A Wearable Gaze Tracking System for Children in Unconstrained Environments

Basilio Noris, Jean-Baptiste Keller and Aude Billard

Learning Algorithms and Systems Laboratory
Tel: +41-21-6935464
Fax: +41-21-6937850
basilio.noris@epfl.ch

Abstract

We present here a head-mounted gaze tracking system for the study of visual behavior in unconstrained environments. The system is designed both for adults and for infants as young as 1 year of age. The system uses two CCD cameras to record a very wide field of view (96° × 96°) that allows to study both central and peripheral vision. A small motor-driven mirror allows to obtain the direction of the wearer’s gaze with no need for active lighting and with little intrusiveness. The calibration of the system is done offline allowing experiments to be conducted with subjects who cannot cooperate in a calibration phase (e.g. very young children, animals). We use illumination normalization to increase the robustness of the system, and eye blinking detection to avoid tracking errors. We use Support Vector Regression to estimate a mapping between the appearance of the eyes and the corresponding gaze direction. The system can be used successfully indoors as well as outdoors and reaches an accuracy of 1.59° with adults and 2.42° with children.

Keywords: Gaze Tracking, Appearance Based, Lay Users, Children, Eye Blinking, Unconstrained Environments, Support Vector Regression

1. Introduction

The study of gaze, be it saccadic eye movements, visual focus of attention or scene exploration is of great importance in a wide field of research topics [1]. In recent years gaze tracking systems have become a key tool for the study of visual attention and its role in various cognitive and social tasks.
Gaze Tracking systems also offer numerous opportunities in engineering by endowing users with new means to interact with machines [2, 3]. The variety in fields of application has led to the development of different means to monitor eyes movements and to infer gaze direction. Gaze tracking systems can be divided into two broad classes: external and wearable.

**External systems** are composed of single- or multi-camera systems that allow to estimate gaze direction in a limited workspace [4]. The main advantage of external systems is their non-intrusiveness. However, measuring gaze from an external point of view implies that the direction of the gaze depends on the direction of the user’s head. Most commercial solutions solve this problem by constraining the movement of the head, either using a specially tailored head or chin mount, or by emitting visual and auditory signals when the head is leaving the acceptable locations. However, some solutions have been proposed that allow free head movement [5, 6, 7] obtaining tracking errors as low as $0.43^\circ$. In most cases these systems use moving cameras to follow the movement of the head, however in some cases [8, 9] motor-driven mirrors are used to allow faster and more accurate tracking. Due to the static nature of external systems, determining the contents of the measured field of view is usually reduced to capturing the image coming from a computer screen or a projector.

**Wearable systems** allow users to move around freely and look around themselves in a less constrained environment. Moreover, wearable systems avoid the problem of absolute head position as their position is fixed with
respect to the user. However, this imposes limitations in terms of the tracking equipment: components weight, volume and ergonomy have to be adapted to the needs or abilities of the users. Moreover, controlling or obtaining an image of what the user is looking at becomes more difficult. A common solution is to add a standard video camera to the system; this, however, increases the weight and volume of the system. Unfortunately this solution usually allows to capture a limited part of the field of view. Despite these drawbacks, several wearable systems have been developed and applied both in psychological studies and to improve human-machine interfaces obtaining accuracies below 2° [10, 11]. Few commercial solutions exist that provide tracking errors below 0.5° [12, 13]. In all cases, these systems have either protruding elements with cameras facing the eyes or semi-transparent surfaces placed in front of one of the eyes.

We propose a wearable device which solves the problem of the limited field of view by combining two wide-angle cameras. Moreover the intrusiveness of the system is kept very low as only a small motor-driven mirror is visible to the wearer. We will describe the device in Section 2.1.

1.1. Estimating gaze direction

The problem of estimating the direction of gaze is usually tackled in two different ways [14]:

**Feature-Based methods** use eye contours, pupils, glints of lights and other features to estimate a geometrical model of the eye direction and infer the direction of the gaze [6, 11]. The calibration of such systems is usually simple as it requires a small number of calibration points to adjust the geometrical parameters for the gaze estimation. These methods usually avail themselves of active lighting to enhance the detection of shapes and the effect of glints on the eyes. This is often used in the form of infrared lights shining on the cornea and retina and providing an accurate and stable reference point. However, this restricts the usage to controlled environments where external factors (e.g. sunlight) do not affect the gaze estimation. Additionally, these methods only work when the features are visible and the eyes are sufficiently open or directed in the direction of the system. This limits the range of operation of these systems or requires such systems to be placed in front of the user.

