CALIBRATION-FREE EYE GAZE DIRECTION DETECTION WITH GAUSSIAN PROCESSES

Basilio Noris, Karim Benmachiche and Aude G. Billard
LASA Laboratory - EPFL, Station 9, CH-1015 Lausanne, Switzerland

Keywords: Eye gaze detection, wireless head mounted camera, gaussian processes, appearance-based.

Abstract: In this paper we present a solution for eye gaze detection from a wireless head mounted camera designed for children aged between 6 months and 18 months. Due to the constraints of working with very young children, the system does not seek to be as accurate as other state-of-the-art eye trackers, however it requires no calibration process from the wearer. Gaussian Process Regression and Support Vector Machines are used to analyse the raw pixel data from the video input and return an estimate of the child's gaze direction. A confidence map is used to determine the accuracy the system can expect for each coordinate on the image. The best accuracy so far obtained by the system is 2.34° on adult subjects, tests with children remain to be done.

1 INTRODUCTION

For several years now, the detection of the direction of gaze has been used in a variety of different domains, from psychological investigations on reading to the study of visual points of interest in advertisement or for Human Computer Interfaces (HCI)(Duchowski, 2002). The measure of gaze patterns, fixations duration, saccades and blinking frequencies has been the topic of many works(Young and Sheena, 1975). The introduction of real-time gaze detection has opened the way to a multitude of user interfaces such as writing systems for disabled people or virtual reality (VR) navigation controls or to measure the focus of attention(Stiefelhagen, 2002).

Gaze detection systems can be separated into non intrusive and head mounted devices. Non intrusive systems usually use an external camera filming the user and detecting the direction of the eyes with respect to a known position. However the task is rendered more difficult by the variability in the user's head direction. Head mounted devices allow to detect the direction of the gaze without having to cope with the pose of the subject's head. Most modern systems use infra red (IR) lighting to illuminate the pupil and then extract the eye orientation by triangulation of the IR spotlights reflections or other geometrical properties(Zhu et al., 2002; Ohno and Mukawa, 2004). When IR lighting is impracticable, image based methods have been used to estimate the direction of the eyes(Baluja and Pomerleau, 1994; Tan et al., 2002). Regardless of the method used, a calibration involving the user looking at known locations is performed beforehand. The most accurate systems can achieve precisions below 0.5° of the visual field(Tan et al., 2002; Hennessey et al., 2006).

Figure 1: A normally developing 14 months old child, wearing the WearCam with eye-mirror.

In this work, we developed a head mounted gaze direction detection system designed for 6 months old children. The system aims at studying the visual behaviour of young children to help in the diagnosis of developmental disorders. Children with developmental disorders such as Autism Spectrum Disorder (ASD) have a strong tendency to avoid eye contact or to avoid looking at people altogether. This behaviour

Noris B., Benmachiche K. and G. Billard A. (2008).
CALIBRATION-FREE EYE GAZE DIRECTION DETECTION WITH GAUSSIAN PROCESSES.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 611-616
DOI: 10.5220/0001082706110616

is often present in very young children already, but it is difficult to measure and analyse. In this work we used a WearCam(Piccardi et al., 2007), a lightweight wireless head mounted camera developed in our lab, to record the field of view of the infants. The recordings could then be analysed to verify and study the children's social interaction aptitude. However, early assumptions that the head direction sufficed to estimate the direction of the child's focus of attention (i.e. the object of interest is always at the center of the field of view) proved unsatisfying and thus we added the detection of the gaze direction into the analysis. The recordings from the WearCam present several challenges: the camera can incur into severe lighting changes as the child moves the head towards bright areas of the environment; the wireless transmission can overlay distortion artifacts over the video image. For these reasons and due to the absence of user calibration, the accuracy we aim at is not as high as other state-of-the-art systems.

