Low Cost Gaze Estimation: Knowledge-Based Solutions

AQ:1 Ion Martinikorena, Andoni Larumbe-Bergera, Mikel Ariz, Sonia Porta, Rafael Cabeza, and Arantxa Villanueva

***Abstract*— Eye tracking technology in low resolution scenarios is not a completely solved issue to date. The possibility of using eye tracking in a mobile gadget is a challenging objective that would permit to spread this technology to non-explored fields. In this paper, a knowledge based approach is presented to solve gaze estimation in low resolution settings. The understanding of the high resolution paradigm permits to propose alternative models to solve gaze estimation. In this manner, three models are presented: a geometrical model, an interpolation model and a compound model, as solutions for gaze estimation for remote low resolution systems. Since this work considers head position essential to improve gaze accuracy, a method for head pose estimation is also proposed. The methods are validated in an optimal framework, I2Head database, which combines head and gaze data. The experimental validation of the models demonstrates their sensitivity to image processing inaccuracies, critical in the case of the geometrical model. Static and extreme movement scenarios are analyzed showing the higher robustness of compound and geometrical models in the presence of user's displacement. Accuracy values of about 3° have been obtained, increasing to values close to 5° in extreme displacement settings, results fully comparable with the state-of-the-art.**

***Index Terms*— Gaze estimation methods, low resolution, eye tracking.**

I. Introduction

DURING the last decades, especially during the last five years, a big effort has been made by the scientific community in order to extend the application of eye tracking systems to other frameworks, such as off-the-shelf systems or low resolution hardware, i.e. eye trackers employing a webcam or the mobile device camera. The application of eye tracking technology, in their high resolution fashion, can be verified in fields such as the analysis of eye movements or human computer interaction for severely disabled people [1]. The high resolution systems are a fact, although further improvements are still pursued in order to increase the accuracy and reduce head movement constraints [2], [3].

Regarding low cost systems, we find some publications in which the accuracies reported are far from being comparable to the ones obtained by high resolution systems. This is partially comprehensive due to the lack of detail in the image and the inaccuracies arisen from the features detection. The employment of more off-the-shelf cameras, such as a webcam, reduces considerably the density of pixels in the pupil area compared to high resolution systems. Consequently, the research related to low cost eye tracking is also named as low resolution eye tracking as it will be considered in this article. Apart from the lower resolution, there are additional factors that can contribute to the inaccurate gazed point estimation. High resolution systems use high focal length lenses with narrow Field of View (FoV) providing an extremely detailed image of the eye area and not allowing large movements of the subject to remain visible (for the camera). Contrarily, the wider FoV of a webcam permits the user to move freely. Additionally, when moving to webcam-based systems, it is reasonable to remove the infrared light sources, the goal being to reach a plug-and-play eye tracking technology. The absence of the infrared light produces, on the one hand, a lower quality image and on the other hand, the lack of a key feature, i.e. corneal reflection (glint), for gaze estimation. In summary, the extrapolation of the know-how obtained in the field of high resolution infrared gaze tracking cannot be applied to low resolution systems straightforwardly [4].

First, the image processing algorithms employed need to be reoriented to low resolution images (obtained using systems with no infrared light). Second, geometrically speaking, regardless of the type of system employed, i.e high or low resolution system, the head position with respect to the camera and the eyeball pose within the head are required to determine the Line of Sight (LoS) with respect to a remote camera. For high resolution systems, the corneal glint is normally assumed to be a reference for the head position. Thus, alternative gaze estimation methods incorporate the head pose information in different manners. When regression based methods are employed for gaze estimation, e.g. a second degree polynomial, the Pupil Center-Corneal Reflection (PC-CR) vector is used as independent variable, assuming its robustness against head movement [5]. On the other hand, the geometrical methods do an explicit modeling of head position based on the information provided by the glints and assuming a simplified eye model [6]. The absence of infrared light reinforces the need of incorporating head information by using alternative

AQ:2 Manuscript received May 14, 2018; revised November 6, 2018 and September 13, 2019; accepted September 25, 2019. This work was supported in part by the Ministry of Economy and Competitiveness under Grant TIN2014-52897-R and in part by the Ministry of Science, Innovation and Universities under Grant TIN2017-84388-R. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Francesco G. B. De Natale. *(Corresponding author: Arantxa Villanueva.)*

I. Martinikorena, A. Larumbe-Bergera, S. Porta, R. Cabeza, and A. Villanueva are with the Department of Electrical, Electronic and AQ:3 Communications Engineering, Public University of Navarre, 31006 Pamplona, AQ:4 Spain (e-mail: avilla@unavarra.es)

M. Ariz was with the Department of Electrical, Electronic and Communications Engineering, Public University of Navarre, 31006 Pamplona, Spain. He is now with IDISNA, Ciberonc and Solid Tumours and Biomarkers Program, Center for Applied Medical Research, University of Navarra, 31009 Pamplona, Spain

Digital Object Identifier 10.1109/TIP.2019.2946452

methods not previously employed in the field of high resolution gaze tracking. Moreover, high accuracy Head Pose Estimation (HPE) methods are required since any HPE error would contribute directly to the gaze estimation error.

Alternative solutions can be found in the literature proposing gaze tracking methods for low resolution systems. One of the first works regarding low resolution is that presented by Valenti *et al.* [7]. In this paper, it is explicitly stated that the head modelling is a requirement in low resolution scenarios. The paper clearly demonstrates that a joint modelling of the head and eye improves gaze estimation. An iterative process is carried out in which "normalized" eye images are obtained from the head position, and the eye position is then employed to correct head information. A couple of years later, in the paper by Wood and Bulling [8], a model-based approach for binocular gaze estimation to be run in a tablet was shown. The accuracy obtained was about 6° but the tolerance to head movement was not clearly demonstrated. The accuracy values obtained in low resolution systems are below those achieved by high resolution gaze trackers, but there are some interesting applications for which no outstanding accuracies are required. In the work by Vicente *et al.* [9], a remote gaze tracking system is presented to be installed in a car to detect "eyes off the road" situations. A complete system is proposed composed by an image processing stage leading to the geometry based estimation of head pose and gaze direction. More details are provided about head pose results than about gaze tracking accuracy. Similar works aimed to detect gazing zones in driving scenarios [10] can be found.

The methods mentioned can be grouped under the term of *feature-based-methods*. Regardless of the gaze estimation method employed, an image processing stage is required to extract specific image features to be used as input for the gaze estimation method. During the last years, alternative works based on deep learning, e.g. Convolutional Neural Networks (CNNs), have been proposed for gaze estimation. CNNs, as supervised learning tools, have demonstrated to be a nice solution for many computer vision problems, such as object detection or scene recognition among others. The methods based on CNNs have common aspects with *appearance-based-methods* [11]. Roughly speaking, it is not required to extract features from the image but it is the network which, automatically, learns the required information from the image to carry out the classification/regression, i.e. the gaze estimation in our case. In other words, when dealing with CNNs there is not a division between eye tracking (i.e. image processing) and gaze estimation, but both stages are performed by the same tool. In the work by Krafka *et al.* [12] CNNs are used to calculate gaze direction. A database of approximately 2.5M images containing faces of individuals gazing points on a screen is used for training the network. Basically, the network is fed using three cropped images of the face and both eyes. Additionally, an empty image in which the face position within the image is marked is employed as input. The network is trained to obtain the head pose with respect to the camera and the position of both eyes with respect to the head. Thus, combining the output data, the gaze direction can be inferred. In the work by Zhang *et al.* [13], the gaze is estimated by means of a two-step procedure based on CNNs. Cropped eye images are used as input to a CNN whose output is combined with data about the head pose to obtain the gaze. The suitability of CNN-based methods relies basically in two aspects: first, the availability of a large scale database that is able to represent the variability of the problem to be solved. Second, its success depends on the trained network ability to generalize, i.e. the capability to obtain a correct output for samples not included in the training stage. The requirement of having a representative database is key to obtain successful results. In fact, during the last few years, interesting efforts have been carried out in order to produce this kind of databases, such as POG Eye Tracking [14], EYEDIAP [15], MPIIGaze [13], [16], Columbia dataset [17] and TabletGaze [18].

The works employing these databases utilize deep learning as gaze estimation method. The main contribution of these works is valuable from the point of view of the regression method employed, more than from the perspective of the results representability. The number of training and testing images of the mentioned databases approximates some thousands, except for the MPIIGaze database containing about 250,000 images. Nevertheless, they are far from being considered large scale databases. The difficulty of obtaining large scale databases in the field of eye tracking is the fact that the data labelling is not straightforward. Eye images have to be linked with the gazed point and this information is not easily available. The most remarkable work in the field is the one developed at the MIT [12] containing 2.5 millions of images from 1450 participants. The method employed for obtaining labelled data is based on *crowdsourcing* by means of a designed application named GazeCapture, installed in subjects' tablets and phones. In this manner, the subjects could activate the application any time and gaze specific points on the screen that could be registered together with the eye images captured by the gadget camera. An alternative solution for overcoming the problem of obtaining tagged data is to use "learning by synthesis" approaches. Employing simulation environments, synthetic images are constructed in which the labels are already known as they have been used to build the image. In this manner, enormous amount of tagged images can easily be obtained. Remarkable works in this area are the ones presenting Multi-view gaze dataset [19] and the proposals made by Świrski and Dodgson [20] and Wood *et al.* [21].

Accuracies reported for low resolution gaze tracking are far from being comparable with the results obtained by other approaches using a geometrical perspective, and highly dependent on the database for which the method has been trained. Reviewing the literature, angular errors in the range of 7°-9° are reported for Columbia dataset, while values in the range of 6°-20° are found for the EYEDIAP, showing a strong dependency on the estimation method used [8], [22], [23]. CNNs show up as a promising technique to be applied to gaze estimation, and could probably provide better results than the ones reported to date if the existing difficulties are overcome in the near future. For MPIIGaze, which is one of the most referenced datasets in the literature, errors in the range of 7°-9° have been reported [16] using appearance-based methods.

Moreover, in a later work of the authors, it is shown that feature-based approaches using explicit landmarks extracted from the image can outperform appearance-based approaches to date, showing errors in the range of 3°-6° [24].

In any case, today, feature-based methods show up as a possible solution for low resolution gaze tracking systems. Moreover, working from a more geometrical perspective permits to obtain a valuable knowledge of the system under study and provides a deeper understanding of the different variables affecting the system accuracy.

In this paper, we review the basics of high resolution systems and we propose novel solutions for low resolution remote eye trackers using the knowledge acquired so far. The know-how constructed in the last decades about the geometry and the key aspects of gaze estimation permits to approach the problem from an advantageous perspective. Therefore, three alternative models are proposed for gaze estimation in the low resolution environment. Moreover, image processing strategies are evaluated and suggested for both head pose estimation and iris detection, which are key for the different gaze estimation methods proposed.

