Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A Bayesian hierarchy for robust gaze estimation in human–robot interaction ☆

Pablo Lanillos [a,b], João Filipe Ferreira [b,*], Jorge Dias [b,c]

[a] Institute for Cognitive Systems (ICS), Technische Universität München, Arcisstrasse 21, 80333 München, Germany
[b] AP4ISR team, Institute of Systems and Robotics (ISR), Dept. of Electrical & Computer Eng., University of Coimbra, Pinhal de Marrocos, Pólo II, 3030-290 Coimbra, Portugal
[c] Khalifa University of Science, Technology, and Research, Abu Dhabi 127788, United Arab Emirates

ARTICLE INFO

Article history:
Received 15 May 2016
Received in revised form 3 February 2017
Accepted 18 April 2017
Available online xxxx

Keywords:
Gaze estimation
Head pose estimation
HRI
Robustness to distance
Robustness to missing data
Bayesian estimation

ABSTRACT

In this text, we will present a probabilistic solution for robust gaze estimation in the context of human–robot interaction. Gaze estimation, in the sense of continuously assessing gaze direction of an interlocutor so as to determine his/her focus of visual attention, is important in several important computer vision applications, such as the development of non-intrusive gaze-tracking equipment for psychophysical experiments in neuroscience, specialised telecommunication devices, video surveillance, human–computer interfaces (HCI) and artificial cognitive systems for human–robot interaction (HRI), our application of interest. We have developed a robust solution based on a probabilistic approach that inherently deals with the uncertainty of sensor models, but also and in particular with uncertainty arising from distance, incomplete data and scene dynamics. This solution comprises a hierarchical formulation in the form of a mixture model that loosely follows how geometrical cues provided by facial features are believed to be used by the human perceptual system for gaze estimation. A quantitative analysis of the proposed framework's performance was undertaken through a thorough set of experimental sessions. Results show that the framework performs according to the difficult requirements of HRI applications, namely by exhibiting correctness, robustness and adaptiveness.

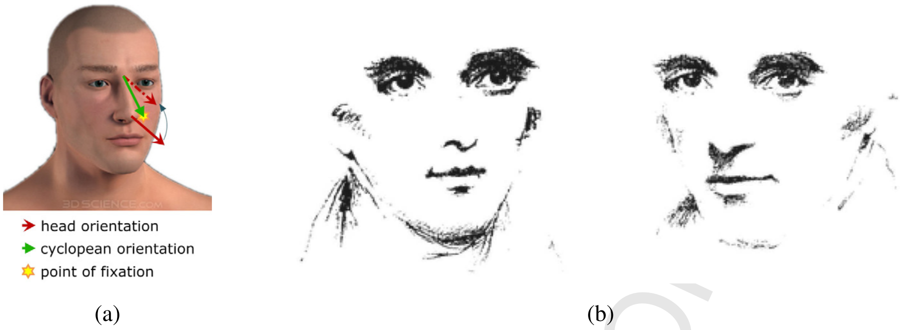
1. Introduction

Head movements are commonly interpreted as a vehicle of interpersonal communication. For example, in daily life, human-beings observe head movements as the expression of agreement or disagreement in a conversation, or even as a sign of confusion. On the other hand, gaze shifts are usually an indication of intent, as they commonly precede action by redirecting the sensorimotor resources to be used. Gaze direction can also be used for directing a person to observe a specific location. As artificial cognitive systems with social capabilities become more and more important due to the recent evolution of robotics towards applications where complex and human-like interactions are needed, basic social behaviours such as *joint attention*, which is the means by which an agent looks at where its interlocutor is looking at by producing an

☆ This paper is part of the Virtual special issue on Special Issue on Unconventional computing for Bayesian inference, Edited by Jorge Lobo and João Filipe Ferreira.

[*] Corresponding author.
E-mail address: jfilipe@isr.uc.pt (J.F. Ferreira).

http://dx.doi.org/10.1016/j.ijar.2017.04.007



(a) (b)

Fig. 1. The gaze estimation problem. (a) Gaze direction is given by the orientation of the line extending from the middle point between the eyes to the point of fixation given by the intersection of both optical axes (i.e. the cyclopean orientation). This direction, although not equal, is closely related to the direction of the head. (b) The Wollaston illusion: although the eyes are the same in both images, gaze direction as perceived by the human brain is dictated by the orientation of the head – facsimile of original drawings from [10].

eye–head movement that attempts to yield the same focus of attention (FOA), have increasingly become important research topics in this field [1]. As a consequence, in the context of human–robot interaction this skill represents an important part of building a social bridge between humans and computers.

In HRI, as in human–human interaction, distance to the interlocutor and the estimation of gaze direction in general and the awareness of gaze lock in particular have been found to be crucial factors in engagement and communication. A great deal of research in this respect has shown that the proximity zones originally suggested by Hall [2] have an important role to play, namely in approach for engagement. For example, Bergstrom et al. [3] classified human behaviour in HRI according to the speed and direction of motion of the interlocutor, but also the distance between the robot and interlocutor. In their solution, interlocutors included within boundaries of what Hall defined as the "social zone" (i.e. within 1.5 to 3 m) are already potentially classified as "interested" in engaging. Other authors have agreed that initial engagement may occur starting from distances substantially over 1.5 m – e.g. [4–7]. Analogously, in the context of video surveillance it may also be important to be able to estimate the gaze of direction of a customer or an intruder from a distance, as to be able to infer intent. Since the performance of gaze direction estimation is directly influenced by distance, dealing with this issue becomes even more important. Another challenge in gaze direction estimation is incomplete data, which is mostly a consequence of occlusion and misdetections. Finally, scene dynamics adds an additional level of perceptual uncertainty which must be dealt with appropriately.

Bayesian theory has recently been applied to many applications, since it inherently deals with perceptual uncertainty, and a variety of *Bayesian inference algorithms* for artificial perception inspired by studies of biological systems have been proposed [8,9]. The specific application of HRI is particularly appropriate for such probabilistic approaches.

In the following text, we will propose a robust solution based on a probabilistic approach capable of adaptively dealing with perceptual uncertainty resulting from distance, incomplete data and dynamics in gaze estimation. Our solution combines multiple Bayesian models feeding from different sources of data in order to always provide a gaze estimate. The expected outcome of this work will be a crucial component of an artificial cognitive system, with the ability of robustly estimating gaze direction of interlocutors within the context of HRI.

2. Related work and contribution

2.1. Gaze estimation in humans

Over the last few years the problem of gaze estimation has been extensively studied in humans. Physiological investigations have brought to light several clues on how the human brain estimates gaze [11,12] — see Fig. 1 (a). By itself, head pose provides an estimate that represents a coarse approximation of gaze direction, that can be used in situations in which the eyes of a person are not visible (e.g when observing a distant person, or in the presence of eye-occluding objects like sunglasses) [13]. When the eyes are visible, head pose becomes an additional cue to accurately predict gaze direction. Although the details of gaze estimation, as it is solved by the human brain, are still an open research issue, it is mostly consensual that it results from the combination of the results of two fundamental estimation subprocesses: *head pose estimation* and *eye gaze estimation* [11,12].

This combination means that partial information can be used to already provide an estimate; however, this happens at the expense of biasing. This bias manifests itself in the form of the *Wollaston illusion*, named after British chemist and natural philosopher William Hyde Wollaston who first reported this phenomenon in the early 19th century [10] (see Fig. 1 (b)). Nevertheless, the error induced by this bias is greatly compensated by the fact that the human brain is able to yield an estimate *even when only presented with partial or incomplete information* – seen as an engineering solution, one can easily argue that it seems that *precision is sacrificed for the sake of robustness*, which in a social interaction context is essential to maintain continuous communication.

Please cite this article in press as: P. Lanillos et al., A Bayesian hierarchy for robust gaze estimation in human–robot interaction, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.04.007

2.2. Gaze estimation in computer vision

In the survey by Murphy-Chutorian and Trivedi [13], solutions for head pose estimation are divided into eight categories: seven represent pure methods, while the remaining are hybrid methods, i.e. combinations of the other methods. As mentioned in this survey, most of the computer vision head pose estimation approaches have diverged greatly from the results of psychophysical experiments as to how the brain tackles this problem. In fact, while the former have focused predominantly on *appearance-based* solutions, the latter consider human perception of head pose to rely on *geometrical cues*, such as the deviation of the nose angle and the deviation of the head from bilateral symmetry. These effects and other factors, such as the location of the face in relation to the contour of the head, strongly influence the human perception of head pose, suggesting that these are extremely salient cues regarding the orientation of the head.

Geometrical approaches attempt to detect head features as accurately as possible in order to compute the pose of the head. An example of a geometrical approach is the method proposed by Kaminski et al. [14] uses a model of the face and the eye deduced from anthropometric features in order to determine head orientation. This method uses only three points (e.g. eye centres and the middle point between the nostrils) to perform the desired task. Another solution for head pose estimation is introduced in [15]. The main idea here is to consider an isosceles triangle, with corners in both eyes and in the centre of the mouth. The direction of the head is computed if we assume that one side of the triangle lies on the image plane, such that applying a trigonometrical function we can estimate the angle between the triangle plane and the image plane [15]. Finally, Canton-Ferrer et al. [16] use multiple cameras with accurate calibration information. Skin colour segmentation is performed on each camera, and then data fusion is performed, resulting in a 3D model of the head. The orientation of the head is estimated based on a particle filter.

As for eye gaze estimation, one can find several intrusive commercial applications to track eyes and compute the gaze direction, which need to be calibrated. For instance Patrão and Menezes [17] devised an eye tracking system to help cognitive impaired people. In terms of non-intrusive solutions, there are geometrical gaze estimation methods that depend on the correct detection of the pupils [18]. There are also some machine learning approaches that use as input labelled data of the segmented region of both eyes. For instance, Sugano et al. [19] propose an appearance-based gaze estimator using the saliency map of the images that the person is looking at and then a Gaussian process to learn the parameters. Conversely, in [20] learning is performed through adaptive linear regression. In another related gaze approach, as shown in [21], first eye contact is detected and subsequently images are labelled in terms of the region containing the eyes, data which is then used by a support vector machine classifier to decide if there is gaze locking (binary classifier).

The majority of state-of-the-art methods tackle the problem of gaze direction estimation using either head pose estimation or eye gaze estimation. However, work such as [18,22–24] represent hybrid approaches, that combine the head pose and the eye gaze estimates to obtain an overall gaze estimate. Valenti et al. [18] presents a solution for eye detection and tracking, combining the results with a cylindrical head model (CHM) approach for head direction, in order to obtain the final gaze direction estimate. The data fusion between the eyes pose (computed in 3D) and the head pose is made by averaging both vectors, when the distance between them is larger than a certain threshold. Conversely, gaze direction estimation in [22] is solved, after a camera calibration process, in two stages: the eye orientation vector, referred to the head coordinate system, is estimated, and subsequently rotated using information about the head orientation (in a world coordinate system), resulting in the final gaze direction estimate. Both articles have limitations in estimating gaze direction when estimates of the eye or the head poses are unreliable. The method proposed by Sung et al. [23] presents a combination between active appearance models (AAM) and CHM to obtain the gaze direction and seems to have good results with off-line data. Another interesting 3D approach can be found in [24], where they use a precomputed 3D model of the head to estimate head pose and subsequently an appearance-based method to estimate eye gaze direction. The segmented eyes region (acquired using a Microsoft Kinect RGB-D sensor and tracked in the 3D model), is used to learn the gaze direction using the adaptive linear regression method [20]. As a follow-up to this work, these authors later proposed a head pose invariant gaze estimation model relying on a semantic segmentation of the eye region using a generative process [25]. Yücel et al. [26], on the other hand, present an image-based method for establishing joint attention between an experimenter and a robot by gaze interpolation and saliency. In a nutshell, the authors estimate gaze direction by using regression-based interpolation of the gaze direction from the head pose of the experimenter, which they claim is easier to track, given that “[t]he precise analysis of the experimenter’s eye region requires stability and high-resolution image acquisition, which is not always available” [26]; they then refine their estimate by using image-based saliency (which in fact assumes that stimulus-driven priority is assigned equally by the human and the robot) to pinpoint the actual fixation target, thereby upgrading head pose estimation to full gaze estimation. They impose a further restriction on their system by assuming that the human interlocutor always starts the experiments by promoting eye contact with the robot (i.e. a “zero” initial condition). Using saliency to refine head pose estimates in order to obtain full gaze estimation was also explored in an earlier work by Hoffman et al. [27]. These authors used a probabilistic method with an ellipsoidal model of a human head is used to estimate pan and tilt angles relative to the camera. To circumvent the problem of the narrow range of detectable head poses offered by typical detectors trained using frontal views (see section 4), their system wait until a frontal view of the face is detected and then tracks the head across different movements using the Meanshift algorithm. More recently, Massé et al. [28] have proposed an on-line Bayesian temporal model for the estimation of the focus of attention as a part of an integrated process of estimating, not only gaze direction and head poses, but also potentially fixated object locations. Finally, Asteriadis et al. [29] propose a technique for estimating visual focus of attention using head rotation, as well as fuzzy fusion of head rotation and eye gaze estimates, in a fully automatic manner, which they claim precludes the need for any special hardware or a priori knowledge regarding the user, the environment or the setup. According to the authors, their solution assumes a close-proximity human–computer interaction scenario, in which the human interlocutor is (frontally) facing a monitor with a web-cam mounted on top.

Most of the methods presented above are unsuitable for open-ended HRI applications. For example, as commented in [21], in HRI applications gaze estimation should be passive and non-intrusive. However, the most challenging problems yet to be adequately addressed, as can be seen in the set of representative examples presented above, are the *distance to the interlocutor* including and beyond the "social zone", *missing data* (due to extreme head poses or distance) and *scene dynamics*. In fact, few of the solutions in the literature explicitly deal with extreme head poses (the *"frontal view assumption"* [13]), while even fewer address dynamics (the *"continuous video assumption"* [13], meaning that gaze direction changes are assumed small or even absent). However, distance to the interlocutor has been the most neglected issue in the literature. For example, all of the solutions presented above have been tested in experimental conditions in which distance to sensor ranges from 50 cm to 1.5 m – the only exception, at first glance, would be [21], since they test their solution at 2 m; however, these authors report using a high resolution camera (5184 × 3456 pixel), the likes of which one would not usually use in an HRI context (in fact, they emulate conditions at 6, 10, 14 and 18 m by down-sampling their original images).