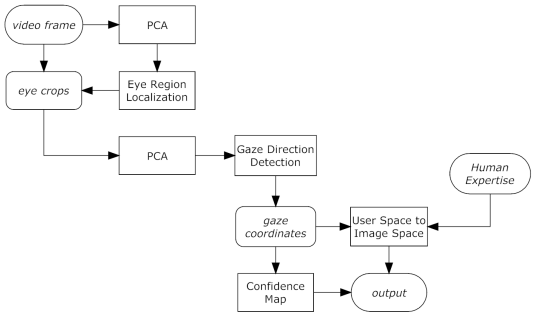

Figure 2: Flowchart of the gaze detection system. The video frame from the camera is first projected to a lower dimensionality with PCA and the result is used for the localization of the right and left eye regions. When these are known, the corresponding pixels from the source image are extracted and projected once again into a lower dimensional space for the gaze direction detection process. The resulting gaze direction coordinates are then modified to obtain the actual coordinates in the image. Additionally, the gaze coordinates are used to read the system confidence on its results. These informations are combined to give the final output.

2 CALIBRATION-FREE GAZE DETECTION

Working with children entails a number of constraints on the hardware that can be used. The system is designed to be used in a freeplay environment. As such a head mounted solution becomes essential. However, to avoid distracting the infant the use of goggles is precluded. Likewise, it is not possible to use IR leds pointing towards the child's eyes. Our system uses a 42g wireless camera mounted on the child's head using adjustable straps. The camera films the frontal field of view of the child. The field of view is of 74° horizontally and 56° vertically. A small mirror protruding from the bottom part of the camera reflects the eye portion of the wearer's face. The small mirror occludes 20% of the bottom region of the image, reducing the vertical field of view to 44°. The camera can be tilted to be aligned with the eyes of the wearer. In the cases where the children are too aware of the mirror, the mirror can be placed on the upper part of the camera, becoming less disturbing to the child.

Most works on gaze tracking systems can benefit from a certain degree of cooperation from the users in terms of calibration processes and predictability. Unfortunately this is not the case with young children. For this reasons, the accuracy of the gaze tracking system had to be sacrificed to enable the system to work with little to no calibration. However, the system as it is intended does not require subpixel precision methods and the rough region estimate that we can expect from a calibration-free system can be be sufficient to suit our needs.

The camera is adjustable to take into account the different eye positions and head shapes of the children. Due to these differences, the first step for gaze tracking is localizing the eyes (Fig. 2 shows a schematic representation of the detection process).

2.1 Eye Region Localization

To be able to localize the eye regions, the mirrored region of the input images was extracted from the camera videos. The 768 by 120 pixels image was first reduced to a 30 by 10 pixels then projected on the first 60 principal components to reduce its dimensionality. The choice of components was made by evaluation of the reconstruction error to explain 99% of the data. To extract the cropping values for the eyes regions (top, bottom, left, right) in the sample images we used Gaussian Process Regression (GPR)(Rasmussen and Williams, 2005). Although not extensively, Gaussian Processes have been used in the computer vision domain for appearance-based approaches (Kim et al., 2006)(Meng et al., 2000). To test GPR against a more established method, Support Vector Machines (SVM)(Osuna et al., 1997) were used. Separate systems were trained for left and right eyes. For GPR the Rational Quadratic (RQ), Squared Exponential(SE), Neural Network (NN) and Matern covariance functions were used (a thorough description of these covariance functions can be found in(Rasmussen and Williams, 2005)). For SVM we used a RBF kernel with variating sigmas. The eye region is evaluated over a short lapse of video frames and averaged to

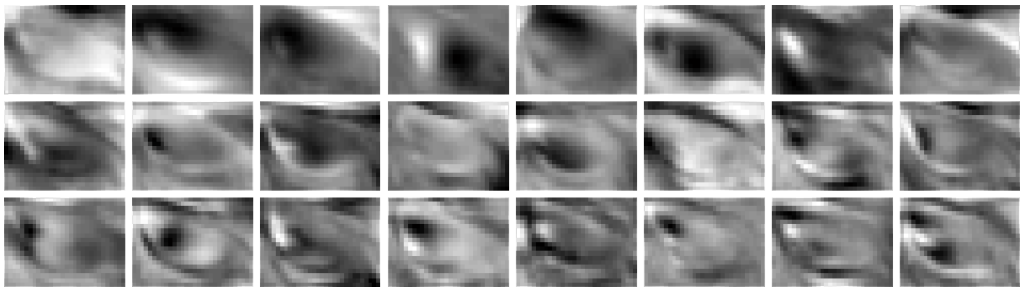

Figure 3: First 24 principal components of the right eye region samples.

obtain the output for the eye localization.