**Appearance-Based methods** use the whole image of the eyes as input and try to implicitly exploit all elements of the image to create a mapping between the appearance of the eyes and the coordinates of the gaze [15, 16, 17]. Such
methods can exploit the shape of the eye, eyelids, eyelashes as well as pupil, iris and glints to obtain an accurate estimate of the gaze direction. This allows the gaze to be estimated even when the users are looking downwards or when glints are not visible anymore (see Figure 2). Appearance-based methods usually require a considerably larger amount of calibration samples than feature-based ones.

Regardless of the choices made, gaze tracking system usually require users to endure a calibration phase to allow the system to adapt to inter-user variability. Some notable exceptions exist nevertheless [18, 19], which use user-independent models for gaze tracking and obtain accuracies between 3° to 7°.

When no active lighting is used, an illumination normalization procedure is often done to increase the robustness to lighting changes and uneven illumination. A common technique to do this is histogram equalization, which increases the contrast of an image by modifying the shape of its histogram [20]. A different approach which models the light as a product of luminance and reflectance is the retinex technique [21]. A drawback of the basic retinex is that halo artifacts appear in regions of high contrast. A solution to this problem consists in weighting the effect of the retinex based on the amount of local contrast [22].

We propose an appearance-based method with passive lighting which can be calibrated offline (i.e. after the recording process). This allows the system to be used with lay users or in situations where a calibration phase is not always possible. Our system is able to work outdoors as well as indoors. We will describe how in Section 3.
1.2. Gaze Tracking for Children

The experimental part of our research involves studies with children from 12 months to 10 years of age [23, 24]. Working with children poses a number of difficulties. Depending on the age of the children the degree of understanding and response that we can expect from our subjects does not allow us to use traditional calibration methods (e.g. ensuring that the child follows a moving target, or glances at specific known locations). Moreover, children can be prone to excitement or tantrums that can result in the equipment being moved or taken off. For most gaze tracking systems, such accidents would require to stop the experiment and proceed to a new calibration phase, or could lead to an early end of the experiment.

In addition to calibration considerations, working with children poses additional limitations in terms of the technology we can use. Pointing active lighting, normal or IR, towards the subject’s eyes is not always a viable solution for studies with children \(^1\), and does not help in strongly-lit situations (e.g. outdoors). For similar reasons, goggles and cameras encumbering the subject’s field of vision can not be used either.

Although some research has been proposed on an external gaze tracker for 6 month old infants [25], to our knowledge, the system we propose is the first wearable system that allows to estimate gaze direction in a free environment on subjects as young as 1 year of age.

In a study in collaboration with the university hospitals of Lausanne and Geneva, we used the WearCam to monitor gaze in children with Autism Spectrum Disorder (ASD). We contrasted typically developing children to children with ASD according to two gaze parameters: fixations to faces and presence of faces in the peripheral view during a social interaction in a free environment. Due to the constraints described above, we would not have been able to obtain these measurements with existing gaze tracking systems.

2. WearCam

To obtain the input images for our gaze tracking system, we have developed a lightweight head-mounted device we call WearCam (see Figure 1).

\(^1\)The effect of invasive techniques such as IR on the eyes of very young infants, whose visual system is still developing, are not known yet.
2.1. WearCam System

The WearCam is designed to be worn by children as young as 1 year of age but can also be used by adult subjects. The whole system weighs 180g (camera plus cap/straps) and uses two miniature CCD cameras (EcoLine TV7105). The cameras provide a 768 × 576 pixel image at 25 frames per second. Using two cameras serves two purposes: firstly, it is possible to
obtain a larger field of view. This allows to monitor both what the wearer is doing with his/her hands and the people and things in the environment. The second purpose is to allow to fit a motor-driven mirror between the two cameras.

The first camera points slightly downwards, capturing the region the user explores while manipulating objects with his/her hands. The second camera points forward capturing the rest of the field of view, where the wearer is most likely to see people and distant objects (see Figure 4). Each camera is equipped with a lens covering a range of $96^\circ \times 72^\circ$. However, the cameras are angled at $30^\circ$ so as to obtain sufficient overlap between the two images to cover the region of the image occluded by the mirror\(^2\). The whole image captured by the system covers an angle of $96^\circ \times 96^\circ$ which is more than 3 times the field of view of other wearable eye trackers (e.g. [11, 19]). The advantage of using such a system is to be able to study both central and peripheral vision at the same time, while most gaze trackers usually only allow to focus on central vision.

A $2.7cm \times 0.7cm$ mirror is mounted between the two cameras. The mirror reflects the eyes of the wearer back into one of the cameras. The reflected image will be used to obtain the direction of the gaze. The mirror occludes part of the field of view of the camera. However, the overlap between bottom and top images allows to recover the missing region. A remote controlled servo motor allows to pivot the small mirror so as to align it with the eyes of the wearer without having to adjust the position of the camera. This allows to extract synchronously the appearance of the eyes and the corresponding field of view.