In order to address these issues, we propose a robust solution for gaze detection based on a hierarchical probabilistic approach, **inherently and adaptively dealing with perceptual uncertainty due to distance, incomplete data and scene dynamics**. Our hierarchical framework, which was designed with human–robot interaction in mind, in particular in the context of social interaction, consequently attempts to loosely follow the characteristics of gaze estimation as performed by the human brain.

Our solution is contingent on the following assumptions, in our opinion without compromising its usability in realistic settings:

- **3D data assumption** – it is presumed that the visual sensor used is capable of yielding 3D coordinates of the facial landmarks on the image that comprise the output of the feature detectors. A side-effect of this assumption is that the system is upper-bounded in precision by the worst performing feature detector.
- **Gaze stability assumption** – it is expected that gaze direction is approximately piecewise-constant in time, and hence gaze direction estimated for instant t will probably be the same as gaze direction for instant $t-1$.
- **Antidrift assumption** – as defined in [13], gaze estimation time is presumed to be short, and therefore during the estimation process there will be no significant distortions in visual information (see section 7 for more details on future work on a real-time implementation of our system, that will allow this assumption to hold in realistic conditions).

The modular configuration of our solution, combining the results of several models following geometrical approaches, promotes robust performance, the main focus of our work, by yielding the best possible estimate with *all available data*, even if provided with a minimum amount of features or if at the expense of maximum precision.

3. Hierarchical Bayesian framework for gaze estimation

The problem of gaze estimation consists in establishing $G \equiv \{\mathbf{se}_X, \mathbf{se}_Y, \mathbf{se}_Z, \theta, \phi\}$, a conjunction of five discrete random variables composing the gaze vector, including the 3D coordinates of its initial point, the anatomical landmark named *sellion* (**se**) corresponding to the midpoint between the eyes, and pitch and yaw direction angles, denoted as θ and ϕ respectively. Even though there is no need for a roll angle in the final estimate vector $\hat{G}$, roll angles of the head and each of the eyes influence the geometry underlying each submodel – in fact, mathematically, it would be straightforward to include this influence in the submodels. However, the exponential increase in memory usage and number of normalisation sums due to the resulting size of the joint distributions led us to consider these angles as free variables, whose influence will be diluted in the uncertainty of the model, given that roll angles in practical settings have a negligible effect comparing to all other factors.[1] To solve this problem we propose a hierarchical framework comprised of one top integration level and several submodels, denoted generically as π_i, each of which representing a specific strategy assuming different levels of incompleteness in the data perceived by the sensors.

As seen in the framework overview shown in Fig. 2, at each time instant, the hierarchy receives inputs including a preliminary estimate of the general orientation of the head, a rough estimate of an origin for a head-centred frame of reference, and 3D positions of detected facial features relating to that frame, yielding as a result a posterior distribution and respective *maximum a posteriori* (MAP) estimate $\hat{G}$. As mentioned earlier, it also uses the posterior of the previous time instant as prior knowledge for a prediction model under a piecewise-smooth motion assumption. The fact that the framework uses 3D coordinates of facial features as inputs makes it flexible enough to be used with any combination of feature detectors of choice and image-based 3D sensors (e.g. stereo rigs, RGB-D sensors or 3D cameras).

[1] A discussion on complexity and its relation to details of variable support and discretisation, including the choice for the convex hull dimensions, are given in section 4.3.

Please cite this article in press as: P. Lanillos et al., A Bayesian hierarchy for robust gaze estimation in human–robot interaction, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.04.007

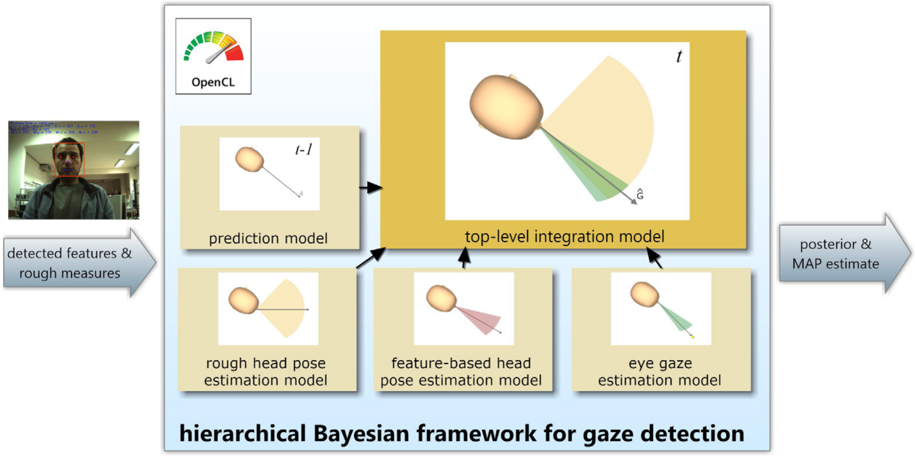


Fig. 2. Overview of hierarchical Bayesian framework.

The final gaze orientation estimate is given by the "consecutive" refinement in the *top-level integration model* (section 3.5) of the output given by the hierarchy's submodels:

- The *rough head pose estimation model* (section 3.1), yielding an approximate measure of head orientation as frontal or left or right profile, which represents the minimum amount of data that can be sensed at a given instant, and the only data expected to be reliable enough at larger distances – the associated MAP estimate is denoted as $\hat{H}_R$.
- The *feature-based head pose estimation model* (section 3.2), that uses the restrictions imposed by whatever facial features are detected at a given instant, which are expected to become increasingly available and reliable at closer distances to the artificial observer – the associated MAP estimate is denoted as $\hat{H}$.
- The *eye gaze estimation model* (section 3.3), that uses the relative positions of the irises within each eye, expected to be available when the interlocutor is in close proximity to the artificial observer, allowing a more fine-grained estimation of a fixation point – the associated MAP estimate is denoted as $\hat{E}$.
- The *prediction model* (section 3.4), that uses information from previous estimation steps, thus imposing the antidrift assumption stated in section, and allows the system to provide an estimate even if the interplay between the likelihoods of all the other models result in multimodal, or in the limit, flat posterior distributions 2.2.

For inference to be tractable, however, it is not wise to use an unconstrained support of both G and each of the random variables denoting the 3D coordinates of each feature, since this would result in a combinatorial explosion which would severely impact memory usage and computational performance. Moreover, the actual facial features and gaze vector are enclosed in a relatively restricted convex hull within the workspace that would correspond to a world reference frame $\{W\}$ or even the artificial observer's egocentric reference frame $\{\mathcal{E}\}$. Therefore, the appropriate course of action is to perform a reference frame shift and define the support space of all random variables as residing within a convex hull capable of withholding facial features within the full scope of facial rotations within that hull, as depicted in Fig. 3. This reference frame shift is accomplished by displacing the egocentric reference frame (Z pointing forward, Y downwards and X to the right, following the right-hand convention) to a rough estimate of the 3D position of the initial point for the gaze vector (given, for example, in a worst case scenario, by the three-dimensional centre-of-mass of the blob of pixels in the image classified as belonging to the head). Then, a convex hull around this point, the origin of the head-centred reference frame, denoted as $\{\eta\}$, is defined, with a size of more than an "average head" diameter.

In section 4.2 we present specific values for all parameters used in our implementation.

3.1. Rough head pose estimation model

This very simple model, formally defined in Fig. 4, takes input from a detector or set of detectors that roughly classify head orientation as either frontal (i.e. facing forwards, towards the general direction of the artificial observer), sideways to the left, or sideways to the right, summarised as the discrete class random variable R_H, yielding a distribution on the gaze vector G resulting from naïve Bayes fusion[2] and by assuming its initial point in all cases as coinciding with the origin of the head-centred reference frame η, within an error margin.

As with all submodels in the following subsections, a free binary variable modelling "false alarm", denoted as D, is used, a technique explained in [9,8]. Besides providing a sensible way of modelling cases in which the detectors used to feed

[2] The hypothesis space under consideration is $G \equiv \{\mathbf{se}_X, \mathbf{se}_Y, \mathbf{se}_Z, \theta, \phi\}$.

Please cite this article in press as: P. Lanillos et al., A Bayesian hierarchy for robust gaze estimation in human–robot interaction, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.04.007

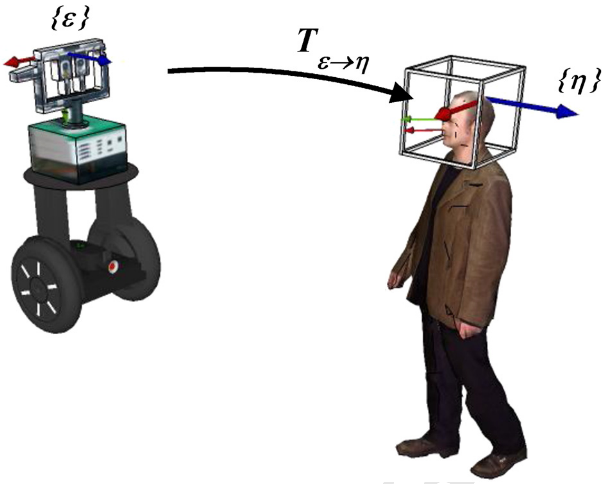


Fig. 3. Shift from egocentric to head-centred reference frame and respective convex hull of support for 3D coordinate random variables. These are used to constrain joint distributions to a manageable size.

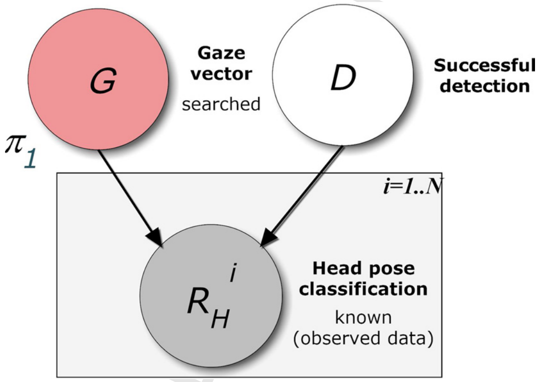


Fig. 4. Bayesian network for rough head pose estimation model (π_1). This model takes as input the outputs of face detectors tuned to each of the possible classifications to be collected as the conjunction represented by $R_H = \wedge_i^N R_H^i$ and yields a posterior on gaze direction that corresponds at its maximum to a rough estimate of head orientation $\hat{H}_R = \operatorname{argmax}_G P(G \mid [R_H = r_H] \wedge \pi_1)$. It includes a free binary variable, denoted as D, which is marginalised out during inference (please refer to main text for explanation). In this sense, $P(R_H^i \mid G \wedge D \wedge \pi_1)$ corresponds to a conditional probability table (CPT): for $[D = 1]$, $P(R_H^i) = f(G)$ for each and every hypothesis for G; for $[D = 0]$, it is a uniform distribution $P(R_H^i) = \mathscr{U}(R_H^i)$. Prior $P(G \mid \pi_1)$ is uniform, while $P(D \mid \pi_1)$ is a conditional probability table (CPT) attributing a probability to “false alarms”.

the framework yield estimates that do not correspond to any valid feature or context, they also prevent deadlocks in the application of the Bayes update rule (i.e. cases when updating the posterior from zero- or one-probability situations, which correspond to absolute certainty, becomes impossible).

3.2. Feature-based head pose estimation model

A great deal of research has been undertaken regarding the anthropometric measures of the human body for the most diverse reasons – medical purposes, anthropological studies, computer graphics, etc. As a byproduct, a reasonable amount of work has been made in surveying anthropometry for populations consisting of groups of dozens of people under study, resulting on average measures for anatomical landmarks of the body, and more specifically of the face. Hence, one may take advantage of these summary statistics to construct the geometry of an “average human face”.[3]

Consider the head pose geometry represented on Fig. 5, assuming a 3D coordinate system, with the origin as the subnasale **sn**, the Z axis orientation given by that same vector, the Y axis as the vertical symmetry line of the trapezium and the X axis orthogonal to both the latter using a right-hand rule. Assume that the hypothesis under consideration is $G \equiv \{\mathbf{se}_X, \mathbf{se}_Y, \mathbf{se}_Z, \theta, \phi\}$.

[3] See section 4.2 for further details.

Please cite this article in press as: P. Lanillos et al., A Bayesian hierarchy for robust gaze estimation in human–robot interaction, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.04.007

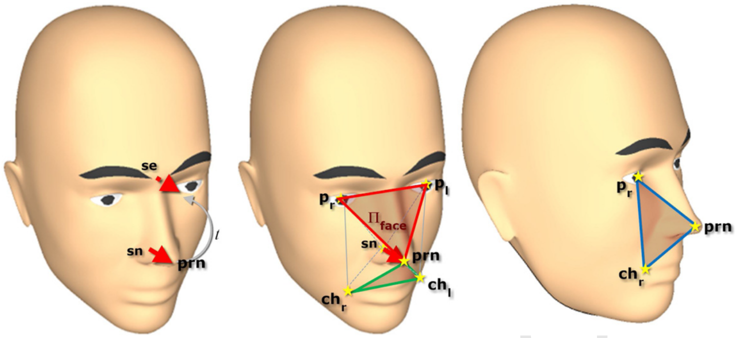


Fig. 5. Facial feature geometry given head pose. Average anthropometric measurements allow relating a specific given head pose to the pyramid with an isosceles (i.e. symmetric) trapezium base, Π_{Face}, formed by the average positions of a subset of the anatomical landmarks defined by Farkas [30], represented centrally on the figure. Anatomical landmarks are denoted as follows: **se**, **prn**, **sn**, $\mathbf{p}_{\{\mathbf{l},\mathbf{r}\}}$ and $\mathbf{ch}_{\{\mathbf{l},\mathbf{r}\}}$ denote the sellion, the pronasale, the subnasale, the left and right pupils and the left and right cheilion, respectively. On the right, a face in profile position is shown to still relay enough information to compute the average positions of a smaller, but sufficient, subset of anatomical landmark positions.