2.2 Gaze Direction Detection

Once the location of the eye is known, the left and right eyes regions of the image are cropped and combined to extract the most likely direction of the gaze in terms of coordinates on the full image. The GPR is trained using the raw pixels intensities of the eye regions as input and the coordinates on the image of the point the wearers were looking at as output. Training was done both using a single eye at a time or by combining both eyes together. A simple setup with a colored object on a black wall was used to generate the data. This yields a mapping between eye appearance and gaze direction in the image. To reduce computational costs, PCA was applied on the input eye regions. The principal components were computed on a subset of the training data. 99% of the data is explained by the first 24 components (See Fig.3). The same amount of significant principal components was found with sample resolution of 60 by 60 pixels down to 5 by 5 pixels. However, due to the high variability of eye shapes, the first principal components contain information manly pertaining to the shape and position of the eye corners. In order to be able to represent in a more detailed way the small changes in the direction of the gaze (i.e. small changes in the iris position), a total of 40 principal components was used. A resolution of 40 by 40 pixels was used to allow enough quality while manually reviewing the eye samples.

2.3 Confidence Map

Regardless of whether a calibration process is available, the problem of detecting the gaze direction is more challenging when the eyes are pointing in certain directions (e.g. detecting the direction of the eyes when the user is looking at extreme angles as opposed to when the user is looking straight forward). To take into account this fact we generated, for each region

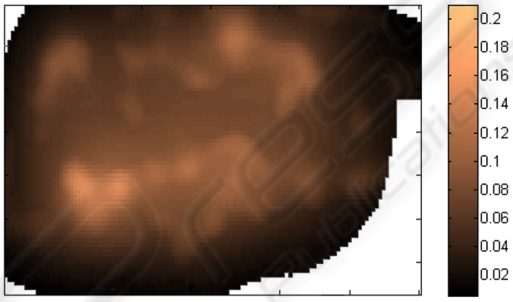

Figure 4: Horizontal confidence map for gaze detection. For each coordinate in the image, the average training error is computed to estimate the precision that can be expected for detections at those coordinates. The brighter the pixels, the bigger the error. The values are computed in fractions of the whole image width. The horizontal and vertical confidence maps are computed separately.

in the field of view, a map of the confidence the system manages to achieve(see Fig. 4). The map was created by segmenting the image space into several regions and computing the mean square errors of all estimation made inside each area. As the training data does not fill every single region of the image, multiple resolutions were used so as to be able to average over regions where the data is too sparse. The vertical and horizontal errors were computed separately, resulting in two different maps. The map can then be used to accept or reject results that fall off a certain confidence. Moreover this can be useful where one of the directions of the gaze (vertical or horizontal) does not play a critical role.

The resulting confidence maps allow for each gaze direction detected to have an estimation of how sure the system is of its decision. As the maximum precision of the system is far inferior to state-of-the-art methods, this method renders the system usable for a number of applications (e.g. determining if an object is inside the wearer's focus of attention).

3 USER SPACE TO IMAGE SPACE

The system is sensitive to the variation in shape and position of the eyes as seen from the reflecting mirror. If the camera is positioned slightly upwards on the forehead, the angle from which the eyes will be seen will be different which might appear as though the wearer is looking downwards. Additionally, depending on the success of the eye localization process, the cropped eye region can result shifted compared to the actual eye position, which results in a shift in the gaze direction detections.

