Signal Processing: *Image Communication* 36 (2015) 29–42

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

# On visual gaze tracking based on a single low cost camera

Evangelos Skodras*, Vasileios G. Kanas, Nikolaos Fakotakis

*Department of Electrical and Computer Engineering, University of Patras, Patras, Greece*

ARTICLE INFO

*Article history:*
Received 26 November 2014
Received in revised form
9 April 2015
Accepted 27 May 2015
Available online 4 June 2015

*Keywords:*
Gaze estimation
Gaze tracker
Eye tracking
Human computer interaction
HCI

ABSTRACT

Gaze tracking technologies provide an unconventional way of human–computer interaction, envisaged to advance practical applications and industrial products in a multitude of fields. The success of such systems depends on selecting the best calibration setup and image features that correspond to a person's line of sight. The purpose of this study is to estimate eye gaze from a single, low cost web-cam, under natural lighting. Facial traits are extracted from the sensory data, from which distance vectors related to gaze are derived. Different experimental setups are studied to evaluate the robustness of the proposed method with respect to various calibration setups, camera position and head movements. The use of new additional features improves the modeling of the subtle eye movements in the vertical direction, while a new calibration setup is proposed that further enhances the performance. The results demonstrate that the proposed framework is able to track gaze with good accuracy, consolidating the use of inexpensive equipment and techniques towards an ever-expanding range of gaze tracking applications.

## 1. Introduction and motivation

The swift growth of functional sophistication in computing over the last decades has inevitably induced a growing interest in improving all aspects of interaction between humans and computers. This emerging field is gaining momentum for scientists across several different disciplines such as computer science, engineering, psychology and neuroscience. Along with speech, eye gaze comprises the most natural and comfortable means for human–computer interaction (HCI), giving rise to the ever-growing interest to develop systems that take advantage of gaze tracking technologies.

Eye gaze is defined as the direction of a person's line of sight, revealing a person's focus of attention. It comprises a significant source of information about the cognitive and affective state of human beings, providing implicit cues of intention and interest. As a control input, in conjunction with the standard input methods, gaze can greatly increase efficiency and usability. Gaze monitoring can be applied in a wide range of applications [1], including non-glasses type 3D technologies [2], monitoring of drivers attention and vigilance [3–5], visual attention analysis (*e.g.* for marketing purposes [6]), interactive gaze-based interfaces for disabled people [7], diagnostic purposes [8,9] and attentive HCI interfaces [10,11].

Despite active research in the field, ubiquitous gaze tracking is beyond the grasp of the current systems. The vast majority of research in academia and industry is directed towards gaze tracking using active light sources, *i.e.* infrared (IR) illumination, achieving high accuracy rates [12,13]. Hitherto, numerous commercial products making use of this well known approach are already on the market. However, active light approaches require dedicated hardware equipment which is usually high priced and the intrusiveness of which is controversial. Moreover, as they usually require a controlled environment to prevent undesired reflections in the eyes, their applicability during day time is precluded. Other common approaches employ 3D techniques [14] (using multiple cameras or depth sensors) and wearable devices such as helmets or glasses [15], being cumbersome for the users. Universal gaze tracking from completely unobtrusive, remotely located low-cost sensors (*e.g.* web-cams) still remains one of the most sought-after goals among researchers.

* Corresponding author. Tel.: +30 6972965909.
*E-mail addresses:* evskodras@upatras.gr (E. Skodras), vaskanas@upatras.gr (V.G. Kanas), fakotaki@upatras.gr (N. Fakotakis).

http://dx.doi.org/10.1016/j.image.2015.05.007
0923-5965/© 2015 Elsevier B.V. All rights reserved.

Although web-cam based gaze trackers have so far demonstrated inferior performance compared to active light approaches, they are apt in applications where accuracy can be traded off for low cost, simplicity and flexibility. In this paper, a feature-based gaze tracking system is proposed, based on a single web-cam. Instead of the most common approach of detecting eye corners in every frame, in our framework, salient feature points serving as anchor points are utilized to extract discriminative features. Additional features corresponding to the vertical position of the eyelids are proposed in order to increase accuracy and robustness. Despite the number of gaze tracking approaches in the literature, a common basis is not yet established, since standard databases for evaluating gaze direction have only recently emerged. For this purpose, different calibration setups and the influence of head movements are extensively evaluated. Furthermore, a new arrangement of the calibration points is presented, contributing towards an even higher accuracy rate. In summary, the characteristics of the proposed system, which also correspond to the desired attributes of the entirety of gaze trackers, are the following:

- It estimates gaze with fairly high accuracy; this is suitable for most gaze tracking applications.
- It presents minimal intrusiveness and obstruction, given that only a single camera is required. Not requiring any special equipment, gaze tracking is not bound to a specific *ad hoc* computer specialized for this purpose but can run universally.
- The setup is straightforward and flexible. The procedure is autonomous (no expectation of manual initialization), any camera available can be used and is free from the need to zoom into the eyes or perform any other special actions.
- It works under different illumination conditions and is also independent of the special characteristics of each subject.
- It is demonstrated to function for lower resolution cameras, thus constituting a low-cost approach and offers the capability for real time processing.

The layout of the paper is organized as follows: in Section 2 a review of the literature in the area of natural lighting approaches (without IR illumination) and not using any wearable equipment is presented. The model used for gaze estimation is also defined in the same section, while in Section 3 an in-depth description of the proposed algorithm is given. Section 4 presents the experimental setup and Section 5 the obtained results, as well as a meticulous analysis of them. Finally, the discussion of the main points in Section 6 is followed by a recapitulation in Section 7 which concludes this work.

## 2. Related work and background

### 2.1. Gaze estimation methodologies

The problem of gaze estimation is an active research topic with several recent publications [16], the overwhelming majority of which are using hardware-based approaches (*i.e.* IR light sources, high-resolution cameras, multiple cameras and wearable equipment). The focus of the current overview is on approaches working under natural illumination, using a single, remotely located camera.

The current subset of gaze estimation methods can be subdivided into two broad categories, *i.e. feature-based* and *appearance-based* methods [16].

*Feature-based* methods use computer vision techniques to extract and track local eye features such as eye centers, corners and contours. The extracted features can be used to derive feature vectors which can be directly related to gaze. According to the approach used for relating the image features with gaze direction, *feature-based* methods can be further divided into *geometric* (or *model based*) and *interpolation based* (or *regression based*). *Geometric* methods [4,17–20] compute directly the gaze direction from the image features based on a geometric model of the eye, while *interpolation based* methods [21–26] built a mapping function between them, using parametric (*e.g.* polynomial) or non-parametric forms (*e.g.* neural networks).

Torricelli et al. [21] use image processing algorithms to extract and track the eye features and perform mapping with gaze direction using neural networks. In [22], Zhu et al. detect and track eye centers and corners using subpixel accuracy and a linear interpolation model for inferring gaze coordinates. The works in [17,18] use geometric models which rely upon facial feature tracking for estimating head pose and eye orientation. The feature vectors between eye centers and eye corners are used in [27,23,24] to derive gaze estimation through 2D interpolation mapping. A comparison between a common polynomial mapping function and a geometric model is performed in [28], giving a slight edge to the latter. Ishikawa et al. [4] use Active Appearance Models (AAM) to detect and track facial points and employ an edge-based algorithm for iris refinement. They subsequently employ geometric models to derive gaze direction combining head pose and eyeball orientation. Authors in [29] also use AAM for iris and eyelid tracking in order to derive gaze information. The gaze angle is geometrically defined combining head pose and eyes position information. In the work of Salam et al. [30] the head pose is derived using a 2.5D global AAM, while a multi-texture AAM is used for iris localization. The contribution of each eye to the final gaze direction is weighted depending on the detected face orientation. In general, global appearance-based methods are very robust in detecting the overall rough positions of the facial features. However, as they depend on the convergence of the full model (*i.e.* by satisfying a minimization function or reaching a maximum number of iterations), they do not ensure localization of each feature with high precision, thereby adversely affecting the gaze estimation accuracy. Chen and Ji in [19] also use a geometric model to localize facial points, manually extract pupil centers and build a 3D gaze estimation model, tailored with person-specific eye parameters. Heyman et al. [20] track the 3D pose of the head using Canonical Correlation Analysis and extract the positions of the irises using blob detection and 4-connected component labeling. Valenti et al. in [25] employ their eye detection approach based on isocenters, to also detect eye corners which are used as *anchor* points for gaze estimation. Given the vectors between the detected eye centers and eye corners they perform 2D linear mapping to screen coordinates. Their work is extended in [31] where pose estimation and eye localization algorithms are combined so that they complement each other, thus increasing the system's performance. In their work in [32] they combine eye gaze information derived from their proposed eye gaze tracker and a commercial gaze tracker with saliency maps (probability maps representing the likelihood of receiving eye fixations). Shape modeling of the iris methods have been often employed, many of which exploit the fact that when the iris orientation changes, the shape of the iris appears to deform from circular to elliptical. However, for such shape modeling approaches, higher resolution images are required. In [33,34] the gaze direction is inferred by estimating the shape of the iris through ellipse fitting. Active contour tracking using particle filters is used in [35] to built a generic system, working on various setups and conditions, investigating also the lower bound calibration requirements.