The nose tip feature, corresponding to the pronasale landmark using this geometry, can be computed as having $X = 0$, $Y = -d_{\mathbf{sn}-\mathbf{se}}$ and $Z = d_{\mathbf{sn}-\mathbf{prn}}$, with **sn**−**prn** denoting the anthropometric measure giving the protrusion of the nose and **sn**−**se** denoting the height of the nose.[4] The notation used next will follow the same rationale.

This model is expected to yield an estimate even when the eyes are not clearly detectable; hence, the simplifying assumption is made that only the approximate location of the centre of the eye is available and that the eyes are gazing forward, making the centre of the eye roughly correspond to the pupil location. The left eye centre feature, which corresponds to the left pupil landmark using this geometry, can be computed as having $X = \frac{d_{\mathbf{pl}-\mathbf{pr}}}{2}$, $Y = 0$, where the Y subscript represents a projection in the Y axis, and $Z = 0$. The right eye centre feature is symmetrical in X.

The left mouth corner, which corresponds to the left cheillion in this geometry, can be computed as having $X = \frac{d_{\mathbf{cl}-\mathbf{cr}}}{2}$, $Y = d_{\mathbf{sn}-\mathbf{prn}} + d_{\mathbf{sn}-\mathbf{sto}}$, where the Y subscript represents a projection in the Y axis, and $Z = 0$. The right mouth corner feature is symmetrical in X.

After computing each of these feature positions in the proposed referential, the corresponding coordinates are then shifted so that the initial point becomes the sellion **se**, and finally transformed so that they are referred to in the head-centred coordinate system.[5] This makes it possible to relate the mean of each feature likelihood, assuming normal distributions, which is taken as being equal to the shifted feature positions, to each and every hypothesis for G, within an error margin bounded by the overall sum of:

- all standard deviations of the involved anthropometric measurements;
- estimation error determined during 3D sensor calibration;
- reprojection of error in pixels incurred by the feature detectors to 3D.

This model is formally defined in Fig. 6. In summary, performing estimation for this model corresponds to evaluating the following expression

$$\hat{H} = \mathrm{argmax}_G P(G \mid [F_H^1 = f_H^1] \wedge \ldots \wedge [F_H^N = f_H^N] \wedge \pi_2), \tag{1}$$

with

$$P(G \mid [F_H^1 = f_H^1] \wedge \ldots \wedge [F_H^N = f_H^N] \wedge \pi_2) \propto \prod_i^N P(D_i \mid \pi_2) P(F_H^i \mid G \wedge D_i \wedge \pi_2). \tag{2}$$

3.3. Eye gaze estimation model

Consider now the cyclopean gaze geometry represented on Fig. 7, assuming a 3D coordinate system, with the origin as the initial point of the gaze orientation vector (corresponding to the selion, **se**), and the X axis corresponding to a line that is related to the line connecting the eyes by a rotation of θ_{eyes}, the Y axis corresponding to the vertical orthogonal direction bisecting the line connecting the eyes, and the Z axis as orthogonal to both the latter, using a right-hand rule. Assume that the hypothesis under consideration G is as in the previous subsection. Let us also add a new variable d_{FP} that denotes the

[4] Anthropometric measures are always positive.
[5] Preferably, the sellion should be as close to the origin of this reference frame as possible.

Please cite this article in press as: P. Lanillos et al., A Bayesian hierarchy for robust gaze estimation in human–robot interaction, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.04.007

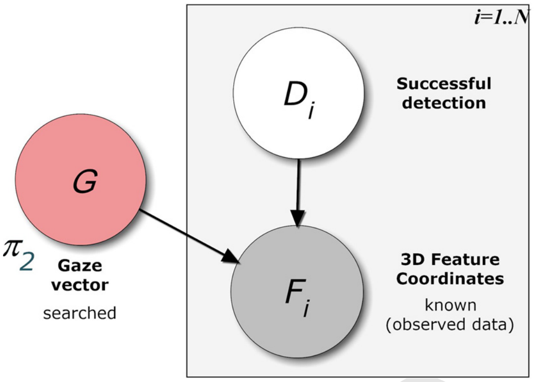


Fig. 6. Bayesian network for feature-based head pose estimation model (π_2). This model takes as input a set of 3D coordinates for facial landmarks F_H^i output by N feature detectors and yields a posterior on gaze direction that corresponds at its maximum to an estimate of head orientation $\hat{H} = \text{argmax}_G P(G \mid [F_H^1 = f_H^1] \wedge \ldots \wedge [F_H^N = f_H^N] \wedge \pi_2)$. Analogously to the model in Fig. 4, it includes a set of free binary variables, denoted as D_i, which are marginalised out during inference. Prior $P(G \mid \pi_2)$ is uniform, while $P(D_i \mid \pi_2)$ are conditional probability tables (CPTs) attributing a probability to "false alarms". The likelihood distributions, $P(F_H^i \mid G \wedge D_i \wedge \pi_2)$, are defined as $P(F_H^i) = \mathcal{N}(f(G), \sigma)$ (normalised to sum to 1) for each and every hypothesis for G for $[D_i = 1]$, and a uniform distribution $P(F_H^i) = \mathcal{U}(F_H^i)$ for $[D_i = 0]$.

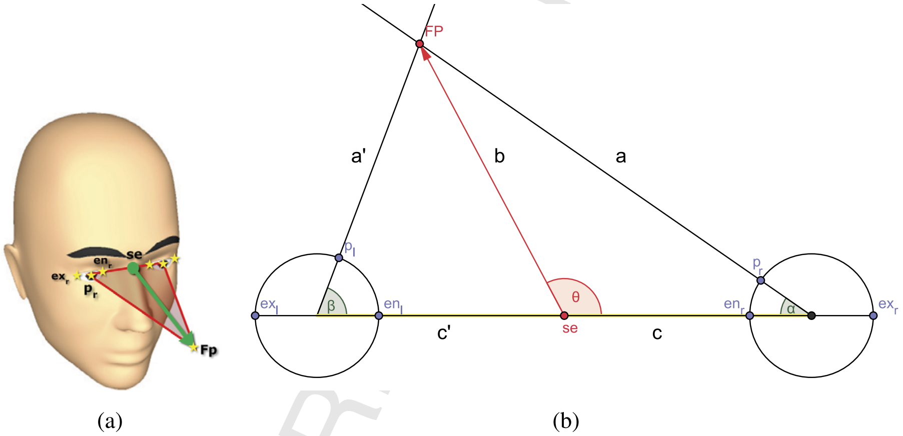


Fig. 7. Eye feature geometry given gaze direction and fixation point. Detected anatomical landmarks allow relating a specific given gaze direction and fixation point to the triangle formed by the optical axes – see (a). Anatomical landmarks are denoted as follows: **se**, $\mathbf{p}_{\{\mathbf{l},\mathbf{r}\}}$, $\mathbf{en}_{\{\mathbf{l},\mathbf{r}\}}$ and $\mathbf{ex}_{\{\mathbf{l},\mathbf{r}\}}$ denote the selion, the left and right pupils, the left and right ectocanthus and the left and right endocanthion (i.e. the corners of the eyes), respectively (only the right landmarks are labelled, for clarity). Note that the two corners of each eye give the diameter of the respective ocular globes (assuming they are perfectly spherical, which is an acceptable approximation). In (b), the corresponding gaze triangulation geometry is shown after compensating for the yaw angle and partial pitch angle that align the "fixation triangle" of (a) with the XZ-plane and the line connecting all the corners of the eyes with the Y-axis, respectively, assuming **se** as the origin.

distance to the fixation point. This variable is irrelevant for the final estimate for G, and as such is to be removed through marginalisation as a free variable [8,9].

The feature detectors feed the model with the 3D positions of the anatomical landmarks corresponding to the eyes, more specifically the left and right pupils, $\mathbf{p}_{\{\mathbf{l},\mathbf{r}\}}$, and the internal and external corners of each eye, $\mathbf{en}_{\{\mathbf{l},\mathbf{r}\}}$ and $\mathbf{ex}_{\{\mathbf{l},\mathbf{r}\}}$. However, not all of this information is needed to establish the fixation triangulation geometry – in fact, as can be seen in Fig. 7 (b), the 3D pupil positions and the horizontal ($\theta_{\text{eyes}} + \alpha$ for the right eye, $\theta_{\text{eyes}} + \beta$ for the left eye) and vertical ($\gamma_\text{l} \approx \gamma_\text{r} \equiv \hat{\phi}$) angles the pupils make with the line joining the corners of the eyes are sufficient.

Bearing this in mind, and in order to simplify the model and also to make it tractable, the data that are actually used from sensor readings are $\mathbf{p}_{\{\mathbf{l},\mathbf{r}\}}$, θ_{eyes}, α, β, γ_l, and γ_r, in addition to a binary random variable Inf that indicates if the fixation point is traceable to a position at a reasonable distance from the artificial observer; if not, the model assumes that α or/and β are direct estimates of θ, therefore modelling the fixation of a point at infinite distance.

In the geometry presented in Fig. 7 (b), the ocular globe radius is given by $r_{\text{eye}} = \frac{d_{\mathbf{ex_r}-\mathbf{ex_l}} - d_{\mathbf{en_r}-\mathbf{en_l}}}{4}$, $c = c' = \frac{d_{\mathbf{ex_r}-\mathbf{ex_l}} + d_{\mathbf{en_r}-\mathbf{en_l}}}{4}$ and $b = d_{FP}$, which allows the gaze triangulation geometry to be used to determine pupil positions and the angular relations between detected eye features for the model likelihood distributions.

In the case of $Inf = 0$ (i.e. feature data for both eyes is available, and the fixation point is within detectable range), this is done as follows:

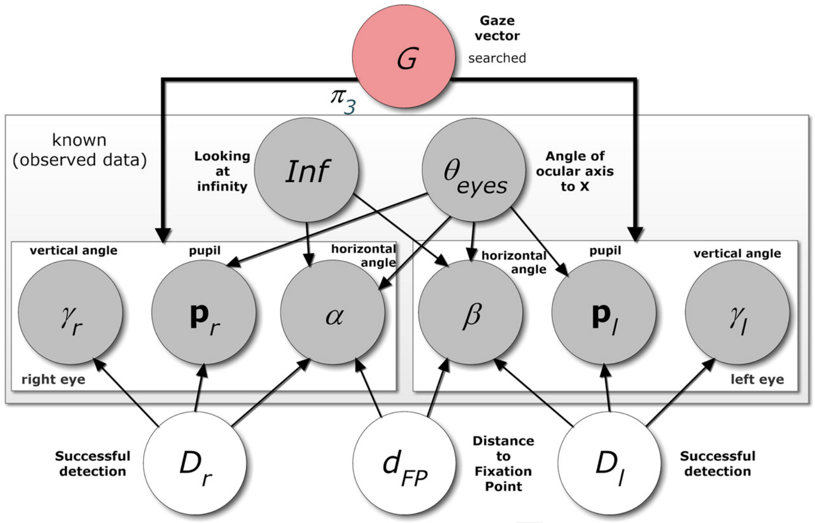


Fig. 8. Bayesian network for eye gaze estimation model (π_3). This model takes as input a set composed of eye parameters resulting from feature detectors, $[\mathbf{p}_{\{l,r\}}, \theta_{\text{eyes}}, \alpha, \beta, \gamma_r, \gamma_l]$, and a decision variable Inf describing the interlocutor as "looking towards infinity" or not, yielding a posterior on gaze direction that corresponds at its maximum to an estimate of eye gaze direction $\hat{E} = \text{argmax}_G P(G \mid \mathbf{p_l} \wedge \mathbf{p_r} \wedge \theta_{\text{eyes}} \wedge \alpha \wedge \beta \wedge \gamma_l \wedge \gamma_r \wedge \pi_3)$. All priors, i.e. $P(G \mid \pi_3)$, $P(d_{FP} \mid \pi_3)P(\theta_{\text{eyes}} \mid \pi_3)$, are uniform. $P(D_l \mid \pi_3)$, $P(D_r \mid \pi_3)$ are conditional probability tables (CPTs) attributing a probability to "false alarms". The remaining distributions are all CPTs defined as follows: for $[D_{\{l,r\}} = 1]$, likelihoods computed as described in main text; for $[D_{\{\,l,r\}} = 0]$, uniform distributions.

1. Since a direct estimate for ϕ, given by $\gamma_{\{l,r\}}$, is available, the first likelihood to be constructed is $P(\gamma_{\{l,r\}} \mid \phi \wedge [Inf = 0] \wedge \pi_3) = P(\gamma_{\{l,r\}} \mid \phi \wedge \pi_3) \equiv \mathcal{N}(\mu = \phi, \sigma_\phi)$.
2. Assuming θ and ϕ are given as a hypotheses and that θ_{eyes} is given by direct computation from sensor readings, the likelihoods $P(\alpha \mid \theta \wedge \gamma_r \wedge \theta_{\text{eyes}} \wedge d_{FP} \wedge [Inf = 0] \wedge \pi_3)$ for the right eye and $P(\beta \mid \theta \wedge \gamma_l \wedge \theta_{\text{eyes}} \wedge d_{\text{FP}} \wedge [Inf = 0] \wedge \pi_3)$ for the left eye can be computed. This is achieved in three steps:
   (a) First, the effects of θ (hypothesis) and $\gamma_{\{\,l,r\}}$ (known) are removed, followed by the effects of θ_{eyes} (known). This makes the "fixation triangle" reside on the XZ-plane and the line connecting the corners of both eyes become coincident with the Y-axis, therefore providing the means to go from the situation depicted in Fig. 7 (a) to the simpler geometry of Fig. 7 (b).
   (b) Next, the Law of Cosines is applied to obtain the unknown sides of the triangles Fig. 7 (b), a and a'. In the case of the right eye (the left eye is obtained in analogous fashion), this results in $a = \sqrt{b + c - 2bc \times \cos\theta}$.
   (c) Finally, the Law of Cosines is applied to obtain the desired angles, α for the right eye and β for the left eye. In the case the former (the process for the latter is analogous), this yields $\alpha = \arccos\left(\frac{c^2 + a^2 - d_{FP}{}^2}{2ca}\right)$.
3. Finally, the likelihoods for each of the pupils can be computed as $P(\mathbf{p_r} \mid \mathbf{se} \wedge \theta_{\text{eyes}} \wedge \alpha \wedge \gamma_r \wedge [Inf = 0] \wedge \pi_3) = P(\mathbf{p_r} \mid \mathbf{se} \wedge \theta_{\text{eyes}} \wedge \alpha \wedge \gamma_r \wedge \pi_3)$ and $P(\mathbf{p_l} \mid \mathbf{se} \wedge \theta_{\text{eyes}} \wedge \beta \wedge \gamma_l \wedge [Inf = 0] \wedge \pi_3) = P(\mathbf{p_l} \mid \mathbf{se} \wedge \theta_{\text{eyes}} \wedge \beta \wedge \gamma_l \wedge \pi_3)$. Again, we will show how the former is computed, since the process to obtain the latter is analogous:
   (a) First, the components of $\mathbf{p}'_\mathbf{r}$ in the two-dimensional referential of Fig. 7 (b) are computed as given by $\mathbf{p}'_{\mathbf{rx}} = c - r_{\text{eye}} \cos\alpha$ and $\mathbf{p}'_{\mathbf{rz}} = r_{\text{eye}} \sin\alpha$.
   (b) Then, the displacement enforced by $\mathbf{se}$ (hypothesis) and the rotations θ_{eyes} and γ_r (all known) are applied to get $\mathbf{p_r}$ related to the desired reference frame η.

