Pervasive Retail Strategy Using a Low-Cost Free Gaze Estimation System

Dario Cazzato(✉), Marco Leo, Paolo Spagnolo, and Cosimo Distante

National Research Council of Italy - Institute of Optics, Arnesano, LE, Italy
{dario.cazzato,marco.leo,paolo.spagnolo,cosimo.distante}@ino.it
http://www.ino.it/en/

Abstract. This paper proposes a pervasive retail architecture based on a free human gaze estimation system. The main aim of the paper is to investigate the possibility to automatically understand the behavior of the persons looking at a shop window: this is done by a gaze estimation technique that uses a RGB-D device in order to extract head pose information from which a fast geometric technique then evaluates the focus of attention of the persons in the scene (even more persons at the same time). The main contribution concerns with the application into this challenging research field of a gaze estimation working without any initial calibration and, in spite of this, able to properly deal with completely unaware persons moving in unconstrained environments. Preliminary experiments were conducted in our lab in order to quantitative validate the accuracy of the gaze estimation on different benchmarks of persons. Then the qualitative evaluation of the effectiveness of the proposed architecture was conducted in a real shop window demonstrating the ability to deal with real challenging environmental conditions.

Keywords: Free gaze estimation · Human-computer interaction · Focus of attention · Pervasive retail · Digital signage

1 Introduction

Gaze tracking plays a fundamental role to understand human attention, feelings and desires [17]. Automatic gaze tracking can open to several application in the fields of human-computer interaction and human behavior analysis, therefore several techniques and methods have been investigated in recent years. When a person is in the field of view of a static camera, gaze can give information about the focus of attention of the subject, allowing for gaze-controlled interfaces for disabled people [22], driver attention monitoring [10], pilot training [33], provision of virtual eye contact in conferences [32] or for the analysis of marketing strategies [25]. Concerning the last, consumers' visual attention estimation can lead to understand underlying display organization, define the optimal disposition of product in shelves, conduct statistics about most interesting products and several other applications. Therefore, several works that make use of computer

C. Distante et al. (Eds.): VAAM 2014, LNCS 8811, pp. 23–39, 2014.
DOI: 10.1007/978-3-319-12811-5_2

vision and pattern recognition algorithms have been presented in the last few years, as well as design principles to reproduce intelligent environments that are sensitive and responsive to the presence of users and their environment on a large scale, like in [9]. In [26], a camera enhanced digital signage display is presented. The system can extract metrics like person's dwell time, display in-view time and attention time. For the last, a multi-view Active Appearance Model (AAM) registration method is used to estimate head orientation. The work of [31] head pose is used to infer people's visual focus of attention in dynamic meeting scenarios. In [29], a study about searching for a target by consumers performed by using infra-red (IR) markers and an head-mounted eye tracking system is proposed. A review of the usage of commercial eye-tracker in marketing analysis can be found in [25].

Most of the works in the state of the art uses only a face detection algorithm to determine whether observers are facing the object of interest, or a discrete head pose estimation procedure to reveal the macro-area of interest. Moreover, works that try to understand the focus of attention for an environment make use of a commercial eye-tracker, making the overall cost of the system prohibitive for the retail market. Instead, works that use low-cost systems try to understand the focus of attention on a single object, like a target or a screen. This paper presents an intelligent shop window architecture for indoor environments able to understand where persons are looking at. The architecture makes use of an RGB-D device in order to extract head pose information as the input for a fast geometric gaze estimation technique that evaluates the focus of attention of the persons in the scene. The contributions of the work under consideration are that:

- an estimation of the gaze ray for users that are looking on a shop window and understand the observed object is proposed;
- the presented system is low-cost and makes use of a commercial depth sensor;
- the system can handle more users at the same time;
- no calibration nor training phases are required;
- privacy principles in the field of ubiquitous computing are followed, based on [6,19];
- a contribution to a computer vision problem (i.e. free gaze estimation) is given by our technique.

Preliminary experiments were conducted in our lab in order to quantitative validate the accuracy of the gaze estimation on different benchmarks of persons. Then the qualitative evaluation of the effectiveness of the proposed architecture was conducted in a real shop window demonstrating the ability to deal with real challenging environmental conditions. The followings of the paper is organized as follows: Sect. 2 gives an overview of the related works about gaze estimation and it highlights the contributions of the proposed solution with respect the leading state of the art methods. Section 3 deeply details the proposed method whereas in Sects. 4 and 5 the experimental setup and results are reported and discussed. Finally conclusions are in Sect. 6.