In the next section, the basics about gaze estimation geometry problems are reviewed. Section III presents alternative gaze estimation methods proposed for low resolution eye trackers. In section IV the framework in which the methods are evaluated is carefully described. Additionally, the I2Head database, which is key for the validation of the gaze estimation methods, is presented. Section V shows the results achieved in the different tests carried out in this work. Finally, the discussion and conclusions of the work are presented in section VI.

II. Gaze Estimation Revisited

In this section, the basics of gaze estimation theory are described and discussed. It is important to analyze the problem by using high and low resolution perspectives with the aim to identify those points that can be applied to both frameworks and to detect their main differences from the gaze estimation point of view.

A. High Resolution Gaze Estimation

Gaze estimation based on remote video-oculography has been around since decades ago. High performance or high resolution eye trackers using infrared light sources, optical filters and high focal length lenses produce high resolution pupil area images. Hence, the detection of the pupil center and corneal glints is feasible.

Different approaches have been proposed to approximate the geometry of the 3D framework composed by the user, the camera, light sources and the screen. A review of the alternative methodologies can be found in [11]. Regarding gaze estimation methods, the most popular ones due mainly to their robustness and accuracy are, on the one hand, the methods based on interpolation models (i.e. using a polynomial) and, on the other hand, geometrical models. All these methods consider as input the information extracted from the image, i.e. image features, and provide as output the 2D gaze position on the screen, named the Point of Regard (PoR) or the 3D Line of Sight (LoS). The Line of Sight can be geometrically determined by knowing the head position and the eye pose within the head model.

According to the literature, eye tracking methods with an acceptable accuracy require a user calibration stage in which the unknown parameters of the gaze estimation model are to be estimated. The calibration consists in asking the subject to gaze specific targets on the gazing area. The number of targets can vary from one to more points, e.g. grids of nine or sixteen points, according to bibliography.

Regarding *geometry-based-models*, the parameters to be deduced in the calibration procedure are individual's parameters such as corneal radius or angular offset between optical and visual axes. The fovea is a small depression of the retina responsible for our most accurate vision. It is the area in which the gazed objects are projected. The fovea is located temporally in the eyeball, meaning that there is an angular offset between our Line of Sight represented by an imaginary axis (named the visual axis) and the symmetry axis of the eye (named the optical axis of the eye). The output of the geometry-based methods is the 3D LoS resulting in the estimation of the 2D PoR when the intersection with the screen plane is calculated. Geometrically, it has been demonstrated that a single camera and two light sources is the minimum hardware required to determine gaze direction with no head movement constraints [6]. Thus, geometry based frameworks present better robustness regarding head movements of the user. The handicap of geometrical methods is the model complexity involving projective relationships and 3D models of the alternative elements, eyeball, camera, light sources and screen. On the other hand, the complete knowledge of the system requires a setup calibration, i.e. calibration of the camera, the screen position and the light sources. In summary, eye trackers using geometrical models are far from being plug-and-play systems.

The alternative is to use *interpolation-based-methods* for which the simplicity is one of their outstanding characteristics. Interpolation based methods can be considered as blind methods in which no knowledge about the system or the user is required. The model is able to adapt to the subject working with the system. It has been shown that a second degree polynomial is sufficient for gaze estimation purposes [5]. Generally, the interpolation based methods output is the 2D PoR. During the calibration, the unknown polynomial coefficients are deduced. Most of this type of approaches take under consideration the head movement in an approximate manner. The infrared light sources employed by high resolution eye tracking systems produce corneal reflections that are visible for the camera and normally named glints. It is assumed that the vector connecting the pupil center and the glint(s) named Pupil Center-Corneal Reflection, PC-CR vector, is approximately stable against head movement. The calibration procedure of the user permits to adapt the model to the specific situation in the calibration position.

In general, the user's displacement from the calibration situation affects the accuracy. Fortunately, due to the high focal lengths used by high performance eye trackers, the allowable head movement is reduced. Thus, the assumption that the

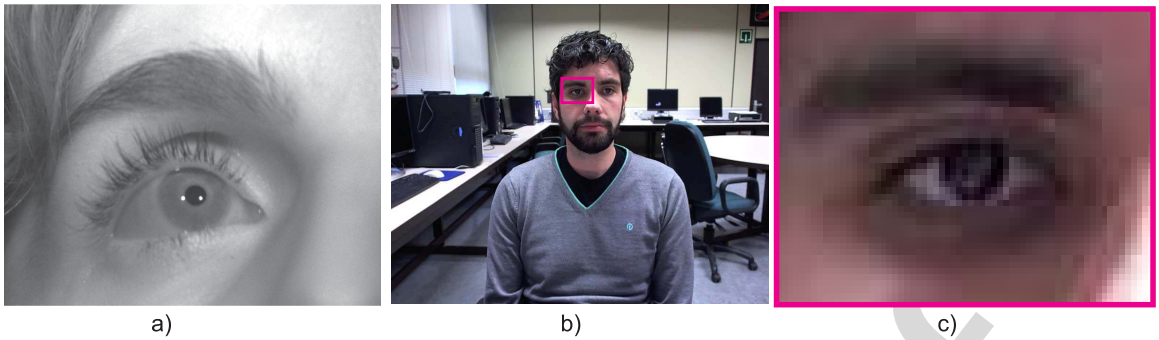

Fig. 1. a) An image with a resolution of 800×600 pixels captured by a high resolution eye tracking system using a focal length value of 35 mm. The glints and the pupil are clearly visible. b) An image with a resolution of 800×600 pixels captured by a low resolution system, i.e. using a webcam with a focal length value of 2.7 mm. c) The eye area shown by the pink frame is extracted from b). The detail level in the eye area is low compared with a). The eye area in b) is limited to an area of about 66×49 pixels size while it covers the whole image resolution in a).

calibration results are stable is partially acceptable within the range of permitted head movements at the expense of losing some accuracy.

B. Low Resolution Gaze Estimation

During the last years a big effort has been made to extend gaze estimation technology to low resolution environments where no infrared light is used and lower focal lengths are employed.

In figure 1, a comparison between the images acquired using a high resolution and a low resolution eye trackers is shown. As it can be seen in the image, the resolution regarding the eye area is not comparable between the two frameworks. The lenses employed by high resolution systems present high focal lengths, e.g. 35 mm, while standard low resolution systems using webcams show lower focal length numbers of about 2 or 3 mm. In this manner, the Field of View (FoV) of high performance systems permits to obtain a more focused image of the eye with higher resolution in the eye area, i.e. more pixels, than lower resolution systems. In fact, strictly speaking, the term resolution when differentiating between high and low resolution systems should be understood as the resolution in the eye region and not as the resolution of the whole image.

The scenario is completely different and affects most of the stages of the gaze estimation procedure. The most obvious one is the task related to image processing. First, the scene lighting is no longer under control, and second, the lower resolution of the image in the eye area makes the pupil/iris center detection more difficult. In terms of gaze estimation, in principle, the basics are still valid, i.e. the Line of Sight can be calculated as a function of the head position and the eyeball pose within the head. However, there are key differences with respect to high resolution systems that make gaze estimation more complicated: first, if a geometrical model is used, the absence of infrared light sources prevents the system from using them as valid features to estimate the head position. In this manner, an alternative method is required to determine the head pose and to complete the geometrical model. Second, if an interpolation model is used, one could think of employing another head-fixed feature as head position indicator, such as the eye corner, and use the Pupil Center-Eye Corner (PC-EC) vector as an alternative. However, in this

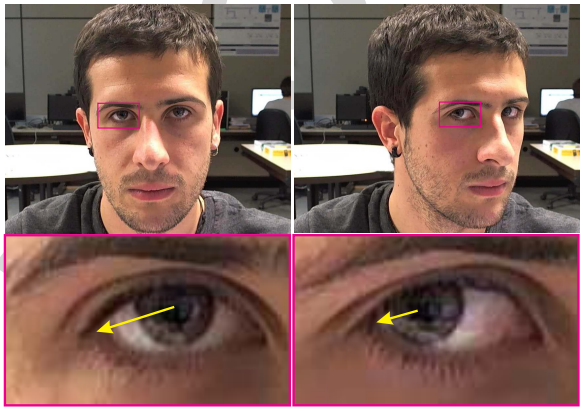

Fig. 2. PC-EC vector (in yellow) when gazing the same point from different head positions. a) In the upper row the images captured by the camera are shown. The user gazes at the same point from different head poses b) In the lower row zoomed versions of the eye region are shown together with the PC-EC vector, i.e. for the same gazed point different values of the vector can be obtained. It has to be taken into account that low resolution scenarios permit larger head movements and this type of situations are potentially more frequent than in high resolution setups.

new scenario in which the range of head movement is larger, the fact of considering the PC-EC vector "stable" in the presence of head movement is less assumable compared to high resolution systems. In figure 2 we observe the PC-EC vector behavior when gazing at the same point, i.e. same PoR, from different extreme head positions in a pure rotation of the head. It can be observed that the PC-EC vector has not a univocal value for the same gazed point, i.e. PoR, when large head movements are allowed.

The objective of this work is to analyze different gaze estimation methods for low resolution scenarios using as departure point the knowledge of the problem geometry. The paper suggests alternative models ranging from interpolation based methods to pure geometrical methods for gaze estimation that, on the one hand, provide a deeper insight about the underlying theory of low resolution systems and, on the other hand, demonstrate the possibilities to adapt part of the know-how acquired to this new paradigm, i.e. the low resolution scenario.

III. Gaze Estimation Methods for Low Resolution

The aim of this section is to propose three models that try to solve the problem of gaze estimation in low resolution

TABLE I
SUMMARY OF SYMBOLS EMPLOYED IN THIS PAPER

Symbol	Description
H	Head system of coordinates
HP	Head pose
C	Real camera system of coordinates
T	Real camera system of coordinates
V	Virtual camera system of coordinates
g	Gaze direction (3D vector)
S	Screen system of coordinates
I	Real image system of coordinates
In	Normalized image system of coordinates
p	Pupil (iris) center in the image (2D)
P	Pupil (iris) center (3D)
n	Pupil (iris) center in the normalized image **In** (2D)
E	Eyeball center (3D)
r	Eyeball radius
κ	Angular offset between optical and visual axes
$PCEC$	Pupil (iris) center-eye corner vector (2D)
q	Point of Regard (PoR)

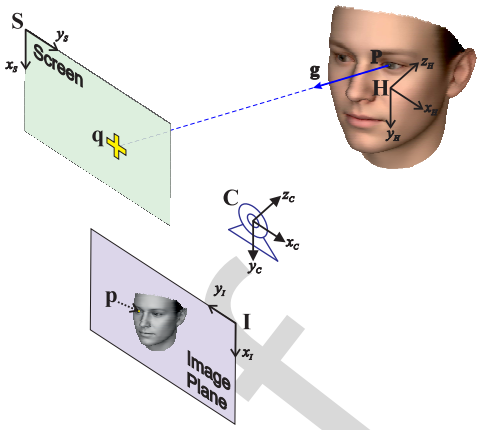

Fig. 3. Elements of the system. The camera, **C**, is considered to be the WCS. The individual's position is defined by the head position, **H**. In addition, reference systems are defined for the gazing area, **S**, and the image, **I**.

scenarios. In order to help readers to follow the explanation of the models, a table of symbols is provided as reference (see table I).