In the case of $[Inf = 1]$ these likelihoods simplify as follows:

1. As before, since a direct estimate for ϕ, given by $\gamma_{\{l,r\}}$, is available, the first likelihood to be constructed is $P(\gamma_{\{l,r\}} \mid \phi \wedge [Inf = 1] \wedge \pi_3) = P(\gamma_{\{l,r\}} \mid \phi \wedge \pi_3) \equiv \mathcal{N}(\mu = \gamma_{\{l,r\}}, \sigma_\phi)$.
2. In this case, both α and β become direct estimates for θ, and therefore $P(\alpha \mid \theta \wedge [Inf = 1] \wedge \pi_3) \equiv \mathcal{N}(\mu = \theta, \sigma_\theta)$ and $P(\beta \mid \theta \wedge [Inf = 1] \wedge \pi_3) \equiv \mathcal{N}(\mu = \theta, \sigma_\theta)$.
3. Finally, the likelihoods for each of the pupils, $P(\mathbf{p_r} \mid \mathbf{se} \wedge \theta_{\text{eyes}} \wedge \alpha \wedge \gamma_r \wedge [Inf = 1] \wedge \pi_3) = P(\mathbf{p_r} \mid \mathbf{se} \wedge \theta_{\text{eyes}} \wedge \alpha \wedge \gamma_r \wedge \pi_3)$ and $P(\mathbf{p_l} \mid \mathbf{se} \wedge \theta_{\text{eyes}} \wedge \beta \wedge \gamma_l \wedge [Inf = 1] \wedge \pi_3) = P(\mathbf{p_l} \mid \mathbf{se} \wedge \theta_{\text{eyes}} \wedge \beta \wedge \gamma_l \wedge \pi_3)$, are computed in the same fashion as with $[Inf = 0]$.

This model is formally defined in Fig. 8. In summary, performing estimation for this model corresponds to evaluating the following expression

$$\hat{E} = \text{argmax}_G P(G \mid \mathbf{p_l} \wedge \mathbf{p_r} \wedge \theta_{\text{eyes}} \wedge \alpha \wedge \beta \wedge \gamma_l \wedge \gamma_r \wedge \pi_3), \tag{3}$$

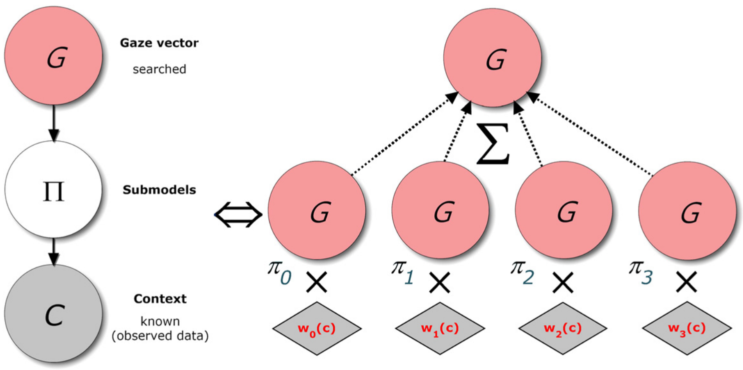


Fig. 9. Bayesian network of top-level mixture model (on the left) that integrates all the results of all submodels for gaze estimation through conditional weighting depending on gaze detection context c, yielding the final estimate $\hat{G} = \text{argmax}_G P(G \mid [C = c]) = \sum_{i=0}^{3} w_i(c) \times P(G \mid [\Pi = \pi_i])$ for instant t (equivalence shown on the right). $[\Pi = \pi_1]$ corresponds to the rough head pose estimation model (section 3.1), $[\Pi = \pi_2]$ – feature-based head pose estimation model (section 3.2), $[\Pi = \pi_3]$ corresponds to the eye gaze estimation model (section 3.3), and $[\Pi = \pi_0]$ corresponds to the prediction (prior) model (section 3.4).

with

$$\begin{aligned}
&P(G \mid \mathbf{p_l} \wedge \mathbf{p_r} \wedge \theta_{\text{eyes}} \wedge \alpha \wedge \beta \wedge \gamma_l \wedge \gamma_r \wedge \pi_3) \propto \\
&\quad P(D_r \mid \pi_3) P(D_l \mid \pi_3) \times \\
&\quad P(\gamma_r \mid D_r \wedge \phi \wedge \pi_3) P(\gamma_l \mid D_l \wedge \phi \wedge \pi_3) \times \\
&\quad P(\alpha \mid D_r \wedge \theta \wedge \gamma_r \wedge \theta_{\text{eyes}} \wedge d_{FP} \wedge Inf \wedge \pi_3) \times \\
&\quad P(\beta \mid D_l \wedge \theta \wedge \gamma_l \wedge \theta_{\text{eyes}} \wedge d_{FP} \wedge Inf \wedge \pi_3) \times \\
&\quad P(\mathbf{p_r} \mid D_r \wedge \mathbf{se} \wedge \theta_{\text{eyes}} \wedge \alpha \wedge \gamma_r \wedge \pi_3) P(\mathbf{p_l} \mid D_l \wedge \mathbf{se} \wedge \theta_{\text{eyes}} \wedge \beta \wedge \gamma_l \wedge \pi_3).
\end{aligned} \tag{4}$$

3.4. Prediction model using prior knowledge

This model uses the gaze estimate computed in the previous time instant $t - 1$ to estimate the gaze at time t, in order to build the prediction prior as $P(G \mid \pi_0) = \mathcal{N}(\mu = G^{t-1}, \sigma)$. If the initial point of the gaze vector at time $t - 1$ is not traceable to inside the operative convex hull at time t, the distribution is made uniform with respect to the initial point, while only the angles estimated at $t - 1$ are used to establish the prior.

3.5. Top-level mixture model for integration

Probabilistic conditional weighting is an important construct for designing hierarchical Bayesian frameworks, often used to model bioinspired perceptual integration processes, in which each submodel π_i is taken as an individual expert whose contribution is weighted against all others depending on a contextual condition [8]. The concrete formulation of the top-level tier of the Bayesian hierarchical model we propose is given in Fig. 9. The top-level mixture model fuses the posterior distributions of the different submodels to provide a more robust gaze estimate.

Random variable C, the value of which is always established at runtime, represents the current gaze detection context, and embodies the core of our solution's adaptivity. It is in fact a concatenation of two discrete variables defining this context, $C = Dist \wedge FeatSuff$, where $Dist \in \{\text{near, average, far}\}$ qualifies the context relating to distance from the observer, and $FeatSuff \in \{\text{very incomplete data, many detected features}\}$ qualifies the sufficiency of the data to construct an "informed guess" through inference by all likelihood-based models. The rationale behind this is as follows:

- Distance from the observer, in conjunction with camera resolution, directly influences the reliability of the data coming from feature detectors. This means that for $Dist = \text{far}$, the rough estimate yielded for head orientation will be more reliable than all others, for $Dist = \text{near}$ features such as the irises and corners of the eyes will become more reliable, and for $Dist = \text{average}$ most probably all models will be equally reliable, and as such their intrinsic uncertainty should be weighted more evenly. Therefore, the set of weights for the submodels in each context should reflect these considerations.
- The availability of a sufficient number of features should influence the weight of the prediction model as opposed to all others (i.e. prediction should take on a more prominent role if there is a lack of sufficient features, for example due to occlusions or extreme facial poses). Again, the set of weights for the submodels should reflect the current context in this perspective.

Please cite this article in press as: P. Lanillos et al., A Bayesian hierarchy for robust gaze estimation in human–robot interaction, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.04.007

Table 1
Mixture model weights per condition for top-level mixture model. (π_0 – prediction; π_1 – rough head pose; π_2 – feature-based head pose; π_3 – eye gaze).

Condition	Models			
	π_0	π_1	π_2	π_3
near ∧ many detected features	0.375	0.125	0.167	0.333
average ∧ many detected features	0.360	0.160	0.160	0.320
far ∧ many detected features	0.297	0.297	0.147	0.259
near ∧ very incomplete data	0.391	0.130	0.130	0.349
average ∧ very incomplete data	0.369	0.180	0.123	0.328
far ∧ very incomplete data	0.327	0.327	0.082	0.264

Given that our solution relies on 3D points alone, this precludes the model directly accounting the effects of conditions such as illumination variations or skin colour. However, the probabilistic nature of the model treats these conditions as latent variables [8], therefore incorporating their effects as part of the uncertainty embedded in the spread of its distributions.

The actual values to define distance ranges and contextual weights used with our specific setup are presented in section 4.2.

4. Implementation details

4.1. Feature detection

In this work, we devised a fairly standard off-the-shelf feature detector framework to serve as a realistic test-bed for the gaze estimation hierarchy.

The rough estimate for head orientation is obtained using a set of Haar classifiers for face detection trained for each of the possible classifications[6] to be collected in the resumed form of R_H. The origin (i.e. central point) of the convex hull of Fig. 3 is taken as positioned at the 3D coordinates corresponding to the centre of the region-of-interest (ROI) resulting from this detection stage. Facial features, comprising nose tip, eye positions are also detected using Haar classifiers. The eye ROIs obtained after applying each of the eye Haar classifiers are then used as inputs for iris detection. Mouth extremities are extracted using the method proposed in [31]. Nose tip detection is based on the same algorithm but completed with additional stages, as described in [32]. Eye extremities and iris detection are performed using Haar classifiers and the method presented in [32]. Inside a reasonable ROI, chosen according to the rough head pose detection result, a Haar classifier is applied to establish an overall ROI for the eye. The iris and the extremities of the eye are then segmented within that region. These extremities are approximately situated on the vertical lines described by the Haar detector, at half distance between left top and right bottom of eye ROI. In our current implementation, Inf is set to 0 if features for both eyes are detected, and in all other cases it is set to 1. Haar-wavelet-based detectors have been reported to find face regions correctly in about 90% of the frames when the images are 30% of their original size [26].

4.2. Building model likelihoods

Table 1 shows the sets of weights corresponding to each condition for the top-level mixture model of section 3.5, which were chosen empirically, by trial-and-error during preliminary testing (see section 5). For example, the eye gaze estimation model (π_3) has less weight in general to avoid overriding the contribution of other models, since the eye model posterior globally presents lower variance. In the case of far and very incomplete data this weight has to be reduced even more as the expected error in the eye feature detection is higher. Conversely, the rough head pose estimation model (π_1) has greater weight, since its variance is larger and it helps to constrain the hypothesis space (e.g. profile detection to the left means that the eyes are unlikely to look to the right at 90 degrees). The conditions "near", "average" and "far" correspond to a distance to the interlocutor of under 1.5 m, between 1.5 m and 2 m, and over 2 m, respectively. This distance is measured using the estimate for the centre of the convex hull as reference. The conditions "many detected features" and "very incomplete data" correspond ≥ 4 and ≤ 3 facial features detected (not including eye extremities), respectively.

For the rough head pose estimation model (section 3.1), the functions $P(R^i_H) = f(G)$ are built as follows:

- $P([R^i_H = \text{left}]) = \mathcal{N}(\mu = 90, \sigma = 90)$;
- $P([R^i_H = \text{right}]) = \mathcal{N}(\mu = -90, \sigma = 90)$;
- $P([R^i_H = \text{frontal}]) = \mathcal{N}(\mu = 0, \sigma = 50)$;
- $P([R^i_H = \text{left-frontal}]) = P([R^i_H = \text{left}]) \times P([R^i_H = \text{front}])$;
- $P([R^i_H = \text{right-frontal}]) = P([R^i_H = \text{right}]) \times P([R^i_H = \text{front}])$.

[6] Frontal view detection has been reported by Hoffman et al. [27] to be successful only for a narrow range of detectable head poses (plus or minus approximately 5°–7° in pan and tilt), which makes having a set of diversely trained classifiers running simultaneously all the more important.

Table 2
Statistics for anthropometric distances used as constants in the computation of feature-based likelihoods.

Symbol	Measurement	References	Mean (mm)	Standard deviation (mm)
$d_{\mathbf{sn-prn}}$	Nose Protrusion	U.S. Army [34]	18.8	2.5
$d_{\mathbf{pr-pl}}$	Interpupillary Breadth	U.S. Army [34]	64.7	3.7
$d_{\mathbf{n-sn}}$	Height of Nose	Farkas et al. [35,36]	53.4	3.5
$d_{\mathbf{sn-sto}}$	Height of Upper Lip	U.S. Army [34]	15.5	2.8
$d_{\mathbf{chl-chr}}$	Mouth Breadth	U.S. Army [34]	56.1	4.3
$d_{\mathbf{ex-en}}$	Eye Fissure Length	Farkas et al. [35,36]	31.2	1.3

Table 3
Feature-based likelihood standard deviations (π_2 – feature-based head pose; π_3 – eye gaze).

variable ↦ model	$[F_H^i \equiv \mathbf{p}_{\mathbf{r},1}] \mapsto \pi_2$	$[F_H^i \equiv \mathbf{ch}_{\mathbf{r},1}] \mapsto \pi_2$	$[F_H^i \equiv \mathbf{prn}] \mapsto \pi_2$	$\mathbf{p}_{\mathbf{r},1} \mapsto \pi_3$
standard deviation (mm)	33.7	35.0	32.5	22.5

Table 2 shows the statistics for the anthropometric measures used as constants in the computation of 3D facial feature-based likelihood parameters (i.e. means and standard deviations) in both the head pose estimation model of section 3.2 and the eye gaze estimation model of section 3.3. Besides these values, the 3D precision specifications of the Kinect sensor at the average operating distance and also a simulated value for the 3D projection of the mean error specifications of each of the feature detectors were added to the overall standard deviations of the likelihoods. The resulting standard deviations for the feature-based likelihoods are summarised in Table 3.