2 Related Works on Gaze Estimation

A survey of existing works and a detailed classification of methods can be viewed in [14]. Most gaze tracking methods are based on Pupil Center Corneal Reflection (PCCR) technique, [13,23]. It obtains the pose of the eye using the center of pupil contour and corneal reflections (glint) on the corneal surface from point light sources, usually one or multiple infrared (IR) lights. This method is not quite appropriate for general interactive applications. Usually a high-resolution camera is needed, and extra IR lights and the camera need to be calibrated carefully. Here, we concentrate on eye-tracking solutions without the usage of the beforehand exposed technique. These solutions can be divided into feature-based and appearance-based. Feature-based gaze estimation methods use extracted local features like contours or eye corners, while appearance-based methods utilize the image contents as an input with the intention of mapping these directly to screen coordinates, without requirements for scene geometry nor a camera calibration, but they need a significative high number of calibration point and, in general, are not head pose invariant. The work of [12] proposes an appearance-based method to achieve gaze estimation from multimodal Kinect data that is invariant to head pose, but it needs a learned person-specific 3D mesh model. In [8], after a one-time personal calibration, facial features are tracked and then used to estimate the 3D visual axis, proposing a 3D geometrical model of the eye. The method needs to accurately detect eye corners in order to create a complete 3D eye model. In [18], gaze tracking is performed using a stereo approach to detect the position and the orientation of the pupil in 3D space. A low-cost system with low-resolution webcam images, allowing for cursor control systems, is presented in [15]. The work of [16] proposes a method to estimate gaze tracking using a single and low-quality webcam. It limits head movements, but assumes that if the head moves, a head pose is estimated by an external program. A calibration phase is required. Valenti et. al [30] combine head pose and eye location information to accurately estimate gaze track. Eyes are located using isophote properties to obtain the center of semicircular patterns; head pose utilizes the cylindrical head model approach [34]. Their result are suitable for several applications, but they need a calibration phase and are tested only at a distance of 750 mm.

Most of state of the art work operates in constrained condition and needs a learning phase, using manually labeled data to train various type of classifiers. Furthermore, often a calibration phase occurs. Even methods that are considered unconstrained can work only in a very short range of head pose variations, and often the allowed translation is of less than 10 cm. Moreover, most of methods that produce the eye-gaze track are evaluated at a distance of 50–75 cm, when reported. Finally, often head-mounted devices are used.

Generally, head-pose is considered as a coarse gaze estimation technique. Authors of [28] assert that the head pose contributes to about 70 % of the visual gaze and focus of attention estimation based on head orientation alone can get an average accuracy of 88.7 % in a meeting application scenario. The perception of eye-gaze direction is also influenced by parameters like head contour and

nose angle [20]. Despite that, the idea of achieving gaze tracking using head pose information only is not new. A work that utilizes the same approach is in [7]. Here, using only the head pose information and given a model of the environment, the system is automatically able to give the estimation of the viewed object. In [27] a method for estimating where a person is looking in images where the head of a person is about 20 pixel high is presented. Here, eye information is not available, and estimation is made over head pose and the general body direction, combining direction and head pose using Bayes' rule to obtain the joint distribution over head pose and direction. The method proposed in [4] introduces the use of a classifier without any hand labelled data but based only on the output from an automatic tracking system in surveillance scenarios. In [11], a method that estimates the gaze direction accurately using information on both head and body pose directions and without using eye information is analyzed. Even in [3] the visual focus of attention is recognized by evaluating head pose information, getting anyway encouraging results. The work of [24] use head pose information to control a mouse, but investigating only the 2D information coming from a consumer camera. Finally, in [35], gaze direction is estimated by considering head posture information and using information that comes from the pupil, but it has been tested at a distance of 40 cm only.

Most of these methods pay attention only to the rough area of interest of the person, and are not seen as a possible technique to obtain the control of a device. Furthermore, no study is performed on the feasibility of an accurate gaze estimator that considers the more precise information that can be achieved using a device like a depth sensor. In the proposed work, a Microsoft Kinect is used to investigate same aspects, not only detecting the area of attention but trying also to achieve the exact position of gaze tracking ray.

In summary, our proposed work differs from the state of the art in the following aspects. First of all, we try to achieve the gaze estimation ray without using information different from head pose. Secondly, our method doesn't need both a training phase and a calibration phase. Thirdly, our work is tested with several different distance ranges going from 70 to 250 cm. Finally, in order to answer to our question, a full experimental setup was created and tested with different people, and all examinations and results are illustrated.

3 Proposed Method

3.1 Head Pose Estimation

The head pose estimation module takes care of supplying the information about rotation angle from the frontal pose, in terms of yaw, pitch and roll and translations, in meters, from the sensor's position.