The setup is composed by a subject, i.e. the head of the user is taken as reference and named **H**, gazing at different points in the gazing surface, **S** (see figure 3). The WCS (World Coordinate System) is assumed to be the camera, named **C**. The gaze direction **g** is defined as the vector pointing in the LoS direction. It can be referenced to the head as $\mathbf{g}^{\mathbf{H}}$ or to the camera, namely, $\mathbf{g}^{\mathbf{C}}$, i.e. superscripts will be used to show the coordinate system an element is referenced to. The position of the head with respect to the camera, the head pose $HP^{\mathbf{C}}$, can be expressed by means of a rigid transformation $(\mathbf{R}^{\mathbf{CH}}, \mathbf{T}^{\mathbf{CH}})$ where $\mathbf{R}^{\mathbf{CH}}$ is the rotation matrix of the head reference system with respect to the camera and $\mathbf{T}^{\mathbf{CH}}$ is the translation vector of the head reference system with respect to the camera.

In this manner, the gaze direction with respect to the camera can be calculated geometrically, knowing the gaze direction with respect to the head, by means of the following expression:

$$\mathbf{g}^{\mathbf{C}} = (\mathbf{R}^{\mathbf{CH}}|\mathbf{T}^{\mathbf{CH}})\mathbf{g}^{\mathbf{H}} \quad (1)$$

The PoR, **q**, can be calculated as the result of the intersection of **g** and the gazing surface, **S**.

On the other hand, in the image, **I**, features such as pupil/iris center is defined as **p** which is approximated by the projection of the 3D iris center **P**, onto the image plane. Figure 3 summarizes the elements involved in the system framework.

Three models are presented: the first model is the geometrical model, which tries to mimic the same principles of high resolution systems but considering the new scenario in which no infrared light sources are employed and larger head movements are possible. The second method presents an interpolation model, i.e. new features are proposed to be extracted from the image and the model output is understood in a geometrical context. Lastly, a compound algorithm is proposed, trying to combine the interpolation model simplicity and the robustness of the geometrical model in the presence of large head movements. Since no infrared lighting is used in the proposed low cost system, alternative HPE techniques are to be used as it will be later explained.

A. Geometrical Model

This model is fully based on the system geometry. The LoS is calculated as a function of head position and eyeball information. Assuming that the head pose is known (see section IV-B), a simplified eyeball model is proposed consisting of a sphere rotating around the eyeball center. This assumption is slightly different from the one considered in high resolution systems [25]. The approach employed in most high resolution systems is to consider the cornea as a sphere rotating around the eyeball center. Hence, the cornea center translates with respect to the center of the eyeball as the eye focuses on alternative points on the screen. In our proposal for the eyeball model, the cornea is not explicitly modeled, i.e. the pupil center moves along a sphere centered at a fixed point with respect to the head, named eyeball center, $\mathbf{E}^{\mathbf{H}}$.

A correct estimation of the angular offset between optical and visual axes has demonstrated to be critical in most high resolution gaze estimation systems. The horizontal angular offset between optical and visual axes is named kappa, κ, and it is an individual's parameter in the range of 3° to 7°. A smaller vertical offset exists between the axes but it is obviated in this work for simplicity. The visual axis is normally approximated by the imaginary line joining the fovea with the cornea center. In the simplified model assumed in this work, the optical axis is defined as the line connecting the pupil center and the eyeball center, and the visual axis is considered to be the line intersecting the eye at the eyeball center forming an angle equal to κ with the eye optical axis. The angle κ and the eye sphere radius named r are estimated for each individual through a calibration procedure. The 3D eye model employed by this method is shown in figure 4.

Once the pupil center is detected in the image, **p**, it is back projected as a line $\in R^3$ with respect to the camera. As a result of head pose estimation, and knowing the head model, the position of the eye sphere is calculated centered at **E** with a radius equal to r. 3D pupil center, **P**, is calculated as the

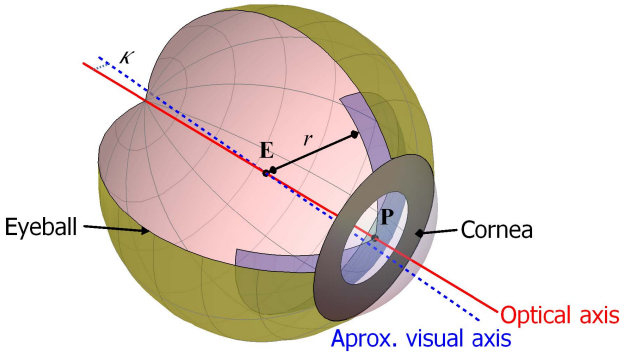

Fig. 4. The 3D eye model employed in this work. The eye is a sphere and the pupil center **P** moves along an imaginary surface (in purple). Most of the models employed by high resolution systems include the cornea as an additional sphere as it is shown in the figure. In our model for low resolution, the cornea is obviated and a single sphere centered at **E** with radius r is considered. The LoS is approximated by the visual axis calculated as the line containing the eyeball center and presenting a horizontal angular offset, κ, with respect to the optical axis of the eye model. Both angle κ and eyeball radius r are to be estimated during the calibration procedure.

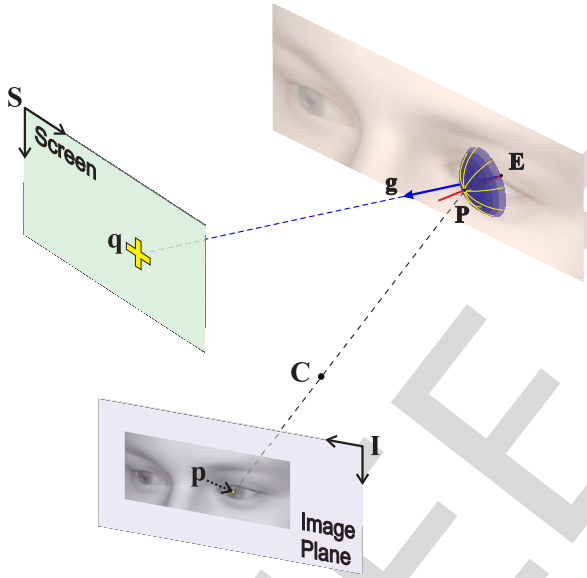

Fig. 5. Geometrical model scheme for a single eye. The pupil center **p** is back projected from the image **I** onto the eyeball modeled as a sphere centered in **E**. The intersection point is considered to be the pupil center position in 3D, **P**. Assuming that κ is known, the visual axis is estimated, **g**. Once the visual axis is estimated the intersection with **S** can be calculated to obtain the PoR, **q**.

intersection of the back-projected line and the eyeball sphere. Thus, the optical axis can be calculated as the line connecting **E** and **P**. The visual axis estimation is straightforward if κ is known from calibration [25]. The gazed point **q** is calculated as the intersection between the visual axis and the gazing area **S** (see figure 5). A simulation tool has been constructed in order to test and evaluate this model, in terms of accuracy and calibration issues, based on the tool designed by Böhme *et al.* [26].

B. Interpolation ~~model~~

This model is based on the interpolation methods employed for high resolution systems. As mentioned before, the use of Pupil Center-Corneal Reflection vector, PC-CR vector, as a reliable feature for gaze estimation is based on the idea that it is robust against head movements. The limited FoV in those systems does not allow large head movements. In that scenario, the assumption regarding PC-CR vector is partially acceptable since the accuracy decreases as the user moves from the calibration position.

In the low resolution scenario no infrared light sources are used, hence, PC-CR vector cannot be calculated. Instead, the eye corner is proposed in this method as anchor point, i.e. as reference point of head position. In any case, according to figure 2, the PC-EC vector does not provide a univocal relationship with the gaze direction, $\mathbf{g}^{\mathbf{C}}$. However, the PC-EC vector provides information about the eyeball orientation with respect to the head univocally. In other words, instead of calculating $\mathbf{g}^{\mathbf{C}}$, the position of the pupil center with respect to the eye corners can be used to estimate gaze direction with respect to the head, $\mathbf{g}^{\mathbf{H}}$. As a remark, the pupil is not easily distinguishable from the iris; in fact, the iris center is pursued assuming it is equivalent to the pupil center. However, the nomenclature PC-EC is maintained to refer to the iris (Pupil) Center-Eye Corner vector. The PC-EC vector is represented by $PCEC$ symbol and is calculated as:

$$PCEC^{\mathbf{I}} = (PCEC_x, PCEC_y)^{\mathbf{I}} = \frac{\mathbf{p}^{\mathbf{I}} - \mathbf{c}^{\mathbf{I}}}{\|\mathbf{c}^{\mathbf{I}}_{left} - \mathbf{c}^{\mathbf{I}}_{right}\|} \quad (2)$$

where $\mathbf{c}^{\mathbf{I}}$ is the eye outer corner in the image coordinate system. In fact, a normalized version of the $PCEC$ is employed, i.e. the vector is divided by the distance between the right and left outer corners of the eye. This type of strategy has demonstrated to work nicely in high resolution systems, making the system more robust against subject's displacements from the calibration position [5].