However the detections for a single subject usually yield coherent results: if the wearer follows the edges of his/her field of view, the resulting detections will feature a similar behaviour, but covering a space that will be offsetted and rescaled compared to the full image. This results in a user-dependent coordinate space that can be transformed into the actual image space by a linear tranformation, reducing significantly the detection error. Since the analysis of the videos is done offline, it is possible to learn the transformation between the two spaces (user-space and image-space) by manually selecting a few reference points.

A reference point in one image is selected and its coordinates are coupled with the detection from the system. If only one reference point is selected, the offset computed as the distance between reference and detection points is applied to all successive detections. If several reference points are available (i.e. it is possible to determine by looking at the image that the wearer is looking at a specific location) the Procrustes analysis(Small, 1996) applied between the reference and detection points (see Fig.5). This gives a linear transformation between the user-space and the actual image-space coordinates which can be used to correct all further detections.

4 EXPERIMENTAL SETUP AND RESULTS

4.1 Dataset

In order to train the system and to obtain ground truth data, we collected a database of video footage from 33 adult subjects[1] evenly distributed between bright and dark eye colors and of different skin tones. The subjects sat in front of a black wall and followed with their eyes a colored object moving in random patterns covering most of the subjects' field of view (see Fig. 5). An average of two minutes per subject were recorded for a total of 67'000 frames of video. Ground truth coordinates of the colored object were obtained by semi-automatic tracking with Adobe After Effects. This forced us to make the assumption that the subjects eyes were constantly following the object, even though empirical analysis of the video footage show that a certain amount of predictive tracking and saccades appears in all subjects.

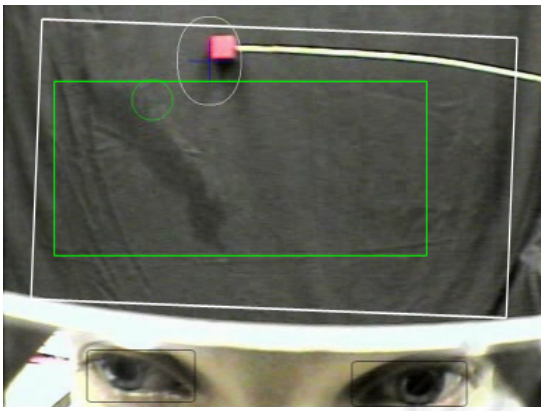

Figure 5: Experimental setup and corresponding system response. The inner circle and rectangle correspond to the raw response of the system in user-space coordinates, the rectangle displays an approximation of the boundaries of the user-space. The outside ellipse and rectangle correspond to the response after the user space to image space conversion. The conversion parameters were extracted from 3 manually set reference points. The axes of the ellipse show the confidence of the system obtained from the confidence map.

The training set and testing set were generated subject-wise using 22 subjects for training and 11 for testing. 11-fold cross-validation was performed to take into account inter-subject variability. GPR was trained by minimizing its hyperparameters on the whole training set. However, SVM parameters were optimized via cross-validation by training the system on two thirds of the training set and validating on the remaining third (14 subjects for training, 8 for validation). Once the optimal parameters were found, the system was retrained on the whole training set. For the gaze direction detection process tests were made using the information from each eye separately or using the information of both eyes together.

4.2 Eye Localization Results

Table 1 shows the results for the eye localization performance of GPR and SVM with different parameters. The error is computed in terms of the mean square distances between the ground truth and estimated boundaries (upper, lower, left and right) of the eye regions.

[1] Obtaining ground truth data from children can be challenging and will be attempted after the system will have been validated with adult subjects

Table 1: Eye Region Localization error with different algorithms computed in terms of mean square distances from the manually set ground truths (as fractions of the whole image width). *(ISO: isotropic, ARD: Automatic Relevance Determination, ONE: single variable parameterization).*

	left eye	right eye	combined
GPR NN ONE	0.1334	0.1259	0.1296
GPR SE ISO	0.1539	0.1358	0.1449
GPR SE ARD	**0.1208**	0.1491	0.1349
GPR Mat3 ISO	0.1403	0.1177	0.1290
GPR RQ ISO	0.1387	**0.1104**	**0.1245**
GPR RQ ARD	0.1692	0.1473	0.1583
SVM RBF	0.1245	0.1490	0.1368

Table 4: Comparison of different eye gaze tracking systems. *(In each case it is noted whether the system requires offline training and user calibration).*

system	accuracy	train	user calib
Baluja-Pomerleau	1.5°	Yes	Yes
Tan-Krieg.-Ahuja	0.38°	No	Yes
Ohno-Mukawa	1°	No	Yes
Hennessey	0.46°	No	Yes
Our System	2.34°	Yes	No

All distances are normalized by the screen resolution.