For the eye gaze estimation model, the standard deviations of the likelihoods corresponding to random variables relating to angles were chosen empirically, by trial-and-error experimentation, resulting in $\sigma_\theta = 20.0°$ and $\sigma_\phi = 50.0°$. However, in future work we plan to use Bayesian learning for these distributions. For the prediction model all parameters were chosen empirically, resulting in $\sigma_{\mathbf{se}} = 40$ mm and $\sigma_\theta = \sigma_\phi = 20°$.

4.3. Probabilistic inference implementation and random variable support design

The Bayesian models were implemented for the experiments described in this text using the ProBT® v2.3.0 by ProBAYES, a generic inference programming tool that facilitates the creation of Bayesian models and their reusability using the Bayesian programming formalism [9]. This library is a powerful programming tool for low-to-medium complexity models, and also allows for easy design, prototyping and testing of high complexity models, such as the framework presented in this paper. Currently, these models are running in real-time using a GPU-based massively parallel solution described in [33].

The complexity of a model can be gauged by examining the cardinality of the conjunction of variables included in the joint distribution, manifesting itself in practice in increased memory usage and processing time. As explained in [8], the conditional independence principle of Bayesian approaches allows, to a certain degree, to keep complexity under control, while the recursive Bayesian updating principle allows to subdivide larger problems into sequential applications of inference computations of the form

$$P(S^i \mid O^i) = \frac{P(S^i)P(O^i \mid S^i)}{\sum_S P(S^i)P(O^i \mid S^i)}, \tag{5}$$

where i denotes the inference computation index, S represents a generic unknown state variable, O represents a generic known observation variable, and $P(S^i) \equiv P(S^{i-1} \mid O^{i-1})$.

These principles were applied in the implementation of the proposed solution to break down computations into tractable sequences, in addition to assuming each 3D feature coordinate as independent, and also to pre-compiling all the likelihoods, since they do not change at runtime. However, this kind of "tractability by design" effort can only go so far. Moreover, while processing time can be greatly improved by using massively parallel computing (see section 7), memory usage is much a more difficult problem to overcome. As a consequence, random variable support design, which reflects on the ranges and resolutions available to the models is of utmost importance and will be described next.

All the random variables used in the models are discrete. Variables specifying the 3D coordinates of points in space (i.e. facial features and the initial point of the gaze vector to estimate) have a precision of 3 mm and ranges (also in mm) of $[-70, 70]$, $[-20, 90]$ and $[-80, 30]$ for X, Y, and Z coordinates, respectively, thereby defining the convex hull of Fig. 3. The angles of G are implemented as follows (related angles α, β and $\gamma_{\mathrm{l,r}}$ use the same support): θ has a precision of 2° and its range is $(-90°, 90°)$, and ϕ has a precision of 4° and its range is $(-45°, 45°)$. This results in an estimated maximum memory usage during an inference step, occurring when the state variable S is G and the observation variable O is a 3D feature X-coordinate, given by $\lfloor 141/3 \rfloor \times \lfloor 141/3 \rfloor \times \lfloor 111/3 \rfloor \times \lfloor 111/3 \rfloor \times \lfloor 181/2 \rfloor \times \lfloor 91/4 \rfloor \times 8 = 48{,}990{,}760{,}200 \approx 49$ GB, if using 64-bit floating point numbers.

Please cite this article in press as: P. Lanillos et al., A Bayesian hierarchy for robust gaze estimation in human–robot interaction, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.04.007

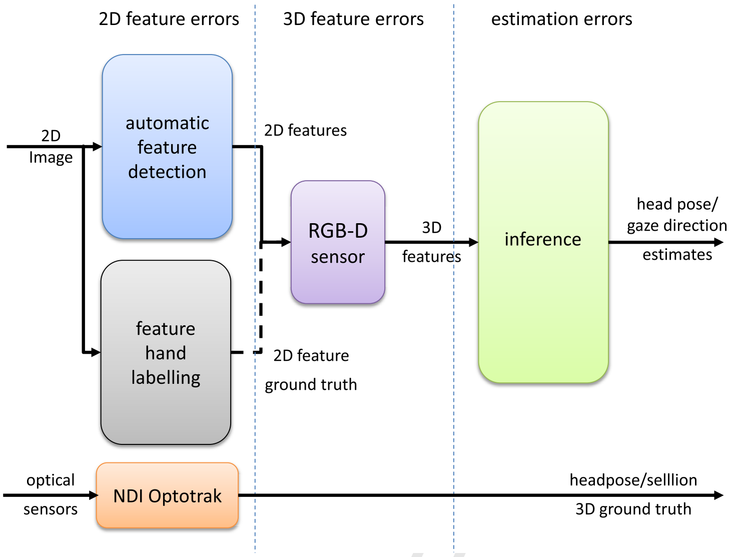


Fig. 10. Processing dataflow. Both 2D and 3D ground truth is obtained, respectively using hand-labelling of 2D features and the optical tracker, to isolate estimation errors from feature detection and RGB-D sensor errors.

5. Experimental results

In the experiments described in the text that follows, a Microsoft Kinect RGB-D sensor was used (640 × 480 resolution); however, any generic 3D camera setting can be used to provide the framework with inputs (e.g. stereo rigs, camera arrays, etc.). Note that the proposed solution is image-independent, since it uses 3D points, and therefore the effect of image resolution or other imaging conditions (illumination variations, etc.) will be diluted in the 2D-to-3D process.[7] Moreover, in real-life applications, feature detectors such as those described [32] or any other off-the-shelf solution may be used together with the 3D camera rig of choice.

The following features are extracted by the detectors described in section 4.1, and are used as inputs for each model in section 3 (Fig. 2):

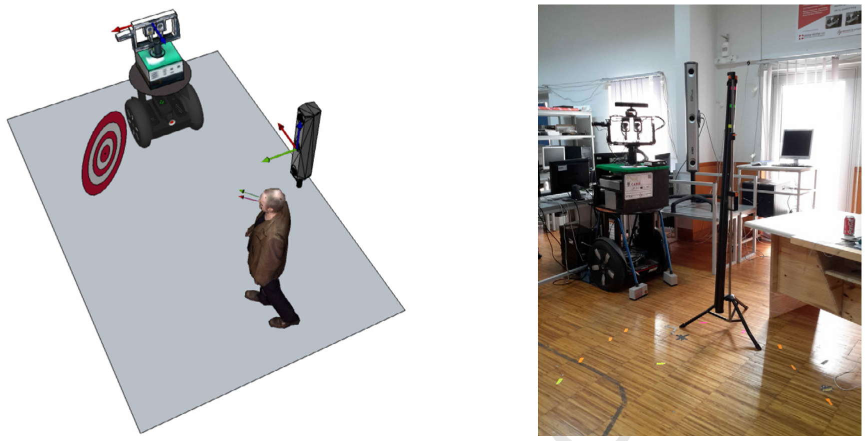
- a rough estimate for head orientation (R_H);
- the 5 features relating to the vector anatomical landmarks $[\mathbf{prn}, \mathbf{p}_{\{\mathbf{l},\mathbf{r}\}}, \mathbf{ch}_{\{\mathbf{l},\mathbf{r}\}}]$ of the geometry presented in Fig. 5, with the exception of the sellion;
- the 2 pairs of eye extremities comprising the vector $[\mathbf{en}_{\{\mathbf{l},\mathbf{r}\}}, \mathbf{ex}_{\{\mathbf{l},\mathbf{r}\}}]$ used to complete the geometry presented in Fig. 7.

Nevertheless, since feature detector development is not the focus of our work, features were also earmarked by hand on the image, so as to allow us to test how system performance in a simulated context where we would have a best-case scenario, with the best possible automatic feature estimation, and also to estimate the error resulting from the automatic eye detection framework, by using the annotated points as 2D ground truth. A Northern Digital Inc. (NDI) Optotrak Certus motion capture system[8] was used to obtain 3D ground truth pertaining the position of the subjects and their head pose throughout the experiment, which was calibrated with the Kinect for registration. The data processing flow is shown in Fig. 10, and can be seen in action in real-time in the attached multimedia file, with illustrative examples of detection errors and missing data.

Details concerning the set-up and the protocol used for these experiments can be seen in Fig. 11. The Optotrak's accuracy and precision specifications are much greater than Kinect, therefore avoiding the errors introduced by the latter for processing if it was to be used to measure ground truth, as in [24,25]. Note, however, that the Optotrak's "cone of visibility" and distances to the interlocutor condition the positioning of the targets in such a way that testing in ϕ is limited – see the description of Fig. 11. Nonetheless, in realistic HRI conditions, most of the gaze angles are sufficiently represented – it is unusual for interlocutors to look at directions beyond ±45° in elevation, and in those extreme cases it is not crucial to have an accurate estimate of gaze direction.

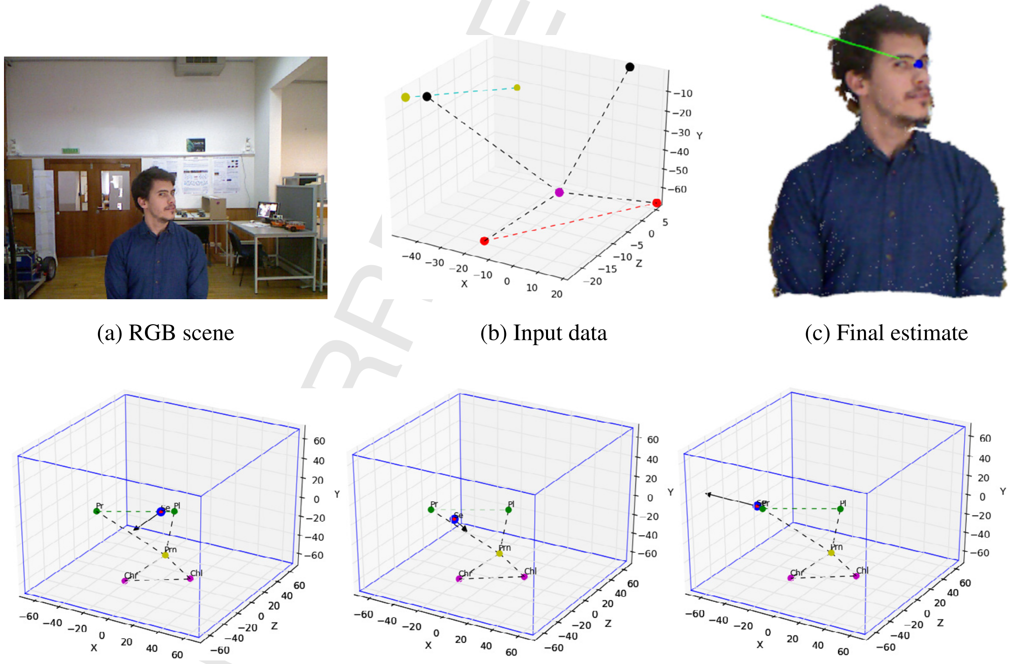
A preliminary set of experiments was conducted first so as to qualitatively evaluate the correctness of the proposed solution under varying conditions, and also to tune the parameters of the hierarchy. More specifically, the distribution of each model was investigated during this preliminary round to understand its importance in the mixture model, and mixture

[7] The impact of decreasing image resolution, in particular, will be analogous to increasing distance-to-robot.
[8] http://www.ndigital.com/msci/products/optotrak-certus/.



(a) (b)

Fig. 11. Experimental setup. In (a), the experimental plan and protocol are presented, showing the sensor rig mounted on a robotic head, the Optotrak Certus used to obtain 3D ground truth and the gaze target. In (b), a photo of the actual experimental setup and the Kinect 3D sensor is shown – note that, as opposed to the schematic in (a), the targets (three coloured stripes, respectively at 1.43, 1.57 and 1.68 m from the ground) are on the same side as the Optotrak. This is because the Optotrak can only track its optical devices within a "cone of visibility" (markings on the floor show the cone's limits), therefore constraining the usable space available for the experiments to take place. The subjects were asked to travel from a predefined distance at about 3 m towards about 1.5 m from the Kinect sensor, following a relatively free path within that usable area always keeping their gaze fixed on one of the targets throughout each trial. They were also instructed to stop in three different midway positions along the path, and at those positions slowly turn their head up, down, left and right, without losing their gaze focus on the designated target.



(a) RGB scene (b) Input data (c) Final estimate

(d) Rough head model (e) Head pose model (f) Eyes gaze model

Fig. 12. Inference process. Facial features are used by the submodels to infer the best feasible gaze estimation and then the mixture model combines them to provide the final estimate. Red dots are the extremities of the lips, purple dot is the nose, black dots are the eyes and cyan dots are the extremities of the eyes. The overall estimate given by the top-level model is shown in (c) as a green vector with a blue initial point superimposed on a 3D reconstruction of the Kinect point cloud. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

model weights were fine-tuned accordingly. An example of the result of the inference process in a given instant is depicted in Fig. 12. The final overall estimate yielded by the top-level model is shown in Fig. 12c, and reflects how the framework outputs a solution that is consistent with the geometry of the features given as input. An additional sequence of results that attest to the correctness of the model is shown in Fig. 13, demonstrating the consistency of the model for a set of very diverse situations, ranging through many different combinations of head poses and eye gaze directions. Even when,

Please cite this article in press as: P. Lanillos et al., A Bayesian hierarchy for robust gaze estimation in human–robot interaction, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.04.007

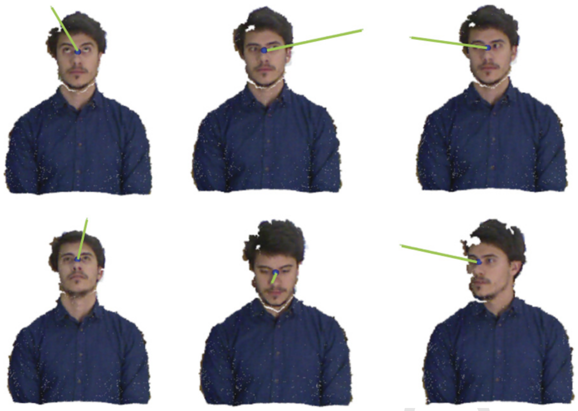


Fig. 13. Results for different combinations of eye gaze and head pose directions. In each case, a 3D reconstruction of the subject and the gaze estimate are shown, using the same convention as in Fig. 12c.

for a specific situation, head pose and gaze directions are in blatant contradiction, the framework still yields a perfectly acceptable result. A slight bias introduced by head pose, which in most cases is more robust in the presence of errors, was found in many situations, hinting towards an effect similar to the Wollaston illusion of Fig. 1 (b).