Adjustable straps allow to fit the device to young children as well as to adult subjects. To suit the needs of the subject the device can be mounted on a cap.

The WearCam is connected to a standard computer (usually a laptop) via two USB video converters. USB bandwidth limitations allow us to record images from the WearCam only at reduced resolution without loss of frames. In order to maintain synchronization between the cameras, the two images are blended into a single vertical image of $384 \times 576$ pixels (that is half the

\(^2\)Due to the parallax between the two cameras it is not possible to define the overlap between the images in terms of angles. It is however possible to indicate the minimum overlapping distance which is of $20cm$. 

7
full resolution available). Figure 3 shows a sample image from the WearCam.

2.1.1. Users wearing glasses

Due to the position of the WearCam on the forehead of the subject, the angle between the mirror and the eyes is such that, in most cases, spectacles do not appear in the reflected image. In some cases however the frame of the eyewear appears in the mirror. As the system does not rely on active lighting and glint detection, its performance is not affected by the presence of eyewear.

2.1.2. Saccades

The recording frame rate of the WearCam is only 25 frames per second. The span of time between each frame is of 40 msec. Assuming a perfect image quality, the fastest movements that could be recorded are movements slower than the span of 2 frames. This means that, in the best of cases, the WearCam can not record fast saccadic movements (which range around the 30 msec) but can detect slow saccadic movements (100 to 120 msec). In practice, however, image noise and motion blur degrade the quality of the detection: only the destination of saccadic movements can be safely recorded and estimations of the velocity of slow saccades should not be considered robust.

3. Appearance Based Gaze Tracking

The gaze tracking process is described by the graph presented in Figure 5. We start by extracting the mirrored region from the image. While the position of the mirror on the WearCam is known, its projection in the image varies depending on its angle. However it does not move throughout a recording. As the camera is head-mounted, small head displacements occur at every frame. Therefore it suffices to extract the unmoving region in the image to obtain a mask of the mirror pixels. We obtain the position mask by computing the image difference over a small number of frames. As the eyes inside the mirror might move, the mask is then cleaned with morphological closing operations.

In order to reduce computational costs, we resize the eyes region into a smaller image of size $w \times h$. The effect of the choice of $w$ and $h$ is discussed in Section 5.
Figure 5: Basic flow of the algorithm. We start by extracting the eyes image from the mirror region on the WearCam. We apply illumination normalization to improve the robustness to lighting changes. If a blink is not detected we estimate the gaze coordinates through \( \nu \)-SVR regression. Finally we combine the gaze coordinates and the WearCam image without the eyes into a single output image.

3.1. Illumination Normalization

We increase the robustness of the system to changes in lighting by applying a weighted retinex filter to the eyes image. By removing the low frequency luminance (e.g. directional shading on the eyes region) we obtain an image which retains details and features such as eyelids, eyelashes, iris, pupil etc. but without the main shading. In order to obtain a better estimate of the luminance and to reduce the impact of digital noise and the influence of automatic white-balancing from the cameras, we convert all images to grayscale and normalize the pixel values prior to the normalization step which we describe next.

Let \( I(x, y) \) be the input image at pixel position \((x, y) \in \{0, \cdots, w\} \times \{0, \cdots, h\}\). \( I(x, y) \in [0, 1] \) We obtain a luminance normalized image \( I'(x, y) \) as follows
\[
I'(x, y) = \ln(I(x, y)) - f(I(x, y))
\] (1)
where \( f(I(x, y)) \) is a weighted low-pass filter. For a neighborhood radius
Figure 6: weighting factor $\lambda$ as a function of local contrast (see Equation 3). The $\tau$ threshold allows to adjust the center point of the downward slope.

Figure 7: Examples of eyes region as obtained from the WearCam before and after homomorphic filtering using different neighborhood sizes. Top to bottom: original, 40, 20, 10, 5 pixels neighborhoods.

We use the pixel dependent weighting factor $\lambda(x, y)$ to reduce the halo effects appearing on high contrast portions of the image. We compute it as follows

$$\eta \in \mathbb{N}^* \text{ and a weighting factor } \lambda(x, y) \text{ we compute}$$

$$f(I(x, y)) = \frac{\lambda(x, y)}{\eta^2} \sum_{i=-\frac{\eta}{2}}^{\frac{\eta}{2}} \sum_{j=-\frac{\eta}{2}}^{\frac{\eta}{2}} \ln(I(x + i, y + j)) \quad (2)$$