Head pose estimation is a problem with 6-DOF, and can be represented with the parameter vector $\boldsymbol{p} = [\omega_x, \omega_y, \omega_z, t_x, t_y, t_z]$, where $\omega_x, \omega_y, \omega_z$ are the rotation parameters and t_x, t_y, t_z are the translation parameters. They define the 3-DOF rotation matrixes $R_{3\times3}$ as:

$$R = \begin{bmatrix} 1 & -\omega_z & \omega_y \\ \omega_z & 1 & -\omega_x \\ -\omega_y & \omega_x & 1 \end{bmatrix} \tag{1}$$

and the 3-DOF translation vector $T_{3\times1}$ as:

$$T = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{2}$$

The rigid motion of a head point $\boldsymbol{X} = [x, y, z, 1]^{\mathrm{T}}$ between time t and time $t+1$ is:

$$\boldsymbol{X}(t+1) = M \times \boldsymbol{X}(t), \tag{3}$$

where M is defined as [21]:

$$M = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \tag{4}$$

Let point $\boldsymbol{X}(t)$ be projected on the image plane in $\boldsymbol{u} = \begin{bmatrix} u_x & u_y \end{bmatrix}^{\mathrm{T}}$. The explicit representation of the perspective projection function in terms of the rigid motion vector parameters and the coordinates of the point at $t+1$ is:

$$\boldsymbol{u}(t+1) = \begin{bmatrix} x - y\omega_z + z\omega_y + t_x \\ x\omega_z + y - z\omega_x + t_y \end{bmatrix} \cdot \frac{f_L}{-x\omega_y + y\omega_x + z + t_z}(t) \tag{5}$$

where f_L is the focal length.

The system has been realized using the Kinect for Windows SDK [1]. The used model is the Candide-3 [2], and 2D coordinates of the key points on the aligned face in video frame coordinates are available. 121 2D feature points are tracked. By the Iterative Closest Point (ICP) [5] technique, by which a 3D point cloud model is iteratively aligned with the available 2D facial features (target), a 3D model on a detected face is built. The algorithm revises the transformation, i.e., combination of translation and rotation, needed to minimize the distance between the model and the target. Yaw, pitch and roll angles are extracted basing on the estimated rotation of the overlapped mask with respect to the Candide-3 model frontal pose. The X, Y, and Z position of the user's head are reported based on a right-handed coordinate system (with the origin at the sensor, Z pointed towards the user and Y pointed up).

For our purpose, both RGB and depth images are at a resolution of 640×480. Figure 1 shows the 3D mask overlapped to the 2D facial image in three different frames. From the figure it is possible to observe that the face tracker works also in presence of non frontal views. Multiple persons in the scene at the same time are also managed by the system.

3.2 Gaze Estimation

Our gaze estimation method works as follows. First of all, the 2D position of the detected eye center points are taken from the face mask. Note that small

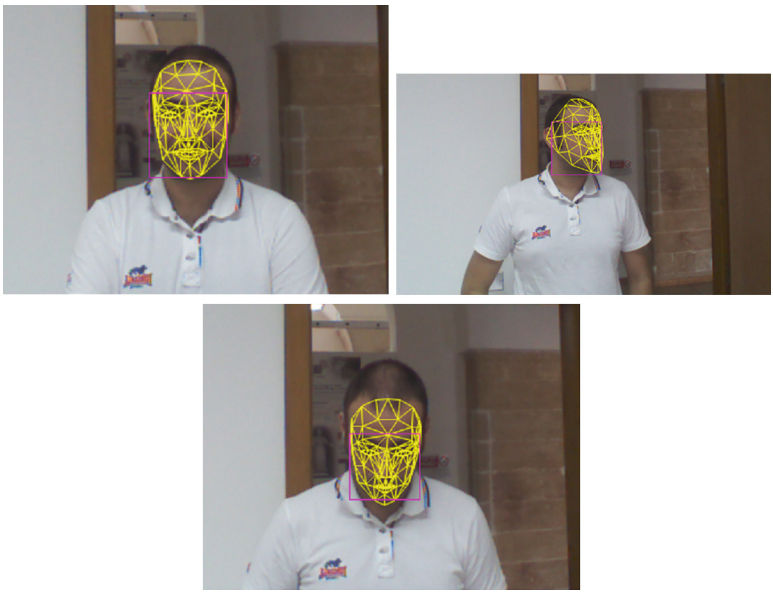

Fig. 1. Three different snapshots of the face tracking module.

occlusions are handled, and the eye center point is always estimated in the used model when the overlapping with the face successes. After that, the average value is taken, in order to take a point corresponding more or less with the nose septum and to use it as the origin of the gaze track. This value is converted into 3D coordinates with regard to a cartesian coordinate system centered inside the sensor. Starting from the head pose information, two angles are taken, i.e. ω_x and ω_y, corresponding respectively to pitch and yaw. Then, the gaze track is computed and its intersection with a plane vertical with regard to the ground and passing from the center of the sensor is calculated. Note that with this method it is possible to achieve also the intersection point with every plane parallel to the considered one, just adding a translation parameter k that will be algebraically added to the depth information, and then using the exposed procedure.

