J Multimodal User Interfaces
DOI 10.1007/s12193-009-0029-1

ORIGINAL PAPER

Investigating shared attention with a virtual agent using a gaze-based interface

Christopher Peters · Stylianos Asteriadis · Kostas Karpouzis

Received: 16 April 2009 / Accepted: 9 November 2009

Abstract This paper investigates the use of a gaze-based interface for testing simple shared attention behaviours during an interaction scenario with a virtual agent. The interface is non-intrusive, operating in real-time using a standard web-camera for input, monitoring users' head directions and processing them in real-time for resolution to screen coordinates. We use the interface to investigate user perception of the agent's behaviour during a shared attention scenario. Our aim is to elaborate important factors to be considered when constructing engagement models that must account not only for behaviour in isolation, but also for the context of the interaction, as is the case during shared attention situations.

Keywords Shared attention · Gaze detection · Embodied agents · Social behaviour

1 Introduction

Interfaces capable of detecting user behaviour and inferring mental states are still not widespread, due to the requirement for specialised and expensive equipment, unavailable to the majority of casual users. Most modern interfaces still use simplistic and explicit means for detecting the presence of a user and obtaining feedback, for example, by waiting for a key press or feedback from a device, such as a mouse, before continuing. These interfaces are not very natural to use, requiring the user to fully adapt to them. Research into the creation of more natural interfaces, capable of adapting to human social signals and accounting for emotional displays is thus desirable, and well under way [1].

Nonetheless, while research has been focused on low-level detection and mid-level inference techniques, many questions remain unanswered in relation to high-level aspects of interaction management. Factors such as engagement and context, for example, must be addressed and modelled if interfaces are to be created that are capable of managing their interactions intelligently, naturally and robustly. Yet any effort to embark on the creation of such models immediately raises questions as to what these apparently well-known concepts actually refer to in practice; creating connections between concrete low-level detection techniques and harder-to-discern high-level factors is a difficult proposition and must be informed by investigation.

We regard shared attention to be such a connection. It refers to the shared experience between two or more participants as they observe objects and events together, and is recognised as a pivotal skill in early social understanding [2, 3]. This mechanism seems crucial for endowing the machine with a fundamental understanding of the social behaviour of the user. It may also help to link low-level aspects, such as gaze and facial expression detection, with contextual details, to achieve a better understanding of the interaction. This is because shared attention not only requires an account of the state of the user in isolation, but additionally a consideration of that state in relation to another interactor and the

Electronic supplementary material The online version of this article (http://dx.doi.org/10.1007/s12193-009-0029-1) contains supplementary material, which is available to authorized users.

C. Peters (✉)
Department of Engineering and Computing, Coventry University, Coventry, UK
e-mail: christopher.peters@coventry.ac.uk

S. Asteriadis · K. Karpouzis
National Technical University Athens, Athens, Greece

S. Asteriadis
e-mail: stiast@image.ece.ntua.gr

K. Karpouzis
e-mail: kkarpou@image.ece.ntua.gr

Published online: 05 December 2009

environment. For example, in many systems, a user may be considered interested if they are looking at the other interactor. During shared attention scenarios however, this may be reversed: continuing to look at the other when they make reference to an object, for example by gazing at or pointing towards it while verbally describing it, could be interpreted as disinterest or lack of understanding. We believe the consideration of situations, such as the aforementioned, can illuminate the factors that contribute to the many concepts of engagement described in the literature [4] and inform the design of more sophisticated computational models.

In this work, we investigate shared attention behaviours between a user, a conversational embodied agent and a number of virtual objects residing on the screen. While the user is engaged in a monologue with the agent, during which it refers to surrounding scene objects, their head movements are tracked in real-time using a standard web-camera. The purpose of the investigation is thus two-fold: (i) to test the method for detecting the gaze of the user using a single, standard web-camera and evaluate its suitability for investigating gaze-related behaviours and (ii) to use the gaze detector to investigate users' perceptions of the agent when engaged in shared attention with it.

In Sect. 3, we describe how the gaze detector operates and how head movements are resolved to screen coordinates, and how the screen coordinates are interpreted with respect to scene objects. Section 4 provides the details of the two experiments conducted and implications of the results, before concluding (Sect. 5). First, we discuss the background of shared attention and gaze detection in Sect. 2.

2 Background

2.1 Interaction and shared attention

Humans are sensitive to the direction and nature of the gaze of others, behaviour that may provide general information about their interests and intentions, and aid interaction management. Attempts to automate the theorising of complex mental and emotional states based on behaviour have met with some success, based not only on the analysis of gaze and facial expression (see for example [5]) but additional factors, such as posture [6].

While such studies have focused on inferring details in relative isolation, others have focused on how to account for the context of the interaction. For example, [7] consider context during a chess game between children and a robot companion, while [8] use a multimodal approach coupled with task state information to classify the interest of children during a game. Context is also of importance when the environment must be accounted for during survival and social situations. In such cases, judging the direction of another's attention may be of critical importance [9]: in terms of survival, gaze-following may allow for the unintentional direction of another's attention towards potential threats and rewards in the environment [10]. During social encounters, objects or events may be purposefully cued with gaze in order to disambiguate or establish shared experience with the focus of discussion, a topic receiving attention in robotics research [11], with important links to imitation [12]. The role of gaze has also been studied in relation to virtual artificial entities, such as embodied conversational agents, or ECA's. Attentive presentation agents [13], for example, rely on the eye gaze of the user to infer their attention and visual interest. This is used to alter the ongoing behaviour of characters in real-time so that they may better adapt to the user. In a similar vein, [14] and [15] have been investigating systems for estimating user engagement based on gaze behaviour during interaction situations with a conversational agent, while [16] have considered the role of gaze for allowing a listening agent to provide feedback. These studies are concerned with asking important questions about higher-level aspects of the role of gaze in interaction, and are necessarily dependent on the appropriate functioning of low-level detection systems, described next.

2.2 Gaze detection

Two major methods for detecting a user's gaze direction have been extensively studied in the literature: head pose estimation, and eye gaze estimation. Various approaches have been adopted for retrieving important facial features from an image sequence.