We use the pixel dependent weighting factor $\lambda(x, y)$ to reduce the halo effects appearing on high contrast portions of the image. We compute it as follows

$$\lambda(x, y) = 0.5 + \frac{1}{2} \tanh \left( 20 \left( \tau - \max \left\{ \frac{\partial}{\partial x} I(x, y), \frac{\partial}{\partial y} I(x, y) \right\} \right) \right) \quad (3)$$

where $\tau$ is a manually set threshold. Figure 6 shows the response of $\lambda(x, y)$ as a function of local contrast values. By varying the neighborhood radius $\eta$ we are able to modulate the amount of low frequency luminance that will be filtered (see Figure 7). In image terms, the radius of the neighborhood is related to the size of the smallest details that will be eliminated.
3.2. Eye Blinks

We detect blinks by computing image differences on both eyes separately. We suppose that the mirror image contains both eyes at all times, therefore split the eyes image in the middle and consider each half of the image separately. Since we are analyzing images through time, let \( I(x, y, t) \) be the intensity of pixel \((x, y) \in \{0, \ldots, w\} \times \{0, \ldots, h\} \) at time \( t \). We compute the image differences \( d_l \) and \( d_r \) (for left and right eyes respectively) over time as

\[
d_l = \frac{1}{w \times h} \sum_{x=0}^{w/2} \sum_{y=0}^{h} |\frac{\partial}{\partial t} I(x, y, t)|
\]

we obtain \( d_r \) similarly by summing over \( x \in \{w/2+1, \ldots, w\} \) instead. We consider a frame as the beginning of a blink if

\[
d_l + d_r > \theta
\]

where \( \theta \) is set empirically and can be adjusted dynamically during the tracking process. Computing the differences \( d_l \) and \( d_r \) separately on each eye allows to avoid misclassifying as blinks sudden variations such as bright lights shining on one of the eyes.

This method unfortunately is not able to detect blinks when the user is looking downwards as there will be little to no visible change in the eyes appearance. In very low-light conditions, the noise due to the camera gain affects this measures, degrading the overall accuracy of the blink detection. In these cases, it is possible to disable the eye blinking detection altogether.

The duration of spontaneous eye blinking in healthy subjects is below 240 msec [26]. This corresponds to 6 frames at 25 fps. When a blink is detected in a frame, that and the following 5 frames are discarded and no output is given by the system.

3.3. Mapping input space to image coordinates

If a frame is not discarded by the eye blinking detection, we collect every pixel from the illumination normalized image of the eyes \( I'(x, y) \) (see Equation 1) into a single vector \( \mathbf{a} \in [0, 1]^m \) with \( m = w \times h \) where each element of \( \mathbf{a} \) is given by \( a^j = I'(j \mod w, \lfloor j/w \rfloor) \).

To obtain an estimate of the gaze direction, we want to find a mapping

\[
f(\mathbf{a}) : [0, 1]^m \rightarrow [0, 1]^2
\]

\[
\mathbf{a} \quad \mapsto \quad \mathbf{p}
\]
from the eyes appearance \( \mathbf{a} \) to a gaze point \( \mathbf{p} = [p_x, p_y]^T \). We model this mapping using Support Vector Regression (SVR) [27]. For practical purposes, we compute the horizontal and vertical coordinates separately, learning two mappings from \( \mathbf{a} \) to \( p_x \) and \( p_y \) respectively. However for the sake of simplicity we will describe the problem only once and we will note \( p \) instead of \( p_x \) and \( p_y \).

For a set of \( N \) calibration samples \( \mathbf{a}_i \) and corresponding coordinates \( p_i \), \( i \in \{1, \ldots, N\} \) we obtain the mapping through

\[
f(\mathbf{a}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{a}_i, \mathbf{a}) p_i + \beta
\]

where the \( \alpha_i \) are weighting coefficients associated to each calibration sample; \( \beta \in \mathbb{R} \) is a bias. \( k(\mathbf{a}_i, \mathbf{a}) \) is a kernel function allowing us to obtain a non-linear mapping. We use a gaussian RBF kernel

\[
k(\mathbf{a}_i, \mathbf{a}) = \exp\left(-\frac{1}{2 \gamma^2} [\mathbf{a}_i - \mathbf{a}]^T [\mathbf{a}_i - \mathbf{a}]\right)
\]

which depends on a parameter \( \gamma \) which will be optimized in Section 5.1. Both \( \alpha_i \) and \( \beta \) are computed during training. We call Support Vectors (SV) the samples whose weights \( \alpha_i \) are higher than zero. We can see from Equation 6 that the amount of samples that are used as SV directly influences the amount of detail the mapping can capture. The same goes for the computational cost for each estimation. We can adjust the amount of SVs through the parameter \( \nu \)

\[
\nu \geq \sum_{i=1}^{N} \alpha_i \quad \nu \in [0, 1]
\]