*Appearance-based* or holistic approaches incorporate eye information implicitly by using the intensity distribution or filter responses of the eye area. They usually classify gaze according to the appearance of the eye area for each direction. The systems described in [36–40] learn the gaze direction by modeling the corresponding eye appearance. In the work of Schneider et al. [40] several regression techniques are evaluated, modeling the appearance of the eyes (in terms of features such as Histogram of Oriented Gradients and Local Binary Patters) when gazing at different directions, in order to build a calibration-free system. Hansen et al. [41] use an active appearance model of the eye using shape and texture properties to be used with an eye typing interface. Saliency information of the displayed images is aggregated with an appearance-based gaze estimator in [42]. The main drawbacks of holistic approaches is that the appearance of the eyes is significantly affected by the head pose and usually only a limited number of discrete eye directions are modeled. These limitations can preclude their use in many applications, giving rise to the more prevalent use of *feature-based* approaches.

### *2.2. Model assumptions*

A person's gaze is determined by the combination of head pose (position and orientation) and the orientation of the eyeball [43]. Head pose determines the direction of gaze in a coarse scale while the orientation of the eyeballs provide information for fine gaze direction. In the general case, when fixating at a specific target, first the head moves to a comfortable position before the eyes are oriented towards it, since extreme eyeball orientation angles cause user discomfort.

Head pose invariance in gaze estimation either requires special hardware configurations (such as helmets and, glasses) or prior knowledge of the camera parameters and geometry. Head pose estimation using monocular information and no special lighting still remains a challenging research topic [44], requiring complex, computationally demanding algorithms for modeling. Inaccurate estimations of head pose may trigger much larger errors in final gaze estimation, reducing the system's overall accuracy and robustness. In view of this, information of head pose in gaze models is more commonly incorporated implicitly through the mapping function. When looking at a computer screen from a typical viewing distance (in our case ≈50 cm), there is no need for the head to be displaced in order to reach a comfortable position since the eyes do not reach extreme orientation angles, even when fixating near the screen borders. Taking the aforementioned facts into consideration, we herein introduce the simplification to discard head pose information and regard the head as stable throughout the eye tracking session. Given that minor head movements do not have a noticeable influence on the outcome, the head does not have to be fixated.

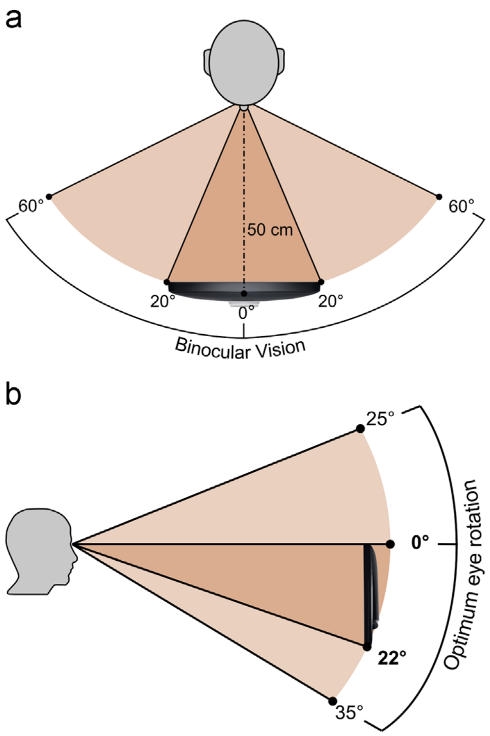

**Fig. 1.** (a) Horizontal and (b) vertical field of view.

Another assumption is that the eyes do not rotate inside the ocular cavities but just shift. This assumption holds true for small rotation angles $\theta$ where $\theta \approx \sin(\theta)$ is valid. While more complex models of the eyeball would require higher resolution images and larger computational times, the current approximation greatly simplifies calculations, introducing negligible errors (for gaze direction within $\pm 15^\circ$ the error is less than $0.17^\circ$ [22]).

### *2.3. Field of view*

According to studies on the human visual field [45], while staring straight ahead keeping the head fixed, there is approximately a horizontal vision span of 180° (90° on each side) and a vertical span of 120° (50–55° looking above and 60–70° looking below). However, the visual field where both eyes are used together, referred to as binocular or 3D field of view, is quite narrower. The binocular field of view spans at 60° on each side on the horizontal direction, 25° up and 35° down on the vertical direction [45].

Considering that the approximate distance from the subject to the computer screen is 50 cm and given the screen size for the specific experiment, the angle span required is 40° horizontally and 22° vertically (Fig. 1), computed as

$$\theta = 2\arctan\frac{W}{2D}, \quad \phi = 2\arctan\frac{H}{2D}$$

where $W$ and $H$ denote the width and height of the screen, respectively, and $D$ the distance of the user from the screen. The computed visual angles indicate that the eyes do not need to move to extreme positions and thus the head rests in a comfortable position, without the urge to move throughout the experiment.

From an eye physiology perspective, high-acuity vision which corresponds to 5.5° of the visual angle is owing to the *fovea*, a cone-concentrated area in the retina. In the center of the *fovea* lies the *foveola*, a rod-free area responsible for the highest visual acuity in the eye which corresponds to $\approx 1.7$–$2°$ of the visual field. Within this range everything can be seen with the highest acuity without requiring any saccade, thus the angle of 2° can be regarded as a good accuracy benchmark for gaze tracking systems.

## 3. System architecture

The proposed gaze estimation pipeline is depicted in Fig. 2. During the calibration phase, a set of known points on the screen is displayed to the user. Upon gaze fixation at each point, an image is captured. The goal of the calibration procedure is for the computer to learn how the eyes look like (in terms of features) so as to derive a mapping function between the image data and screen coordinates. At the testing phase, the inverse procedure is followed; given the (unknown) image data and the regression model built, the system is predicting how the extracted test image features are translated into screen coordinates. Initially, the information of the sensory device is transformed from pixel level to a higher representation level, where specific facial characteristics are recognized *i.e.* the location of the face and the positions of the eye centers and the eyelids. Subsequently, the distance vectors between the located *moving* points and stable facial points (*anchor* points) are utilized to form the feature space, while a regression model is finally applied to derive a mapping function between features and the corresponding screen coordinates.

### 3.1. Moving points

### 3.1.1. Detection of eye centers

The high-precision localization of eye centers constitutes the cornerstone for successfully tracking the gaze, and is inextricably linked to the overall system accuracy. To this end, an algorithm for accurate eye center localization in low resolution images based on our previous work [46,47] is used. In order to robustly and accurately detect eye centers, the system builds upon a synergy of chrominance information and radial symmetry.

In a given image, the face is detected using the well-known Viola–Jones face detector [48] and regions of interest containing the eyes are heuristically defined. Chrominance information is used to build an *eye map* that optimally distinguishes the eye areas from the rest of the skin area. The image is at first transformed to *YCbCr* colorspace and the proposed *eye map* is given as

$$EyeMapI = \frac{EyeMapC \oplus B1}{(Y \ominus B2)} \tag{1}$$

where $\oplus$ and $\ominus$ denote gray-scale dilation and erosion, respectively, with the flat circular structuring elements $B1$ and $B2$. $Y$ represents the grayscale image and $EyeMapC$ is an intermediate *eye map* which combines the chrominance information in the *Cb* and *Cr* components of *YCbCr* as follows:

$$EyeMapC = \frac{1}{3}\{(Cb)^2 + (1-Cr)^2 + (Cb/Cr)\} \tag{2}$$

Radial symmetry is calculated using a fast and highly efficient radial symmetry transform, first introduced in [49]. The transform is a gradient-based interest operator that works by considering the contribution each pixel makes to the symmetry of pixels around it. The radial symmetry transform is applied both to the grayscale component of the eye image and to the calculated *EyeMapI*. The cumulative result of the transforms indicates the precise positions of the eye centers.

The proposed eye localization algorithm accurately detects the centers of the irises even when they reach their circular shape limits (*i.e.* due to occlusions by the eyelids or when reaching extreme positions inside the eye socket). This is mainly attributed to the use of *eye maps* and the convolution

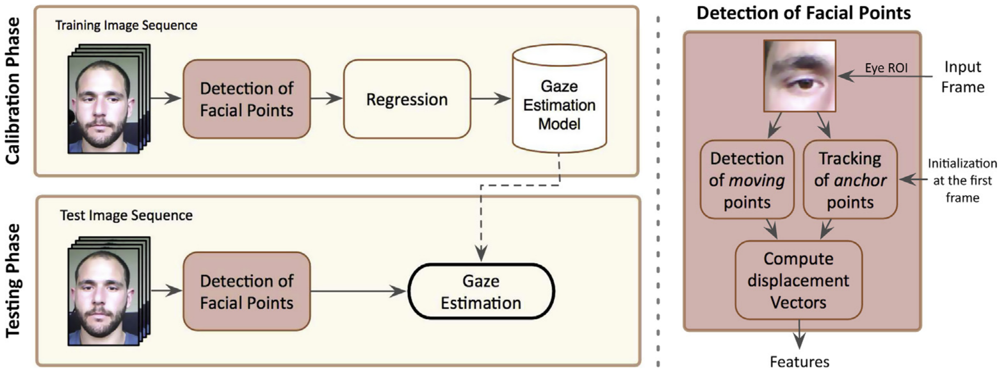
Calibration Phase
Training Image Sequence
Detection of Facial Points
Regression
Gaze Estimation Model
Testing Phase
Test Image Sequence
Detection of Facial Points
Gaze Estimation
Detection of Facial Points
Eye ROI
Input Frame
Detection of *moving* points
Tracking of *anchor* points
Initialization at the first frame
Compute displacement Vectors
Features

**Fig. 2.** Overview of the gaze estimation model.

with round structuring elements, which further emphasize the circular pattern of the iris. In addition, given that the iris always preserves some level of symmetry, regardless of the eye state, the choice of a low radial strictness parameter in the radial symmetry transform also highlights less radially symmetric patterns; *i.e.* it assigns big values at the center of the iris even when only the vertical edges of the eyes are visible.

