function is computed on a set of \( n \) conics. The good initial quadric is the one for which the mean error, on these \( n \) conics, is significantly smaller than the mean error of any other quadrics. The mean error gives a good estimation of the initial quadric because the envelope is effectively very well approximated by a quadric and because the amount of outliers relative to the measured conics is actually small.
4. Lines (4-9) perform the minimisation to remove the ambiguity, which is removed in Section 7.

6. Optical axes

Given a point \( x \) on the envelope \( \Gamma \), the pupil axis, namely the optical axis, is, therefore, the line \( l(\lambda) = x + \lambda \Pi \), with \( I = \text{diag}(1, 1, 1, 0) \). In Plücker coordinates the line is defined as \( L = x \Pi^\top I - \Pi x^\top \).

Given the left and right pupil axes the 3D Point of Regard is estimated. We express the two axes in the common coordinate frame of the world cameras, this is specified in the following Section 7. Then, assuming that there is a functional dependence of the left and right eye torsion with respect to both the pupil axes, we have estimated the POR by a suitable learning method not described here. We just note that the assumption of the functional dependence is far more general than the generalised Listing Law which is commonly employed in several Gaze Trackers as reported in [5].

7. The POR: estimating the camera poses

In order to estimate the 3D location of the gazed point the relative pose of each camera in the system must be recovered. We introduce a camera pose estimation procedure that uses the data from the already described calibration step and assumes no prior knowledge. Estimating the geometry of the multi-camera system can be easily accomplished by matching corresponding image features [17, 11, 12]. This does not apply to the views collected by the Gaze Machine (GM) and we have to rely on a calibration procedure and some assumptions.

To calibrate the Gaze Machine the subject is asked to gaze at the centre of a marker\(^1\), whose 3D dimensions are known, while panning and tilting her/his head. Data used for the Gaze Machine calibration thus consist of a sequence of quadruples of synchronously acquired frames: the calibration marker imaged by the left and right scene cameras and the left and right eyes imaged while gazing at it. Having computed the pupil envelopes for both eyes (Section 4), to complete the GM calibration we need to recover the relative pose of every camera building up the system and the unknown scale factors of the envelopes.

The scene cameras form a stereo rig whose extrinsic parameters can be estimated straightforwardly from the calibration sequence [22].

The poses of the eye cameras are estimated in the reference frame of one of the scene cameras, making use of the trifocal relation induced by the projection of the 3D calibration point and the projections of the optical axes, which pass through it (see Figure 5). The envelope describing the positions of the pupils of the entire sequence is estimated according to the method described in the previous sections.

---

\(^1\)from ARToolkit
up to an unknown scale factor. As a consequence, the bundles of compatible optical axes describe the planes $\Pi$ and $\Pi'$, whose intersection is the line $L$, which passes through the 3D point $X_{3D}$. Assuming that

1. the subject is gazing at $X_{3D}$ and
2. the optical axes intersect in $X_{3D}$

the estimated 3D point of regard belongs to the line $L$, whose projections on the eye camera image planes coincide with the projections of the 3D optical axes.

Thus, a correspondence can be established between the left and right optical axes, imaged by the left and right eye cameras, and the $X_{3D}$ point as imaged by the left scene camera. Let $x$ be the image of the $X_{3D}$ point framed by the left scene camera, $l'$ the image of the left optical axis framed by the left eye camera and $l''$ the image of the right optical axis framed by the right eye camera. The pose of the left and right eye cameras relative to the left scene camera (Figure 6) can be estimated by means of the correspondent image features $x$, $l'$ and $l''$. The geometry of the system is described by the Trifocal Tensor [14, 18, 7], which relates the involved features according to the incidence relation [8]:

$$x^i l'_j l''_k T_{jk}^{i} = 0$$

We estimate the trifocal tensor $T$ from the sequence of correspondences $x \leftrightarrow l' \leftrightarrow l''$ collected during the Gaze Machine calibration. Equation (16) provides a single constraint for the linear estimation of $T$. Finally, in the calibrated case, eye camera poses can be consistently estimated from $T$ up to a common similarity transformation of 3-space [3]. Knowledge about the 3D position of the observed point during the calibration phase is then applied to recover the unknown scale factors. The solution obtained is then used as starting point for a non-linear Levenberg-Marquardt based optimisation to remove the previously stated assumption of intersecting optical axes which, indeed, does not hold [2].

8. Experiments

The main goal of this section is to quantitatively analyse the sensitiveness of the proposed gaze estimation model, when dealing with inaccurate measurements, provided by real sensors and demonstrate how we overcome some of the limitations characterising many available systems. Changing in light conditions plagues those systems relying on glint detection for the estimation of the parameters given some a priori model for the pupil and cornea shape. Robustness against extremely varying light conditions has been achieved by choosing not to rely on eye modelling but instead to describe the locus of pupil centres. At the same time the estimation of the camera poses outperforms also those purely regression-based approaches, that do not take

Figure 4. Frames from an on-field Gaze Machine calibration sequence.

Figure 5. The trifocal relation involving the image of the calibration point, $X_{3D}$, and the two optical axes passing through it.

Figure 6. Estimated eye camera poses w.r.t. the left scene (reference) frame. Here the Gaze machine is worn by a fire fighter in a training facility.
into account the exact pose of the cameras and thus suffer from poor estimation when the distance from the observed point changes considerably. This allows for application to a wide range of indoor/outdoor scenarios with different lighting conditions and distances.