The intersection point is computed separately for each angle and using the same method, and the euclidean distance from the sensor can hereafter be computed. The procedure is showed for one angle in Fig. 2 and described in the followings. The Kinect sensor can give the length of the segment $\overline{AC}$ as the component t_z of the translation vector T. It follows that, knowing a side and an angle, we can completely solve the right-angled triangle $\widehat{ABC}$. In particular, $\overline{AB} = \frac{\overline{AC}}{\cos \omega_y}$ and $\overline{BC} = \sqrt{\overline{AB}^2 - \overline{AC}^2}$. Using the same coordinate system, it is possible to compute also the cartesian equation of the gaze ray as the straight line passing for points $A = (x_A, y_A, z_A)$ and $B = (x_B, y_B, z_B)$ expressed as:

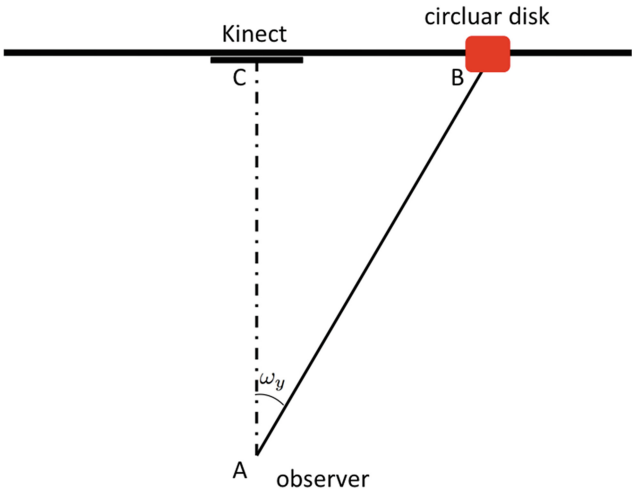

Fig. 2. A scheme of the gaze estimation solution.

$$r : \begin{cases} \frac{x-x_A}{x_B-x_A} = \frac{y-y_A}{y_B-y_A} \\ \frac{y-y_A}{y_B-y_A} = \frac{z-z_A}{z_B-z_A} \end{cases} \tag{6}$$

with $z_A = 0$ for the particular plane under consideration.

In case of translations on the x and y axes, the vector can be algebraical summed up with the computed value, in order to translate the gaze vector to the right position. Finally, in order to represent the real intersection point with the environment and to realize experimental tests, coordinates are normalized to image plane coordinates with the generic formula, valid for both coordinates x and y of the image plane:

$$c_{norm} = c - \frac{\text{bound}_{\text{low}}}{\text{bound}_{\text{upp}} - \text{bound}_{\text{low}}} \cdot \text{size}(I) \tag{7}$$

where $\text{bound}_{\text{upp}}$ and $\text{bound}_{\text{low}}$ are the two bounds, in meters, of the space, and $\text{size}(I)$ is the width (or height, depending on the coordinate in exam), expressed in pixels.

4 Experimental Setup

The environment for validation was defined as follows: a Microsoft Kinect device was positioned in front of the person, at a distance of 150 cm from the ground. Just behind the sensor, a panel with a set of 14 circular markers was positioned on its surface. Disks were distributed at a distance of 50 cm among x,y or both axes. Figure 3 shows one quarter of the panel, exactly the upper-leftmost. All their distances from the sensor are known, and they are used as ground truth

Fig. 3. A portion of the used panel for testing.

information. Finally, the normalization Eq. 7 is used to bring back a distance range of $[-2\,\text{m}, 2\,\text{m}]$ into a square of extension of 1000 pixels. In Fig. 4, the reproduced window is showed. Here, the gaze point is automatically drawn as a red circle. Even the markers have been reported into the same image plane and are represented with a green circle; this way, the window realizes a possible cursor device. Even the small displacement between the sensor and the plane with ground truth data is managed by our system, using a k value of 4 cm so that the total length of the side $\overline{AC}$ of the right-angled triangle will be $t_z + k$.

In both experimental setups, persons can be in the view angle of the sensor, i.e. 43 ° vertical by 57 ° horizontal field of view, and at a distance from the sensor in the range [40 cm, 300 cm] since we are operating in near mode (see Fig. 5 for more details). Head rotations are allowed in the three axes (also in the z-axes, because we are not using the Viola-Jones face detector), as also head translations.

All these working conditions are very suitable for a fully unconstrained system. Furthermore, using only the 3D information coming from the sensor to transform image coordinates into 3D camera coordinates and solving the gaze estimation as a three dimensional geometric problem, the system does not need any calibration phase. Finally, our algorithm has been tested with a frame rate of 30fps even on a common PC, i.e. an Ultrabook Intel i3 CPU @ 1.8 GHz with 4 GB of RAM, and was easily able to work in real-time. The usage of our technique on a common Ultrabook was made in order to facilitate for raw installations like in shelves or in wilder and nontraditional environments.