In high resolution scenarios, second degree polynomials using PC-CR vector as input are generally employed to estimate 2D gaze position (PoR). For our low resolution framework, the interpolation-based approach is to propose two second degree polynomials to estimate the gaze direction with respect to the head, using as independent variable the $PCEC^{\mathbf{I}}$ vector and as dependent variable the unity norm 3D vector representing the gaze direction, $\mathbf{g}^{\mathbf{H}} = (\mathbf{g}^{\mathbf{H}}_x, \mathbf{g}^{\mathbf{H}}_y, \mathbf{g}^{\mathbf{H}}_z)$. In this manner, we can construct an interpolation model to estimate gaze as:

$$\begin{aligned}\mathbf{g}^{\mathbf{H}}_x &= a_1 \cdot PCEC_x^2 + a_2 \cdot PCEC_y^2 + a_3 \cdot PCEC_x \cdot PCEC_y \\ &\quad + a_4 \cdot PCEC_x + a_5 \cdot PCEC_y + a_6 \\ \mathbf{g}^{\mathbf{H}}_y &= a_7 \cdot PCEC_x^2 + a_8 \cdot PCEC_y^2 + a_9 \cdot PCEC_x \cdot PCEC_y \\ &\quad + a_{10} \cdot PCEC_x + a_{11} \cdot PCEC_y + a_{12}\end{aligned}$$

The previous expressions can be more simply expressed using matrix notation as:

$$\mathbf{g}^{\mathbf{H}} = \begin{pmatrix} \mathbf{g}^{\mathbf{H}}_x \\ \mathbf{g}^{\mathbf{H}}_y \end{pmatrix} = \mathbf{A} \begin{pmatrix} PCEC_x^2 \\ PCEC_y^2 \\ PCEC_x \cdot PCEC_y \\ PCEC_x \\ PCEC_y \\ 1 \end{pmatrix}; \|\mathbf{g}^{\mathbf{H}}\| = 1 \quad (3)$$

where $\mathbf{A}$ is a 2×6 matrix containing the unknown coefficients, $[a_1, ..., a_6; a_7, ..., a_{12}]$, of the second degree polynomials to be solved during the calibration stage [5]. In the equation $PCEC^{\mathbf{I}} = PCEC$ has been used for simplicity. The calibration procedure conducted by the user will permit to fit the polynomials, i.e. to calculate $\mathbf{A}$, to the calibration situation in which the gaze direction will be "learnt" as a function of $PCEC$ vector extracted from the image.

In order to determine the LoS with respect to the camera, this model requires to know the value of $HP^{\mathbf{C}}$. Combining both head position, $HP^{\mathbf{C}}$, and gaze direction with respect to the head, $\mathbf{g}^{\mathbf{H}}$, the gaze direction with respect to WCS, $\mathbf{g}^{\mathbf{C}}$, is obtained using equation 1.

The proposed method can encounter some limitations in the fact that the eye corner does not stay completely stable as the eyeball rotates [27]. However, more importantly, the method can fail in the presence of strong head translations and rotations that force the eyeball to rotate to poses not covered during the calibration process, in which the polynomial obtained as result of the calibration can behave slightly worse.

C. Compound Model

The last model proposed tries to take advantage of the interpolation model simplicity and the robustness of the geometrical model in terms of head movement, trying to combine the benefits of both approaches. The main limitation of the interpolation model in low resolution scenarios is that the calibration procedure, in the way it is conducted, is not able to cover all the possible eyeball rotations gazing at different points from any head position. Extending the calibration procedure to cover as many head positions as possible is not a feasible option.

The proposal made in this model is to conduct the calibration procedure carried out by the interpolation model in a virtual normalized camera with respect to the head of the user named as $\mathbf{V}$. Starting from the image obtained by the real camera, the objective is to infer the image that would be obtained by a camera placed in front of the user for any head position. In other words, the user's head remains static in the virtual normalized camera framework, i.e. the paradigm of high resolution systems is fairly approximated in this manner since no head movements take place with respect to the virtual camera. A simplified eyeball model is used for all the users (r is equal to 8 mm and κ is assumed to be 0), i.e no calibration is performed for the eyeball.

First, the pupil center in the image $\mathbf{p}^{\mathbf{I}}$ is back projected from the real image onto the simplified eyeball, using the $HP^{\mathbf{C}}$ information. Once the intersection is calculated, $\mathbf{P}^{\mathbf{C}}$, this point is projected onto the virtual camera, $\mathbf{V}$, fixed with respect to the head, obtaining the normalized pupil center in the virtual image defined as $\mathbf{n}^{\mathbf{In}}$. The value of the head position with respect to the virtual camera is defined as $HP^{\mathbf{V}} = (\mathbf{I}_3, (0, 0, 500)^T)$, where $\mathbf{I}_3$ is the 3×3 identity matrix. Figure 6 summarizes the components of the compound model.

In this manner, during the calibration process the normalized gaze direction with respect to the virtual camera, $\mathbf{g}^{\mathbf{V}}$, is adjusted using the information of the normalized iris center, $\mathbf{n}^{\mathbf{In}}$. As in the interpolation model, a second degree polynomial is employed using the normalized iris center as independent variable and the normalized gaze direction as the dependent one. Note that the PC-EC vector ($PCEC$) is no longer employed in this model. In the normalized image the eye corners remain static, thus they do not provide any useful information about the head, which is considered to be fixed with respect to the virtual camera. Therefore, equation 3 is modified accordingly as:

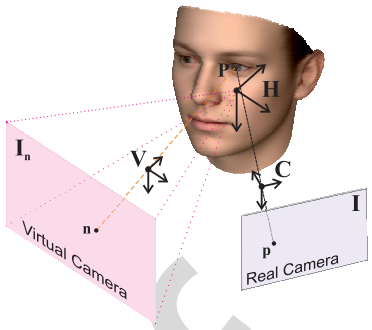

Fig. 6. The pupil center, $\mathbf{p}^{\mathbf{I}}$, is backprojected knowing $HP^{\mathbf{C}}$. Once the intersection point with the eyeball $\mathbf{P}^{\mathbf{C}}$ is calculated, it can be projected onto the virtual image of the virtual camera, which is fixed with respect to the head to calculate $\mathbf{n}^{\mathbf{In}}$. The calibration is performed using the normalized data in $\mathbf{In}$.

$$\mathbf{g}^{\mathbf{V}} = \begin{pmatrix} g_x^{\mathbf{V}} \\ g_y^{\mathbf{V}} \end{pmatrix} = \mathbf{B} \begin{pmatrix} (n_x^{\mathbf{In}})^2 \\ (n_y^{\mathbf{In}})^2 \\ n_x^{\mathbf{In}} \cdot n_y^{\mathbf{In}} \\ n_x^{\mathbf{In}} \\ n_y^{\mathbf{In}} \\ 1 \end{pmatrix}; \|\mathbf{g}^{\mathbf{V}}\| = 1 \quad (4)$$

where $\mathbf{B}$, is a 2×6 matrix containing the unknown coefficients, $[b_1, ..., b_6; b_7, ..., b_{12}]$, of a second degree polynomial to be solved during the calibration stage.

Once the normalized gaze direction is obtained, a de-normalizing process is conducted to calculate the Line of Sight, i.e., LoS, with respect to the WCS, $\mathbf{g}^{\mathbf{C}}$. Using head pose information, HP, this transformation is straightforward according to equation 1. In the same manner as in the case of the geometrical model, a simulation environment has also been designed in order to test the model under controlled conditions before employing real data.

IV. Framework

In order to evaluate the models presented in the previous section in a real scenario, essential elements are required. First, an annotated database is needed to study the gaze estimation methods. The proposed models use head pose information to estimate gaze, i.e. head pose needs to be estimated. Additionally, the proposed models use key image features as inputs, such as pupil and corners centers. In the following sections these questions are addressed.

A. I2Head Database

With the aim to evaluate the different models, a consistent framework is required. As mentioned in the introduction,

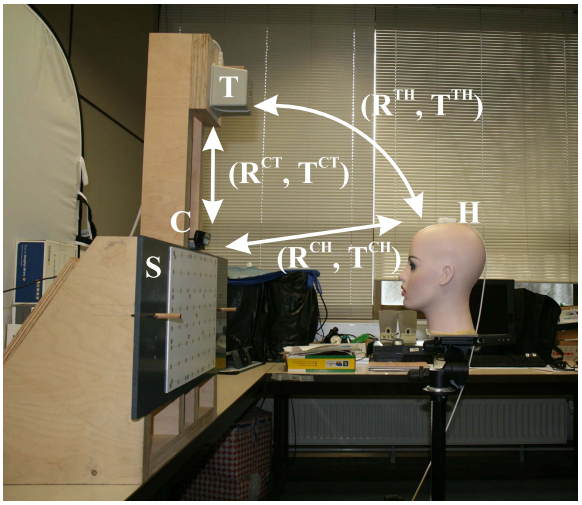

Fig. 7. In the photograph, the mannequin represents the user with the sensor attached to the head. The camera, the transmitter and the gazing surface are placed in the same wood structure in order to fix their relative poses. The framework is sketched showing its main elements. The relative position between the transmitter and the camera is carefully calibrated to obtain $(\mathbf{R^{CT}}, \mathbf{T^{CT}})$.

several databases devoted to gaze estimation can be found in the bibliography. In comparison with other datasets, I2Head provides not only images and data about gaze but also accurate head pose data [28]. In addition, partial information of head models for each user is provided, more specifically the position of four eye corners and the nose tip with respect to the head are included for each participant. In this manner, it is not only a valid framework to test gaze estimation methods but also to evaluate HPE techniques.

Moreover, since ground truth data of the head pose is provided, the contribution to the error from different sources can be more easily determined. It is already known that the robustness of the gaze estimation method against head movements is one of the cornerstones of the technology. Hence, any database intended to be a framework to evaluate gaze estimation methods should consider head movements. The I2Head database contains sessions performing controlled movements of the subjects: the subject is displaced to specific positions to measure the effect of translation in gaze estimation. In some of the sessions the user is asked to remain static while in others the user is able to move the head freely.

One of the criticisms that can be made to several articles in the field is the fact that they measure the accuracy using the same grid employed for calibration. Due to the fact that the calibration procedure is the result of an optimization process, we can expect a better behavior when the accuracy is tested using the calibration points, especially in the interpolation and compound methods. As in any other learning process it is not convenient to employ training data as test data. In order to measure the generalization capacity of the gaze estimation method, sessions including different grids of points are included. Finally, the coordinates of the gaze points and the intrinsic camera parameters are provided.

I2Head provides gaze and head pose data of twelve users performing different head movements in a controlled

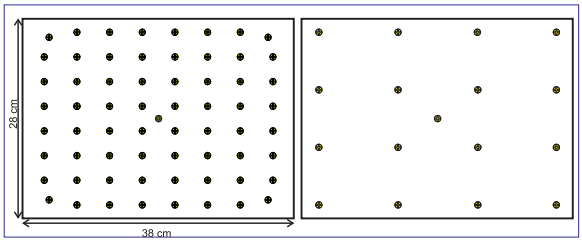

Fig. 8. Grids of points used for sessions recordings. Left) 65-point grid. Right) 17-point grid.

procedure. The hardware employed to construct the database consists of the Flock of Birds (Ascension Technologies) magnetic sensor for 3D pose estimation and a camera. The sensor is used to register the head pose with respect to the transmitter $\mathbf{T}$. To this end, the sensor is attached to the head of the user while performing head movements and registers 240 samples per second. The sensor output is a 6D vector containing translation information and rotation information, i.e. roll, yaw and pitch angles. The system can register the position of the sensor with an accuracy value of 1.4 mm rms and 0.5° rms as provided by the manufacturer.