### *3.1.2. Detection of eyelids*

The positions of the upper and lower eyelids provide information about the degree of eye opening [50] and greatly contribute in defining the gaze along the vertical axis. The *y-positions* of the eyelids correspond to the horizontal boundaries between two homogeneous areas, *i.e.* the iris area and skin area. The *x-position* of the eyelids is regarded the same as those of the corresponding eye centers. Starting from the localized eye center we define a rectangular *Region of Interest (ROI)* in which the eyelids are searched. The distance between the eyeball centers, also known as interocular distance, is used as the reference distance. Assuming that the iris diameter roughly corresponds to 10% of the interocular distance, the width and height of the *ROI* is defined as $0.1{*}d_{ioc}$ and $0.3{*}d_{ioc}$ correspondingly ($d_{ioc}$ stands for the interocular distance); each vertical side is at a distance of $0.05{*}d_{ioc}$ from the eye center so that only the iris area (not the sclera) is enclosed, thus constituting a homogeneous area, and the distance of each horizontal side is $0.15{*}d_{ioc}$ from the eye center, so that the eyelid boundary is certainly included, regardless of the eye state (Fig. 3(a)).

In order to detect the boundary of these distinct regions, *integral projection functions* are used. Image projection functions have been proven to be effective methods for extracting boundaries between different areas, representing the image by 1-dimensional orthogonal projections usually along the vertical and horizontal axes [51,52]. However, in view of the specific application, head rotations may change the boundary orientation on other directions rather that the horizontal one. To this end, the *integral projection function* is generalized to detect projections on different angles. Suppose $I(x,y)$ is the intensity of a pixel at the location $(x,y)$. The *integral projection* along a direction $\vartheta$ for a rectangular area is defined as

$$IP(I,\rho)=\int_{-H/2}^{H/2} I\,(x_0+\rho\cos\vartheta-h\sin\vartheta,\; y_0+\rho\sin\vartheta+h\cos\vartheta)\,dh \qquad (3)$$

where $(x_0,y_0)$ is the rectangle center (eye center), $\rho=0,1,\ldots,W$, with $W$ being the width of the rectangle, and $H$ represents the height of the rectangle or, equivalently, the number of pixels to be integrated for each $\rho$.

Given the search *ROI* for the eyelids, denoted hereafter as $I(x,y)$, we first perform gray-scale erosion with a rectangular structuring element $B$ to remove artifacts (such as glimpses), enhancing the homogeneity of the areas:

$$I_e(x,y)=(I\ominus B)(x,y) \qquad (4)$$

The *integral projection function* of Eq. (3) is computed for $I_e$ with $\vartheta$ being the inclination of the line connecting the two detected eye centers, which represents the rotation of the head. Determining the derivative of the projection result, peak values are reported in the boundary between the two areas (Fig. 3(c)).

Subsequently, a gradient image is computed by performing convolution of the image with the vertical Prewitt operator. The resulting edge map $E$ presents large values in areas of vertical abrupt changes of pixel intensity (Fig. 3(d)).

The $y-$positions of the eyelids are computed as the positions of each of the two global maxima bilaterally of the eye center, in the cumulative result of the aforementioned $I_e$ integral projection derivative and the edge map integral projection. The upper side maximum position is thus defined as

$$y_{uel}=\underset{\rho\in[0,W/2]}{\arg\max}\left\{\left|\frac{\partial IP(I,\rho)}{\partial\rho}\right|+|IP(E,\rho)|\right\} \qquad (5)$$

The lower side maximum position is calculated in a similar manner.

## *3.2. Anchor points*

The appearance of the *moving* points, notably the eyes, alters considerably throughout the frame sequence, depending on their orientation, occlusion by the eyelids, *etc.* and are difficult to track based on appearance trackers. Therefore, their localization on every frame is crucial and tracking can only assist in ensuring their spacial continuity in the temporal dimension (*i.e.* throughout the frame sequence). Unlike *moving* points, *anchor* points do not need to be specific (their role is to serve as reference points) and can be arbitrarily chosen at any face location. The most common approaches [25,22,23] consider inner and outer eye points as *anchor* points, detected in every frame. Aiming to a low computational cost, we do not regard explicit facial points as *anchor* points, but rather points that have the same appearance in every frame and are

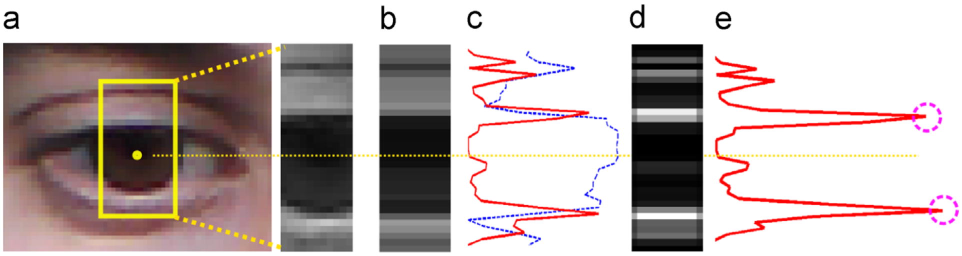

**Fig. 3.** Pictorial representation of the proposed eyelid detection. (a) Original eye image and the defined search *ROI*, (b) dilated search ROI, (c) the *integral projection* of the original image along the vertical direction (blue dotted line) and its derivative (red line), (d) the computed edge map and (e) the final cumulative result with the positions of the global maxima superimposed.

easy to locate and track (*i.e.* located in highly textured places where edges are present). Contrary to *moving* points, because the chosen *anchor* points exhibit very similar appearance throughout the frame sequence, they comprise very robust features to track. Using an appearance-based tracking approach rather than a detection in every frame approach, high precision detections are achieved.

Instead of tracking isolated points, we consider the most efficient alternative of tracking an image patch and consider as *anchor* point its center coordinates. The patch contains the inner eye corners and eyebrow edges, therefore comprising a highly textured area containing edges which are easy to track robustly. The patch to be tracked is initialized only once at the beginning of the image sequence: considering the captured frame which corresponds to the first calibration image, and having as starting points the detected eye centers, we define the dimensions and position of the patch with respect to the interocular distance as in Fig. 4. Another benefit of the proposed approach is that the absolute positions of the *anchor* points have no effect on the system's precision, meaning that the only prerequisite is the targets to be tracked consistently in every frame. Thereby, the current approach is insusceptible to appearance variations across different subjects (*e.g.* wearing eyeglass frames which occlude eye corners). The method utilized for tracking the image patch (reference image) is a variation of the well established Lucas–Kanade algorithm [53], denoted as Lucas–Kanade inverse affine transform [54].

### 3.3. Feature extraction

The movement of the eyes inside the eye socket is represented using distance feature vectors. In order to precisely represent gaze, the feature vectors require consistent and accurate localization of the points that move and contribute to gaze direction, *i.e.* eye centers and eyelids (which determine eye opening), and the *anchor* points which constitute stable face locations serving as reference points throughout the image sequence (so that we are able to measure distance vectors independent of the position of the face). Given the assumption of independence of gaze estimation in the two axes, two separate feature vectors are constructed for each direction. The feature vectors are formed as horizontal and vertical distances between *moving* and *anchor* points. The redundancy of information stemming from considering information from both eyes (as opposed to approaches using only a single eye for gaze tracking, *e.g.* [55]) are a major enhancement for the system's accuracy and robustness. The information from the two separate sources (eyes) is combined at feature level by creating an aggregate vector.

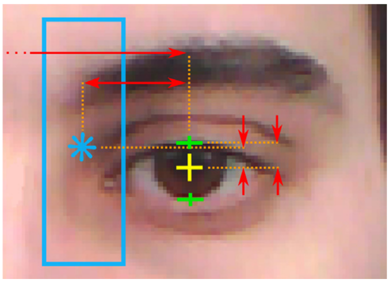

**Fig. 4.** Detected facial points and the extracted features. The *moving* points are indicated with the yellow cross (iris center) and the green crosses (eyelids). The blue rectangle denotes the image patch which is tracked throughout the image sequence and from which the *anchor* point (blue asterisk) is derived. The extracted features are depicted as red arrows.