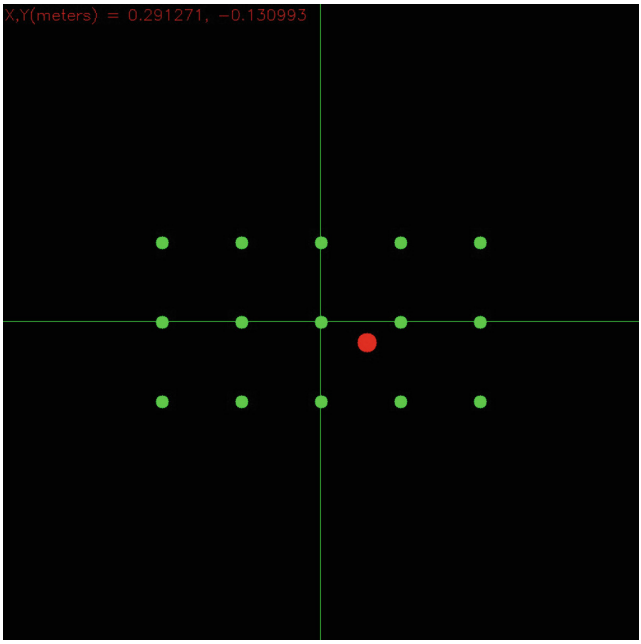

Fig. 4. The reproduced panel that realizes a cursor device.

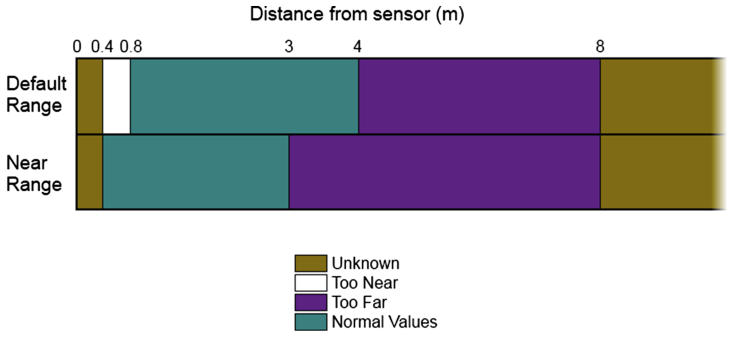

Fig. 5. The depth sensor has two depth ranges: the default range and the near range. The image shows the sensor depth ranges in meters.

5 Experimental Results

The proposed method has been tested with 9 different people. In order to get a comprehensive study, people are divided into three groups, three persons for each group. First group is composed by experienced people, i.e. people that know how the method works and that already tried the system before the test session. Second group is composed by people that try the system for the first time but that are well-informed about how the system works. Therefore, if they want to

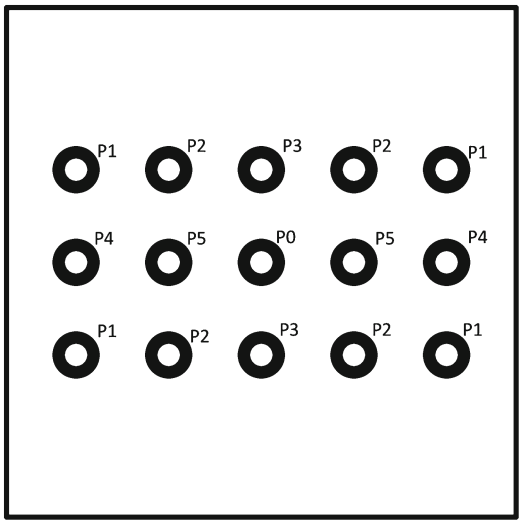

Fig. 6. The used grouping scheme for target points during tests.

point the attention on a given target, they will move the head in that direction without, for example, falling into the temptation to move only the eyes if the new point is very close to the previous one. Finally, in the last group there are unaware people that are just placed in front of the camera and are asked to point the markers. No constraints are given to the participants in terms of eyeglasses, beard or hairstyle and, in order to allow for wild settings, no panel or uniform background color has been put behind the participants.

The experiment is made as follows. People are asked to look at each of the placed markers, in a fixed order, using our real environments instead of a screen. Errors are measured as the angle between the ground truth gaze and the estimated gaze. Ground truth gaze is given by an oral feedback from the person, that stops moving when it is focused on a marker. For a given angle, the difference between the estimated gaze and the ground truth information increases if the distance grows. Differently from most of state of the art methods, in real environments the angle of error is considered also as a function that depends on the distance between the user and the sensor, because the error of the upstream method increases, unless the method is based on tools like head mounted devices. Even our head pose estimation tends to be inaccurate at the growing of the distance.

Results are showed in Tables 1, 2 and 3. Markers are divided into subset like in Fig. 6 in order to group together points that present the same distance from the sensor in terms of x, y or both axes, from P1 to P5, while P0 corresponds to the depth sensor position. For example, P4 are the points with a distance of 1 m from the sensor along the x axes and aligned along y axes, and so on. The second column shows the tested distance, i.e. 70, 150 and 250 cm. Errors are computed

Table 1. Experiments with the first group.