In head pose estimation, many of the approaches proposed in the literature require more than one camera, or extra equipment [17–21], making the final system expensive, complex or intrusive. Furthermore, algorithmically, some methods require a set of facial features to be detected and tracked with a very high degree of accuracy [22, 23]. These techniques are usually sensitive to even small displacements of the features, which may cause the system to fail. Other techniques input the facial area and compare it against training sets of facial images [24, 25]. These methods suffer from the problem of alignment, especially in natural environments, where it usually is not easy to achieve good alignment between training and test images. Motion recovery is also effective for recovering head pose parameters. In this group of methods, the face is tracked and mapped onto a *3D* model [26] and motion parameters are extracted from it. These methods are often very accurate; however, they require knowledge regarding camera parameters or/and approximate knowledge of the distance between the user and the camera. Non-rigid models have also drawn much attention in recent years. In this group of methods, a series of transformations take place on a trained mesh of nodes and connections, in order to match with the shape or/and texture of the face region. Active Appearance Models [27] is

a characteristic paradigm and the parameters of such networks have been used to extract pose information. A major drawback of such methods, however, is the need for accurate initialisation, as non-rigid models are prone to the effects of local minima. A multitude of solutions, however, have been proposed in the literature to tackle this problem, as in [28], where the authors combine the local character of AAMs with the global motion parameters of cylinder models. The pose range reported in the paper, however, does not perform beyond $\pm 45^{\circ}$ for yaw angles. Many authors also use hybrid techniques [29, 30] in order to take advantage both from holistic and local features. While results have been promising for these approaches, the aforementioned problems also tend to be more apparent.

In the method employed in this paper, our goal has been to develop a real-time system that is reasonably robust to various lighting conditions, image resolution and specific training. To this end, although our system depends on feature tracking, it is not dependant on the geometrical relations between the tracked features, but on the relative motion between them with regards to key frames. This restriction had to be satisfied in order to allow the system to accommodate inaccurately tracked points, since the algorithm runs under natural and spontaneous conditions. Furthermore, facial feature detection during initialisation depends on a method utilising an edge map of the face area, and the geometrical relation of each feature with the closest edge [31] as an attempt to avoid issues of lighting variations between training and testing data. Additionally, in order to avoid error accumulation, our system re-initialises when a number of conditions are not met, as detailed in Sect. 3.1.

In our work, non-intrusive conditions are possible, allowing the user to behave in a spontaneous manner. The system does not need to be trained according to the user or background and although it uses facial feature detection and tracking, it is not highly dependant on accurate and exact localisation of the facial points, as both head pose and eye gaze are functions of relative movements among facial features and not their positions or *3D* relative positions.

3 Gaze-based interaction

The gaze detector (see Sect. 3.1) employs facial feature analysis of the images captured from a standard web-camera in order to determine the direction of the user's gaze. This information allows the user's gaze inside or outside the screen to be calculated, so that metrics relating to the user's attention and interest can be processed (Sect. 3.2). Based on the interpreted metrics, an assessment of the state of the interaction can be made in order to support shared attention behaviour and infer the state of the interaction.

3.1 Gaze detection from the user

The purpose of the gaze module is to detect the raw user gaze direction details from the web-camera in real-time. It is based on facial feature detection and tracking, as reported in [32], and follows a variant of this method for head pose and eye gaze estimation. More specifically, starting from the eye centers, which are easily detected [31], the eye corners and eyelids are detected, as well as two points on each eyebrow, the nostrils' midpoint and four points on the mouth. These features are subsequently tracked using an iterative, 3-pyramid Lucas Kanade tracker [33]. Lucas-Kanade tracking is one of the most widespread trackers cited in the literature: the choice of this tracker was based on the fact that it can accurately and effectively track features under a large variety of conditions. However, as is the case in real world conditions, a series of rules has to be adopted in order to tackle constraints imposed by natural lighting and motion conditions. Here, we assume an orthographic projection at successive frames, so that, for such small periods of time, the motion vectors of all features can be considered to be almost equal. Features whose motion vector length m_i is much larger or smaller than the mean motion of all features m_{mean} ($m_i > t_1 * m_{mean}, m_i < t_2 * m_{mean}$, here we considered $t_1 = 1.5$ and $t_2 = 0.5$) are considered as outliers and their position is recalculated based on their previous position and the recalculated mean motion of the other features. In our tests, this step proved to be very important at improving the tracker's performance under difficult lighting conditions and occlusions.

Head pose estimation Head pose is estimated by calculating the displacement of the eye centers' midpoint, with regards to its position at a frame where the user faces the screen frontally, referred to in the remainder of this paper as the *frontal reference frame*. This displacement produces the head pose vector which is a reliable index of where the user's head is currently oriented towards (see Fig. 1). Normalisation with the inter-ocular distance, in pixels, at start-up guarantees that the head pose vector is scale independent. In order to distinguish between displacements caused by head rotations and by translations, the triangle formed by the triplet of the eyes and the mouth is monitored and the head pose vector is only calculated when the inter-ocular distance to the eyes-mouth vertical distance changes significantly with regards to a frame where the person is looking frontally.