By adjusting \( \nu \) it is possible to limit the amount of SV the system will use. Additionally, to adjust the sensitivity of the mapping function to noise and outliers, we can introduce a cost parameter \( C \) such that

\[
0 \leq \alpha_i \leq C \quad \forall i, i \in \{1, \ldots, N\}
\]

By setting high values of \( C \), we penalize errors during the optimization process. We optimize the SV parameters \( \alpha_i \) and \( \beta \) under the constraints set by \( C \) and \( \nu \) through Sequential Minimal Optimization (SMO) [28].
3.4. Visualizing the gaze direction

Once we obtain the SVR estimation of gaze coordinates for a frame, we simply combine the original image and the gaze coordinates for visualization purposes. We combine the upper and lower parts of the input image from the WearCam together. This is done by removing the overlapping part which contains the mirror region. We then blur the whole image except for a circular region centered on the estimated gaze direction. We limit this region to a 5° radius circle, corresponding to the span of central vision, or macula vision.

3.5. Offline Calibration

All the steps described above are done in real time at each frame. However this assumes that the system has been calibrated. The calibration of the system itself (i.e. collecting calibration samples and computing the SV parameters) is done offline. The task of collecting calibration samples is performed by the experimenter instead of the gaze tracker user. As the computation times for calibration and processing are quite low (see Section 5.2), we are able to perform the calibration process incrementally as we describe now:

A simple interface allows the experimenter to review the recorded images. A timeline allows to scrub rapidly through the video. When the experimenter can make a safe assumption on the point in the image the wearer is looking at (see Section 4.3), she places a calibration point and the system collects the current appearance of the eyes. SVR training is performed automatically after each calibration point is added (when training time raises over 1 second, the calibration is done automatically only every couple of new samples). The system then displays the gaze direction estimation at each frame in the input video. The experimenter can then improve the estimation by adding further calibration points.

As the accuracy of the system increases, it also becomes easier to estimate additional calibration points (e.g. when an estimation is only slightly off an object in an empty area we can speculate to it being the object gazed at by the user). The calibration data is automatically saved after each calibration point is added/modified and the process can be stopped and resumed at any time. Once the process is completed to satisfaction, an output rendering of the video and gaze estimation can be saved for further analysis.

Depending on the recording, obtaining 100 to 200 calibration samples can take up to 10 minutes.
4. Experimental Setup

We validated the performance of our system in separate calibration experiments with adults and with children\(^3\). We now describe how.

\(^3\)A video example of the experimental protocols and the results of gaze tracking is available for reviewing purposes as supporting material
4.1. Accuracy Studies with Adults

Experiments with adults were conducted in our lab, while the subjects sat on a chair facing a blank wall. The chair was fixed to the ground at 1.5 m from the wall. A colored cube fixed to a pole was moved along regular patterns (lines, circles) covering the whole visual field of view of the subject. The experimenter moved the colored object while looking at the output video from the WearCam to ensure the object was not leaving the field of view. The subjects were asked to follow the object with their eyes (see Figure 8a). The experiment was conducted with 21 adults of different gender (15 males, 6 females), eye color (11 dark eyes, 10 bright eyes), ethnicity and age; 4 subjects wore glasses. On average $3 \pm 0.5$ minutes of data were recorded per subject.

To assess the accuracy of our system under sunlight, we conducted the same experiment outdoors with 4 additional subjects (3 males, 1 female, 1 person wearing glasses). The location presented a cluttered background and sunlight coming from the side. On average $1.8 \pm 0.16$ minutes were recorded per subject.

4.2. Accuracy Studies with Children

In order to obtain viable calibration data with children, and to capture their attention for a long enough period of time, we developed a simple computer game interfaced with a 19-inch touch-screen. The game presented objects of different kinds and colors (vehicles, fruits, animals, etc.) scanned from children’s books. One object at a time appeared at a random position and with a random scale on the screen. If the child touched the screen where the object was, the game made a funny sound (taken from animated cartoons), the object disappeared, and a new object appeared at a new location. If the child did not interact with the game for five seconds, the object jiggled a bit to reclaim the child’s attention. The variations on the scale of the objects ranged from 200 to 300 pixels.

In order to estimate the accuracy of the system over the whole field of view, the children sat on a chair at approximately 25 cm from the monitor. This resulted in the monitor covering most (>75%) of the field of view captured by the WearCam (see Figure 8b). The size of the objects presented on the monitor, once captured by the WearCam, ranged from 50 to 150 pixels depending on their position.