Regarding the horizontal direction, three features are computed: the distance between the $x-$coordinates of each of the eye centers and the ipsilateral *anchor* point, as well as the distance of one of the eye centers and the anchor point at the opposite side (Fig. 4). The respective distance for the remaining eye center and its opposite anchor point is redundant as it occurs as a linear combination of the foregoing distances, thus is not used as an additional feature. In the vertical direction, except from the eye centers' $y-$coordinates, the $y-$position of the eyelids constitutes a significant source of information. The opening of the eye is primarily determined by the position of the upper eyelid rather that the lower eyelid, whose contribution can be regarded as negligible for the purposes of gaze tracking. Therefore the feature vector for the vertical direction is composed of four features, namely, the distances between the $y-$coordinates of each of the eye centers and the respective *anchor* points, as well as the distance between the upper eyelids and the *anchor* points (Fig. 4). The addition of extra features stemming from the position of the lower eyelid was proven experimentally not to improve overall accuracy (Section 5.4), incurring also the drawback of requiring a larger number of training images for building a regression model.

### 3.4. Interpolation model

Given the image data $--$screen positions correspondences, a mapping function between them is established. Second$-$order polynomial equations are most commonly used for 2D mapping of image data to the screen plane, using except from the linear terms, squared terms and interactions between them [23]. The non$-$linear terms are in general useful for correcting curved distortions and obtaining smoother scaling along the screen. However, non$-$linear terms of the mapping function may introduce errors; when approaching the screen borders the squared terms become quite large and as a result, small head movements or errors during the calibration procedure evoke much larger errors on screen coordinate estimations. Moreover, as the number of coefficients of the mapping function that need to be computed enlarges, so is the number of training (calibration) examples required. The requirement of keeping the training sessions as fast and straightforward as possible (so as not to evoke fatigue to the user) is more important than introducing more terms with dubious improvement in accuracy [56]. Therefore linear regression was employed for deriving the mapping function. Given that gaze estimation is calculated independently along the two axes, two mapping functions are derived.

## 4. Experimental setup

This section summarizes the experimental setups performed to evaluate the performance of the proposed gaze estimation system. In general, the quality of gaze estimation determines its appropriateness for different gaze interaction applications. The concept of 'quality' in gaze tracking is not well established, lacking also standard norms for measuring it. Several measures are lately gaining consensus [57], being the most commonly encountered in the literature. In this study, accuracy (also known as offset) is used as evaluation metric. Accuracy refers to the spatial deviation between the actual and the measured fixation point on the screen (gaze direction). It is usually measured on a sample to sample basis and it is averaged to characterize the overall performance. Except from the mean, accuracy is also calculated in terms of standard deviation of the gaze error. Accuracy is initially measured in pixel accuracy; given the screen dimensions and resolution as well as the distance of the user from the screen, it can also be expressed in spatial accuracy (cm or mm) and, most commonly, in visual degrees.

### 4.1. *Test dataset creation*

The absence of a publicly available database which could support the realization of the entity of the above−mentioned experiments has led us to the construction of a new dataset, which can be made available upon request. It consists of twelve (12) male and female users, with and without glasses, under uncontrolled, natural lighting conditions. For the needs of the dataset, the participants were asked to place themselves in a comfortable position in an approximate distance of $\approx 50$ cm from the computer screen. The dimensions of the screen are $35 \times 19.5$ cm while the resolution was set to $1920 \times 1080$ pixels. The web-cam used had the standard resolution of $640 \times 480$ pixels. Given these specifications, the face region roughly corresponded to $250 \times 250$ pixels with the iris radius being $\approx 9$ pixels.

In order to obtain the *optimal configuration scenario* regarding the calibration setup and camera placement extensive experiments with 4 of the participants were performed. Using the *optimal configuration scenario* derived (*i.e.* $9_{+1}$ calibration points and camera placed on top of the screen, as presented in Section 5.1), the accuracy of the proposed system was measured for all the participants of the dataset.

#### 4.1.1. *Calibration and testing phase*

During the calibration phase, the participants were asked to look at several known points which appeared successively on the screen. The arrangement and number of points depended on the calibration scheme, varying from 5 to 16. After fixating at one point the users pressed any button to proceed to the next point. The image which corresponded to the specific fixation was captured and the subsequent point appeared. The captured images correspond to fixations of the user on known locations on the screen, serving as the training data from which the coefficients of the mapping functions are derived. The time to complete the current calibration phase was very limited, only lasting for a few seconds, making the procedure effortless for the participants.

During the testing phase, the participants were asked to look at 111 points which appeared successively on the screen and followed the patterns as in Fig. 7 (red dots). The known locations of the points served as the ground truth in order to measure accuracy by determining the displacement between the estimated and actual gaze positions.

#### 4.1.2. *Description of different calibration configurations*

The first experiment investigated the influence of the number of calibration points as well as their arrangement on screen. Another parameter examined was the position of the camera *i.e.* above and below the screen.

The number of calibration points reported in the literature spans the range from 3 to 25 points, depending on each system's specifications and required accuracy [58]. In this work, 5 different calibration setups were examined. In the first setup, 5 points were presented to the user at each of the four corners of the screen and in the center of the screen, which correspond to points 1, 3, 5, 7, and 9 in Fig. 5(a). Given the 3-feature vector for the horizontal and 4-feature vector for the vertical direction, this number of points is the borderline case for deriving the mapping function coefficients. In the second setup the same number of points, plus the twice-counted center point, was presented to the users at a different arrangement, following a 'cross' pattern, as shown in Fig. 5(b) (points 1, 3, 5, 6, 8, and 10). The most typical scenario of 9 points, as shown in Fig. 5(a), and in the proposed different arrangement (Fig. 5(b)) comprise the third and the fourth setup, respectively. Finally, a 16 point scenario (Fig. 5(c)) was evaluated. In the setups which follow the proposed 'cross' pattern, the middle point is counted twice to ease eye movement continuity; these arrangements are denoted hereafter as $5_{+1}$ and $9_{+1}$, respectively.

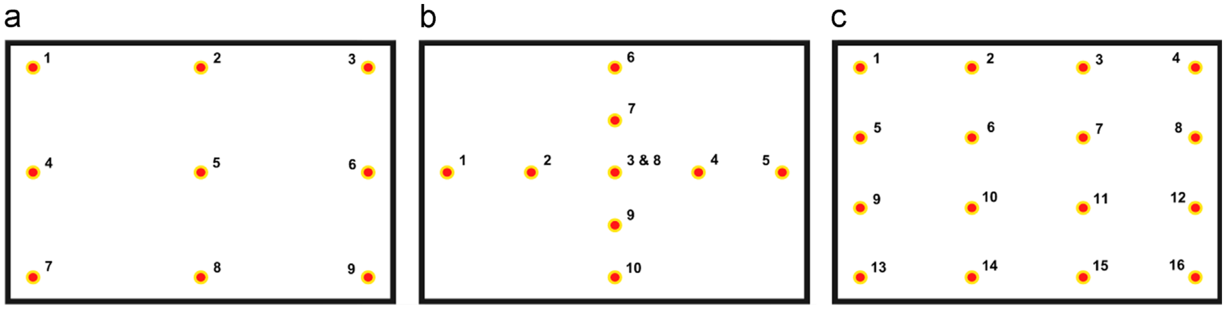

**Fig. 5.** The positions and arrangement of calibration points for different scenarios: (a) 9 points, (b) $9_{+1}$ points, and (c) 16 points.

The proposed 'cross' arrangement of points intends to separate the eye movements in the horizontal and vertical dimensions, so that the mapping coefficients are calculated independently. As opposed to the standard scenarios where the horizontal and vertical eye movements are correlated to each other (Fig. 6(a)), the proposed arrangement separates them into horizontal-only and vertical-only movements, as shown in Fig. 6(b).

To study how the position of the camera affects the proposed system's performance, experiments with the two most common setups were conducted. In the first setup the camera was placed at the top of the screen, at approximately the same level as the eyes of the user, while in the second the camera was positioned below the screen.

### 4.1.3. *Influence of head movements*

The effect of head movements in the performance of the proposed gaze estimation was investigated in three different experiments. In the first one the head was kept fixed using a chin rest. In the second the movement of the head was constrained, *i.e.* the user was asked to keep his head still during the experiment. The last case involved unconstrained head movement; the user was asked to move his head naturally. The fixed head case was only examined to establish the lower borders of the proposed system. It cannot be used in our system as it lists among the approaches using dedicated equipment (chin rest) in addition to being cumbersome and uncomfortable for the user. The parameter of number and arrangement of calibration points was also included in order to be further investigated.

## 4.2. *Public databases*

Aiming to systematically evaluate the proposed gaze tracking system's performance in reference databases, giving also the potential for direct comparison with other counterparts, two publicly available databases were considered.

The Columbia gaze dataset was recently published by Smith et al. [59]. It contains 5880 high-resolution images of 56 subjects that look at 21 gaze points on a 7 × 3 grid for different head poses ($0^o$, $\pm 15^\circ$, $\pm 30^\circ$ yaw angles). The Columbia dataset was chosen for our evaluation as it provides a reasonable amount of gaze directions and subjects; the subjects are ethnically diverse and 21 of them wear glasses. Given the amount of gaze direction arranged on a grid we were able to evaluate the performance on the database using four different training–testing schemes. These involved leave-one(gaze point)-out cross validation and splitting in training–testing sets; training with 9 points (arranged as in Fig. 5(a)), training with 5 points (corner points and middle point) and training with 5 points at cross arrangement, using the remainder of the points in each case as the testing set.