		Errors			
		x (cm)	x (deg)	y (cm)	y (deg)
P0	70 cm	1.50	1.22	2.66	2.18
	150 cm	3.50	1.33	4.83	1.84
	250 cm	6.00	1.37	8.50	1.94
P1	70 cm	n.a.	n.a.	n.a.	n.a.
	150 cm	6.00	1.61	8.00	2.79
	250 cm	2.60	0.52	4.00	0.88
P2	70 cm	8.77	5.03	7.66	4.37
	150 cm	0.16	0.05	11.83	4.15
	250 cm	4.33	0.95	5.83	1.29
P3	70 cm	5.61	4.58	3.50	1.94
	150 cm	6.83	2.60	8.83	3.08
	250 cm	4.66	1.06	1.83	0.40
P4	70 cm	n.a.	n.a.	n.a.	n.a.
	150 cm	0.50	0.13	9.33	3.56
	250 cm	0.66	0.13	15.66	3.47
P5	70 cm	3.66	2.03	3.83	3.13
	150 cm	0.33	0.11	4.16	1.59
	250 cm	4.33	0.95	8.16	1.87
Total averages	70 cm	4.88	3.22	4.41	2.90
	150 cm	2.88	0.97	7.83	2.83
	250 cm	3.77	0.83	7.25	1.64

separately for each group and for yaw and pitch angle, in order to evidence where inaccuracies are located. In order to look at all the markers on the panel from the three positions, head pose of the users during the experiment with the three group can vary in the range $[-56.0\,°, +56.0\,°]$ in terms of yaw and in the range $[-35.5\,°, +35.5\,°]$ concerning pitch. Considering that a small angle from a wide distance corresponds to a bigger displacement, also errors in cm are reported. Note that "n.a." stands for a not available data, in our experiment exclusively due to the excessive rotation of the head to see objects at a far distance compared to the distance from the sensor, such that the Candide-3 model was not able to be overlapped on the face image from the system.

As can be observed, results for the first group are very accurate and, considering that all state of the art constraints are being removed, comparable to the state of the art methods in gaze estimation. Subsequently, the first group is perfectly able to control our device as it was a classic cursor device. The second group shows the same results, with some short outliers depending on the speed of the people to become familiar with the system. Anyway, they perform almost

Table 2. Experiments with the second group.

		Errors			
		x (cm)	x (deg)	y (cm)	y (deg)
P0	70 cm	2.50	2.04	2.33	1.90
	150 cm	6.24	2.38	7.41	2.83
	250 cm	28.5	6.50	13	2.97
P1	70 cm	n.a.	n.a.	n.a.	n.a.
	150 cm	2.00	0.53	21.58	7.70
	250 cm	19.00	3.84	27.00	6.05
P2	70 cm	10.16	5.89	17.16	10.40
	150 cm	5.83	1.89	19.08	6.78
	250 cm	3.00	1.50	0.65	0.33
P3	70 cm	2.83	2.31	8.33	4.77
	150 cm	15.83	6.02	13.33	4.69
	250 cm	18.5	4.23	20.5	4.58
P4	70 cm	n.a.	n.a.	n.a.	n.a.
	150 cm	4.75	1.27	15.08	5.74
	250 cm	33.00	6.79	12.50	2.86
P5	70 cm	11.16	0.83	6.51	0.68
	150 cm	7.50	2.53	3.00	1.14
	250 cm	22.50	5.03	19.00	4.34
Total averages	70 cm	6.66	4.19	7.16	4.44
	150 cm	6.98	2.44	13.25	4.81
	250 cm	20.75	4.51	15.58	3.52

the same result as the first group, because the given information was easy to be assimilated. Even this group is able to use the device. The third group shows that some error can occur, but results are encouraging for applications where the focus of attention is the preponderant measure to be estimated. Finally, results are satisfactory even at a distance of 150 cm.

For informed users, during the experiments also the ability to monitor a possible device is tested in a dual way: first of all, the set of gaze points was registered and evaluated. The usage of the head pose information and the tracking algorithm applied with the facial mask model, has not shown outliers nor flickering effects. Finally, after each experiment, the virtual panel has been shown to the participants, and it was asked them to try to touch with the "gaze cursor" all the drawn fixed points from a distance of about 80–90 cm and to give a feedback. All of the participants felt comfortable and able to use our control device.

After validation, an intelligent shop window has been realized, as can be observed in Fig. 7. It has size of 3.70 m width and 2.80 m height. Between the user and the sensor, there is the window glass. This does not degrade overall

Table 3. Experiments with the third group.