Ten typically developing children with no known visual impairment took part in the experiments (six girls and four boys with an average age of $29 \pm 6$
months). The children were recruited at the university campus daycare, where the recordings were made using a standard laptop. The session lasted until the children showed disinterest in the game. On average, $4.1 \pm 1.5$ minutes of data were collected for each child.

4.3. Data Labelling

The data was labeled in a different manner for the first (adult) and second (children) study. In the first study, we tracked the moving object throughout the whole video sequence for each subject by detecting the object color with 3d histograms over the YCbCr colorspace and tracking the histogram’s back-projection centroid using mean-shift tracking. The data was collected under the assumption that the subjects were constantly following the tracked object with their gaze. To reduce potential errors induced by predictive tracking of the gaze, the object was moved at a slow pace.

The data labeling for the children experiment was performed manually by a group of undergraduate students using a custom made software. The labelers scrubbed through the input videos (field of view image and eyes image) and added calibration points where it was possible to estimate the direction of the gaze, for example when the child reached for the touch-screen. Studies by [29] show that children will look at their target at the end of a reaching action. Therefore, whenever the children touched the objects on the screen in the calibration game, the labelers could place a ground-truth point. As the position of the screen in the input image was not known automatically, the labelers had to manually click on the image where the child’s finger was pointing. Moreover, the spans of time just before and after the pointing action were also considered as valid calibration candidates by looking at the shifts in the children’s eyes. The inter-experimenter correlation for the children experiment (computed over 20 minutes of input footage redundantly labelled) was of 0.91. On average we obtained 2800 ground-truth points per subject in the adult experiment and 1600 ground-truth points per subject in the children experiment.

4.4. Parameters Optimization and Tests

Overall we had to examine several parameters for the system, namely the size of the training set, the resolution of the input eyes images, the SVR $C$ cost function, the RBF kernel $\gamma$ and the $\nu$ fraction. The size of the training set is a fundamental problem in so far as it is directly related to the effort the experimenters have to put in the offline analysis. While we
collected sufficient amounts of calibration points to perform tests up to 800 samples, in the practical use only smaller amounts (50 to 200) are likely to be available\textsuperscript{4}. When working with adults, it is always possible to include a calibration phase in the experiments and so this limitation is lessened.

Instead of optimizing the training parameters for each participant, we optimize them globally for each study. While the accuracy may not be optimal for each user independently, this allows to provide an estimate of the ability of the system to cope with the individual eye differences among all the adults (or among all the children), with no need for an ulterior optimization for each new user. We however optimize separately the parameters for the children group and for the adult group as each group presents very different features.

In all experiments, the neighborhood radius $\eta$ for illumination normalization was set to 20 pixels. The threshold $\tau$ for illumination normalization weighting was set to 0.7 and the threshold $\theta$ for the eye blinking detection was set to 0.65.

5. Results and Discussion

5.1. Grid searches

We performed a 10-fold cross-validated grid search of the $\nu$ fraction, RBF kernel $\gamma$ and $C$ cost factor using a fixed training set size of 100 samples and estimating on 900 samples. The selection of samples among all data available was done randomly. Results for the adult and children experiments are displayed separately on Table 1. In the adults experiment the optimal precision ($2.23^\circ$ horizontally and $1.51^\circ$ vertically) was obtained with $\nu = 0.9$, $C = 100$ and $\gamma = 0.005$. In the children experiment the optimal precision ($3.29^\circ$ horizontally and $2.49^\circ$ vertically) was obtained when $\nu = 0.5$, $C = 25$ and $\gamma = 0.01$. We also performed 10-fold cross-validated grid search of training set size and input samples resolution $w \times h$ using the optimal parameters found on the first phase (see Figure 9). The optimal resolution we found depended on the amount of samples available: for training sets of less than 100 samples the best results were obtained with input sizes of $32 \times 24$ pixels; for bigger training sets (200 to 800) better results were obtain with

\textsuperscript{4}Obtaining more than 200 calibration points in a 5-10 minutes video without a calibration phase from the user can be challenging and time-consuming (this amounts to one point every 37-75 frames).
40 × 30 samples. The best accuracy obtained using the maximum amount of samples is shown in Table 2. As we can see on Figure 12 the region of parameter space that yields the optimal results is quite large. This might be due to the amount of support vectors that is used. With a limited amount of training samples (< 400), most if not all of the samples are selected as support vectors as long as the $C$ and $\nu$ parameters do not have extreme values. Tests with a greater number of calibration samples would refine the informations about the optimal parameters but would not reflect the real usage of the system (where the number of samples available is small).

We used the optimal parameters found in Table 1 to perform accuracy tests using increasing amounts of calibration samples. We performed 10-fold cross-validation with increasing amount of training samples, ranging from 10 to 800. Figure 10 shows the results obtained for both adults experiments (indoors and outdoors) and for the experiment with children. We can observe that above 200 samples the gain in accuracy is very limited. Table 2 shows the best results obtained both with 800 (the maximum amount of calibration samples tested) and with 200 samples.