Re-initialisation To further suppress error accumulation, the system re-initialises when certain conditions regarding head pose vector length are met. Rapid head rotations, for example, may cause some features to be occluded and thus, when the user returns to a frontal position, one of the two eye

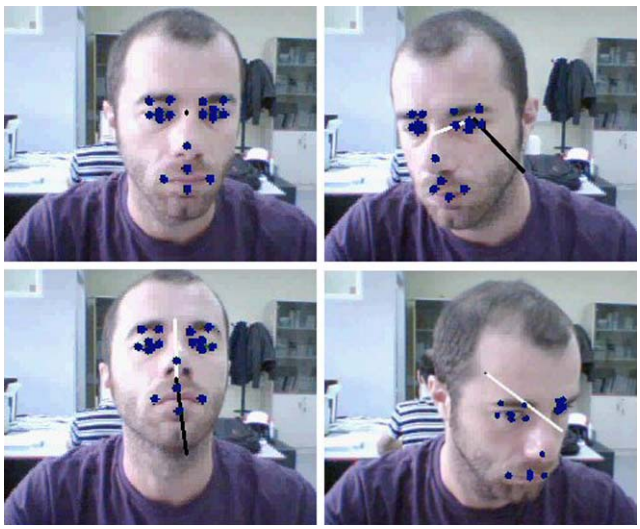

Fig. 1 Gaze direction detection is based on a number of tracker features (shown here as *black dots*) in order to calculate a final head pose (*white line*) and eye gaze (*black line*) vector

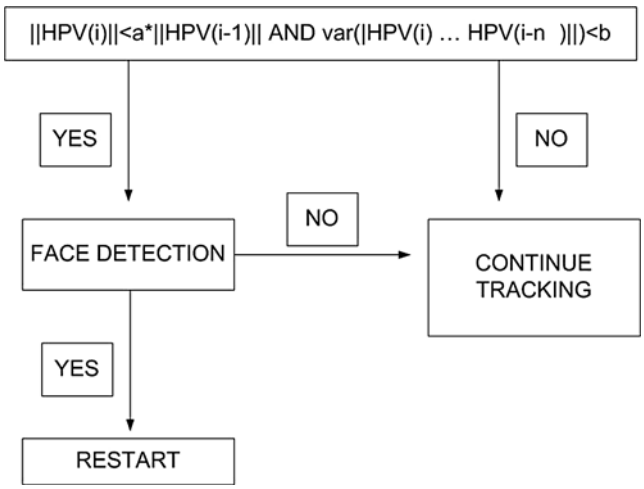

Fig. 2 Diagram depicting the conditions and process for system reinitialisation

centers might be erroneously tracked while the second continues to follow the movement of the head. In such cases, the head pose reduces in length and stays fixed when the person is facing the camera frontally. This allows the system to reinitialise by detecting the facial features again and restarting the tracker. The above process is depicted in Fig. 2.

Here $\|\mathbf{HPV}(i)\|$ is the head pose vector length at current frame i, $a = 0.7$, $b = 0.07$, $n = 10$. As face detection and facial feature detection update slower than the tracker, video streaming continues normally and the second frame is processed by the camera in real time. However, initialisation normally runs at ∼3 fps: thus pose, expressions and the fluency of tracking are not significantly effected when initialisation takes place. Re-initialisations, whenever they occur, require 330 ms.

Eye gaze estimation During the process of eye gaze estimation, relative displacements of the iris center with regards to the points around the eye provide a good indication of the directionality of the eyes with regards to the frontal reference frame. These displacements correspond to the eye gaze vectors (see Fig. 1). To reinforce correct eye center tracking, the tracked eye centers' positions are updated by searching for the darkest neighbourhoods around them and placing the eye center in the midpoint of this neighbourhood. This helps to alleviate the effects of blinking and saccadic eye movements. These displacements are normalised by the inter-ocular distance at start-up and, thus, are scale independent.

The computational complexity of the method permits real-time operation and requires only a simple web-camera to function. Tracking the features takes 13 milliseconds per frame on average for a resolution 288 × 352 pixels of the input video, using a Pentium 4 CPU, running at 2.80 GHz.

While the detector is capable of head and eye gaze estimation, for the purposes of the experiments, gaze was based solely on head direction. Although eye gaze would also have been desirable, robustness problems were encountered when attempting to obtain both at the same time from the tracker due to the image resolution required to encapsulate the head. Nonetheless, head direction is regarded as having a significant contribution towards computing another's direction of attention [9].

3.2 Attention and engagement

The module described in Sect. 3.1 detects the user's head direction. For this information to be of use, it must first be mapped onto screen coordinates to determine where the user is looking at any specific time: further abstractions are required to be able to infer higher-level details related to interaction over longer time spans. Each level of abstraction provides relationships to a greater amount of information over longer periods of time [34], starting for example with the ability to cluster regions of interest together to resolve what objects are being looked at, and later, inferring engagement by considering sets of objects with respect to the content of the ongoing dialogue.

Conversion to screen coordinates This step involves transforming the user's head (or eye) directions into 2D screen coordinates. To do this calculation, the raw direction information must be converted into 2D coordinates allowing them to reference the screen. The procedure for converting head direction into screen coordinates uses a calibration process, during which head movement to screen extents is considered. It is invoked at the beginning of each interaction scenario with the user in order to find the corresponding maximum and minimum extents of the screen boundary

in terms of raw head direction values. During this process, the user directs their head towards a cursor that moves between eight screen extents: top, bottom, left, right and each of the screen corners. At each update of the gaze detector, the absolute head direction detected is mapped to screen (x, y) coordinates based on comparison with the maximum and minimum screen border extents taken during the calibration process. The user may need to rerun the calibration process if their head position changes significantly, e.g. by moving their seat, or leaning to the left or right. In practice, however, we found that small head position changes did not greatly effect the robustness of the system.

A list of 2D coordinates are stored for each update of the gaze detector, referring to the screen coordinates that gaze is considered to have targeted at that specific time. There are two general possibilities: the user is either looking inside or outside of the screen area containing the 3D scene. This area can be considered as the *action space*, within which the events and objects relating to the interaction are based. If the user is not looking at the action space, we might presume that they are either disengaged or uninterested in the interaction to some degree. The data structure containing the final coordinates includes a flag signalling if gaze fell inside or outside of the screen, in addition to the 2D coordinate. If gaze fell inside the screen area, the 2D coordinate corresponds to the (x, y) screen position.

We also consider the boundary regions, beyond which gaze may extend when it falls outside of the screen area. This was modelled in order to provide the system with the ability to collect more precise data regarding user disengagement. For example, a user detected as looking above the top boundary of the screen may be thinking, an endogenous disengagement behaviour that may not be considered in the same category of disengagement as when a user looks away to attend to an exogenous distraction. However, we could not collect data in this instance relating to events taking place outside the screen, making it difficult to provide a context for potential correlation with user disengagements.