The employed camera is a Logitech webcam with a resolution of 1280×720 pixels working at 30 fps. The hardware has been calibrated, i.e. the camera and the position of the transmitter have been accurately calculated, and thus, the head pose obtained with respect to the transmitter $HP^{\mathbf{T}} = (\mathbf{R^{TH}}, \mathbf{T^{TH}})$ can be transformed into camera coordinates, $HP^{\mathbf{C}}$. Moreover, the camera has been calibrated and the positions of the target points in the gazing surface $\mathbf{S}$ have also been calculated. In the database the camera projection center is taken as the origin of the WCS. In figure 7 a detailed scheme of the recording framework is presented.

Two different patterns of gaze points are employed in the surface area of size 28×38 cm. The first one is composed by 17 points, i.e. a 4×4 regular grid plus the central point. The second one consists of 65 points, i.e. a 8×8 regular grid plus the central point (see figure 8).

For each user, eight videos are recorded under controlled movements. In a centered position four sessions are recorded. The user is asked to keep the head static in the first two sessions, during which the 17-point grid (static) and the 65-point grid (static) are recorded. During the next two sessions the user is allowed to move the head in a free fashion while the 17- and 65-point grids are recorded. In the remaining four sessions the 17-point grid is exclusively employed changing the position of the user. The user is moved approximately 5 cm in forward, backward, leftward and rightward directions. During these sessions the user is asked to remain static. Table II summarizes the main I2Head dataset features. Additionally, in table III the recorded sessions are summarized.

No chin rest is employed in any of the sessions. While the head pose is registered employing the main sensor, a second sensor is used to mark the eye corners and the nose tip using a dedicated tool. In this manner, 3D face information is recorded, which is useful to create the simplified head and eyeball models.

TABLE II
MAIN FEATURES OF I2HEAD DATASET

Feature	Value
No. of images per point	10 images
No. of subjects	12 subjects
No. of images per user	2,320 images
Total no. of images	27,840 images
Head tracker precision	1.4 mm (rms) and 0.5° (rms)
Range of eye movements	$\sim \pm$ 20°
Image resolution	1280×720 pixels
Camera focal length	2.7 mm

TABLE III
THE FOLLOWING TABLE SUMMARIZES THE EIGHT SESSIONS RECORDED FOR EACH USER. THE CHARACTERISTICS FOR EACH SESSION ARE PROVIDED IN THE COLUMNS. THE FIRST COLUMN SHOWS THE NAME OF THE SESSION, THE SECOND ONE INDICATES THE GRID, THE THIRD ONE DESCRIBES THE FREE OR STATIC HEAD CONDITION WHILE THE LAST ONE SHOWS THE POSITION OF THE USER

name	No. of points	static/free	position
17_points_free	17	free	centered
17_points_static	17	static	centered
65_points_free	65	free	centered
65_points_static	65	static	centered
17_points_bwd	17	static	5 cm backwards
17_points_fwd	17	static	5 cm forwards
17_points_left	17	static	5 cm to the left
17_points_right	17	static	5 cm to the right

The sensor registers the user's position during all the sessions together with the time stamp. In the same manner, for any gazed point 30 images are recorded for which the registration times are saved. Hence, employing a careful synchronization procedure, user images and the head pose information can be paired.

Light conditions were not controlled, thus different light intensities can be observed in the database. However, no complex variations of lights or wild images have been considered.

The objective pursued with this database is to obtain solid conclusions based on real data about gaze estimation methods for low cost systems using controlled head movements. The database provides the perfect framework to test HPE and gaze estimation methods in a reliable manner. Ground truth (GT) values for the head position and the Point of Regard (PoR) are available together with the corresponding images, making it possible to evaluate the contribution of each source of error to the final LoS estimation.

B. Head Pose Estimation

The gaze estimation algorithms proposed in this paper largely rely on the knowledge of the head pose. One of the most effective and computationally assumable algorithms for HPE is POSIT (Pose from Orthography and Scaling with ITerations) method [29]. The method is based on knowing the correspondence between the 2D landmarks in the face image and the corresponding 3D landmarks in the head model assumed for the user, using the camera calibration parameters. If this knowledge is available, the 3D pose of the user with respect to the camera is obtained by means of POSIT [30]. This method assumes a scaled orthographic projection of the object, i.e. head, instead of using perspective projection. This assumption permits to find rotation and translation parameters by solving a linear system. Consequently, considering a calibrated camera, two inputs are required to apply POSIT for HPE: 2D landmarks in the image and their corresponding 3D points in a head model.

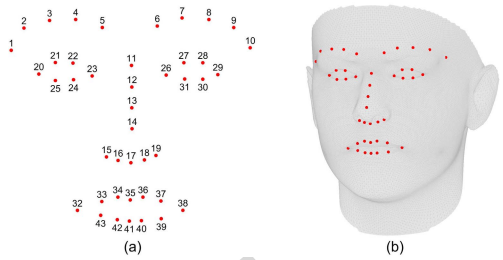

Fig. 9. a) Our HPE method considers the first 43 landmarks detected by IntraFace software, i.e. it obviates the inner landmarks of the mouth area. b) This figure shows the 43 corresponding 3D points in the BFM mean head model.

With the aim to obtain 2D landmarks in the image, IntraFace [31] software is used. IntraFace is a commercial software employing Supervised Descent Method (SDM) in which face tracking is provided together with HPE and gaze direction among others. The authors do not provide detailed information about the implementation of the training procedure. However, it is known that a proprietary version of Scalar Invariant Feature Transform (SIFT) is employed. The detection of 2D landmarks corresponding to characteristic face points resulting from IntraFace is highly accurate and robust. IntraFace detects 49 points from which the first 43 are used (see figure 9a) in our HPE method. Characteristic face points are selected as tracking points assuming that they are the best features to be tracked.

Regarding the 3D head model, alternative options can be chosen. In our method the Basel Face Model (BFM) has been selected [32]. It is a publicly available 3D morphable face model. The model was built based on training data obtained from the 3D scans of 200 subjects, 100 females and 100 males, between 8 and 62 years old, most of them Caucasian. All the scans contained a neutral facial expression and were registered using an Optimal Step Nonrigid ICP Algorithm [33] to ensure an optimized anatomical point correspondence between faces. The faces were parameterized as triangular meshes after registration, resulting in 53,490 vertices described by a coordinate vector $(x_i; y_i; z_i)^T \in R^3$ with an associated colour $(r_i; g_i; b_i)^T \in [0; 1]^3$. Principal component analysis (PCA) was then applied to create an orthonormal basis of 199 principal components of texture and shape, which permits to generate new observations as linear combinations of those components. The average head is obtained from the model as standard for all the users in our database. The 3D landmarks of the model have been carefully identified in order to be associated with the 2D landmarks obtained from IntraFace (see figure 9b).

Thus, once the corresponding points have been identified, POSIT is applied for every acquired frame to obtain the head pose with respect to the camera, named as $HP^{\mathbf{C}}_{est}$,

and to be compared with the ground truth obtained by the sensor transformed into camera coordinates using I2Head database [34].

C. Accurate Iris Detection

The accurate iris center estimation is a key point for the gaze estimation algorithms presented in our work. Different methods can be found in the literature regarding iris center estimation [35], [36]. With the aim to measure the performance of any iris center detection method two alternative approaches are possible. On the one hand, there are several public databases, such as GI4E [37] and others [13], in which iris centers have been manually labeled with varying accuracy. Some of these databases, such as GI4E, contain images from low cost gaze tracking scenarios while others such as LFPW [38] present images from users facing at a camera but not performing an eye tracking session. For those databases in which the landmarks corresponding to iris centers are provided, the accuracy of the iris detection method can be easily measured by comparing the labeled values with the outputs of the detection method. However, the tedious procedure of labeling images makes it difficult to find gaze tracking databases with an acceptable number of images and accurate landmarks.

On the other hand, we find those datasets, such as I2Head or the MIT database [12], devoted to gaze tracking in which no information about the image is provided except for data of gazed points on a screen. The subject is asked to gaze specific points on the screen while the camera is recording. In this manner, the obtained images can be easily correlated with the gazed points. In those cases, the performance of the iris detection algorithm can be potentially determined as its ability to estimate the gazed points correctly.

In our proposal, two methods have been evaluated in order to select the best iris tracking algorithm. First, the aforementioned IntraFace algorithm has been selected because it provides the iris center together with the rest of the face points as output. Second, a method based on Radial Symmetry Transform (RST) has been used [39]. The RST method tries to detect the point in the image with the highest radial symmetry value. The points in the image vote according to their gradient direction and magnitude for varying radii. Assuming that the iris can be approximated by a circle and that the range in which the radius may vary can be standardized, the RST is applied to detect the iris center as the point with the highest number of votes for the correct radius. Both methods assume that the face has been correctly identified and that the eye area has been detected. In the case of IntraFace this is straightforward since all the points are numbered and easily identifiable. In the case of the method based on radial symmetry, the Viola-Jones face detector is applied to detect the eye region [40].

V. Results

The results section is organized as follows: first, the results obtained by our HPE method are shown. To follow, the iris center detection algorithms are evaluated using alternative databases. Finally, the main results obtained by the proposed gaze estimation methods are shown.

A. HPE Results

In order to evaluate the performance of our algorithm, the head pose value obtained, $HP^{\mathbf{C}}_{est}$, is compared with the ground truth stored in the I2Head database, $HP^{\mathbf{C}}$, for every single frame. The proposed method (see section IV-B) to obtain $HP^{\mathbf{C}}_{est}$ has been tested on different datasets, showing a performance improvement of about 60% with respect to state-of-the-art methods [34].

In the database, the sensor origin placed on the top of the head is considered to be the head model origin. However, the POSIT algorithm devoted to estimating the head pose considers the origin of the BFM as the reference point of the head coordinate system, which is located approximately in the midpoint between the ears. In order to carry out a fair comparison, the coordinate system of the head model, **H**, has to be the same. To this end, the relative poses with respect to the pose in the first frame are compared instead of using absolute values. In table IV, the average differential errors for rotation and translation are provided.