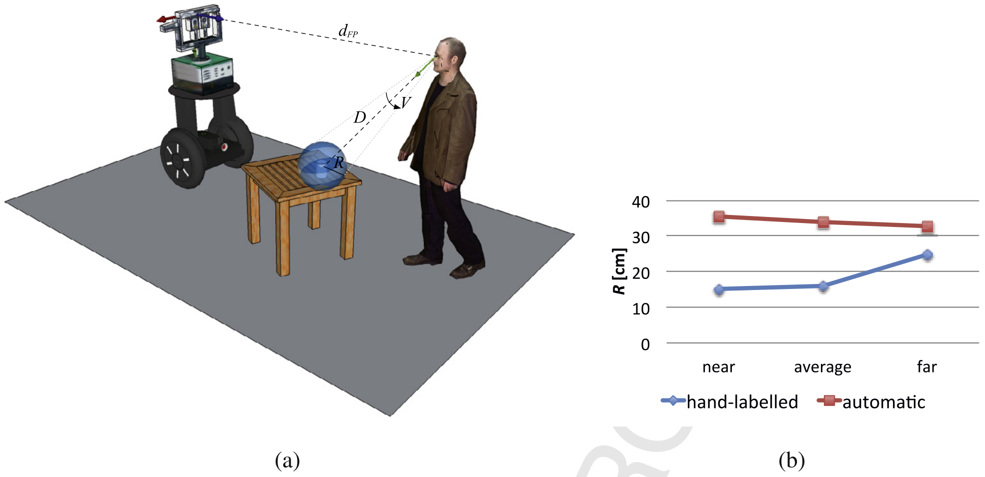
Next, the main set of experiments was conducted so as to thoroughly and quantitatively evaluate the performance of the hierarchy. During the main set of experiments, *all model parameters as determined during the preliminary round were fixed*. The experiments were devised to again test system performance for many different combinations of head poses and eye gaze directions, but also for different distances in a dynamic setting context. Eleven subjects (2 female and 9 male, 8 with white complexion and 3 with dark skin colour), with ages ranging from 8 to 43 and heights ranging from 1.3 m to 1.89 m, were instructed to fixate a target placed in a position with known 3D coordinates throughout each trial (see Fig. 11). Since publicly available datasets for gaze estimation are currently limited to distances from the observer under 2 m, as described in section 2.2, which would make it impossible to test our solution under the conditions it is supposed to be tested, we finally decided in publishing the data we collected during the main set of experiments as a public dataset open to the research community, hosted at http://mrl.isr.uc.pt/experimentaldata/public/gaze-casir.

Ground truth data obtained with the Optotrak were then compared with the estimates for G^t yielded by the system for each time instant t (closest timestamps are used for temporal matching), and position and angular errors computed for performance analysis. Two types of estimate errors were measured:

- gaze orientation error: $E_G^t = |\mathbf{G}^t - \hat{\mathbf{G}}^t|$ (for top-level model);
- head orientation error: $E_H^t = |\mathbf{H}^t - \hat{\mathbf{H}}^t|$ (for head pose estimation models, with particular emphasis on the feature-based model).

Unfortunately, as described in section 2.2, it is infeasible to make a fair comparison between the accuracy of our solution and what is reported in related work, given the difference between experimental conditions in terms of distance and/or resolution, even in the case of the work that is most related to ours due to the use of the Kinect sensor [24,25]. It is also impracticable to compare to the accuracy of gaze estimation in humans, since quantitative studies in this context have also relied on close proximity (typically 50 to 100 cm) – e.g. [39,40]. Therefore, the only way to evaluate the usefulness of our solution is to define the requirements of our HRI target application, and roughly relate accuracy performance to the ability of pinpointing the interlocutor's fixation to a circle, with radius function of distance from the interlocutor. This is done by knowing that $R = D\tan\left(\frac{V}{2}\right)$, where R represents the gaze radius, D represents distance to fixation, and V represents the visual angle delimited by the maximum average error in $\hat{\theta}$ and $\hat{\phi}$ – see Fig. 14. The "cone of uncertainty" delimited by the estimated visual angle V defined by average error ratings, refined and constrained by use of saliency (similarly to what is done in works such as [26]) can be seen in use in HRI applications in [37,38].

Overall results summarising data for all subjects are presented on Tables 4 to 7. Results for hand-labelled features show that the hierarchy's performance is quite satisfactory in ideal conditions of feature detection. *More specifically, for an error of* 21.05° *(absolute average across all subjects), the system is able to pinpoint interlocutor gaze to within a radius of approximately* 20 cm *for an interlocutor fixation target at* 1 m. This result is all the more impressive considering a consumer grade 3D sensor such as the Kinect, with random errors and depth resolution errors escalating exponentially that combine to yield overall errors that can sum up to several centimetres at operating distances of over a couple of meters [41]. Moreover, for hand-labelled features, the effect of having larger discrepancies between head pose and expected gaze direction is generally noticeable, causing increasing gaze estimation errors. On the other hand, using the off-the-shelf automatic feature detection system, performance degrades significantly due to the introduction 2D detection errors (recall Fig. 10), but still allows the system to produce sufficiently accurate estimates to be used in an HRI setting. *More specifically, for an error of* 36.04° *(absolute average*



(a) (b)

Fig. 14. Robot's perception of interlocutor's fixation point at $D = 1$ m within a cone delimited by visual angle V under uncertainty bounds defined by average gaze estimation errors as lying within a sphere of radius R, when the interlocutor is at $d_{FP} > 1.5$ m meters from the robot – illustrative diagram shown in (a). This demonstrates the capability of the robot of pinpointing the FOA of a human in a typical HRI scenario (see also [37,38]). It is expected that R increases (i.e. pinpointing capability precision decreases) in proportion to gaze estimation errors, namely with increasing distance d_{FP} and missing or noisy data. In (b), R is plotted for several distances to the robot d_{FP}. While the proportionality between these two values is clear when using hand-labelled features, it seems to be lost when using off-the-shelf detectors. This is due to the bad performance of these detectors (see discussion in main text), which is compensated by the more conservative estimates of the system for farther distances.

Table 4
Overall performance for gaze estimation model using hand-labelled features.

Experimental condition	Average errors			Maximum errors		
	se (m)	θ (°)	ϕ (°)	**se** (m)	θ (°)	ϕ (°)
near $\wedge$ $\lvert\mathbf{G}(\theta,\phi) - \mathbf{H}(\theta,\phi)\rvert < 30°$	0.069	11.41	9.44	0.31	20.78	16.09
near $\wedge$ $\lvert\mathbf{G}(\theta,\phi) - \mathbf{H}(\theta,\phi)\rvert \geq 30°$	0.041	17.22	10.14	0.05	33.55	20.30
average $\wedge$ $\lvert\mathbf{G}(\theta,\phi) - \mathbf{H}(\theta,\phi)\rvert < 30°$	0.066	11.53	8.02	0.43	25.73	25.74
average $\wedge$ $\lvert\mathbf{G}(\theta,\phi) - \mathbf{H}(\theta,\phi)\rvert \geq 30°$	0.070	18.27	8.77	0.32	34.95	24.80
far $\wedge$ $\lvert\mathbf{G}(\theta,\phi) - \mathbf{H}(\theta,\phi)\rvert < 30°$	0.033	26.31	10.97	0.06	57.24	31.37
far $\wedge$ $\lvert\mathbf{G}(\theta,\phi) - \mathbf{H}(\theta,\phi)\rvert \geq 30°$	0.059	27.90	10.97	0.41	51.45	33.71
overall	**0.051**	**21.05**	**10.03**	**0.43**	**57.24**	**33.71**

Table 5
Overall performance for gaze estimation model using features detected automatically.

Experimental condition	Average errors			Maximum errors		
	se (m)	θ (°)	ϕ (°)	**se** (m)	θ (°)	ϕ (°)
near $\wedge$ $\lvert\mathbf{G}(\theta,\phi) - \mathbf{H}(\theta,\phi)\rvert < 30°$	0.155	37.16	12.37	0.74	56.65	14.72
near $\wedge$ $\lvert\mathbf{G}(\theta,\phi) - \mathbf{H}(\theta,\phi)\rvert \geq 30°$	0.033	39.01	11.11	0.05	61.60	22.86
average $\wedge$ $\lvert\mathbf{G}(\theta,\phi) - \mathbf{H}(\theta,\phi)\rvert < 30°$	0.066	34.95	10.90	0.44	59.73	25.47
average $\wedge$ $\lvert\mathbf{G}(\theta,\phi) - \mathbf{H}(\theta,\phi)\rvert \geq 30°$	0.027	37.50	10.60	0.05	66.00	19.86
far $\wedge$ $\lvert\mathbf{G}(\theta,\phi) - \mathbf{H}(\theta,\phi)\rvert < 30°$	0.054	34.48	8.38	0.85	55.52	24.09
far $\wedge$ $\lvert\mathbf{G}(\theta,\phi) - \mathbf{H}(\theta,\phi)\rvert \geq 30°$	0.027	36.27	13.70	0.09	54.69	31.97
overall	**0.053**	**36.04**	**10.69**	**0.85**	**66.00**	**31.97**

across all subjects), the system is able to pinpoint interlocutor gaze to within a radius of approximately 33 *cm for an interlocutor fixation target at* 1 *m.* Also noticeable in both cases is the effect of angle eccentricity, whereby ground truth gaze direction θ angles, being globally larger than ϕ angles in absolute value, results in greater estimation errors.

Results for a representative subject are presented from Figs. 15 to 18, showing the typical evolution of system performance through time (and distance) – see also attached multimedia file. The effect of distance and missing data can be readily seen in these results, with estimates becoming increasingly more accurate as the subject approaches the sensor, and also with extreme head turns resulting in the occlusion of features in several instants (Fig. 15). Also noticeable is the adaptivity and robustness of the integration model, as it combines the submodels differently to yield the best result possible as the estimation context evolves (from far to close, and between many features available and severely missing data).

Table 6
Overall performance for feature-based head pose estimation model using hand-labelled features.

Experimental condition	Average errors		Maximum errors	
	θ (°)	ϕ (°)	θ (°)	ϕ (°)
near	8.71	9.12	27.69	18.74
average	10.35	8.67	28.29	17.75
far	23.12	12.25	61.46	37.91
overall	**16.39**	**10.56**	**61.46**	**37.91**

Table 7
Overall performance for feature-based head pose estimation model using features detected automatically.

Experimental condition	Average errors		Maximum errors	
	θ (°)	ϕ (°)	θ (°)	ϕ (°)
near	24.06	14.89	85.08	36.94
average	30.06	14.44	99.10	38.80
far	34.47	17.04	109.14	54.73
overall	**30.95**	**15.84**	**109.14**	**54.73**

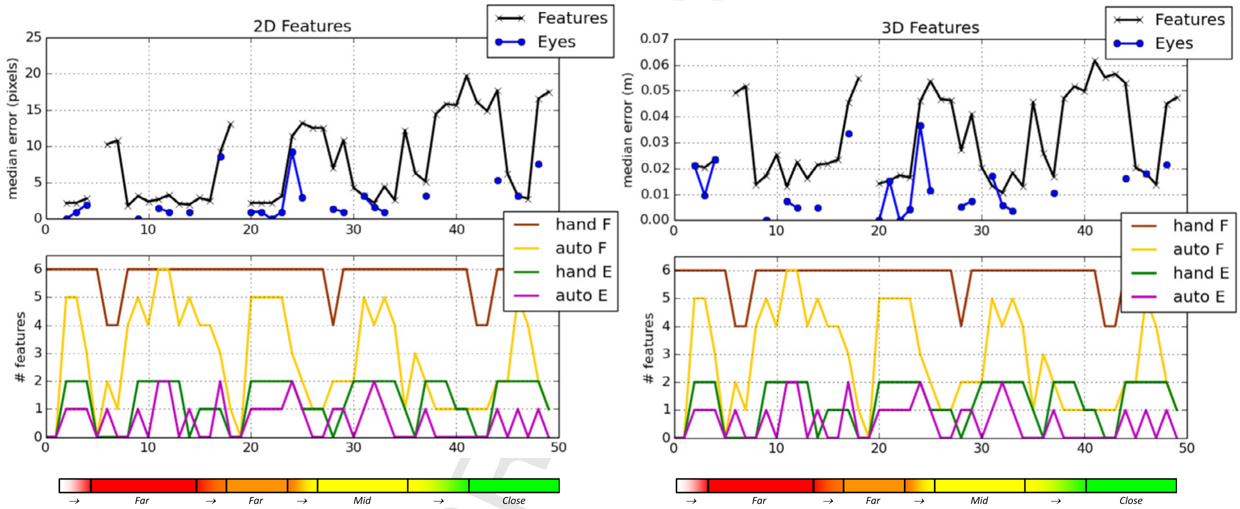


Fig. 15. Subject #3 – evolution of errors (top row) relating to the features detected automatically for the feature-base head pose model and the eye gaze estimation model, if considering the features labelled by hand on the image as ground truth (i.e. perfect detection), and evolution of usable features (bottom row) for both automatic detection and hand-labelled cases. On the left, median of absolute errors in pixels, on the right median of absolute Euclidean errors in 3D (meters). The coloured bar underneath the graphs shows the relative positions of the subject path waypoints throughout the trial, with the arrows indicating periods of time when the subject is travelling between waypoints. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

Finally, since, as mentioned in section 4.2, a few of the parameters used were empirically chosen (namely, and most importantly, the top-level mixture model weights) and others were drawn from measurements taken in anthropometric studies, we performed an additional experimental analysis to test the response of our models against changes in these parameters.