Directedness and level of attention We use a directedness metric to refer to the momentary orienting of the user's body parts with respect to another entity or object from the perspective of that entity or object. This is based on Baron-Cohen's eye, head and body direction detectors [2] and related work [9, 10] based on neurophysiological evidence (see [35] for example). This may include details of the user's eyes, head, body and even locomotion directions. For example, if the user orients their head and eyes directly towards an object, they could be considered as having a high degree of directedness towards it. In these studies the user is constrained in a static position, sitting in front of the monitor and agent, directly facing the screen; since eye direction is not considered in the scenario, head direction is mapped directly onto the directedness metric.

Directedness is a momentary concept. Alone, it is a highly unreliable indicator of attention; for example, if head direction is sampled while the user is in the process of a gaze change, results would be misleading. Level of attention is therefore used to refer to gaze within certain regions over multiple samples. An important issue in this respect relates to the clustering of the foci of interest of the user—in our system, this is achieved using virtual attention objects, or VAO's, described next.

Virtual attention objects In terms of screen coordinates, the user's focus of attention appears transient as it shifts around a scene. A higher level of representation is needed however, as a user may in fact be attending to a single object for the duration of these shifts. In order to simplify the analysis of what is being looked at in the scene, we define virtual attention objects (see Fig. 3). A single VAO is attached to each scene object for which we wish to accumulate attention information—for example, in the shared attention sce-

Fig. 3 Depiction of (*top*) the scenario, containing the Greta agent [16], acting in the role of a salesperson, and a number of objects and (*below*) the contents of the scene depicted in terms of virtual attention objects (VAO's). Each VAO records when and to what degree the user has been looking at it

nario, one VAO is defined for the agent, one for each scene object, one for the scene background, and one to represent the area outside of the screen. Resolving screen coordinates into VAO details is trivial: if the screen-coordinate of the gaze fixation is located inside a VAO, then its corresponding level of attention is increased. Thus, as the user's gaze moves around the screen, each VAO maintains a history of how much and when the user has fixated it. The agent has access to the information of all VAOs in the scene. Since the agent is itself a VAO, it therefore has a full assessment of the user's gaze towards specific objects.

Level of interest Over a larger time-frame, a *level of interest* metric is calculated. Unlike the previous metrics, this is calculated based on the stored attention levels for each member of a set of VAO's. Each member is categorised according to whether it is a scene object, the agent, the background, or a special object representing the area outside of the screen.

It is at this level that specific forms of context can be accounted for: By dynamically defining a set of VAO's containing only those objects relevant to the current interaction, such as recently pointed to or discussed objects, the attention of the user can be compared with this set to obtain a measurement of their level of interest in the interaction itself, referred to here as the *level of engagement*.

Engagement Engagement has been described as "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake" [36] and also as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and continuing the interaction" [37, 38]. We regard engagement as being facilitated by both attentive and emotional processes between the interactors. Although we are attempting to construct a shared attention model, we view engagement as being a complementary related topic underlying this aim. An important factor underlying shared attention through gaze behaviour that may be regarded as differentiating it from pure gaze-following, is that both participants are engaged to some degree with each other before the onset of the shared attention behaviour and there is an explicit goal on behalf of the sender to signal the object of interest to the other. For example, one may consider the case where a mother establishes prolonged mutual eye contact with her infant before providing a gaze cue towards a cuddly toy to be attended to.

In relation to engagement, the level of engagement details how much the user has been looking at the *relevant* objects in the scene at *appropriate times*. These will be recently referenced objects in the interaction, e.g. those looked at, pointed to and/or verbally described. These measures are made possible by considering the specific set of VAO's corresponding to currently and recently referenced objects in the interaction. When the agent is talking, but does not refer to anything in the environment, it will be the only VAO in the set, and when it stops talking, this VAO set will be empty.

Not all attention paid to the scene necessarily indicates engagement in the interaction with the agent. In addition to the level of engagement, a *quality of engagement* may also be defined to account for this. It provides to a slightly more detailed assessment of the type of engagement that the user has entered into. For example, a user who is not engaged in the interaction may not necessarily be looking outside of the scene. Instead, they may be attending to the scene in a superficial manner, looking at objects of interest that are irrelevant to the ongoing interaction. We therefore define three broad quality levels: (i) engaged in the interaction (ii) superficially engaged with the scene and action space and (iii) uninterested in the scene/action space. In this way, the behaviour of the user is not being considered in isolation, but in the context of what the agent is doing. If the agent is describing something important for example, disengagement on behalf of the user can be considered more serious than if the agent is not doing anything at all.

4 Experiments

Two experiments were conducted in order to (i) test the effectiveness of the gaze detector with a standard web-camera for the shared attention scenario (Sect. 4.1) and (ii) use the gaze detector in the shared attention scenario to investigate user behaviour in order to elaborate the metrics described (Sect. 4.2).

System details The system used for the experiments is comprised of two key modules: a gaze detector module and a player module. These modules communicate via a *Psyclone* connection—a blackboard system for use in creating large, multimodal A.I. systems. The gaze detector module comprises the capabilities described in Sect. 3.1, employing facial feature analysis of images captured from a standard web-camera in order to determine the direction of the user's gaze. The player displays the interactive graphics required, including the embodied conversational agent called Greta [16] and the scene. It receives updates of the user's gaze from the gaze module, and interprets and records the results (Sect. 3.2).