The obtained results are fully comparable with the state-of-the-art values which are summarized in the work by Chutorian and Trivedi [41]. As mentioned before the performance of the head pose algorithm employed has been validated in a previous work [34]. However, the tests were carried out using datasets different from I2Head. In order to complete the analysis, our results are contrasted with the ones obtained by IntraFace [31] which can be considered a good performance head tracker for comparison. Head pose is one of the outputs that IntraFace retrieves as result of the tracking. The results obtained by IntraFace for I2Head are (0.92°±0.63°, 2.19°±1.07°, 1.45°±0.45°) for roll, yaw and pitch angles, respectively. It can be easily observed in table IV that our results are significantly better. The average error obtained by Intraface is 1.52°, whereas our method obtains an average error of 0.92°. This supports the results observed in [34], as the improvement given by our algorithm is again of about 60%.

B. Iris Detection

As mentioned before, two algorithms have been selected to detect the iris center, $\mathbf{p^I}$, namely, IntraFace and Radial Symmetry Transform (RST). Two databases are employed to measure the performance of the methods. GI4E database provides accurate labels for the iris center, thus this dataset is used to compare both algorithms in terms of detection error in the image. On the other hand, I2Head is used to evaluate the accuracy, precision and robustness of the algorithms regarding gaze estimation.

The first experiment consists in using the pre-labelled GI4E database in which the center of the irises have been annotated in 1,236 images of users gazing at different points on the screen. The Euclidean distances between the points given by the detection algorithm and the labelled iris centers are calculated for left and right eyes and normalized with respect to the distance between them. The maximum normalized distance is considered to be the detection error for the image. The global accuracy is computed as the mean percentage of images for which the error is below the following thresholds: 0.025,

TABLE IV

THE TABLE SHOWS THE AVERAGE HEAD POSE ESTIMATION ERRORS OBTAINED FOR ALL THE USERS ACCORDING TO THE SESSION. THE TRANSLATION ERROR IN x, y AND z COORDINATES IS PROVIDED TOGETHER WITH THE ORIENTATION ERRORS ACCORDING TO ROLL, YAW AND PITCH ANGLES. IN THE LAST ROW MEAN ERRORS ARE PROVIDED TOGETHER WITH STANDARD DEVIATION VALUES

Session	x (mm)	y (mm)	z (mm)	roll (°)	yaw (°)	pitch (°)
17_points_free	9.17	14.53	2.95	1.35	1.89	1.51
17_points_static	4.31	3.82	4.00	0.33	0.43	0.36
65_points_free	5.31	26.60	5.73	1.45	0.93	2.81
65_points_static	7.69	7.42	2.83	0.59	0.88	0.78
17_points_bwd	4.48	14.96	2.64	0.25	0.35	1.31
17_points_fwd	5.72	5.12	2.08	0.26	0.56	0.54
17_points_left	6.98	16.02	3.85	0.51	0.67	1.46
17_points_right	8.69	14.32	4.76	0.45	0.86	1.60
mean±std	6.54±1.86	12.85±7.36	3.61± 1.21	0.65±0.47	0.82±0.48	1.30± 0.77
	8.91±3.52			0.92±0.59		

TABLE V

AVERAGE VALUES OF GAZE ESTIMATION ERROR FOR CENTERED SESSIONS. THE VALUES TO THE LEFT OF THE SLASH SHOW THE ERROR WHEN GROUND TRUTH VALUES FOR THE HEAD POSE ARE USED WHILE THE VALUES TO THE RIGHT REPRESENT THE ONES OBTAINED IF HEAD POSE ESTIMATIONS ARE EMPLOYED

Session	Geometrical (°)	Interpolation (°)	Compound (°)
17_points_free	9.08/14.90	2.96/2.85	3.85/5.80
17_points_static	8.82/11.57	1.28/1.26	1.29/1.98
65_points_free	12.98/11.44	3.97/3.91	3.02/4.34
65_points_static	11.28/18.73	2.61/2.52	2.36/3.60
mean	**10.54**±3.37/**14.16**±5.98	**2.70**±1.37/**2.64**±1.33	**2.63**±1.37/**3.95**±1.89

0.05 and 0.1. The obtained values show a better performance of IntraFace in comparison to RST. IntraFace presents a global accuracy value of 98.5% whereas it decreases to 90.07% for RST.

Gaze estimation information is used to evaluate the algorithms on the I2Head dataset. The calibration procedure performed for all the gaze estimation methods largely compensates for inaccuracies, not only produced by biases from the gaze estimation method but also for systematic errors of the image processing algorithm. On the other hand, most gaze estimation methods perform an averaging stage, using all the images corresponding to each gazed point, devoted to compensating for the noise inherent to the image. Hence, the accuracy regarding the PoR is not considered to be the only reliable selection criteria for the iris detection method. Alternatively, the method robustness is analyzed based on precision measurements using the interpolation method already described in section III-B. First, the number of outliers is calculated for each method. Thirty images per point are captured and a separate statistical analysis is performed for left and right eyes. An estimation is considered to be an outlier when the distance from the average gazed point on the screen, $\bar{\mathbf{q}}^{\mathbf{S}}$ is larger than the standard deviation of the distribution, $\sigma(\mathbf{q}^{\mathbf{S}})$.

The results show that the method based on RST presents more outliers than IntraFace. In addition, RST presents larger values of $\sigma(\mathbf{q}^{\mathbf{S}})$ and there is less coherence in terms of left and right eye compared to IntraFace. Moreover, the outliers do not present any specific pattern but they are arbitrarily distributed around the average. Once the outliers from both methods are eliminated, the standard deviation values are comparable for both methods. Using the 17-point static session, an analysis is performed trying to identify the best ten images among the thirty for each point to compare IntraFace and RST estimations. Those images with the lowest gaze estimation error are selected. The error is calculated as the sum of errors for both eyes using the Euclidean distance between the estimated PoR and the calibration point position as cost function. As expected, there is a nice coherence between both algorithms regarding the ten best images in both cases, i.e. before and after the removal of outliers. IntraFace method is selected as the most robust and accurate iris detection algorithm among the ones analyzed. It will be used to detect the iris center for the experiments in the next section.

C. Gaze Estimation

In this section the results obtained by each method in terms of gaze estimation are summarized. To this end, data contained in the I2Head database are employed. The head pose with respect to the camera is calculated as shown in section V-A. As mentioned before, I2Head database contains a simplified model for each subject using a reduced number of points, i.e. eye corners and nose tip that are annotated in 3D with respect to the head. Separate models are calculated for left and right eyes, thus, a binocular gaze estimation is performed by averaging the samples obtained for both eyes.

The three methods proposed, i.e. geometrical, interpolation and compound methods, require a user calibration procedure in which alternative parameters for each model are calculated. To this end, the *17_point_static* session is employed for calibration. Then, once the parameters for each model are estimated, gaze accuracy is calculated for the rest of the sessions (see table V). Two different scenarios are evaluated: first, ground truth values are used as head pose by means of the sensor, $HP^{\mathbf{C}}$. Second, the head pose values obtained by our HPE method, $HP^{\mathbf{C}}_{est}$, are used as input to the gaze estimation methods. During the calibration stage, gaze estimation accuracy is optimized. Accuracy is calculated as the

TABLE VI

AVERAGE VALUES OF GAZE ESTIMATION ERROR FOR EXTREME MOVEMENTS SESSIONS. THE VALUES TO THE LEFT OF THE SLASH SHOW THE ERROR WHEN GROUND TRUTH VALUES FOR THE HEAD POSE ARE USED WHILE THE VALUES TO THE RIGHT REPRESENT THE ONES OBTAINED IF HEAD POSE ESTIMATIONS ARE EMPLOYED

Session	Geometrical (°)	Interpolation (°)	Compound (°)
17_points_bwd	12.32/16.10	4.85/4.59	2.27/5.16
17_points_fwd	9.61/15.33	3.53/3.31	2.46/ 7.24
17_points_left	9.46/13.53	5.41/5.24	5.00/ 8.99
17_points_right	10.46/12.68	5.95/5.43	6.11/7.91
mean	**10.46**±1.90/**14.41**±2.41	**4.94**± 2.26/**4.64**±1.19	**3.96**±2.69/**7.33**±4.28

angular absolute difference between real and estimated gaze directions. In the case of interpolation and compound methods, the coefficients of a polynomial are calculated. Moreover, for the compound model, the labelled eye corners are used to calculate the eyeball center, $\mathbf{E}$, as the mean point between the corners. In the geometrical model, angle κ, eyeball center and radius r are the unknown model parameters. During the experiments, it has been observed that the calibration of the geometrical model is highly sensitive to the initial conditions, especially to the initial value of the eyeball center. For this reason, calibration is carried out using two stages for this model. In the first step, a simulated annealing algorithm is employed in which the initial eyeball center is calculated as the average value of the 3D corners obtained from the simplified eyeball model. Additionally, the initial value of the radius, r, is obtained as the half distance between the estimation of the initial value of the eyeball center and eye corner in 3D. Once the simulated annealing is concluded, a further more precise minimization algorithm is employed. A Levenberg-Marquadt procedure is carried out using as initial values the outputs obtained in the previous step.

Tables V and VI show the average accuracy values of the alternative methods. The columns contain the values obtained by each method in the different sessions. The value to the left of the slash is the error obtained by the corresponding method when ground truth head pose, $HP^{\mathbf{C}}$, is used, while the value to the right represents the accuracy if the estimated head pose value, $HP^{\mathbf{C}}_{est}$, is employed. Table V shows the accuracy values for sessions carried out in a centered position of the head and table VI provides the results obtained when the user performs significant translation movements from the calibration position. Thus, the influence of large head movements can be more clearly appreciated. The accuracy is shown in degrees since this is the standard way of representing it, so independent from the screen resolution and the working distance.

As expected, the gaze estimation errors obtained are not comparable to the ones obtained by high resolution systems, but they are fully comparable with the ones obtained by alternative approaches devoted to low resolution eye tracking [7], [13], [24]. It is straightforward to observe that the smallest errors are obtained for the calibration session, i.e. *17_points_static*. Consequently, in general, lower errors are observed for the *65_points_static* session compared to the free sessions in the centered position (see table V). Errors shown in table V are generally lower than the ones reported in the literature and described in the introduction, i.e. 4°-6°. However, in order to perform a fair comparison, free head movements sessions and sessions in which extreme movements from the calibration position are performed, summarized in table VI, should be taken under consideration. Except for the geometrical model, compound and interpolation models present fully competitive results when compared to the state-of-the-art.