		Errors			
		x (cm)	x (deg)	y (cm)	y (deg)
P0	70 cm	4.00	3.27	0.00	0.00
	150 cm	15.00	5.71	10.00	3.81
	250 cm	71.00	12.00	15.85	2.74
P1	70 cm	n.a.	n.a.	n.a.	n.a.
	150 cm	17.00	4.73	29.00	10.46
	250 cm	62.00	13.15	34.00	7.64
P2	70 cm	24.99	23.13	34.61	23.13
	150 cm	1.78	0.61	10.90	3.82
	250 cm	27.00	6.05	17.00	3.79
P3	70 cm	24.61	19.37	10.10	5.11
	150 cm	33.20	12.48	17.10	6.06
	250 cm	90.23	19.84	12.6	2.80
P4	70 cm	n.a.	n.a.	n.a.	n.a.
	150 cm	3.90	1.01	4.80	1.83
	250 cm	12.54	2.52	24.67	5.63
P5	70 cm	19.80	4.10	12.20	3.35
	150 cm	2.78	0.96	17.38	6.60
	250 cm	4.44	0.98	24.24	5.53
Total averages	70 cm	18.35	12.67	12.20	7.90
	150 cm	12.27	4.25	14.86	5.43
	250 cm	44.61	9.75	20.75	4.69

Fig. 7. The realized shop window.

Fig. 8. The 2D projection of item occupancy in order to reveal the observed items.

performances, since Kinect generates IR light to determine an object's depth (distance) from the sensor, and this stream can pass over pane. Even due to the hardware under investigation, this system cannot work in outdoor environments, since the sensor cannot directly point towards direct sunlight.

To implement this system, the shop window has been reproduced on the screen. In order to define when an item should be considered observed by a user, the following method is used: considering Fig. 3, the gaze ray intersection with the virtual panel (the red circle) still continue to move in both horizontal and vertical direction inside an area that represent the shop window, using again Eq. 7. To each item of interest inside the shop window, a square area onto the screen, that represent a 2D projection of its space occupancy, has been manually defined. An item is considered observed if at least one point of the gaze ray is lying inside its square. Figure 8 shows our modeling. Finally, all results about gaze ray and observed items are stored. This way, data can be further used to realize decision making support system or to simply get useful stats, for example to detect most/least observed objects.

6 Conclusions

With this work, pervasive retail architecture based on a free gaze estimation system that can be used, for example, in a shop window to detect which items are observed from people has been proposed. The system exploits a depth sensor in order to extract head pose information, from which a fast geometric technique then evaluates the focus of attention of the persons in the scene (even more persons at the same time). Preliminary experiments were conducted in our lab in order to quantitative validate the accuracy of the gaze estimation on different benchmarks of persons. Then the qualitative evaluation of the effectiveness of

the proposed architecture was conducted in a real shop window demonstrating the ability to deal with real challenging environmental conditions. Future works will deal with the massive tests of the system in real shopping centers and on the definition of digital signage metrics for the exploitation of the extracted information for the definition of decision making support systems oriented to the creation/modification of the retail strategies.

References

1. Apr 2014. http://msdn.microsoft.com/en-us/library/hh855347.aspx
2. Ahlberg, J.: Candide-3-an updated parameterised face (2001)
3. Ba, S.O., Odobez, J.M.: Recognizing visual focus of attention from head pose in natural meetings. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **39**(1), 16–33 (2009)
4. Benfold, B., Reid, I.: Unsupervised learning of a scene-specific coarse gaze estimator. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2344–2351. IEEE (2011)
5. Besl, P., McKay, N.D.: A method for registration of 3-d shapes. IEEE Trans. Pattern Anal. Mach. Intell. **14**(2), 239–256 (1992)
6. Brey, P.: Freedom and privacy in ambient intelligence. Ethics Inf. Technol. **7**(3), 157–166 (2005)
7. Brolly, X.L., Stratelos, C., Mulligan, J.B.: Model-based head pose estimation for air-traffic controllers. In: Proceedings of the 2003 International Conference on Image Processing, ICIP 2003, vol. 2, pp. II-113. IEEE (2003)
8. Chen, J., Ji, Q.: 3d gaze estimation with a single camera without ir illumination. In: 19th International Conference on Pattern Recognition, ICPR 2008, pp. 1–4. IEEE (2008)
9. van Doorn, M., van Loenen, E., de Vries, A.P.: Deconstructing ambient intelligence into ambient narratives: the intelligent shop window. In: Proceedings of the 1st International Conference on Ambient Media and Systems, p. 8. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2008)
10. Doshi, A., Trivedi, M.M.: Attention estimation by simultaneous observation of viewer and view. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 21–27. IEEE (2010)
11. Funatsu, N., Takahashi, T., Deguchi, D., Ide, I., Murase, H.: A study on gaze estimation using head and body pose information. In: International Workshop on Advanced Image Technology (2013)
12. Funes Mora, K., Odobez, J.M.: Gaze estimation from multimodal kinect data. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 25–30. IEEE (2012)
13. Guestrin, E.D., Eizenman, M.: General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Trans. Biomed. Eng. **53**(6), 1124–1133 (2006)
14. Hansen, D., Ji, Q.: In the eye of the beholder: A survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell. **32**(3), 478–500 (2010)
15. Ince, I.F., Kim, J.W.: A 2d eye gaze estimation system with low-resolution webcam images. EURASIP J. Adv. Signal Process. **2011**(1), 1–11 (2011)