4.1 Experiment 1

To assess the effectiveness of the head pose estimation module for allowing a user to interact during the shared attention scenario (Sect. 4.2), the following experiment was set up. We seated seven participants (4M, 3F) in front of a computer

Fig. 4 Depiction of the scenario, where the agent makes a number of different gaze behaviours towards objects as it refers to them while speaking. An image of the participant is shown in the center—during the scenario, their gaze direction is detected in real-time as (*left*) they look towards the center of the screen, or towards the an object being discussed (*center*). The head movements of the participant are depicted in the rightmost image. The current focus of attention in each case is shown as a *red circle* encapsulating a crosshair

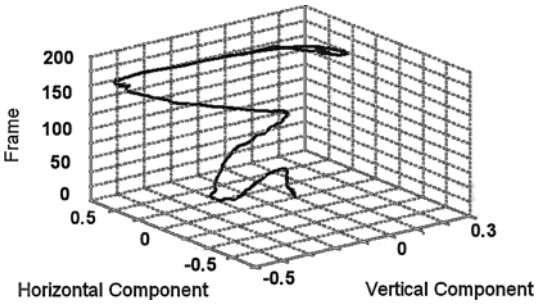

Fig. 5 Manifold depicting average head rotations in space

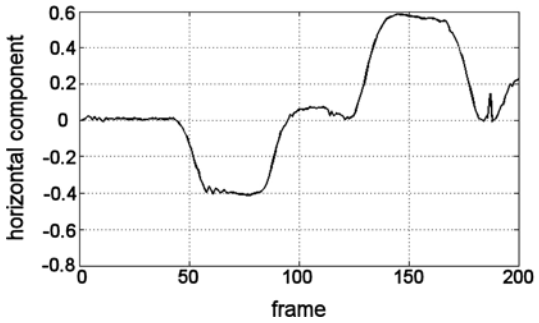

Fig. 6 Average horizontal component of head pose vector

screen and the head pose estimation module followed their head movements. The web-camera was placed on the top of the computer monitor, at a distance of 60–70 cm from the eyes of each participant. Lighting conditions were typical for an office environment, although reliability of the system would not be guaranteed in the case where heavy shadows appeared. For each participant, we monitored their movements for a period of 200 frames, asking them to track, with their heads, a moving object displayed on the screen. At the start of the procedure, a small icon appeared in the centre of the screen, and moved to a number of screen extents (e.g. lower left of the screen) moving back to the center of the screen each time.

Given the head pose vector $\mathbf{HPV} = [hpv_x \; hpv_y]$, with hpv_x and hpv_y being the horizontal and vertical components respectively, the mean values of each component, along all participants, during the period of the experiment are shown in Fig. 5. The mean horizontal component, averaged for all participants, is shown in Fig. 6. It can be seen that, although the icon changed position every 40 frames, there was a small delay in participants' gaze responses ($\simeq$10 frames). This was taken into consideration during the evaluation. In order to evaluate the appropriateness of our head pose estimation scheme in a test scenario, the following rules were defined:

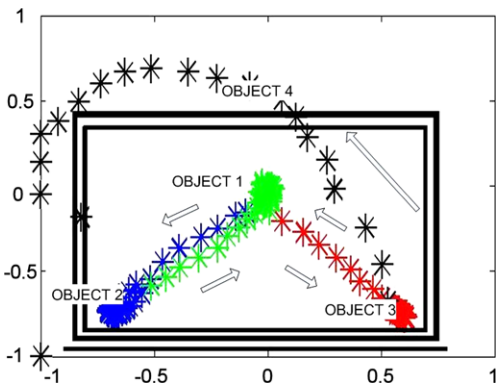

Fig. 7 Gaze patterns of a participant during Experiment 1; different colours correspond to different objects as they appear in the reference data

OBJ1 = screen centre, OBJ2 = lower left corner, OBJ3 = lower right corner, OBJ4 = everything else. A typical example of a participant's gaze path on a monitor, following the pattern OBJ1 → OBJ2 → OBJ1 → OBJ3 → OBJ4 can be seen in Fig. 7.

From patterns like the one in Fig. 7, the following rules were extracted, allowing the identification of the object corresponding to certain values of the head pose vector.

Table 1 Error confusion matrix for all participants, P1 to P7

	OBJ1	OBJ2	OBJ3	OBJ4	PA(%)	UA (%)
OBJ1	493	0	0	13	95.17	97.43
OBJ2	0	143	0	0	70.44	100
OBJ3	12	54	173	1	85.22	72.08
OBJ4	13	6	30	154	91.67	75.86
Total	518	203	203	168		
Overall accuracy:	88.2%	$Khat = 0.83$				

IF $-thr_1 < hpv_y < thr_1$ AND $-thr_3 < hpv_x < thr_3$
THEN OBJ1

IF $hpv_y > thr_1$ AND $hpv_y < thr_2$ AND
$hpv_x < -thr_3$ AND $hpv_x > -thr_4$
THEN OBJ2

IF $hpv_y < -thr_1$ AND $hpv_y > -thr_2$ AND
$hpv_x < -thr_3$ AND $hpv_x > -thr_4$
THEN OBJ3

ELSE OBJ4

In our experiments, we considered $thr_1 = 0.2$, $thr_2 = 0.7$, $thr_3 = 0.2$ and $thr_4 = 0.7$ (**HPV** is normalised with the interocular distance). Since every object in the experiment would appear immediately after the previous one disappeared, the participants needed some transition time from each object to the other. For this reason, at the evaluation stage, we excluded the appearances of every object for the first ten frames from the reference data. Thus, for every participant, there were 160 valid reference frames. Table 1 shows the confusion matrix of the classification of each head pose vector instance to objects for all participants, using the previously defined thresholds.

The last two columns of the above matrix show the Producer's Accuracy (PA) and User's Accuracy (UA) percentages. The former stands for the capability of a classifier to classify a pattern correctly with regards to all of its instances, while the later is the total number of correct classifications with regards to the total number of its identification. As can be seen, the overall accuracy of the method was 88.2%. However, here we also calculated the *Khat index*, as a measure of the difference between the actual agreement between the data and the classifier and the agreement of the data with a random classifier. In this experiment, we achieved a Khat parameter greater than 0.8, which is indicative of strong agreement [39] between the ground truth data and the classification shown in Table 1. The above measures for each participant individually can be seen in Table 2.