The localization accuracy of the systems for different eyes is not constant. This might be due to a difference in the environment illumination. Overall the tested methods perform similarly. The GPR with an isotropic Rational Quadratic covariance function achieves the best performances.

4.3 Gaze Direction Detection Results

The gaze detection performance can be seen in Tables 2 and 3. As could be expected, combining the information from both eyes provides a greater accuracy than using one single eye at a time (see Table 2). Gaussian Processes obtain good results with most covariance functions ranging from 3.33° of average error up to 4.15°. Support Vector Machines achieve better results; the best SVM trained used a RBF kernel ($\sigma = 2.7$) and 1010 support vectors. The best accuracy obtained is 2.97° horizontally and 2.34° vertically. Table 4 shows a comparison of the best performances of existing methods.

5 CONCLUSIONS AND FUTURE WORKS

We have presented a wearable gaze direction detection system based on image appearance, Support Vector Machines and Gaussian Process Regression. The accuracy obtained is inferior to the existing eye trackers, attaining on average 2.34°. However our system does not need any prior calibration and can be used in cases where the cooperation of the user cannot be expected (e.g. working with infants).

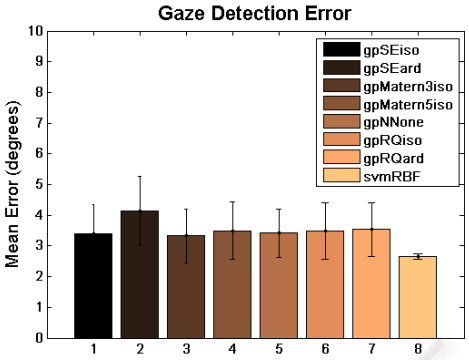

Figure 6: Gaze detection error with different algorithms. While all algorithms perform comparatively, svm outperforms gaussian processes by more than half a degree of error. Moreover the detection performance is less sensitive to changes in the training set.

The system is designed to be used with children, therefore the next step will be to collect a training database with children. Although only preliminary tests have been made, the eye region of children's faces presents less variability as eyelashes, eyebrows and other facial features are not as pronounced as in adults. As such the performance of the system may increase when used by children.

Currently, no special processing is done for eye blinks, the current results (see Table 3) just consider all frames as valid. On average a person blinks 16 times per minute; although measures havent been done, the blinking frequency during a gaze tracking exercice is probably higher, however the amount of blinking during the whole recordings is negligeable. Nevertheless, a method for detecting eye blinking and discarding the corresponding results will have to be developed to increase the fiability of the system.

ACKNOWLEDGEMENTS

This work was supported by the Thought in Action (TACT) project, part of the European Union NEST-Adventure Program, and by the Swiss Science Fundation within the National Center for Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2).

Table 2: Mean gaze angle error (in degrees) using the left eye only, the right eye only or both eyes together.

	left eye	right eye	both eyes
GPR SE ISO	4.4187 (±0.0443)	4.5344 (±0.2956)	**3.4118** (±0.9208)
GPR SE ARD	5.0673 (±0.2388)	4.6751 (±0.2908)	**4.1476** (±1.1279)
GPR Matern3 ISO	4.3318 (±0.4674)	4.1894 (±0.2728)	**3.3326** (±0.8804)
GPR Matern5 ISO	4.7746 (±0.1925)	4.3421 (±0.0967)	**3.5014** (±0.9427)
GPR NN ONE	4.4431 (±0.2189)	4.5548 (±0.2448)	**3.4194** (±0.7898)
GPR RQ ISO	4.4050 (±0.1574)	4.0865 (±0.1970)	**3.4965** (±0.9157)
GPR RQ ARD	4.6159 (±0.3592)	4.7742 (±0.1114)	**3.5334** (±0.8798)
SVM RBF	3.4511 (±0.1060)	3.6634 (±0.0963)	**2.6547** (±0.0941)