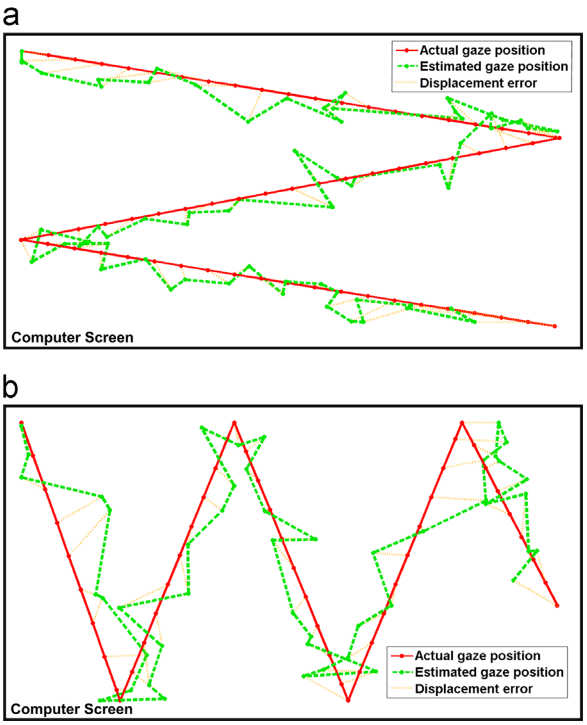

**Fig. 7.** Estimated and actual gaze positions for the scenario of $9_{+1}$ points and camera placed above the screen for subject 1 in the dataset. The red dots depict the 111 points used for testing.

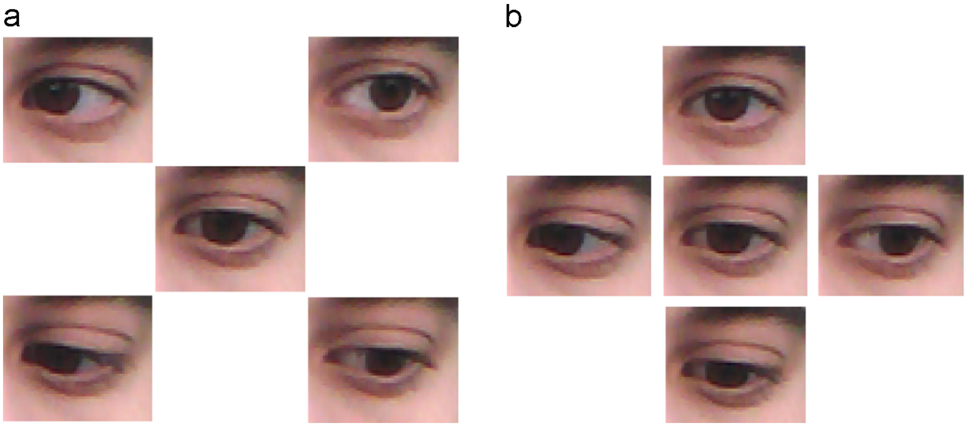

**Fig. 6.** Appearance of eyes during calibration in (a) 5 points scenario and (b) $5_{+1}$ points scenario.

The UUlm Head Pose and Gaze Database (UulmHPG) [60] contains images of 20 subjects in various combinations of head poses and eye gazes. For each subject, images which correspond to 9 horizontal (0°, ±15°) and 3 vertical gaze directions (0°, ±20°) for different head poses (ranging from 0° to 90° with a step of 10° for the yaw angle and 0°, ±20° elevation for the pitch angle) are captured. From the different image resolutions provided in the database, the 800 × 600 resolution which corresponds to the resolution of a standard web-cam was used. With the purpose of being in accordance with the other methods which report results on the same database, three subjects were considered (3, 12, 16), for head rotations restricted to [−30°, 30°] and gaze angles restricted to [−40°, 40°].

## 5. Experimental results

### 5.1. Optimal configuration scenario

#### 5.1.1. Evaluation of different calibration configurations

The obtained average accuracy (across the 4 participants) using the different calibration setups and camera positions are shown in Table 1, with the accuracy measured in pixel, distance and angle units, along each axis. Regarding the number of calibration points in the typical arrangement (5, 9 and 16), we observe that the two later present almost similar results, outperforming the borderline case of 5 points.

One of the most significant observations is that the results of the setups using the proposed arrangement ($5_{+1}$ and $9_{+1}$) excel over the typical arrangements in every case. The separation of eye movements in each of the axis proves to better calculate the coefficients of the mapping function, thus granting superior system performance.

Pertaining to the influence of camera position several observations can be made. In the horizontal direction accuracy results are more or less the same, presenting marginal differences. Conversely, the performance in the vertical dimension when the camera is located below the screen is significantly worse, falling behind in accuracy by almost a half. This difference is attributed to the perspective from which the *moving* points are viewed. Although a larger fraction of the eye is always visible from that position, the movements of the eyes and the eyelids appear to be much more subtle. These very limited movements in the vertical direction cause small errors in detection to be translated in much larger errors in the final position estimate. In practice, the displacement of the eyelids is so imperceptible that the features related to them do not contribute but minimally to the final result. The proposed method's performance for above-screen placed cameras comprises a significant asset, given that laptops with built-in cameras commonly present this configuration.

#### 5.1.2. Influence of head movements

This section studies the results of the influence of head movements in the accuracy of the system, examining three different cases, as described in Section 4.1.3.

Drawing from the results in Table 2, we observe that small head movements *i.e.* out-of-plane rotations and translations do not have a great impact on the system's performance. The performance starts to deteriorate at the point where the head starts moving freely, given that head pose information is not considered.

**Table 1**
Accuracy results (*mean* ± *standard deviation*) for different calibration scenarios and camera positions (above and below screen) for the 4 test subjects. The results are expressed in pixel, distance and angle units in the horizontal ($X$) and vertical ($Y$) dimensions.

| Calib. | Camera above screen | | Camera below screen | |
|---|---|---|---|---|
| | $X$ | $Y$ | $X$ | $Y$ |
| 5 | 134 ± 101px<br>2.44 ± 1.84 cm<br>Á Á<br>2.79 ± 2.11° | 152 ± 127 px<br>2.74 ± 2.29 cm<br>3.14 ± 2.63° | 154 ± 136 px<br>2.80 ± 2.48 cm<br>3.21 ± 2.84° | 317 ± 206 px<br>5.73 ± 3.72 cm<br>6.54 ± 4.26° |
| $5_{+1}$ | 121 ± 92px<br>2.21 ± 1.68 cm<br>Á Á<br>2.53 ± 1.92° | 87 ± 80 px<br>1.57 ± 1.44 cm<br>1.8 ± 1.66° | 94 ± 76 px<br>1.71 ± 1.39 cm<br>1.96 ± 1.59° | 161 ± 139 px<br>2.90 ± 2.51 cm<br>3.32 ± 2.87° |
| 9 | 107 ± 87px<br>1.96 ± 1.58 cm<br>Á Á<br>2.24 ± 1.81° | 85 ± 72 px<br>1.52 ± 1.30 cm<br>1.75 ± 1.49° | 96 ± 70 px<br>1.75 ± 1.27 cm<br>2.01 ± 1.46° | 144 ± 133 px<br>2.60 ± 2.39 cm<br>2.98 ± 2.74° |
| $9_{+1}$ | 94 ± 81px<br>1.72 ± 1.77 cm<br>Á Á<br>1.97 ± 1.69° | 77 ± 63 px<br>1.39 ± 1.13 cm<br>1.59 ± 1.30° | 90 ± 75 px<br>1.64 ± 1.37 cm<br>1.88 ± 1.57° | 120 ± 108 px<br>2.17 ± 1.95 cm<br>2.49 ± 2.23° |
| 16 | 97 ± 79px<br>1.77 ± 1.43 cm<br>Á Á<br>2.03 ± 1.64° | 80 ± 63 px<br>1.44 ± 1.14 cm<br>1.65 ± 1.31° | 86 ± 73 px<br>1.56 ± 1.33 cm<br>1.79 ± 1.52° | 95 ± 80 px<br>1.72 ± 1.45 cm<br>1.97 ± 1.66° |

**Table 2**
Accuracy results in terms of angular deviations (*mean* ± *standard deviation*) for different head movement constraints and different calibration setups for the 4 test subjects.