Table 2 Error confusion matrix for all participants

	Khat measure	Overall accuracy (%)
P1	0.96	97.44
P2	0.96	97.44
P3	0.63	73.72
P4	1.00	100.0
P5	0.62	74.36
P6	0.65	75.00
P7	0.99	99.36
Overall accuracy:	88.2%	$Khat = 0.83$

From the above, it can be seen that Khat measures are very high for participants 1, 2, 4 and 7, while for participants 3, 5 and 7, they are of moderate agreement [39]. For participants 3 and 6, this is mainly due to the fact that OBJ2 was misclassified as OBJ4, and for participant 5, OBJ3 was misclassified as OBJ4. In general, our head pose estimator achieved a total of 88.2% success, which was satisfactory for a non-intrusive system, without any prior knowledge of participants' gaze patterns, in an indoors, uncontrolled environment. The thresholds set here are moderately strict in the sense that, for the problem of discriminating between three objects on the screen, the borders of each object could be wider (see Fig. 7). Thus, when setting $thr1 = 0$, OBJ2 and OBJ3 occupy a larger space. In this case, our algorithm would achieve in total $Khat = 0.88$ and overall accuracy 92.2%. These rates are not much higher than those achieved by setting stricter thresholds, demonstrating that the head pose estimator can provide satisfactory results for relatively fine screen space values.

4.2 Experiment 2

The second experiment focused on a more in-depth shared attention scenario. The purpose was to investigate how subtle changes in the non-verbal cueing behaviour of the agent could affect gaze-following of human participants. Seven participants (4M, 3F) were seated, individually, facing a web-camera and 1280 × 1024 display.

Scenario During the scenario, the virtual agent plays the role of a salesperson in a computer store, presenting a number of different computer accessories to the user for potential purchase (see Fig. 4). Each of the accessories is displayed graphically on screen beside the agent, and represented internally as a VAO (see Sect. 3.2). The agent engages in a monologue with the user, choosing an object randomly and providing a short predefined description of it. The user fulfills the role of passive listener in this case, and their head pose is determined and recorded to account for their interest in the scenario. Since head pose was used as the only determinant of gaze direction in the scenario, a cross-hair was displayed on the screen to provide feedback to the participants about the screen position that their head was oriented towards.

Participants were shown a total of four trials, randomly ordered between participants. Each trial took just over 3 minutes, for a total scenario time of 12 minutes for each participant. Trials differed according to the presence or type of gaze motion made by the agent when referring to the objects according to the following conditions:

1. Condition C1: Gaze forward. In this condition, the agent continually gazes forward, towards the user, at all times.
2. Condition C2: Congruent continual gaze. In this condition, the agent gazes forward while it is talking and gazes towards an object that it is describing for the duration of the description.
3. Condition C3: Congruent temporary gaze. This condition is as above, but the agent does not continue to gaze at the object for the total duration over which it talks about it; instead, it looks back at the user after 6 seconds.
4. Condition C4: Incongruent gaze. In this condition, the agent stares at any object, chosen randomly, apart from the object that it is currently describing.

The differences between the non-verbal gaze cues made by the agent were subtle. In cases where gaze was incongruent with the items being verbally described, the purpose was to investigate if participants were using the verbal or the non-verbal modality to determine where to direct their gaze.

Since the purpose of the study was to account not only for episodes of great interest, but also to investigate variety in a range of related states, the scenario was not intended to maintain stimulation over the course of the trials: Although the object was chosen randomly at each step, the position of each object did not change in the scene, nor did any of the other scene details or the object descriptions, and the behaviour of the agent changed only subtly, in terms of gaze direction as outlined by the condition type. The role of the agent i.e. a salesperson trying to sell items to the participants that they were familiarised with and uninterested in, was also chosen so as not to be particularly stimulating over the length of the interaction. Since the experiments took place in a natural, busy lab environment, it was expected (and indeed hoped) that participants may become distracted or may disengage momentarily from the scenario.

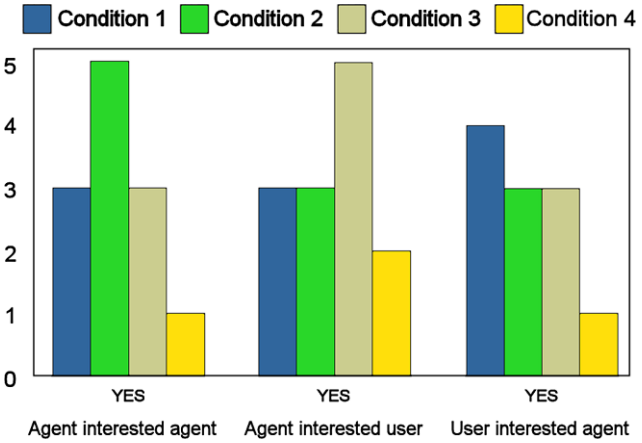

Fig. 8 Averaged results of reports from the study participants, for questions relating to the participants' ratings of (*left*) the agents interest in what it says, (*center*) the agents interest in the user and (*right*) the user's interest in the agent. Results are provided for each of the four conditions in the experiment

In each trial, five objects were randomly chosen to be described by the agent, according to the condition currently in effect, as described above. The ordering of trials was randomised over participants. During each trial, a visual recording was made of the participants as they engaged in gaze behaviours, for allowing validation of the results of the gaze detector. At the end of each trial, participants were asked to fill out a simple questionnaire, answering three YES/NO questions relating to each of the trials they participated in:

1. Agent interested agent: the participant thought the agent was interested in what it was saying.
2. Agent interested user: the participant thought the agent was interested in the user during the description.
3. User interested agent: the participant was interested in what the agent had to say.

The results of these questionnaires are presented in Fig. 8 and discussed next.