To test the sensitivity of the feature-based head pose estimation model to changes of the anthropometric measurement-based parameters, we varied each parameter while keeping all others fixed and tested model performance in terms of mean angular estimation errors. In general, we confirmed that the anthropometric measurement-based parameter set listed in section 4.2 produces robust results overall, and corresponds to comparably low absolute error values when comparing to alternative parameter sets resulting of varying each parameter. We also found that the feature-based head model is also generally robust to changes of the mean parameters (see Table 2) as long as a sufficient amount of features is available – the most relevant specific findings and respective discussion are given in Figs. 19, 20 and 21. As for standard deviation parameter values (Table 3), we found they are only critical for model performance in conjunction with the top-level model weights – as long as relative proportions between parameters are maintained, the resulting estimates are mostly consistent throughout. In any case, small standard deviation values (high certainty), while in some cases increasing precision, do this at the expense of robustness in the presence of outliers; on the other hand, large standard deviation values (low certainty), render the model too weak to be useful.

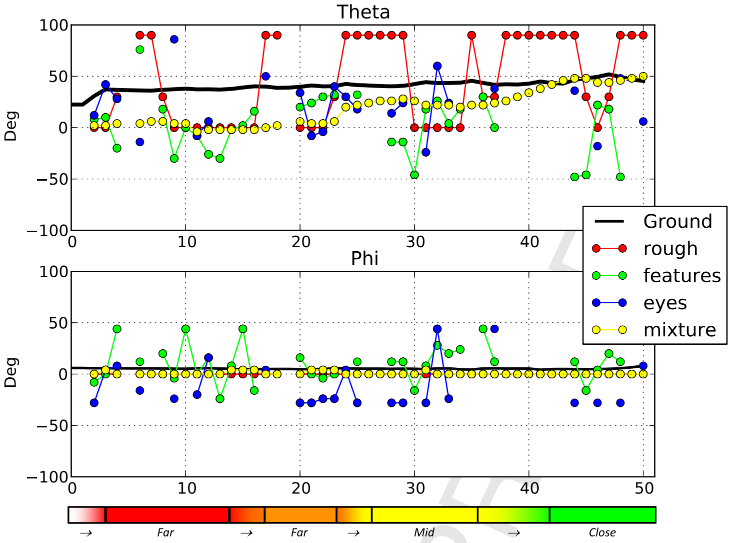


Fig. 16. Subject #3, automatic feature detection – evolution of orientation angles estimates, $\hat{\mathbf{G}}(\theta)$ and $\hat{\mathbf{G}}(\phi)$, against gaze direction ground truth, $\mathbf{G}(\theta)$ and $\mathbf{G}(\phi)$, considering the gaze target and head position. The coloured bar underneath the graph follows the same convention as Fig. 15. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

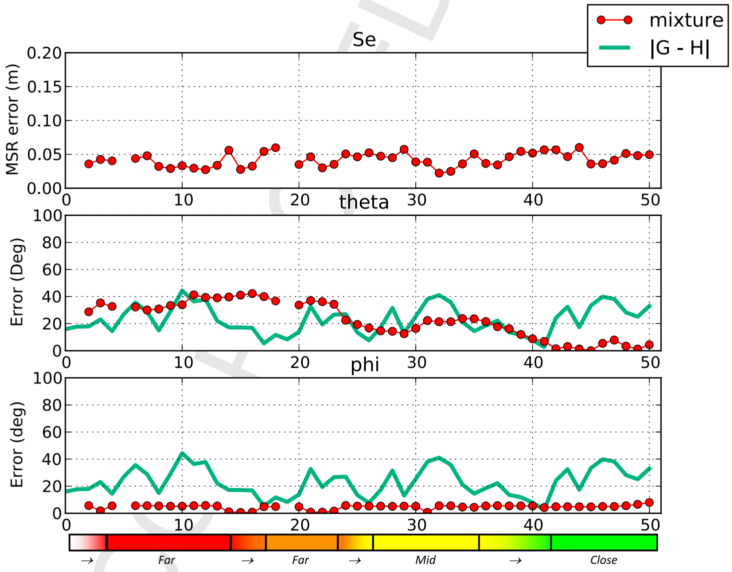


Fig. 17. Subject #3, automatic feature detection – evolution of absolute estimate errors of the top-level model for the sellion, $|\mathbf{G}(\mathbf{se}) - \hat{\mathbf{G}}(\mathbf{se})|$, and for orientation angles, $|\mathbf{G}(\theta, \phi) - \hat{\mathbf{G}}(\theta, \phi)|$, against the discrepancy between the expected gaze direction and the actual head orientation at each frame $|\mathbf{G}(\theta, \phi) - \mathbf{H}(\theta, \phi)|$, demonstrating how this displacement might be influencing the final estimate outputs. The coloured bar underneath the graph follows the same convention as Fig. 15. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

As for the eye gaze estimation model, the role of anthropometric measurement-based mean parameters is insignificant comparing to the influence of α, β and γ. On the other hand, standard deviation parameter values play a similar role as in the feature-based head pose estimation model, a situation that is also true for the rough head pose estimation and prediction models.

Lastly, sensitivity tests were conducted to test the response of the top-level model to changes in the weight parameters – representative results of these tests are presented in Fig. 22. We found, in agreement with our expectations concerning the usefulness of taking estimation conditions into consideration, that weighting the sub-models according to these conditions decreases overall errors when compared to using the sub-models individually[9] or to weighting them equally **under all**

[9] Even though they might be better than the overall model under specific conditions.

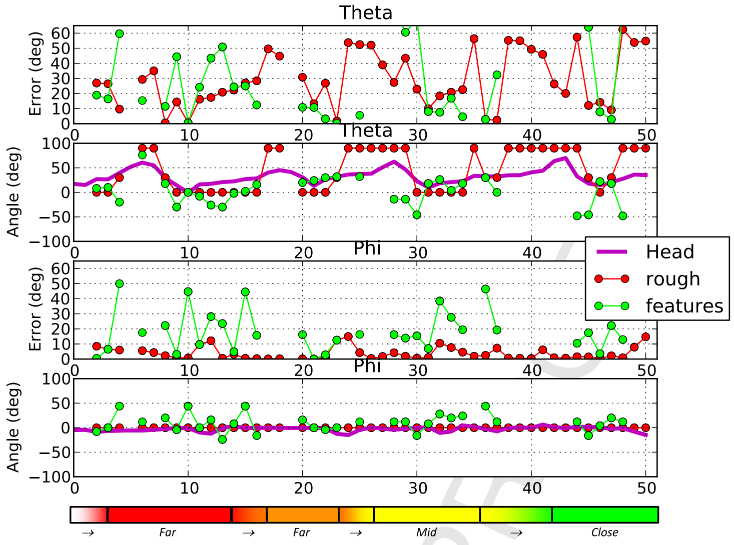
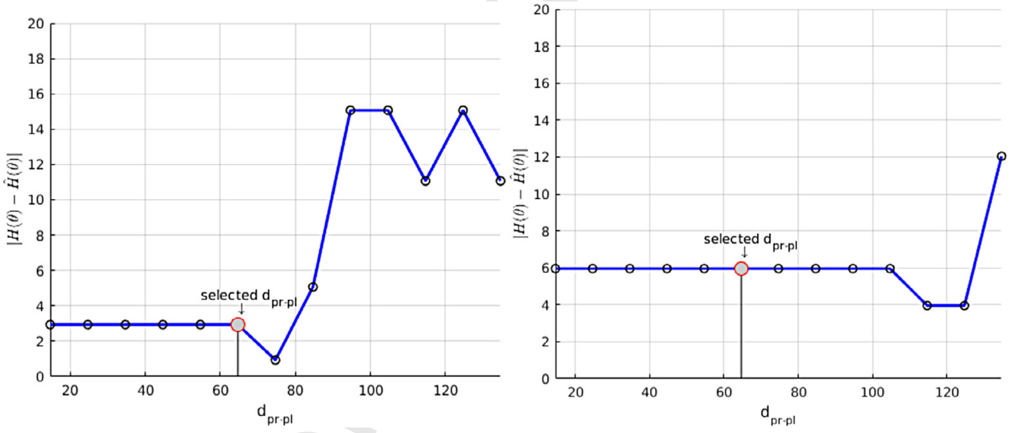


Fig. 18. Subject #3, automatic feature detection – evolution of orientation angles estimates for the rough and feature-based head pose estimation models against head direction ground truth, together with respective evolution of absolute estimate errors $|\mathbf{H}-\hat{\mathbf{H}}|$. The coloured bar underneath the graph follows the same convention as Fig. 15. Note the head turns at waypoints, showing how the subject followed the instructions from the experimental script throughout the trial. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)



(a)

(b)

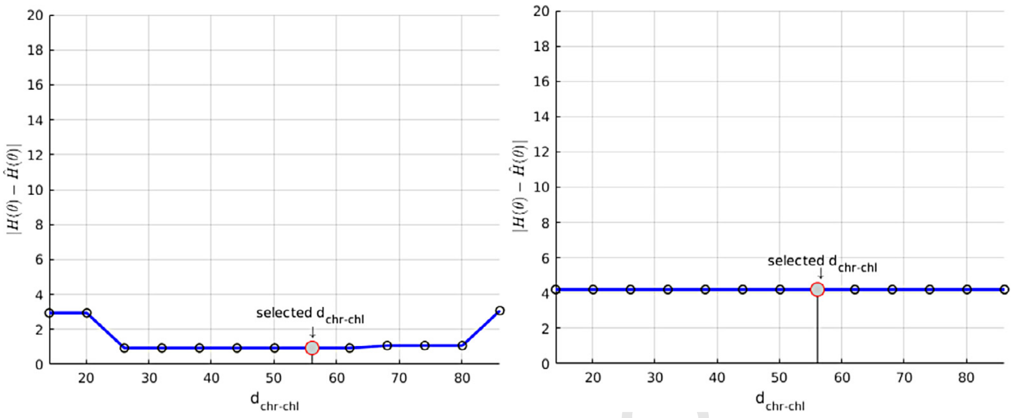
Fig. 19. Sensitivity tests to anthropometric parameter changes for feature-based head model: varying interpupillary breadth mean, in mm – see Table 2. Plot (a) corresponds to an image of subject #3 with head turned at $\theta \approx 9°$ (quasi-frontal), while the plot (b) corresponds to an image of the same subject at $\theta \approx 23°$ (clearly slanted). Both plots present the absolute angular error associated to the offline estimation of head pose direction vector $\mathbf{H}$ in θ using only $\mathbf{p}_{\mathbf{l},\mathbf{r}}$ (eyes). In this extreme case, we found that the algorithm is robust to changes of the mean parameters until a cap is reached after which the likelihood distributions drift too far from each other, and a bias towards one of the eyes appears.

conditions. We are certain that future work described in section 7 concerning the expansion of our context-dependent top-level mixture model will improve performance even further.

6. Discussion

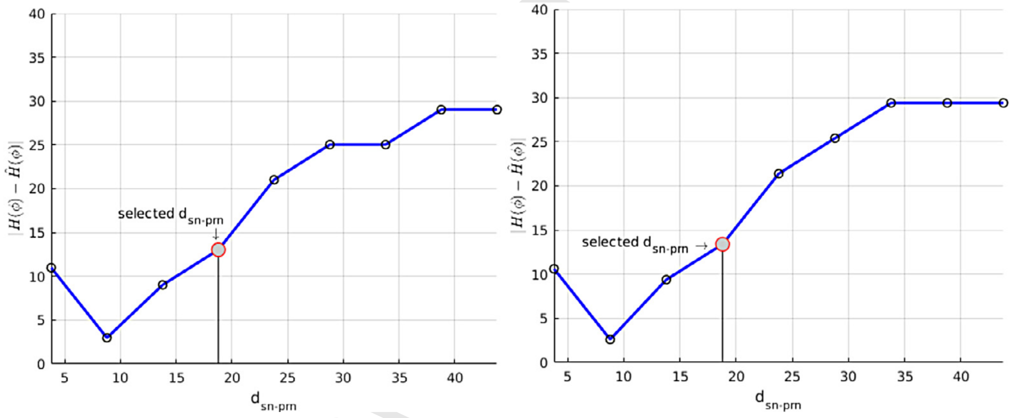
The results presented in the previous section demonstrate the advantages of our system comparing to related work:

- It is robust to extreme conditions – it always provides a “best guess” estimate even when data is incomplete and/or unreliable.
- Its robustness allows it to be used for distances to the interlocutor greater than 2 m and/or low camera resolutions, which is an essential feature for HRI applications. Moreover, the appropriateness of our solution for this target application was demonstrated in the previous section.



(a) (b)

Fig. 20. Sensitivity tests to anthropometric parameter changes for feature-based head model: varying mouth breadth mean, in mm – see Table 2 (plots are arranged as in Fig. 19). Both plots present the absolute angular error associated to the offline estimation of head pose direction vector **H** in θ using only $\mathbf{p}_{\mathbf{l},\mathbf{r}}$ and $\mathbf{ch}_{\mathbf{l},\mathbf{r}}$ (lip extremities). The contribution of the additional features when comparing to Fig. 19 improves model accuracy and robustness, and as a side effect the model is less sensitive to parameter changes.



(a) (b)

Fig. 21. Sensitivity tests to anthropometric parameter changes for feature-based head model: varying nose protrusion mean, in mm – see Table 2 (plots are arranged as in Figs. 19 and 20). Both plots present the absolute angular error associated to the offline estimation of head pose direction vector **H** in ϕ using only $\mathbf{p}_{\mathbf{l},\mathbf{r}}$ and **prn** (nose tip). In the absence of lip extremity features, we found that the estimation of ϕ is highly sensitive to the value chosen for the nose protrusion mean parameter, as can be seen in both plots.