5.2. Discussion

When we compare the accuracy obtained with adult and children subjects we have to be aware of two things: children’s eyes appear smaller in the WearCam image; children have less features in their eyes region (e.g. eyelashes, wrinkles) than adults. In both cases this results in a reduced amount of image details. Unlike in geometric-based approaches where a spe-
Figure 10: Gaze tracking error as a function of calibration samples for adults (outdoors and indoors experiments) and children. Maximum accuracy is obtained using 800 samples. However in keeping with the practical use of the calibration procedure, the results obtained with 200 samples are reported in our final accuracy estimation.

Table 1: Optimal parameters after grid search. Search performed on the cost function $C$ (0.1 to 100) which determines the amount of error we can allow during the optimization, the $\nu$ (0.1 to 1.0) factor governing the amount of support vectors that will be kept, and the gaussian kernel variance $\gamma$ (0.001 to 0.75). All tests were run using 100 calibration samples.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Adults</th>
<th>Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C$</td>
<td>100</td>
<td>25</td>
</tr>
<tr>
<td>$\nu$</td>
<td>0.9</td>
<td>0.5</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.005</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 2: Gaze tracking accuracy using optimal parameters (quartiles represent the 25% and 75% percentiles respectively)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>error (x)</th>
<th>quartiles(x)</th>
<th>error (y)</th>
<th>quartiles(y)</th>
<th>samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adults (in)</td>
<td>1.64°</td>
<td>0.76°; 2.93°</td>
<td>1.26°</td>
<td>0.58°; 2.24°</td>
<td>800</td>
</tr>
<tr>
<td>Adults (out)</td>
<td>2.28°</td>
<td>1.07°; 4.16°</td>
<td>1.78°</td>
<td>0.83°; 3.26°</td>
<td>800</td>
</tr>
<tr>
<td>Children</td>
<td>2.50°</td>
<td>1.16°; 4.49°</td>
<td>2.02°</td>
<td>0.93°; 3.57°</td>
<td>800</td>
</tr>
<tr>
<td>Adults (in)</td>
<td>1.85°</td>
<td>0.85°; 3.36°</td>
<td>1.34°</td>
<td>0.61°; 2.47°</td>
<td>200</td>
</tr>
<tr>
<td>Adults (out)</td>
<td>2.60°</td>
<td>1.19°; 4.87°</td>
<td>2.08°</td>
<td>0.93°; 3.96°</td>
<td>200</td>
</tr>
<tr>
<td>Children</td>
<td>2.65°</td>
<td>1.21°; 4.87°</td>
<td>2.19°</td>
<td>1.00°; 3.98°</td>
<td>200</td>
</tr>
</tbody>
</table>

cific feature (e.g. corneal/retinal glints) is used to extract the gaze direction, appearance-based systems exploit all eye details to learn the mapping be-
Figure 11: Error vectors in the complete field of view using different amounts of calibration samples. The accuracy in outer regions suffers the most from reduced amounts of calibration data. The image was obtained from a single recording with an adult subject, for this reason some regions are blank (where no data was obtained during the recording).

It is important to discuss the amount of calibration samples when we compare our system with other gaze trackers (both external and wearable) such as [7, 11] which only require 9 to 25 calibration samples (see Table 3).
Figure 12: Grid search on cost function $C$, $\nu$ and $\gamma$ parameters for adults (a) and children (b). The convexity of the error function allows the space of optimal parameters to be relatively wide. Moreover, with values of $C$ higher than 1, the impact of $C$ is not so evident. Error is counted in degrees.
Obtaining 200 calibration samples by tracking a colored object at 25 frames per seconds takes about 10-15 seconds. This is a very reasonable time one can spend for calibrating the system. We could obtain a higher amount of samples, but the gain in accuracy with more than 200 calibration points is very little (see Figure 10). This reduces the interest of collecting a higher amount of calibration points. In an offline calibration process such as the one we propose for our experiments with children, it is more common to only be able to obtain 100 calibration points. The average accuracy that can be expected in these cases is slightly lower than the best values we obtained \((2.84^\circ \times 2.37^\circ)\).

A note must be made about the differences between horizontal and vertical errors. During the training of the system we create two separate mappings for estimating the horizontal and vertical gaze directions. The changes of the eyes shape during horizontal movements is lesser than the variations due to vertical shifts in the gaze (one of the major features being the row of eyelashes moving upwards). This is even more true when the subject is looking downwards (and little detail except eyelid shadowing and eyelashes are visible). This can explain the important difference between horizontal and vertical errors.