4.3 Discussion

During the scenarios, it was observed from the data that participants' head movements followed a discernible pattern, consistently directed either towards the agent or towards the objects being discussed by the agent. When the agent was not talking about a particular object, head direction tended to be consistently oriented towards the agent. Although this could suggest a lack of exploratory behaviour, even in the early stages of the experiment, given the distance (60–70 cm) and size of the screen, head motions were

not necessary for inspecting the scene. Careful inspection of the video data shows that in most cases, eye movements were used; peripheral vision and covert attention were also likely to have been employed. Overall, participants seemed exceptionally sensitive to how they made head and eye motions in the presence of the gaze tracker and cross-hair. In relation to head movements, participants seemed occupied with ensuring the cross-hair was targeting correctly, for example, at an object that was being described, and although they were instructed only to listen to the agent and no other task was specified, participants seemed to infer this as the implicit task of the experiment. It is possible that this may have detracted from their observation of the non-verbal behaviours of the agent, leading them to concentrate more on the verbal descriptions. Notably, although only participants' head-motions were tracked and used to provide feedback through the cross-hair positioning, as mentioned above, it is evident from the video recordings that participants also controlled their eye movements, suggesting that it was not only the presence of the feedback mechanism, but possibly also an awareness of the recording of their movements and inference of an implicit task that contributed to their inability to interact in a more flexible way with the system. Thus, although the scenario was constructed in an attempt to elicit a range of interest-related behaviours, results show that all of the participants performed exceptionally well in terms of attending to the object of discussion or agent under most circumstances. This is surprising; the experiment took place in a busy lab environment where there were distractions, and most participants reported afterwards that the experiment became somewhat tedious. Yet, after a consideration of both the tracked data and visual inspection of the recordings, this was not generally reflected in their gaze behaviour. In conjunction with a limited participant population, this made it difficult to obtain data on the somewhat broarder range of attentive behaviours that were hoped for.

In relation to participants' questionnaire responses, in most cases, no significant differences were found between the four conditions. Generally, this seems to suggest that participants did not notice any difference between cases when the agent did and did not cue the object of discussion using subtle gaze motions. However, in one instance a significant difference ($p < 0.05$) was found for conditions C2 and C4 when participants' reported the apparent interest the agent had in what it was describing. Both C2 and C4 are opposite cases, in the sense that C2 relates to the agent gazing at the object it is discussing, while C4 represents the condition where the agent gazes at a different object. This result is of importance, as it provides evidence that participants noticed a difference in the agent's behaviour during these opposing conditions.

In relation to this, a number of interesting behaviours were also observed in some of the resulting videos. In C4, the condition where the agent gazed at an object that was not being described, we observed that participants tended on occasion to follow the gaze of the agent to the incorrect object, before correcting their gaze to orient towards the object being described (see Fig. 9). In a number of other cases, it was observable from video that participants were more cautious, seemingly employing eye movements to carefully locate the object before moving their heads to the correct position. As described in Sect. 4.1, and highlighted in this study, alternative eye-head movement strategies form an integral part of natural human interaction and must be accounted for. Ideally, this entails a visual system capable of deriving both eye and head directions robustly at the same time, something our detection system is not yet capable of.

Fig. 9 An example case, in condition C4, where it appears that the participant followed the agents gaze towards an object other than that being described verbally. In this case, the participant follows the agents gaze towards the joystick object (*center bottom*) when in fact the agent verbally refers to the mouse object (*bottom left*). The participant later directed their head to the correct object

This also raises the important issue of the degree to which the user must control the interface, rather than the interface naturally interpreting the user's state: at present, the gaze detector does not seem robust enough, at least under the conditions in the experiment, to robustly read gaze direction under all conditions and circumstances. It requires for the user to ensure correct tracking by providing feedback in the form of the tracking cross-hair, which may become an obstacle to the interaction, a problem witnessed in these experiments.

5 Conclusions

We have presented a non-intrusive gaze interface and scenario for investigating shared attention behaviours with a virtual agent. The agent conducts a monologue with the user,

during which it references scene objects; user gaze is detected using a standard web-camera in real-time. An important issue for us has been to attempt to improve the naturalness of the interaction by permitting gaze recording in non-intrusive situations using cheap, widely available equipment. We have demonstrated that our gaze detector is effective in allowing head motions to be recorded to the accuracy required for these scenario's, removing the need for a head-mounted tracking device. Much work still remains, nevertheless, to improve robustness and achieve more natural conditions in other respects: the type of interaction conducted in the study presented here is limited and unnatural in many other ways, as the agent is engaged in monologue and does not gracefully open or end interactions with users. Our experiments have also investigated issues related to human interaction with a virtual agent during shared attention situations: while participants appeared to rely primarily on the agent's verbal exposition for directing their gaze to objects, our experiments provide evidence that they were nonetheless aware of differences in its gaze behaviour when it correctly and incorrectly cued objects. Further experimentation is needed to elucidate the results, particularly a consideration of how high-arousal situations may be avoided during experiments so that participants may display a wider range of behaviour, such as disengaging from the scenario. In this respect, another issue of great importance is the consideration of modalities in addition to gaze direction; particularly posture, facial expressions and verbal and non-verbal feedback.

Acknowledgements We would like to thank Etienne de Sevin of Institut Telecom ParisTech for his assistance with work concerning agent motivations.