In order to validate our method with other state-of-the-art database, we have also tested it on the MPIIGaze dataset [13]. This database contains images from fifteen users gazing at their gadgets during different everyday tasks. The MPIIGaze was conceived to be used as a large scale dataset for learning-based approaches, such as CNNs and many works devoted to using machine learning techniques for gaze estimation employ MPIIGaze as a testing benchmark. Therefore, the aim of the captured images is to provide the largest possible variability and representability, i.e. including images of varying quality, illumination and blurring degree. The *annotation subset* of the dataset is used for this evaluation because it is the only set for which iris center and eye corner landmarks have been manually annotated. The dataset provides these data for more than ten thousand images and the accuracy of the labelling procedure is not homogeneous through the annotated subset. There are additional factors that make this comparison a challenging task. No ground truth values for head pose are provided but estimated values for rotation and translation of the user with respect to the camera. Head pose is calculated using a method based on a six-point face model that is described in their paper [13], but no accuracy values are provided. Many of the annotated images are cropped, i.e. only the eye area is provided. Consequently, we cannot apply our HPE method, since it requires the whole face image as input. Contrarily to I2Head database, a single image per gazed point is given. It is of general practice to employ several images per point to average the result and make gaze estimation more robust in the presence of noise. Taking into account the characteristics of the images included in the database, it would be desirable to have more images per point available. Moreover, since the recordings of the database are conducted in everyday situations, it is not feasible to select those images belonging to a regular grid covering the whole screen that are normally required for calibration purposes. Since the database was constructed to be principally used by other methods under different premises the comparison represents a critical task. However, being the main reference database of the state-of-the-art, an evaluation is performed. The interpolation method is selected to be applied to the MPIIGaze dataset due to its simplicity and lower dependency on head pose values.

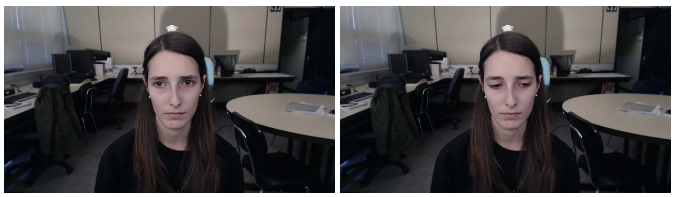

Fig. 10. The image on the left shows a user gazing to a point in the upper part of the grid, while the image on the right shows the same user gazing to a lower point.

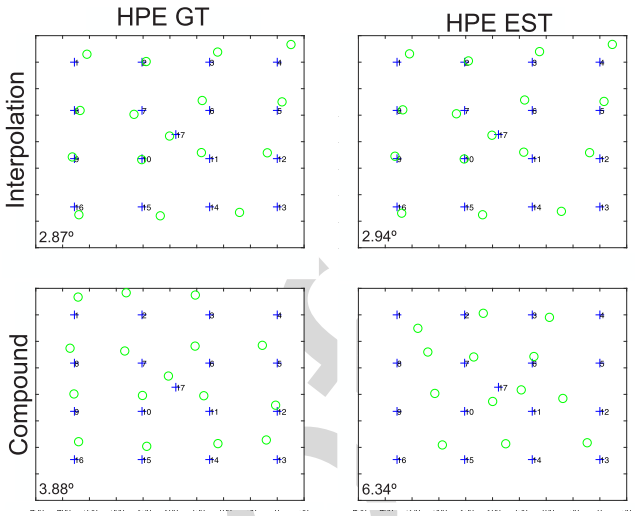

Fig. 11. Interpolation and compound models are compared for the same user using the *17_points_free* session. The blue crosses are the ground truth positions of the gazed points while the green circle shows the estimated PoR. In the upper part of the figure the results of the grid for the interpolation model are shown, while the ones arisen for the compound model are provided below. The left column refers to the results obtained using ground truth values for head pose, and the right column shows the results when estimated head pose values are employed. It can be observed that the effect on the interpolation model is negligible, while it is more significant on the compound model. Average accuracy errors are provided.

It is assumed that the results obtained for the interpolation model can be extrapolated for the rest of the models. Since no calibration grid is available, for each user half of the data are used for calibration purposes and the other half for the testing stage. The results should be compared with the ones obtained for I2Head in moving scenarios (table VI). The average gaze estimation error obtained is 7.49°±0.76°. Since no averaging process is available, an outlier removal stage is included to neglect outliers (values greater than the 0.8 quantile are considered to be outliers) and carry out a fair comparison. The results obtained after outlier removal are 6.07°±1.32°. They are slightly higher values than the ones obtained for I2Head but, taking into account the type of images of the database, this increment could be expected. Additionally, this result is fully comparable with the reference values described in the introduction.

It has been observed in the experiments carried out on the I2Head database that the most important source of error are inaccuracies arisen in the image processing stage. The use of a high number of images per point leads to reduce the noise regarding the iris center estimation. However, for several images, it is not noise the issue affecting the accuracy, but the failures of the algorithm in certain circumstances. It is worth mentioning that, due to the position of the camera, there are more frequent tracking errors in images in which the user gazes at the lower part of the grid, i.e. when the eyelids occlude part of the eye it is more difficult to conduct an accurate tracking of the iris center (see figure 10).

The design of the models proposed in this work is based on the knowledge acquired in high resolution systems where their validity has been demonstrated. The assumptions made for the alternative models also contributed to some extent to the error, but it is negligible compared to the one resulting from the landmarks tracking in the image. The inaccuracies in the landmarks detection affect both, the head pose and gaze direction estimation, being the accurate detection of the iris center key for all the models. It is remarkable to observe that those approaches having higher geometrical modeling, such as geometrical and compound methods, present higher errors due to an inaccurate tracking as it can be observed in the errors arisen for the geometrical model for which non-admissible errors are obtained. The same error in landmark tracking produces extremely higher errors in the gaze value for this model. However, the compound and interpolation models present more assumable errors in the same scenario. Moreover, this hypothesis is reinforced if we focus on the centered position, i.e., table V, and we compare the errors using the ground truth value of the head pose and those using the estimated head position. It is observed that the geometrical and compound models are the ones presenting the highest increments while the interpolation model presents a lower sensitivity to errors in the head position. In figure 11 the behavior of the compound and interpolation models is compared when ground truth and estimated values for head pose are used. It can be observed that the compound model increases the error when estimated head values are used while for the interpolation model no significant increments can be distinguished.

The same effect can be observed in table VI, in which the sessions having strong displacements from the calibration position are shown, i.e. introducing the estimated value of the head position leads to an increment of the error in similar proportion for geometrical and compound models.

Contrarily, the geometrical model presents a higher robustness in the presence of head movements. Average error values in tables V and VI are comparable for the geometrical model, meaning that it presents a higher tolerance to user's extreme translation movement. This conclusion resembles partially the behavior of geometrical models in high resolution systems. In contrast, the interpolation model almost duplicates the error in the presence of extreme movements compared to the values in the centered position. Probably, the compound model is the one presenting the best balance between accuracy and robustness against head movements. An ideal estimation of the head pose, i.e. ground truth, for the compound model would lead to errors comparable to the ones in the centered position, especially for the sessions presenting forward and backward movements.

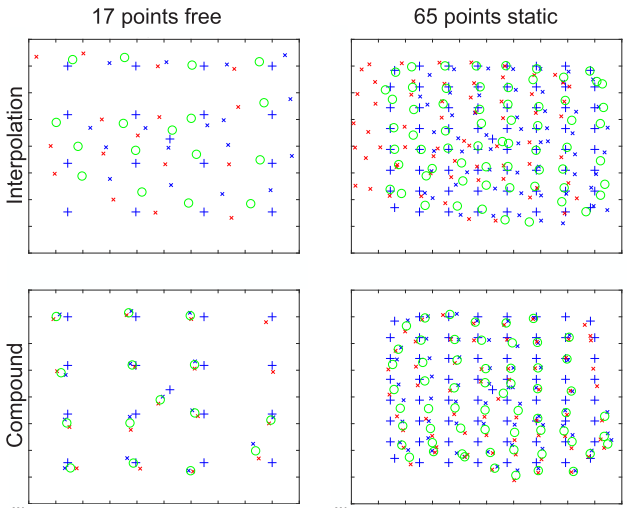

Fig. 12. The first row shows the estimations obtained by the interpolation model while the second row shows the result for the same user and sessions using the compound model. The blue crosses are the ground truth positions of the gazed points while the red and blue x-s show the estimations for the left and right eyes, respectively. Finally, the green circle shows the average between both eyes. Sessions *17_points_free* and *65_points_static* have been selected as example.

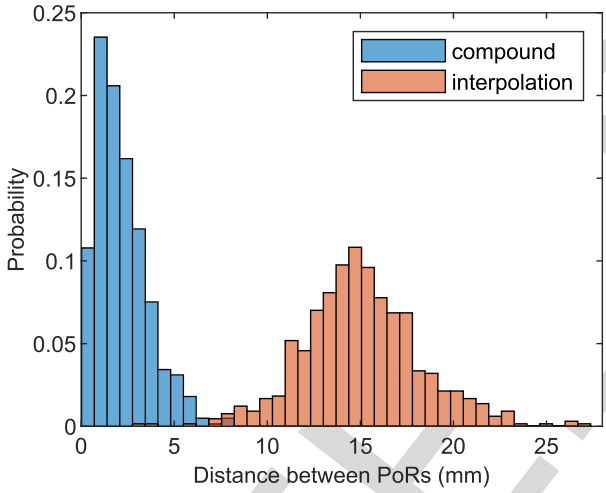

Fig. 13. The consensus between left and right eyes is shown for compound and interpolation models. The difference between the PoRs estimated for left and right eyes is smaller when the compound model is used.

From the average errors it cannot be observed a remarkable property of the compound method in comparison with the interpolation one. The compound model presents a considerably higher consensus between the left and right eyes. In other words, as the estimated PoR is calculated as the average between left and right eyes, a further step is required to evaluate the goodness of the models for each eye separately. In figure 12 the output for two sessions, i.e. *17_points_free* and *65_points_static*, can be observed for the same user using compound and interpolation models. The figures on the left are the grids obtained for the *17_points_free* session using interpolation and compound models while the figures on the right show the estimations for the *65_points_static* sessions. From the figure, it can be clearly seen that the consensus between left and right eyes is significantly higher for the compound model, which is a valuable property to take under consideration. In figure 13 the distribution of the difference between left and right eye estimations can be observed for compound and interpolation models. In the case of the compound model the mean consensus is about 2.5°, increasing significantly in the case of the interpolation model.