| Calib. | Fixed head | | Constrained movement | | Free movement | |
|---|---|---|---|---|---|---|
| | *X* | *Y* | *X* | *Y* | *X* | *Y* |
| 5 | 2.41 ± 1.70° | 1.16 ± 0.71° | 2.79 ± 2.11° | 3.14 ± 2.63° | 8.75 ± 5.33° | 4.82 ± 3.51° |
| $5_{+1}$ | 1.95 ± 1.41° | 1.22 ± 0.94° | 2.53 ± 1.92° | 1.80 ± 1.66° | 5.62 ± 4.24° | 4.16 ± 2.74° |
| 9 | 2.05 ± 1.57° | 1.16 ± 0.97° | 2.24 ± 1.81° | 1.75 ± 1.49° | 7.81 ± 4.80° | 5.49 ± 4.30° |
| $9_{+1}$ | 1.74 ± 1.37° | 1.12 ± 0.91° | 1.97 ± 1.69° | 1.59 ± 1.30° | 4.80 ± 3.75° | 3.79 ± 2.32° |
| 16 | 1.71 ± 1.22° | 1.07 ± 0.79° | 2.03 ± 1.64° | 1.65 ± 1.31° | 4.07 ± 3.53° | 3.01 ± 2.47° |

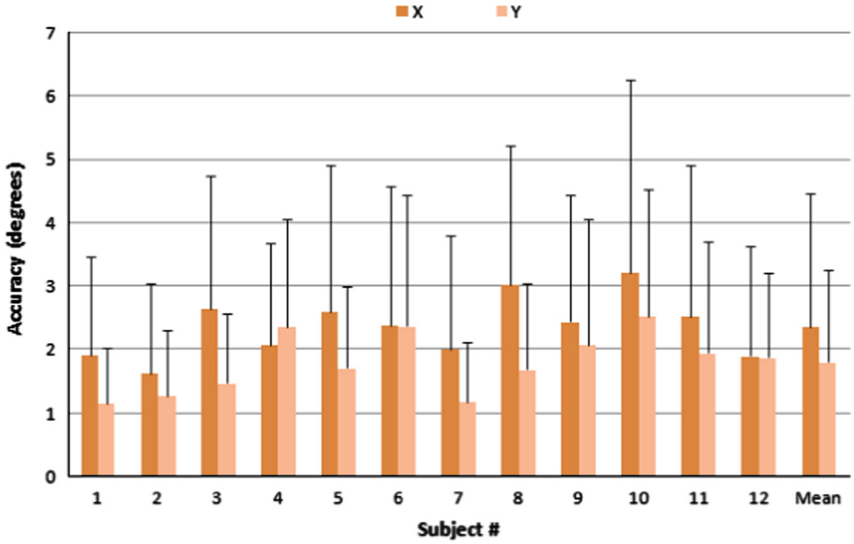

**Fig. 8.** Accuracy (mean ± standard deviation) for all subjects in the test dataset The red dots depict the 111 points used for testing.

In the fixed head case we observe only a slight improvement in performance over the constrained movement case; this can be attributed to the system reaching its borderline in accuracy, due to noise in detection of the *moving* points and tracking of the *anchor* points. The accuracy of the proposed system is bound by the number of pixels that the eye is allowed to move inside the eye socket. This allowed displacement (in terms of pixels) depends in turn on the resolution of the camera and the distance of the user from the camera. Considering the specific experimental setup, a close examination of the feature vectors reveals a permissible movement of the eye center inside the eye socket of ≈ 8–10 pixels in the horizontal direction and ≈ 4–5 pixels in the vertical direction. The fact that limited movements of a few pixels are mapped on high resolution screen coordinates demonstrates that minor misplacement of detections can trigger much larger errors in the final gaze estimation. Increasing the resolution or positioning one's head closer to the camera can be a straightforward manner for improving accuracy. The results for the free head movement scenario are just indicative of the system's accuracy for unconstrained movements, as they greatly depend on the level of head movements; greater head movements can lead to a further reduction in the system's performance.

Pertaining to the number and arrangement of calibration points, we observe that the $9_{+1}$ points scenario demonstrates the best overall performance. From the aforementioned parameters studied, those that yielded the optimal results were considered for conducting experiments on all the subjects of the dataset. Those involve the $9_{+1}$ points setup with the camera positioned on top of screen and keeping the head movement constrained. A visualization of the results for subject 1 and using the *optimal configuration scenario* is shown in Fig. 7.

### 5.2. Evaluation on the test dataset

The results across all the subjects in the dataset for the *optimal configuration scenario* are depicted in Fig. 8. The mean accuracy and mean standard deviation results were computed by averaging the measurements for all the users. The proposed system yields a mean accuracy of 2.36 ± 2.11° in the *X* direction and 1.80 ± 1.45° in the *Y* direction. Measured in pixel displacement, this error corresponds to 113 pixels in the horizontal direction and 87 pixels in the vertical direction. Expressed as error percentage, they correspond to errors of 5.9% and 8.1% in the respective axis. The relatively large values of standard deviations observed are expected in the context of gaze tracking (*e.g.* see [31,21]). The differences in performance between the test subjects are mainly attributed to the degree of head movements during the experiment. The presence of glasses (in subjects 3, 4 and 11) did not have an unfavorable effect on the system's performance, since no reflections from the lens were present. Moreover, different eye colors (subjects 4 and 10) did not hinder the precise localization of eye centers.

The average error reported demonstrates that the proposed gaze tracking approach is appropriate for most gaze interaction applications, provided that, as explained in Section 2.3, in a visual field of ≈ 1.7–2° everything can be seen without requiring a saccade.

### 5.3. Evaluation on publicly available databases

The results of the proposed gaze tracking system for the Columbia gaze dataset are shown in Table 3. The different training schemes *i.e.* leave-one-out cross validation and splitting into training/testing sets with 9 training points exhibit very similar performances. The 5 point calibration scenario with the points located in the corners presents subordinate accuracy compared to the 5 points in the cross arrangement scenario, which is in line with the findings of Section 5.1.1.

Given that the database was only recently published, Schneider et al. [40] are the only authors to report results on Columbia gaze dataset. Using an appearance based approach and evaluating many feature descriptors and dimensionality reduction techniques they achieve at best a 3.53° accuracy for frontal-only faces. However, as their approach is person independent, their results cannot be directly compared to the proposed method's results.

The performance of the proposed gaze tracking algorithm was evaluated in UulmHPG database and compared with the approaches of Heyman et al. [20] and Salam et al. [30]. As can be drawn from the results in Table 4 the proposed method presents lower performance to the rest of the methods for gaze accuracy in frontal faces and comparable performance for gaze accuracy in the case of rotated faces. However, both the methods of [20,30] consider pose information for estimating the final gaze position while our method draws information only from the eye positions. In the scenario where the pose information was also taken into consideration (which at this point is beyond the scope of the current paper), much improved results would be expected. This is also evident by examining the case where the training and testing of the proposed method takes place separately for each pose, with the proposed system presenting $1.42 \pm 1.35^\circ$ total gaze accuracy, $1.05 \pm 0.90^\circ$ gaze accuracy for frontal faces and $1.58 \pm 1.48^\circ$ for rotated faces. This performance reported is indicative of the results that the proposed system would yield if the head pose was flawlessly compensated.

### 5.4. Effectiveness of eyelid features

In order to evaluate the effectiveness of eyelid-based features, we performed tests on the above-mentioned datasets. At first, we examined the mean vertical accuracy for the 12 subjects in the created dataset and the mean vertical accuracy for different poses and subjects in Columbia gaze dataset. The UulmHPG database was not considered in the current experiment as it contains very few images which correspond to vertical gaze directions for each of the subjects. The gaze accuracy was examined in our experiments in three cases: using only upper eyelid features, using both upper and lower eyelid features and without using any eyelid features. The results are shown in Table 5. At this point, aiming to provide a statistical basis among the two feature vectors for interpreting the accuracy results, the Wilcoxon signed-rank test was applied to reject the null hypothesis that the regression model using the two distinct feature vectors performed equally well on the whole collection of data sets. The null hypothesis was rejected with the Wilcoxon statistic being smaller than the critical value for a two-tailed test at the significance level of 0.05 ($p < 0.05$).

The gaze accuracy results demonstrate that, as also mentioned in Section 3.3, upper eyelid information has a considerable influence on the vertical gaze accuracy. Contrariwise, lower eyelid features usually have a negative impact on performance, introducing noise in the feature vector, in addition to increasing dimensionality.