References

1. Picard RW (1997) Affective computing. MIT Press, Cambridge
2. Baron-Cohen S (1994) How to build a baby that can read minds: cognitive mechanisms in mind reading. Cah Psychol Cogn 13:513–552
3. Scassellati B (1996) Mechanisms of shared attention for a humanoid robot. In: Embodied cognition and action: papers from the 1996 AAAI fall symposium. AAAI, Menlo Park
4. Peters C, Castellano G, de Freitas S (2009) An exploration of user engagement in HCI. In: Proceedings of the affective-aware virtual agents and social robots (AFFINE) workshop, international conference on multimodal interfaces (ICMI'09). ACM, Cambridge
5. El Kaliouby R, Robinson P (2005) Generalization of a vision-based computational model of mind-reading. In: ACII 2005: proceedings of the first international conference on affective computing and intelligent interaction, pp 582–589
6. Mota S, Picard RW (2003) Automated posture analysis for detecting learner's interest level. In: Computer vision and pattern recognition workshop, vol 5, p 49. IEEE Comput Soc, Los Alamitos
7. Castellano G, Pereira A, Leite I, Paiva A, McOwan PW (2009) Detecting user engagement with a robot companion using task and social interaction-based features. In: International conference on multimodal interfaces. ACM, Cambridge
8. Kapoor A, Picard RW (2005) Multimodal affect recognition in learning environments. In: ACM conference on multimedia, November 2005
9. Langton S, Watt R, Bruce V (2000) Do the eyes have it? Cues to the direction of social attention. Trends Cogn Sci 4:50–59
10. Emery NJ (2000) The eyes have it: the neuroethology, function and evolution of social gaze. Neurosci Biobehav Rev 24(6):581–604
11. Hoffman MW, Grimes DB, Shon AP, Rao RPN (2006) A probabilistic model of gaze imitation and shared attention. Neural Netw 19(3):299–310
12. Breazeal C, Scassellati B (2002) Challenges in building robots that imitate people. In: Dautenhahn K, Nehaniv CL (eds) Imitation in animals and artifacts. MIT Press, Cambridge, pp 363–390
13. Prendinger H, Eichner T, André E, Ishizuka M (2007) Gaze-based infotainment agents. In: Advances in computer entertainment technology, pp 87–90
14. Ishii R, Nakano YI (2008) Estimating user's conversational engagement based on gaze behaviors. In: Prendinger H, Lester JC, Ishizuka M (eds) Intelligent virtual agents, 8th international conference, IVA. Lecture notes in computer science, vol 5208. Springer, Berlin, pp 200–207
15. Peters C, Asteriadis S, Karpouzis K, de Sevin E (2008) Towards a real-time gaze-based shared attention for a virtual agent. In: International conference on multimodal interfaces (ICMI), workshop on affective interaction in natural environments, AFFINE, Chania, Greece
16. Bevacqua E, Mancini M, Pelachaud C (2008) A listening agent exhibiting variable behaviour. In: Intelligent virtual agents (IVA), Tokyo
17. Voit M, Nickel K, Stiefelhagen R (2005) Multi-view head pose estimation using neural networks. In: Second Canadian conference on computer and robot vision (CRV), Victoria, BC, Canada. IEEE Comput Soc, Los Alamitos, pp 347–352
18. Mao Y, Suen CY, Sun C, Feng C (2007) Pose estimation based on two images from different views. In: Eighth IEEE workshop on applications of computer vision (WACV). IEEE Comput Soc, Washington, p 9
19. Beymer D, Flickner M (2003) Eye gaze tracking using an active stereo head. In: IEEE computer society conference on computer vision and pattern recognition (CVPR), vol 2, Madison, WI, USA, 2003. IEEE Comput Soc, Los Alamitos, pp 451–458
20. Meyer A, Böhme M, Martinetz T, Barth E (2006) A single-camera remote eye tracker. In: Lecture notes in artificial intelligence. Springer, Berlin, pp 208–211
21. Hennessey C, Noureddin B, Lawrence PD (2006) A single camera eye-gaze tracking system with free head motion. In: Proceedings of the eye tracking research & application symposium (ETRA), San Diego, California, USA, 2006. ACM, New York, pp 87–94
22. Gee A, Cipolla R (1994) Non-intrusive gaze tracking for human-computer interaction. In: Int conference on mechatronics and machine vision in pract, pp 112–117, Toowoomba, Australia
23. Gourier N, Hall D, Crowley J (2004) Estimating face orientation from robust detection of salient facial features. In: International workshop on visual observation of deictic gestures (ICPR), Cambridge, UK
24. Seo K, Cohen I, You S, Neumann U (2004) Face pose estimation system by combining hybrid ica-svm learning and re-registration. In: 5th Asian conference on computer vision, Jeju, Korea
25. Stiefelhagen R (2004) Estimating head pose with neural networks—results on the pointing, 04 ICPR workshop evaluation data. In: Pointing 04 workshop (ICPR), Cambridge, UK, August 2004
26. Cascia ML, Sclaroff S, Athitsos V (2000) Fast, reliable head tracking under varying illumination: an approach based on robust registration of texture-mapped 3d models. IEEE Trans Pattern Anal Mach Intell 22:322–336

27. Cootes T, Walker K, Taylor C (2000) View-based active appearance models. In: Fourth IEEE international conference on automatic face and gesture recognition, pp 227–232
28. Sung J, Kanade T, Kim D (2008) Pose robust face tracking by combining active appearance models and cylinder head models. Int J Comput Vis 80(2):260–274
29. Morency L-P, Whitehill J, Movellan J (2008) Generalized adaptive view-based appearance model: integrated framework for monocular head pose estimation. In: Proceedings IEEE international conference on face and gesture recognition
30. Whitehill J, Movellan JR (2008) A discriminative approach to frame-by-frame head pose tracking. In: Proceedings IEEE international conference on face and gesture recognition, pp 1–7
31. Asteriadis S, Nikolaidis N, Pitas I, Pardàs M (2007) Detection of facial characteristics based on edge information. In: Second international conference on computer vision theory and applications (VISAPP), vol 2, Barcelona, Spain, pp 247–252
32. Asteriadis S, Tzouveli P, Karpouzis K, Kollias S (2007) Non-verbal feedback on user interest based on gaze direction and head pose. In: 2nd international workshop on semantic media adaptation and personalization (SMAP), London, United Kingdom, December, 2007
33. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision (IJCAI). In: Proceedings of the 7th international joint conference on artificial intelligence (IJCAI '81), pp 674–679, April 1981
34. Peters C (2006) A perceptually-based theory of mind model for agent interaction initiation. In: International journal of humanoid robotics (IJHR), special issue: achieving human-like qualities in interactive virtual and physical humanoids. World Scientific, Singapore, pp 321–340
35. Emery NJ, Perrett DI (1994) Understanding the intentions of others from visual signals: neurophysiological evidence. Curr Psychol Cogn 13:683–694
36. Sidner CL, Kidd CD, Lee C, Lesh N (2004) Where to look: a study of human-robot interaction. In: Intelligent user interfaces conference. ACM, New York, pp 78–84
37. Conte R, Castelfranchi C (1995) Cognitive and social action. University College London, London
38. Poggi I (2007) Mind, hands, face and body. A goal and belief view of multimodal communication. Weidler, Berlin.
39. Congalton RG, Green K (1999) Assessing the accuracy of remotely sensed data: principles and practices. Lewis Publishers, Boca Raton