Table 3: Mean gaze angle error (in degrees) and corresponding standard deviation using the information from both eyes. The error has been computed separately on vertical and horizontal directions.

	horizontal	vertical	combined
GPR SE ISO	3.7896 (±1.1611)	3.0340 (±0.6805)	3.4118 (±0.9208)
GPR SE ARD	4.7569 (±1.5538)	3.5382 (±0.7021)	4.1476 (±1.1279)
GPR Matern3 ISO	3.6601 (±1.1772)	3.0050 (±0.5835)	3.3326 (±0.8804)
GPR Matern5 ISO	3.8068 (±1.2408)	3.1959 (±0.6446)	3.5014 (±0.9427)
GPR NN ONE	3.7568 (±1.1092)	3.0820 (±0.4704)	3.4194 (±0.7898)
GPR RQ ISO	3.9195 (±1.2942)	3.0734 (±0.5371)	3.4965 (±0.9157)
GPR RQ ARD	3.9179 (±1.2527)	3.1490 (±0.5070)	3.5334 (±0.8798)
SVM RBF	**2.9700** (±0.1059)	**2.3394** (±0.0824)	**2.6547** (±0.0941)

REFERENCES

Baluja, S. and Pomerleau, D. (1994). Non-intrusive gaze tracking using artificial neural networks. Technical report, Pittsburgh, PA, USA.

Duchowski, A. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments and Computers*, 34(4):455–470.

Hennessey, C., Noureddin, B., and Lawrence, P. (2006). A single camera eye-gaze tracking system with free head motion. In *ETRA '06: Proceedings of the 2006 symposium on Eye tracking research & applications*, pages 87–94, New York, NY, USA. ACM Press.

Kim, H.-C., Kim, D., Ghahramani, Z., and Bang, S. Y. (2006). Appearance-based gender classification with gaussian processes. *Pattern Recogn. Lett.*, 27(6):618–626.

Meng, L., Nguyen, T. Q., and Castanon, D. A. (2000). An image-based bayesian framework for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 302–307.

Ohno, T. and Mukawa, N. (2004). A free-head, simple calibration, gaze tracking system that enables gaze-based interaction. In *ETRA '04: Proceedings of the 2004 symposium on Eye tracking research & applications*, pages 115–122, New York, NY, USA. ACM.

Osuna, E., Freund, R., and Girosi, F. (1997). Training support vector machines: an application to face detection. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 130, Washington, DC, USA. IEEE Computer Society.

Piccardi, L., Noris, B., Schiavone, G., Keller, F., Von Hofsten, C., and Billard, A. G. (2007). Wearcam: A head mounted wireless camera for monitoring gaze attention and for the diagnosis of developmental disorders in young children. In *RO-MAN '07: Proceedings of the 16th International Symposium on Robot and Human Interactive Communication*.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Small, C. (1996). *The statistical theory of shape*. Springer, New York.

Stiefelhagen, R. (2002). Tracking focus of attention in meetings. In *ICMI '02: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, page 273, Washington, DC, USA. IEEE Computer Society.

Tan, K.-H., Kriegman, D. J., and Ahuja, N. (2002). Appearance-based eye gaze estimation. In *WACV '02: Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision*, page 191, Washington, DC, USA. IEEE Computer Society.

Young, L. and Sheena, D. (1975). Survey of eye movement recording methods. *Behavior Research Methods and Instrumentation*, 7:397–429.

Zhu, Z., Fujimura, K., and Ji, Q. (2002). Real-time eye detection and tracking under various light conditions. In *ETRA '02: Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 139–144, New York, NY, USA. ACM Press.