## 6. Discussion

The proposed gaze estimation system (categorized under 2D interpolation-based methods) presents the advantage of straightforward implementation, without requiring camera calibration (*i.e.* finding the camera's *intrinsic* and *extrinsic* parameters). Eye physiology, optical properties and geometry are indirectly modeled through the mapping function; there

**Table 4**
Comparison of different methods in the UulmHPG database (*mean* ± *standard deviation*).

| Method | Gaze accuracy | | |
|---|---|---|---|
| | Total | Frontal face | Facial rotation |
| Proposed | $7.53 \pm 7.18^\circ$ | $6.95 \pm 8.67^\circ$ | $7.78 \pm 6.51^\circ$ |
| Heyman [20] | $5.64 \pm 3.95^\circ$ | $3.45 \pm 2.00^\circ$[a] | $7.85 \pm 5.90^\circ$[a] |
| Salam [30] | $7.07 \pm 5.85^\circ$ | $5.00 \pm 3.80^\circ$[a] | $7.85 \pm 6.25^\circ$[a] |

[a] The value is estimated from author's graph

**Table 3**
Accuracy results in terms of angular deviations (*mean* ± *standard deviation*) for Columbia gaze dataset.

| Train/Test | −15° yaw angle | | 0° yaw angle | | 15° yaw angle | |
|---|---|---|---|---|---|---|
| | *X* | *Y* | *X* | *Y* | *X* | *Y* |
| Leave-one-out cross validation | $1.41 \pm 1.98^\circ$ | $1.34 \pm 1.81^\circ$ | $1.71 \pm 2.72^\circ$ | $1.33 \pm 1.90^\circ$ | $1.52 \pm 2.16^\circ$ | $1.16 \pm 1.60^\circ$ |
| 9 point training | $1.55 \pm 2.01^\circ$ | $1.45 \pm 2.25^\circ$ | $1.84 \pm 2.35^\circ$ | $1.43 \pm 2.17^\circ$ | $1.75 \pm 2.37^\circ$ | $1.22 \pm 1.74^\circ$ |
| 5 point training (corner points) | $2.25 \pm 3.20^\circ$ | $4.34 \pm 5.61^\circ$ | $3.30 \pm 4.41^\circ$ | $4.26 \pm 5.49^\circ$ | $2.69 \pm 3.66^\circ$ | $3.79 \pm 5.24^\circ$ |
| 5 point training (cross arrangement) | $2.10 \pm 2.89^\circ$ | $3.35 \pm 4.72^\circ$ | $2.65 \pm 3.96^\circ$ | $4.02 \pm 5.82^\circ$ | $2.22 \pm 2.87^\circ$ | $3.72 \pm 5.85^\circ$ |

**Table 5**
Influence of eyelid features in vertical gaze accuracy rates (*mean* ± *standard deviation*).

| Dataset | Upper eyelids | Both eyelids | Without eyelids |
|---|---|---|---|
| 12 subjects | $1.80 \pm 1.45^\circ$ | $2.98 \pm 2.63^\circ$ | $2.19 \pm 1.81^\circ$ |
| Columbia gaze | $1.28 \pm 1.64^\circ$ | $1.33 \pm 1.67^\circ$ | $1.40 \pm 1.97^\circ$ |

is no need to explicitly define them. Moreover, the proposed framework does not depend on anthropometric assumptions which are burdensome to compute and may often introduce additional errors to the system; in addition, differences across humans or special vision conditions (such as strabismus) are implicitly modeled in the mapping function, without affecting the system's performance. While in *model-based* approaches the requirement of high resolution images for building geometric models is an impediment, our system maintains a good level of performance even for images of lower resolution. One limitation of the proposed method is that it does not handle pose changes well, yet it retains its performance under small head movements.

As the experiments of the proposed gaze estimation system revealed, the most influential factor which may greatly affect accuracy is the movement of the head. The impact on the performance is analogous to the level of head movements. Addressing the problem of head pose invariance in gaze estimation constitutes an extensive research issue which cannot be easily resolved. Inaccurate estimations of head pose may in fact deteriorate the performance instead of alleviating the influence of head movements. A more practical solution for compensating for head pose would be to provide visual feedback of the estimations to the user in real time. Currently, the users perform the entire experiment without receiving any cues regarding the estimated gaze positions; the incorporation of real time visual feedback in the final application is believed be a great asset in compensating for the pose and increasing the overall accuracy.

Changes in illumination conditions are among the most common issues from which gaze trackers suffer. Gaze trackers that depend on appearance are usually more prone to errors due to their dependency on the texture and the intensity distribution of the training data. Regardless of the illumination intensity and type, as long as the contrast of the eye area is adequate, our eye localization and anchor point tracker can consistently yield correct detections, thus not perturbing the final gaze estimation accuracy.

Despite the number of gaze tracking approaches in the literature, a direct comparison between reported accuracies is most of the times not feasible. This is attributed to the fact that standard databases for evaluating gaze are not yet well established and researchers usually have to construct anew data sets for the specific needs of their experiments. Moreover, the accuracy results are bounded by the experimental setup (calibration procedure, number of calibration points, *etc.*) and also heavily depend on the equipment used, the distance of the users from the camera and the degree of head movements. Except from the direct comparison results reported in Section 5.3, we present the results reported for approaches in the same context (using natural lighting and a single camera setup) in Table 6.

**Table 6**
Accuracy results in terms of angular deviations for different gaze tracking methods in the literature.

| Method | Accuracy average | | Head movement | Dataset subject |
|---|---|---|---|---|
| | *X* | *Y* | | |
| Torricelli [21] | 1.7° | 2.4° | Constrained | 9 subjects |
| Chen [19] | 1.8° | 2.0° | Fixed | N/A |
| Ishikawa [4] | | 3.2° | Free | 3 subjects |
| Valenti [31] | 1.9° | 2.2° | Free | 11 subjects |
| Asteriadis [61] | | 4.1° | Free | 8 subjects |
| Sesma [23] | 2.4° | 2.2° | Constrained | 4 subjects |
| Yamazoe [18] | 5.3° | 7.7° | Free | 5 subjects |
| Baek [24] | | 2.4° | Constrained | 6 subjects |
| Alnajar [38] | | 4.3° | Free | 15 subjects |
| Sugano [39] | | 3.5° | Fixed | 7 subjects |

Drawing from the different approaches used, we can conclude that the proposed system achieves comparable accuracy in the horizontal direction to the best accuracies reported and performs better at estimating gaze in the vertical direction.

## 7. Conclusions

The proposed system aims to improve accuracy, robustness, universality and usability in gaze tracking under natural conditions. By working in the natural lighting spectrum, specialized equipment requirements are eliminated, rendering the system suitable for universal use in any application and using any type of cameras. The exponential growth of mobile, hand-held devices which embody very specific hardware and software capabilities further favors the development of such kind of approaches.

The proposed gaze tracking system extracts facial points in order to derive gaze-related features which are subsequently mapped into screen coordinates using interpolation. The calibration procedure is effortless for the users and the system works fully automatically without requiring any user intervention. The results obtained after conducting a number of different experiments reveal that a different arrangement of calibration points in a ‘cross’ pattern gives a slight edge compared to traditional setups. The main asset of the proposed system is its accuracy in gaze estimations along the vertical direction, which is mainly attributed to the additional features providing information about the position of the eyelids. Moreover, the salient tracking of *anchor* points instead of detecting them in every frame contributes in increasing the system's accuracy and robustness. Given the good accuracy rates as well as its characteristics of usability and automation, the proposed system could comprise a step towards further bridging the interaction gap between humans and computers.

## References

[1] A.T. Duchowski, A breadth-first survey of eye-tracking applications, Behav. Res. Methods Instrum. Comput. 34 (4) (2002) 455–470.

[2] B. Kim, H. Lee, W.-Y. Kim, Rapid eye detection method for non-glasses type 3D display on portable devices, IEEE Trans. Consum. Electron. 56 (4) (2010) 2498–2505.

[3] P. Smith, M. Shah, N. da Vitoria Lobo, Determining driver visual attention with one camera, IEEE Trans. Intell. Transp. Syst. 4 (4) (2003) 205–218.

[4] T. Ishikawa, S. Baker, I. Matthews, T. Kanade, Passive driver gaze tracking with active appearance models, in: Proceedings of the 11th World Congress Intelligent Transportation Systems, 2004.

[5] L. Fletcher, G. Loy, N. Barnes, A. Zelinsky, Correlating driver gaze with the road scene for driver assistance systems, Robot. Auton. Syst. 52 (1) (2005) 71–84.

[6] R. Pieters, A review of eye-tracking research in marketing, Rev. Mark. Res. 4 (2008) 123–147.

[7] T.E. Hutchinson, K.P. White Jr, W.N. Martin, K.C. Reichert, L.A. Frey, Human–computer interaction using eye-gaze input, IEEE Trans. Syst. Man Cybern. 19 (6) (1989) 1527–1534.

[8] N. Emery, The eyes have it: the neuroethology, function and evolution of social gaze, Neurosc. Biobehav. Rev. 24 (6) (2000) 581–604.

[9] M. Elsabbagh, A. Volein, G. Csibra, K. Holmboe, H. Garwood, L. Tucker, S. Krljes, S. Baron-Cohen, P. Bolton, T. Charman, et al., Neural correlates of eye gaze processing in the infant broader autism phenotype, Biol. Psychiatry 65 (1) (2009) 31–38.

[10] M. Schiessl, S. Duda, A. Thö, R. Fischer, Eye tracking and its application in usability and media research. MMI-interakt. J. 6 (2003) 41–50.

[11] M.C. Pretorius, A.P. Calitz, D. van Greunen, The added value of eye tracking in the usability evaluation of a network management tool, in: Proceedings of the 2005 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries, South African Institute for Computer Scientists and Information Technologists, White River, South Africa, 2005, pp. 1–10.

[12] C.H. Morimoto, M.R. Mimica, Eye gaze tracking techniques for interactive applications, Comput. Vis. Image Underst. 98 (1) (2005) 4–24.

[13] Z. Zhu, Q. Ji, Novel eye gaze tracking techniques under natural head movement, IEEE Trans. Biomed. Eng. 54 (12) (2007) 2246–2260.