VI. Conclusions

From the gaze estimation results obtained, several conclusions can be drawn. Regarding head pose estimation, the algorithm shows outstanding values compared with the other state-of-the-art algorithms in the literature outperforming the results by 60%. As expected, the lower resolution of the image makes it difficult to obtain an accurate detection of face landmarks resulting in higher errors in the gaze estimation stage. Moreover, those models employing a higher geometrical content present a significantly higher sensitivity to errors in the tracking stage, resulting in non-admissible errors for the case of the geometrical model. The interpolation model, which is the one with the least geometrical information, is more robust against image inaccuracies; however, it doubles the error in presence of severe translation head movements, from 2.70° in the calibration position to almost 5° when severe head movements are performed. In contrast, the geometrical models present better robustness in presence of user's movement. These conclusions firmly support one of the most well-known ideas of eye tracking technology, largely validated in high resolution settings: being the compound model the one with the best balance between accuracy and robustness. Both interpolation and compound models have shown results in the range of 2°-5° assuming an accurate HPE, i.e. significantly good when compared to the state-of-the-art.

Above all, the main conclusion obtained is that improving the accuracy of landmark detection in the image, particularly the tracking of the iris center, is one of the main obstacles to overcome when approaching low resolution scenarios. The error arisen due to the models is negligible compared to the one produced by inaccuracies in the image. Obtaining more accurate and precise image processing methods for low resolution systems is a challenge. Thus, further investigations in low resolution gaze estimation are required to analyze techniques oriented towards artificial intelligence or geometry-based among others.

References

[1] P. Majaranta, H. Aoki, M. Donegan, D. W. Hansen, and J. P. Hansen, *Gaze Interaction and Applications of Eye Tracking: Advances in Assistive Technologies*, 1st ed. Hershey, PA, USA: IGI, 2011.

[2] W. Fuhl, M. Tonsen, A. Bulling, and E. Kasneci, "Pupil detection for head-mounted eye tracking in the wild: An evaluation of the state of the art," *Mach. Vis. Appl.*, vol. 27, no. 8, pp. 1275–1288, Nov. 2016.

[3] Z. Zhu and Q. Ji, "Eye gaze tracking under natural head movements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Washington, DC, USA, vol. 1, Jun. 2005, pp. 918–923.

[4] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," Aug. 2017, *arXiv:1708.01817*. [Online]. Available: https://arxiv.org/abs/1708.01817

[5] J. J. Cerrolaza, A. Villanueva, and R. Cabeza, "Study of polynomial mapping functions in video-oculography eye trackers," *ACM Trans. Comput.-Hum. Interact.*, vol. 19, no. 2, p. 10, Jul. 2012.

[6] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 1124–1133, Jun. 2006.

[7] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 802–815, Feb. 2012.

[8] E. Wood and A. Bulling, "EyeTab: Model-based gaze estimation on unmodified tablet computers," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*. New York, NY, USA, Mar. 2014, pp. 207–210.

[9] F. Vicente, Z. Huang, X. Xiong, F. D. L. Torre, W. Zhang, and D. Levi, "Driver gaze tracking and eyes off the road detection system," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2014–2027, Aug. 2015.

[10] S. J. Lee, J. Jo, H. G. Jung, K. R. Park, and J. Kim, "Real-time gaze estimator based on driver's head orientation for forward collision warning system," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 254–267, Mar. 2011.

[11] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 478–500, Mar. 2010.

[12] K. Krafka *et al.*, "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Washington, DC, USA, Jun. 2016, pp. 2176–2184.

[13] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4511–4520.

[14] C. D. McMurrough, V. Metsis, J. Rich, and F. Makedon, "An eye tracking dataset for point of gaze detection," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*. New York, NY, USA, Mar. 2012, pp. 305–308.

[15] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "EYEDIAP: A database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*, New York, NY, USA, Mar. 2014, pp. 255–258.

[16] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162–175, Jan. 2019.

[17] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: Passive eye contact detection for human-object interaction," in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol. (UIST)*. New York, NY, USA, Oct. 2013, pp. 271–280.

[18] Q. Huang, A. Veeraraghavan, and A. Sabharwal, "TabletGaze: Dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets," *Mach. Vis. Appl.*, vol. 28, nos. 5–6, pp. 445–461, 2017.

[19] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3D gaze estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Washington, DC, USA, Jun. 2014, pp. 1821–1828.

[20] L. Świrski and N. A. Dodgson, "Rendering synthetic ground truth images for eye tracker evaluation," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*. New York, NY, USA, Mar. 2014, pp. 219–222.

[21] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling, "Learning an appearance-based gaze estimator from one million synthesised images," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*, P. Qvarfordt and D. W. Hansen, Eds., Mar. 2016, pp. 131–138.

[22] X. Xiong, Z. Liu, Q. Cai, and Z. Zhang, "Eye gaze tracking using an RGBD camera: A comparison with a RGB solution," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Adjunct*, New York, NY, USA, Sep. 2014, pp. 1113–1121.

[23] K. Wang and Q. Ji, "Real time eye gaze tracking with 3D deformable eye-face model," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1003–1011.

[24] S. Park, X. Zhang, A. Bulling, and O. Hilliges, "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*, Warsaw, Poland, Jun. 2018, pp. 1–10.

[25] A. Villanueva and R. Cabeza, "A novel gaze estimation system with one calibration point," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 1123–1138, Aug. 2008.

[26] M. Böhme, M. Dorr, M. Graw, T. Martinetz, and E. Barth, "A software framework for simulating eye trackers," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*. New York, NY, USA, Mar. 2008, pp. 251–258.

[27] L. Sesma, A. Villanueva, and R. Cabeza, "Evaluation of pupil center-eye corner vector for gaze estimation using a Web cam," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*. New York, NY, USA, Mar. 2012, pp. 217–220.

[28] I. Martinikorena, R. Cabeza, A. Villanueva, and S. Porta, "Introducing I2Head database," in *Proc. 7th Int. Workshop Pervasive Eye Tracking Mobile Eye Based Interact. (PETMEI)*, Jun. 2018, p. 1.

[29] D. F. Dementhon and L. S. Davis, "Model-based object pose in 25 lines of code," *Int. J. Comput. Vis.*, vol. 15, no. 1, pp. 123–141, Jun. 1995.

[30] M. Ariz, J. J. Bengoechea, A. Villanueva, and R. Cabeza, "A novel 2D/3D database with automatic face annotation for head tracking and pose estimation," *Comput. Vis. Image Understand.*, vol. 148, pp. 201–210, Jul. 2016.

[31] X. Xiong and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Washington, DC, USA, Jun. 2013, pp. 532–539.

[32] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. 6th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*. Washington, DC, USA, Sep. 2009, pp. 296–301.

[33] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid ICP algorithms for surface registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Washington, DC, USA, Jun. 2007, pp. 1–8.

[34] A. Larumbe, M. Ariz, J. J. Bengoechea, R. Segura, R. Cabeza, and A. Villanueva, "Improved strategies for HPE employing learning-by-synthesis approaches," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1545–1554.

[35] W. Fuhl, D. Geisler, T. Santini, W. Rosenstiel, and E. Kasneci, "Evaluation of state-of-the-art pupil detection algorithms on remote eye images," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct*. New York, NY, USA, Sep. 2016, pp. 1716–1725.

[36] O. Ferhat and F. Vilariño, "Low cost eye tracking: The current panorama," *Comput. Intell. Neurosci.*, vol. 2016, Feb. 2016, Art. no. 8680541. AQ:5

[37] A. Villanueva, V. Ponz, L. Sesma-Sanchez, M. Ariz, S. Porta, and R. Cabeza, "Hybrid method based on topography for robust detection of iris center and eye corners," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 9, no. 4, pp. 25-1–25-20, 2013.

[38] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Washington, DC, USA, Jun. 2011, pp. 545–552.

[39] E. Skodras and N. Fakotakis, "Precise localization of eye centers in low resolution color images," *Image Vis. Comput.*, vol. 36, pp. 51–60, Apr. 2015.

[40] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Washington, DC, USA, Dec. 2001, pp. 511–518.

[41] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

Ion Martinikorena was born in Spain, in 1986. He received the B.E. and M.E. degrees from the Public University of Navarre, Spain, in 2013 and 2017, respectively. He was a member of the Gaze Interaction for Everybody (GI4E) Group, from 2016 to 2017. His current research interests include computer vision, machine learning, deep learning, and gaze estimation.

Andoni Larumbe-Bergera was born in Spain, in 1993. He received the B.E. degree in telecommunications engineering and the M.E. degree in biomedical engineering from the Public University of Navarre, Spain, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Gaze Interaction for Everybody (GI4E) Research Group. He joined the Public University of Navarre in 2016. His current research interests include computer vision, machine learning, deep learning, and gaze estimation.

AQ:6

Mikel Ariz received the B.Eng. degree in telecommunication engineering from the Public University of Navarre, Pamplona, Spain, in 2008, the master's degree in biomedical engineering from The University of Melbourne, Australia, in 2010, and the Ph.D. degree in engineering (image processing and computer vision) from the Public University of Navarre, in 2016. Since 2015, he has been a Biomedical Image Analyst and a Researcher with the Imaging Platform and also with the Laboratory of Preclinical Models and Analysis Tools, Center for Applied Medical Research, Pamplona, Spain. He is currently a part-time Lecturer with the Department of Pathology, Anatomy and Physiology, and also with the School of Engineering (TECNUN), University of Navarra. His current research interest includes biomedical image processing.

Sonia Porta received the B.Sc. and Ph.D. degrees in physics from the University of Zaragoza, Spain, in 1986 and 1993, respectively. She was with local company of loudspeakers quality control and PCB design for acoustic applications. She was a Research Fellow with the European Union and spent two years in Oxford Brookes University, U.K., collaborating with the 1994 ISCAS Symposium held at London, where she co-edited the book *Circuits and Systems Tutorial*. Since 1995, she has been an Associate Professor with the Department of Electrical Engineering and Electronics, Technical School of Telecommunications Engineering, Public University of Navarra (UPNA). Her current research interests include image and signal processing.

Rafael Cabeza was born in Soria, Spain, in 1967. He received the M.Sc. degree in physics from the University of Zaragoza, Zaragoza, Spain, in 1990, and the Ph.D. degree (Hons.) in telecommunications engineering from the Public University of Navarre, Pamplona, Spain, in 1996. From 2000 to 2010, he was the Head of the Signal Processing, Microelectronic, and Instrumentation Research Group, Public University of Navarre, where he has been an Associate Professor with the Department of Electrical and Electronics Engineering. His current research interests include image processing and computer vision.

Arantxa Villanueva was born in Pamplona, Spain, in 1974. She received the M.Sc. and Ph.D. degrees in telecommunications engineering from the Public University of Navarre, in 1998 and 2005, respectively. She is currently an Associate Professor with the Electrical and Electronic Engineering Department, Public University of Navarre. She is also a Researcher with the Gaze Interaction for Everybody (GI4E) Group. Her current research interests include image processing and gaze estimation.