16. Janko, Z., Hajder, L.: Improving human-computer interaction by gaze tracking. In: 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom), pp. 155–160. IEEE (2012)
17. Jellinger, K.: Cognitive processes in eye guidance. Eur. J. Neurol. **13**(9), e9 (2006)
18. Kohlbecher, S., Bardinst, S., Bartl, K., Schneider, E., Poitschke, T., Ablassmeier, M.: Calibration-free eye tracking by reconstruction of the pupil ellipse in 3d space. In: Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, pp. 135–138. ACM (2008)
19. Langheinrich, M.: Privacy by design – principles of privacy-aware ubiquitous systems. In: Abowd, G.D., Brumitt, B., Shafer, S. (eds.) UbiComp 2001. LNCS, vol. 2201, pp. 273–291. Springer, Heidelberg (2001)
20. Langton, S.R., Honeyman, H., Tessler, E.: The influence of head contour and nose angle on the perception of eye-gaze direction. Percept. Psychophysics **66**(5), 752–771 (2004)
21. Li, Z., Sastry, S.S., Murray, R.: A mathematical introduction to robotic manipulation (1994)
22. Mateo, J.C., San Agustin, J., Hansen, J.P.: Gaze beats mouse: hands-free selection by combining gaze and emg. In: CHI'08 Extended Abstracts on Human Factors in Computing Systems, pp. 3039–3044. ACM (2008)
23. Morimoto, C.H., Mimica, M.R.: Eye gaze tracking techniques for interactive applications. Comput. Vis. Image Underst. **98**(1), 4–24 (2005)
24. Nabati, M., Behrad, A.: Robust facial 2d motion model estimation for 3d head pose extraction and automatic camera mouse implementation. In: 2010 5th International Symposium on Telecommunications (IST), pp. 817–824. IEEE (2010)
25. Pieters, R.: A review of eye-tracking research in marketing. Rev. Mark. Res. **4**, 123–147 (2008)
26. Ravnik, R., Solina, F.: Audience measurement of digital signage: Quantitative study in real-world environment using computer vision. Interact. Comput. **25**(3), 218–228 (2013)
27. Robertson, N., Reid, I., Brady, J.: What are you looking at? gaze estimation in medium-scale images. In: Proceedings of the HAREM Workshop (in assoc. with BMVC) (2005)
28. Stiefelhagen, R., Zhu, J.: Head orientation and gaze direction in meetings. In: CHI'02 Extended Abstracts on Human Factors in Computing Systems, pp. 858–859. ACM (2002)
29. Tonkin, C., Ouzts, A.D., Duchowski, A.T.: Eye tracking within the packaging design workflow: interaction with physical and virtual shelves. In: Proceedings of the 1st Conference on Novel Gaze-Controlled Applications, p. 3. ACM (2011)
30. Valenti, R., Sebe, N., Gevers, T.: Combining head pose and eye location information for gaze estimation. IEEE Trans. Image Process. **21**(2), 802–815 (2012)
31. Voit, M., Stiefelhagen, R.: 3d user-perspective, voxel-based estimation of visual focus of attention in dynamic meeting scenarios. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, p. 51. ACM (2010)
32. Waizenegger, W., Atzpadin, N., Schreer, O., Feldmann, I., Eisert, P.: Model based 3d gaze estimation for provision of virtual eye contact. In: 2012 19th IEEE International Conference on Image Processing (ICIP), pp. 1973–1976. IEEE (2012)
33. Wetzel, P.A., Krueger-Anderson, G., Poprik, C., Bascom, P.: An eye tracking system for analysis of pilots' scan paths. In: The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), vol. 1996. NTSA (1996)

34. Xiao, J., Moriyama, T., Kanade, T., Cohn, J.F.: Robust full-motion recovery of head by dynamic templates and re-registration techniques. Int. J. Imaging Syst. Technol. **13**(1), 85–94 (2003)
35. Zhang, W.Z., Wang, Z.C., Xu, J.K., Cong, X.Y.: A method of gaze direction estimation considering head posture. Int. J. Signal Process. Image Process. Pattern Recogn. **6**(2), 103–111 (2013)

http://www.springer.com/978-3-319-12810-8

Video Analytics for Audience Measurement
First International Workshop, VAAM 2014, Stockholm, Sweden, August 24, 2014. Revised Selected Papers
Distante, C.; Battiato, S.; Cavallaro, A. (Eds.)
2014, X, 159 p. 67 illus., Softcover
ISBN: 978-3-319-12810-8