[14] C.-C. Lai, Y.-T. Chen, K.-W. Chen, S.-C. Chen, S.-W. Shih, Y.-P. Hung, Appearance-based gaze tracking with free head movement, in: 2014 22nd International Conference on Pattern Recognition (ICPR), IEEE, Stockholm, Sweden, 2014, pp. 1869–1873.

[15] A. Duchowski, Eye Tracking Methodology: Theory and Practice, 373, Springer, Springer Science & Business Media, 2007.

[16] D.W. Hansen, Q. Ji, In the eye of the beholder: a survey of models for eyes and gaze, IEEE Trans. Pattern Anal. Mach. Intell. 32 (3) (2010) 478–500.

[17] J. Heinzmann, A. Zelinsky, 3D facial pose and gaze point estimation using a robust real-time tracking paradigm, in: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, IEEE, Nara, Japan, 1998, pp. 142–147.

[18] H. Yamazoe, A. Utsumi, T. Yonezawa, S. Abe, Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions, in: Proceedings of the 2008 symposium on Eye tracking research & applications, ACM, Savannah, Georgia, 2008, pp. 245–250.

[19] J. Chen, Q. Ji, 3D gaze estimation with a single camera without ir illumination, in: The 19th International Conference on Pattern Recognition, 2008. ICPR 2008. IEEE, Tampa, Florida, USA, 2008, pp. 1–4.

[20] T. Heyman, V. Spruyt, A. Ledda, 3D face tracking and gaze estimation using a monocular camera, in: The Second International Conference on Positioning and Context-Awareness (POCA-2011), 2011, pp. 23–28.

[21] D. Torricelli, S. Conforto, M. Schmid, T. DAlessio, A neural-based remote eye gaze tracker under natural head motion, Comput. Methods Prog. Biomed. 92 (1) (2008) 66–78.

[22] J. Zhu, L. Yang, Subpixel eye gaze tracking, in: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, Washington, DC, USA, 2002, pp. 124–129.

[23] L. Sesma, A. Villanueva, R. Cabeza, Evaluation of pupil center-eye corner vector for gaze estimation using a web cam, in: Proceedings of the Symposium on Eye Tracking Research and Applications, ACM, Santa Barbara, California, USA, 2012, 217–220.

[24] S.-J. Baek, K.-A. Choi, C. Ma, Y.-H. Kim, S.-J. Ko, Eyeball model-based iris center localization for visible image-based eye-gaze tracking systems, IEEE Trans. Consum. Electron. 59 (2) (2013).

[25] R. Valenti, J. Staiano, N. Sebe, T. Gevers, Webcam-based visual gaze estimation, in: Image Analysis and Processing—ICIAP 2009, Springer, Vietri sul Mare, Italy, 2009, pp. 662–671.

[26] Y.-T. Lin, R.-Y. Lin, Y.-C. Lin, G.C. Lee, Real-time eye-gaze estimation using a low-resolution webcam, Multimed. Tools Appl. 65 (3) (2013) 543–568.

[27] P.M. Corcoran, F. Nanu, S. Petrescu, P. Bigioi, Real-time eye gaze tracking for gaming design and consumer electronics systems, IEEE Trans. Consum. Electron. 58 (2) (2012) 347–355.

[28] G. Shao, M. Che, B. Zhang, K. Cen, W. Gao, A novel simple 2D model of eye gaze estimation, in: 2010 Second International Conference on Intelligent Human–Machine Systems and Cybernetics (IHMSC), vol. 1, IEEE, 2010, pp. 300–304.

[29] J. Orozco, F.X. Roca, J. Gonzàlez, Real-time gaze tracking with appearance-based models, Mach. Vis. Appl. 20 (6) (2009) 353–364.

[30] H. Salam, R. Seguier, N. Stoiber, et al., Integrating head pose to a 3D multi-texture approach for gaze detection, Int. J. Multimed. Appl. 5 (4) (2013) 1–22.

[31] R. Valenti, N. Sebe, T. Gevers, Combining head pose and eye location information for gaze estimation, IEEE Trans. Image Process. 21 (2) (2012) 802–815.

[32] R. Valenti, N. Sebe, T. Gevers, What are you looking at, Int. J. Comput. Vis. 98 (3) (2012) 324–334.

[33] J.-G. Wang, E. Sung, Study on eye gaze estimation, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 32 (3) (2002) 332–350.

[34] E. Wood, A. Bulling, Eyetab: model-based gaze estimation on unmodified tablet computers, in: Proceedings of the Symposium on Eye Tracking Research and Applications, ACM, Safety Harbor, Florida, USA, 2014, 207–210.

[35] D.W. Hansen, A.E. Pece, Eye tracking in the wild, Comput. Vis. Image Underst. 98 (1) (2005) 155–181.

[36] L.-Q. Xu, D. Machin, P. Sheppard, A novel approach to real-time non-intrusive gaze finding, in: BMVC, 1998, pp. 1–10.

[37] S. Baluja, D. Pomerleau, Non-Intrusive Gaze Tracking Using Artificial Neural Networks, Technical Report, DTIC Document, 1994.

[38] F. Alnajar, T. Gevers, R. Valenti, S. Ghebreab, Calibration-free gaze estimation using human gaze patterns, in: 2013 IEEE International Conference on Computer Vision (ICCV), IEEE, Sydney, Australia, 2013, pp. 137–144.

[39] Y. Sugano, Y. Matsushita, Y. Sato, Learning-by-synthesis for appearance-based 3D gaze estimation, in: International Conference on Computer Vision and Pattern Recognition (CVPR'14), IEEE, Columbus, Ohio, USA, 2014.

[40] T. Schneider, B. Schauerte, R. Stiefelhagen, Manifold alignment for person independent appearance-based gaze estimation, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, IEEE, Stockholm, Sweden, 2014.

[41] D. W. Hansen, J.P. Hansen, M. Nielsen, A.S. Johansen, M.B. Stegmann, Eye typing using Markov and active appearance models, in: Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision, 2002 (WACV 2002), IEEE, Orlando, Florida, USA, 2002, pp. 132–136.

[42] Y. Sugano, Y. Matsushita, Y. Sato, Appearance-based gaze estimation using visual saliency, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2) (2013) 329–341.

[43] S.R. Langton, H. Honeyman, E. Tessler, The influence of head contour and nose angle on the perception of eye-gaze direction, Percept. Psychophys. 66 (5) (2004) 752–771.

[44] E. Murphy-Chutorian, M.M. Trivedi, Head pose estimation in computer vision: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 31 (4) (2009) 607–626.

[45] D.B. Henson, Visual Fields, Oxford University Press, New York, 1993.

[46] E. Skodras, N. Fakotakis, An accurate eye center localization method for low resolution color imagery, in: 2012 IEEE 24th International Conference on Tools with Artificial Intelligence (ICTAI), vol. 1, IEEE, Athens, Greece, 2012, pp. 994–997.

[47] E. Skodras, N. Fakotakis, Precise localization of eye centers in low resolution color images, Image Vis. Comput. 36 (2015) 51–60.

[48] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.

[49] G. Loy, A. Zelinsky, Fast radial symmetry for detecting points of interest, IEEE Trans. Pattern Anal. Mach. Intell. 25 (8) (2003) 959–973.

[50] L. Bour, M. Aramideh, B.O. de Visser, Neurophysiological aspects of eye and eyelid movements during blinking in humans, J. Neurophysiol. 83 (1) (2000) 166–176.

[51] G.-C. Feng, P.C. Yuen, Variance projection function and its application to eye detection for human face recognition, Pattern Recognit. Lett. 19 (9) (1998) 899–906.

[52] Z.-H. Zhou, X. Geng, Projection functions for eye detection, Pattern Recognit. 37 (5) (2004) 1049–1056.

[53] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision., in: IJCAI'81, 1981, pp. 674–679.

[54] S. Baker, I. Matthews, Lucas–Kanade 20 years on: a unifying framework, Int. J. Comput. Vis. 56 (3) (2004) 221–255.

[55] J.-G. Wang, E. Sung, R. Venkateswarlu, Estimating the eye gaze from one eye, Comput. Vis. Image Underst. 98 (1) (2005) 83–103.

[56] J.J. Cerrolaza, A. Villanueva, R. Cabeza, Study of polynomial mapping functions in video-oculography eye trackers, ACM Trans. Comput.-Hum. Interact. 19 (2) (2012) 10.

[57] K. Holmqvist, M. Nyström, F. Mulvey, Eye tracker data quality: what it is and how to measure it, in: Proceedings of the symposium on eye tracking research and applications, ACM, Santa Barbara, California, USA, 2012, 45–52.

[58] D.M. Stampe, Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems, Behav. Res. Methods Instrum. Comput. 25 (2) (1993) 137–142.

[59] B.A. Smith, Q. Yin, S.K. Feiner, S.K. Nayar, Gaze locking: passive eye contact detection for human-object interaction, in: Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, ACM, St. Andrews, Scotland, United Kingdom, 2013, 271–280.

[60] U. Weidenbacher, G. Layher, P.-M. Strauss, H. Neumann, A Comprehensive Head Pose and Gaze Database.

[61] S. Asteriadis, K. Karpouzis, S. Kollias, Visual focus of attention in non-calibrated environments using gaze estimation, Int. J. Comput. Vis. 107 (3) (2014) 293–316.