In terms of performance, the training of the SVR mapping using 200 calibration samples takes less than 1 sec on a Intel Core2 2.4GHz. Training using 800 calibration samples takes around 17 seconds. Illumination normalization and SVR estimation for each frame takes on average 8 msec with 200 samples and 24 msec with 800 samples. This means that while we are currently using the system in an offline paradigm, it would suffice to implement a suitable calibration procedure in order to utilize the system online.

The accuracy we obtained with children and adults, both indoors and outdoors, is acceptable for tasks such as determining the object of attention in the field of view or for studies of central vision. Thanks to its design the system does not hinder or distract the wearer’s field of view, as no elements are placed directly in front of the eyes. Furthermore, the range in which the gaze tracking is possible is very wide, allowing to study the gaze behavior in the whole field of view instead of being contained in a limited region. Its very wide field of view also allows to record the elements in the periphery of the visual field: this allows indeed to measure what the wearer is not looking at.

With these considerations in mind we believe our system to present a novel solution to the gaze tracking problem.
Table 3: Comparison of gaze trackers in terms of calibration samples, accuracy and field of view. (Note: some systems report multiple accuracy values. In those cases we noted the results obtained in as close a manner as possible to our evaluation method.)

<table>
<thead>
<tr>
<th>System</th>
<th>Samples</th>
<th>Accuracy</th>
<th>FOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>WearCam-Adult</td>
<td>200</td>
<td>1.85° × 1.34°</td>
<td>96° × 96°</td>
</tr>
<tr>
<td>WearCam-Child</td>
<td>200</td>
<td>2.65° × 2.19°</td>
<td>96° × 96°</td>
</tr>
<tr>
<td>Baluja94[15]</td>
<td>2000</td>
<td>1.5°</td>
<td>Monitor</td>
</tr>
<tr>
<td>Tan02[16]</td>
<td>252</td>
<td>0.38°</td>
<td>13&quot; Monitor</td>
</tr>
<tr>
<td>Torricelli08[5]</td>
<td>75</td>
<td>1.7° × 2.4°</td>
<td>17&quot; Monitor</td>
</tr>
<tr>
<td>Noureddin05[9]</td>
<td>49</td>
<td>2.9° × 2.7°</td>
<td>Monitor</td>
</tr>
<tr>
<td>Beymer03[8]</td>
<td>22</td>
<td>0.6°</td>
<td>Monitor</td>
</tr>
<tr>
<td>Williams06[17]</td>
<td>16</td>
<td>0.68°</td>
<td>Monitor</td>
</tr>
<tr>
<td>tobi[30]</td>
<td>9</td>
<td>0.5°</td>
<td>17&quot; Monitor</td>
</tr>
<tr>
<td>Hennessey06[7]</td>
<td>5</td>
<td>0.45°</td>
<td>Monitor</td>
</tr>
<tr>
<td>Ohno04[6]</td>
<td>2-9</td>
<td>0.68°</td>
<td>18&quot; Monitor</td>
</tr>
<tr>
<td>Yamazoe08[19]</td>
<td>0</td>
<td>5.3° × 7.7°</td>
<td>60° × 30°</td>
</tr>
<tr>
<td>Morimoto02[18]</td>
<td>0</td>
<td>2.5°</td>
<td>-</td>
</tr>
</tbody>
</table>

6. Conclusion

We have presented a head-mounted gaze tracking system for children. The system can be used even when the wearer is not able to participate in its calibration. The task of obtaining calibration points is postponed and is performed by the experimenters. Using an appearance based approach, we are able to track the gaze of the wearer on the extremely wide field of view of the WearCam (96° × 96°). We use ν Support Vector Regression with a RBF kernel to create the mapping between eyes appearance and gaze direction. To reduce the influence of the environmental lighting, a high-pass homomorphic filter is applied to the image of the eyes as a pre-processing step. In the studies we made with adult subjects we obtained an average accuracy of 1.59°, while the studies we made with 2 to 3 years old children resulted in a 2.42° accuracy. Tests outdoor under a strong sunlight resulted in a 2.34° accuracy.

The system was designed to be used with children, providing both a hardware that is compliant with their needs and constraints, and a method that unburdens the users of calibration actions. At the time of the writing of this
paper, no other system allows to conduct similar experiments. At the same time, the results obtained with adult subjects are competitive with other gaze tracking systems with passive lighting and can therefore be considered as a viable solution for gaze tracking applications. Moreover, the field of view provided by the WearCam is more than three times that of other wearable systems, allowing to study the contents of peripheral vision as well as central vision.


[12] Sr research eyelink ii.
URL http://www.eyelinkinfo.com

URL http://www.a-s-l.com