8.1. Experimental setup

We validated the proposed approach by testing the calibrated Gaze Machine in real experimental setups. Subjects range from twenty to fifty years old, with normal or corrected to normal vision (contact lenses) and they volunteered for the experiments. We asked them to gaze at the centre of the calibration marker while moving their head and freely roaming around 1) inside a lab, 2) an area of 6 × 7 squared meters, and 3) at an outdoor training facility, as shown in Figure 4. Measured error consists in the mean Euclidean distance between the 3D POR and the centre of marker, whose position is measured by the Gaze Machine stereo rig.

More precisely, for every subject we recorded a test sequence. Every test sequence consists in two phases: calibration and validation. Calibrating the Gaze Machine resides in estimating the pupil envelopes for both eyes (Section 5) and the relative pose for every camera building up the system (Section 7). Indeed, pupil envelopes are peculiar for every subject. Similarly, camera poses changes when the GM chassis adapts to different subject heads.

We performed the calibration by asking the subject to gaze at the centre of the calibration marker while standing and freely moving her/his head. A 26 × 26 centimetres calibration marker (Figure 4) has been placed at a distance of about 1.5 meters. Typical calibration phases have durations within 20 and 25 seconds. Given a rate of 30 frames per second for the synchronised acquisition, this yields more than 600 pupil-marker quadruples from the four cameras. The cameras used for scene acquisition are PGR DR2, with a focal length of 263 pixels and the resolution is set to 320 × 240 pixels. Thus, from a position in which the marker is projected at the centre of the image, the maximum panning and tilting in order to maintain the marker in the camera field of view are about 30 and 25 degrees respectively. Those frames in which the marker is not detected are discarded in both the calibration and validation phase. Since they are used for measuring, cameras are calibrated for intrinsic and radial distortion parameter estimation. Pre-processing thus involves image un-distortion for the entire sequence. Camera calibration and the estimation of the homography between the marker plane and its image, used to detect the marker and compute the extrinsic parameter for the GM stereo rig, is mainly based on the work in [22].

Here the pupil envelope Π has been estimated to be approximately a quadric for each eye dynamics, and the pose of each camera making up the system has been recovered as a result of the GM calibration. Every new observation yields the pupil representation \( \Theta = (\psi, \pi) \) providing all the information needed to estimate the 3D pupil axis \( \ell \) (Section 3). The 3D POR is then estimated by a Gaussian Regression Process on the pupil axes, for both eyes, after transforming them to the reference frame of the estimated camera poses.

The validation phase comprises 12 experiments: subjects are again asked to gaze at the centre of the calibration marker while panning and tilting but 1) light conditions are changed 2) the distance of the subject from the marker is changed (Table 1).

![Figure 7. Error variance on a synthetic dataset of 5000 pupil pairs. Pupil angles have been corrupted with the addition of Gaussian noise with zero mean and increasing variance (horizontal axis). Error variance on the optical axes is shown on the vertical axis. Measures are in degrees.](image)

<table>
<thead>
<tr>
<th></th>
<th>1 m</th>
<th>2 m</th>
<th>3 m</th>
<th>6 m</th>
</tr>
</thead>
<tbody>
<tr>
<td>indoor</td>
<td>2.1 ± 0.3</td>
<td>3.6 ± 0.3</td>
<td>5.7 ± 0.5</td>
<td>16.9 ± 1.2</td>
</tr>
<tr>
<td>daylight</td>
<td>2.3 ± 0.5</td>
<td>3.5 ± 0.4</td>
<td>6.1 ± 0.6</td>
<td>21.4 ± 1.7</td>
</tr>
<tr>
<td>twilight</td>
<td>2.2 ± 0.4</td>
<td>3.7 ± 0.6</td>
<td>5.9 ± 1.0</td>
<td>19.1 ± 2.2</td>
</tr>
</tbody>
</table>

Table 1. Average errors in centimetres measured as difference between the estimated POR and the 3D position of the marker centre computed by the GM stereo rig for 1) different light conditions (rows) 2) different distance from the marker (columns). Indoor refers to light conditions in the lab (600 lux); daylight and twilight refer to outdoor settings (11,200 and 28 lux respectively).

8.2. Discussion

Measurement noise affects the ellipse fitting for pupil detection and the estimation of the 3D marker position. In order to analyse the sensitivity to inaccurate pupil detections we tested the proposed system on a synthetic dataset of 5000 pupil pairs with the addition of Gaussian noise with zero mean and increasing variance (Figure 7), measuring the error variance on the optical axis orientation. The results demonstrate that the estimation of the pupil envelopes attenuate the undesirable effects of bad pupil detections. Human failures in the GM calibration procedure introduce outliers. Experiments on synthetic datasets, in which a percentage of outliers is manually introduced, proved that our approach...
is able to cope with a 5% of outliers, which are rejected at the different steps in which consensus-based strategies are employed.

9. Conclusions

We have presented a novel approach to 3D gaze estimation for wearable multi-camera devices and demonstrated its effectiveness both theoretically and empirically. The proposed approach introduces a calibration procedure that is efficient, accurate, highly innovative but also practical and easy. The overall gaze estimation model is general, as no particular complex model of the human eye is assumed in this work. Based on our simplifying assumptions, a small set of parameters describing the locus of pupil centres can be conveniently learned by a practical calibration procedure which is operable directly on field; this calibration sequence also allows the geometry of the multi-camera system to be contextually estimated or adjusted. We account for errors due to the simplifying assumptions using regression on data from the calibration sequence. The resulting system is characterised by a data driven approach to estimate gaze in 3d, based on very few assumptions: no complex model for the eye is assumed and all the information needed to estimate the gaze is efficiently extracted from the calibration sequence; IR glints are not needed to estimate eye parameters and IR light sources can just be used to enhance image contrast.

References