- Fusion of estimates resulting from different data sources using appropriate weights in the mixture model confers adaptivity to our solution, allowing it to adjust the integration of all submodels according to overall circumstances of an interaction scenario.

However, robustness was achieved at the expense of precision, which represents the main limitation of our solution when comparing to related work, especially at closer distances to the artificial observer. Another drawback is the computational complexity involved in performing inference due to the substantial size of the search space. Moreover, being a geometrical approach, the amount of different detectors needed slows down image processing considerably – for example, face detection for all the different poses takes over half a second [42]. Additionally, our solution relies heavily on the piecewise-smooth gaze assumption; however, this will be addressed in the future. Finally, the eye gaze estimation model relies on a geometry that may not be the most robust approach – this will also be addressed in future work (see following section).

As a final remark, our solution, as shown in Fig. 10, only deals with 3D features; feature detection on the image is beyond the scope of our work. In fact, this makes it dependent on the quality of the feature detector suite used in the processing dataflow. This dependence has been shown in results by comparing the best-case scenario of hand-labelled features with off-the-shelf feature detector framework.

Please cite this article in press as: P. Lanillos et al., A Bayesian hierarchy for robust gaze estimation in human–robot interaction, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.04.007

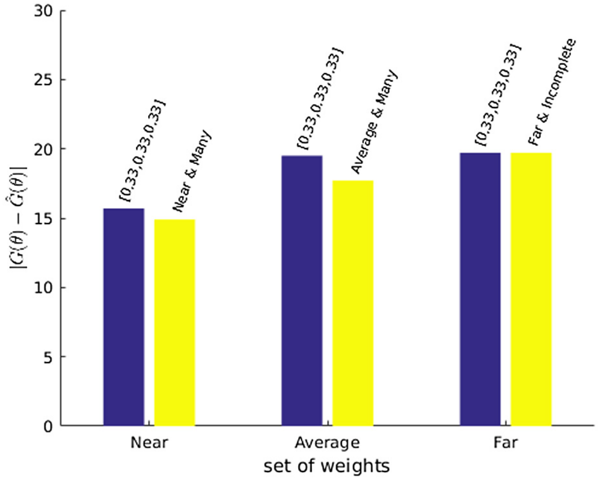


Fig. 22. Sensitivity tests to top-level model weight parameter changes – results corresponding to subject #3. Pairs of bars in the plot give average errors for specific sets of weights at each distance range (the weight for the prediction model was set to zero so as to remove its smoothing effect and also the temporal dependence of the results). In each pair at each distance, the left bar corresponds to a set of uniform weights, while the bar on the right conforms to the sets in Table 1 corresponding to the conditions labelled on top.

7. Conclusions and future work

In this paper, we have proposed a novel robust solution for gaze detection based on a hierarchical probabilistic approach, inherently dealing with perceptual uncertainty and incomplete data. Our hierarchical framework was designed with human–robot interaction in mind, and consequently attempts to loosely follow the characteristics of gaze estimation as performed by the human brain. Moreover, our solution adaptively fuses estimates resulting from complementary data sources using contextually-appropriate weights in the top-level mixture model. Results show that the framework performs according to expectations, appropriately addressing the challenges identified in section 2. This framework is usable in a variety of HRI applications, including scenarios with multiple robots such as presented by Portugal and Rocha [43], and is currently integrated in an artificial cognitive system for HRI developed by our team [42,44].

As future work, we are planning on improving the automatic feature detection framework (for example, by testing out a more robust alternative eye gaze model using the relative position of the irises within the visible portion of the sclera, as several researchers hypothesise to happen in human perception [11,12]), and also the prediction model by refining the respective estimate using information on the most probably attended object given the previous gaze direction estimate (as the human brain most certainly does, and exploiting an idea similar to what is presented in [28]). We will also include an additional condition for the top-level model, which will allow our framework to deal with sudden gaze shifts and therefore explicitly address sudden inflection instants due to the piecewise-constant nature of gaze changes.

Acknowledgements

The authors would like to thank Beatriz Oliveira for her inestimable help in developing the labelling software and also with the experimental sessions, and Gigel Macesanu for contributing to the development of the automatic feature detectors. We would also like to thank Cláudio Araújo for having implemented the experimental dataset website. We would also like to express our gratitude to all the subjects for kindly accepting to participate in the experiments.

This work was supported by the Portuguese Foundation for Science and Technology (FCT) and by the European Commission via the COMPETE programme [project grant number FCOMP-01-0124-FEDER-028914, FCT Ref. PTDC/EEI-AUT/3010/2012].

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.ijar.2017.04.007.

References

[1] J.F. Ferreira, J. Dias, Attentional mechanisms for socially interactive robots – a survey, in: Special Issue on Behavior Understanding and Developmental Robotics, IEEE Trans. Auton. Ment. Dev. 6 (2) (2014) 110–125, http://dx.doi.org/10.1109/TAMD.2014.2303072.

[2] E.T. Hall, The Hidden Dimension, Anchor Books, New York, NY, ISBN 0385084765, 1990.

[3] N. Bergstrom, T. Kanda, T. Miyashita, H. Ishiguro, N. Hagita, Modeling of natural human–robot encounters, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2008, 2008, pp. 2623–2629, http://dx.doi.org/10.1109/IROS.2008.4650896.

[4] S. Satake, T. Kanda, D.F. Glas, M. Imai, H. Ishiguro, N. Hagita, How to approach humans? – strategies for social robots to initiate interaction, in: 2009 4th ACM/IEEE International Conference on Human–Robot Interaction, HRI, IEEE, 2009, pp. 109–116.

[5] Z. Henkel, C. Bethel, R. Murphy, V. Srinivasan, Evaluation of proxemic scaling functions for social robotics, IEEE Trans. Human-Mach. Syst. (ISSN 2168-2291) 44 (3) (2014) 374–385, http://dx.doi.org/10.1109/THMS.2014.2304075.

[6] C. Shi, M. Shimada, T. Kanda, H. Ishiguro, N. Hagita, Spatial formation model for initiating conversation, in: Robotics: Science and Systems, vol. 11, 2011.

[7] E. Pacchierotti, H.I. Christensen, P. Jensfelt, Evaluation of passing distance for social robots, in: The 15th IEEE International Symposium on Robot and Human Interactive Communication, ROMAN 2006, IEEE, 2006, pp. 315–320.

[8] J.F. Ferreira, J. Dias, Probabilistic Approaches for Robotic Perception, Springer Tracts in Advanced Robotics (STAR), vol. 91, Springer, ISBN 978-3-319-02006-8, 2014.

[9] P. Bessière, E. Mazer, J.-M. Ahuactzin, K. Mekhnacha, Bayesian Programming, Machine Learning & Pattern Recognition, Chapman & Hall/CRC Press, 2014.

[10] W.H. Wollaston, On the apparent direction of eye in a portrait, Philos. Trans. R. Soc. Lond. B 114 (1824) 247–256.

[11] S.R.H. Langton, H. Honeyman, E. Tessler, The influence of head contour and nose angle on the perception of eye-gaze direction, Atten. Percept. Psychophys. 66 (5) (2004) 752–771, http://dx.doi.org/10.3758/BF03194970.

[12] D. Todorovic, Geometrical basis of perception of gaze direction, Vis. Res. (ISSN 0042-6989) 46 (21) (2006) 3549–3562, http://dx.doi.org/10.1016/j.visres.2006.04.011.

[13] E. Murphy-Chutorian, M. Trivedi, Head pose estimation in computer vision: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 31 (4) (2009) 607–629.

[14] J. Kaminski, D. Knaan, A. Shavit, Single image face orientation and gaze detection, Mach. Vis. Appl. 21 (3) (2009) 85–98.

[15] A. Nikolaidis, I. Pitas, Facial feature extraction and pose determination, Pattern Recognit. 33 (11) (2000) 1783–1791.

[16] C. Canton-Ferrer, J. Casas, M. Pardas, Head Orientation Estimation Using Particle Filtering in Multiview Scenarios, Multimodal Technologies for Perception of Humans, vol. 4625, Springer, Berlin, ISBN 978-3-540-68584-5, 2008, pp. 317–327.

[17] B. Patrão, P. Menezes, An interactive system for people suffering from cerebral palsy, Int.J. Reliable Qual. E-Healthcare (IJRQEH) 2 (3) (2013) 1–14, IGI Global.

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

[19] Y. Sugano, Y. Matsushita, Y. Sato, Appearance-based gaze estimation using visual saliency, IEEE Trans. Pattern Anal. Mach. Intell. (ISSN 0162-8828) 35 (2) (2013) 329–341, http://dx.doi.org/10.1109/TPAMI.2012.101.

[20] F. Lu, Y. Sugano, T. Okabe, Y. Sato, Inferring human gaze from appearance via adaptive linear regression, in: 2011 IEEE International Conference on Computer Vision (ICCV), 2011, pp. 153–160, http://dx.doi.org/10.1109/ICCV.2011.6126237, ISSN 1550-5499.

[21] B.A. Smith, Q. Yin, S.K. Feiner, S.K. Nayar, Gaze locking: passive eye contact detection for human–object interaction, in: Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, UIST '13, ACM, New York, NY, USA, ISBN 978-1-4503-2268-3, 2013, pp. 271–280, http://dx.doi.org/10.1145/2501988.2501994.

[22] R. Ronsse, O. White, P. Lefevre, Computation of gaze orientation under unrestrained head movements, J. Neurosci. Methods 159 (2007) 158–169.

[23] J. Sung, T. Kanade, D. Kim, Pose robust face tracking by combining active appearance models and cylinder head models, Int. J. Comput. Vis. (ISSN 0920-5691) 80 (2008) 260–274.

[24] K. Funes Mora, J. Odobez, Gaze estimation from multimodal Kinect data, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 25–30, http://dx.doi.org/10.1109/CVPRW.2012.6239182, ISSN 2160-7508.

[25] K. Funes Mora, J.-M. Odobez, Geometric generative gaze estimation (G3E) for remote RGB-D cameras, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1773–1780, http://dx.doi.org/10.1109/CVPR.2014.229.

[26] Z. Yücel, A.A. Salah, Çetin Meriçli, T. Meriçli, R. Valenti, T. Gevers, Joint attention by gaze interpolation and saliency, IEEE Trans. Cybern. 43 (3) (2013) 829–842.

[27] M.W. Hoffman, D.B. Grimes, A.P. Shon, R.P. Rao, A probabilistic model of gaze imitation and shared attention, Neural Netw. 19 (3) (2006) 299–310.

[28] B. Massé, S. Ba, R. Horaud, Simultaneous estimation of gaze direction and visual focus of attention for multi-person-to-robot interaction, in: 2016 IEEE International Conference on Multimedia and Expo (ICME), 2016, pp. 1–6, http://dx.doi.org/10.1109/ICME.2016.7552986.

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

[30] L.G. Farkas, Anthropometry of the Head and Face in Medicine, Elsevier, New York, 1981.

[31] E. Skodras, N. Fakotakis, An unconstrained method for lip detection in color images, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing, 2011, pp. 1013–1016, ISSN 1520-6149.

[32] G. Macesanu, V. Comnac, F. Moldoveanu, S.M. Grigorescu, A time-delay control approach for a stereo vision based human–machine interaction system, J. Intell. Robot. Syst. (ISSN 0921-0296) (2013) 1–17, http://dx.doi.org/10.1007/s10846-013-9994-4.

[33] J.F. Ferreira, P. Lanillos, J. Dias, Fast exact Bayesian inference for high-dimensional models, in: Workshop on Unconventional Computing for Bayesian Inference (UCBI), IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015.

[34] Anthropometry of U.S. Military Personnel (Metric), ARMY: U.S. Army Soldier Systems Center, 1991, DOD-HDBK-743A.

[35] L.G. Farkas, B. Tompson, J.H. Phillips, M.J. Katic, M.L. Cornfoot, Comparison of anthropometric and cephalometric measurements of the adult face, J. Craniofac. Surg. 10 (1) (1999) 18–25.

[36] L.G. Farkas, M.J. Katic, C.R. Forrest, International anthropometric study of facial morphology in various ethnic groups/races, The Journal of Craniofacial Surgery 16 (4) (2005) 615–646.

[37] B. Oliveira, P. Lanillos, J.F. Ferreira, Gaze tracing in a bounded log-spherical space for artificial attention systems, in: ROBOT'2015 – Second Iberian Robotics Conference, Advances in Intelligent Systems and Computing, Springer-Verlag, 2015.

[38] J. Dias, P. Lanillos, J.F. Ferreira, Designing social interaction with robots – towards an artificial attention system, in: Workshop on Improving the Quality of Life in the Elderly Using Robotic Assistive Technology: Benefits, Limitations, and Challenges, Paris, 2015.

[39] S.W. Bock, P. Dicke, P. Thier, How precise is gaze following in humans?, Vis. Res. (ISSN 0042-6989) 48 (7) (2008) 946–957, http://dx.doi.org/10.1016/j.visres.2008.01.011.

[40] L.A. Symons, K. Lee, C.C. Cedrone, M. Nishimura, What are you looking at? Acuity for triadic eye gaze, J. Gen. Psychol. (ISSN 0022-1309) 131 (4) (2004) 451–469.

[41] K. Khoshelham, S.O. Elberink, Accuracy and resolution of kinect depth data for indoor mapping applications, Sensors 12 (2) (2012) 1437–1454, http://dx.doi.org/10.3390/s120201437.

[42] P. Lanillos, J.F. Ferreira, J. Dias, Designing an artificial attention system for social robots, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015.

[43] D. Portugal, R.P. Rocha, Distributed multi-robot patrol: a scalable and fault-tolerant framework, Robot. Auton. Syst. 61 (12) (2013) 1572–1587, http://dx.doi.org/10.1016/j.robot.2013.06.011.

[44] J.F. Ferreira, J. Lobo, P. Bessière, M. Castelo-Branco, J. Dias, A Bayesian framework for active artificial perception, IEEE Trans. Syst. Man Cybern., Part B, Cybern. (ISSN 1083-4419) 43 (2) (2013) 699–711, http://dx.doi.org/10.1109/TSMCB.2012.2214477.

Please cite this article in press as: P. Lanillos et al., A Bayesian hierarchy for robust gaze estimation in human–robot interaction, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.04.007