Towards a Real-time Gaze-based Shared Attention for a Virtual Agent

Christopher Peters
Coventry University
Coventry CV1
United Kingdom
christopher.peters@coventry.ac.uk

Stylianos Asteriadis
National Technical University
Athens
Greece
stiast@image.ece.ntua.gr

Kostas Karpouzis
National Technical University
Athens
Greece
kkarpou@image.ece.ntua.gr

Etienne de Sevin
INRIA Paris-Rocquencourt
78153 Le Chesnay Codex
France
etienne.de_sevin@inria.fr

ABSTRACT

This paper investigates work towards a real-time user interface for testing shared-attention behaviours with an embodied conversational agent. In two-party conversations, shared attention, and related aspects such as interest and engagement, are critical factors in gaining feedback from the other party and allowing an awareness of the general state of the interaction. Taking input from a single standard web-camera, our preliminary system is capable of processing the users eye and head directions in real-time. We are using this detection capability to inform the interaction behaviours of the agent and enable it to engage in simple shared attention behaviours with the user and objects within the scene in order to study in more depth some critical factors underpinning engagement.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Intelligent agents

General Terms

Algorithms, human factors, theory

Keywords

Shared attention, interest, engagement, embodied conversational agent, socially intelligent behaviour

1. INTRODUCTION

A fundamental capability for any agent whose purpose is to interact within an environment, real or virtual, is to be able to sense, interpret and reason about signals of relevance emanating from that environment and adapt its behaviour appropriately to those assessed as being significant. This is no less true for an embodied conversational agent, or *ECA*, whose task is to interact with a human user. The goal of creating such a capability for an agent may be viewed as an attempt to endow the agent with the most basic form of *awareness* (or modalities of awareness) regarding its environment and inhabitants, at least in the sense that there exists somewhere an appraisal of the relevant and significant signals, which may effect processing to varying degrees, be it emotional, planning or otherwise.

Here we describe a work-in-progress towards creating a shared attention scenario between a single user and an ECA. It represents an integration of previous work spanning a number of different domains [1], [11], [17]. Shared attention is a pivotal skill in early social understanding [2] [15]. We focus on shared attention as it relates to interest in the other interactant, the scene, and particularly, the interaction itself, as signalled by gaze motions and gaze following. We aim to investigate shared attention in the context of engagement between user and agent. The agent attempts to track the state of the interaction, based on its interest and the theorised interest of the user, to decide, for example, to halt ongoing behaviour if the user is not interested, or explain an object in detail if the user is paying a lot of attention to it. Our aim is to outline the most important interconnected components, capabilities and metrics that will form the basis of the system to be used for a set of experiments investigating shared attention and engagement between a user and agent. Our model consists of an eye-gaze detector, coupled with a number of interpretation stages relating to user interest, allowing the agent to assess the state of the interaction and conduct shared-attention behaviours, either reactively by following the users gaze, or pro-actively by directing their gaze towards objects of interest (Section 3). We have implemented a prototype of the gaze interaction model in a shared attention scenario (Section 4) in order to test the preliminary system.

2. BACKGROUND

Two major methods for gaze detection have been extensively studied in the literature: *head pose estimation*, and *eye gaze estimation*. Various approaches have been adopted for retrieving features related to these methods from an image sequence. In head pose estimation, many of the approaches proposed in the literature require more than one camera, or extra equipment ([21],[9],[4],[10],[8]), making the final system either complex or intrusive. Furthermore, algorithmically, some methods require a set of facial features to be detected and tracked with very good accuracy ([6],[7]). These techniques are usually sensitive even to small displacements of the features and can cause the algorithm to fail. Other techniques take the facial area as input and compare it against training sets of facial images ([16],[19]). These methods suffer from the problem of alignment, especially in natural environments, where it is not usually easy to achieve good alignment between training and test images. In our work, non-intrusive conditions are possible, in order to allow the user to be as spontaneous as possible. Furthermore, the system does not need to be trained according to the user or background and although the system 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 3-D relative positions.

A number of researchers have considered eye gaze for HCI, either to communicate through a robot or computer with other humans or with virtual agents. Vertegaal et al. [20] considered the significance of gaze and eye contact in the design of GAZE-2, a video conferencing system that ensures parallax-free transmission of eye-contact during multiparty mediated conversation. In work using conversational agents, some approaches have cast the ECA in the primarily role of a listener, for example, as a SAL, or *sensitive artificial listener* [3], that provides feedback to a discourse conducted primarily by the user. In a similar vein to the current work, attentive presentation agents [13] rely on the eye gaze of the user to infer attention and visual interest, based on an algorithm in [14], in order to alter their ongoing behaviour in real-time.

3. GAZE-BASED INTERACTION

The gaze detector (Section 3.1) employs facial feature analysis of images captured from a standard web-camera in order to determine the direction of the users gaze. This information allows the users gaze inside or outside the screen to be calculated, so that metrics relating to the users attention and interest can be applied to the scene (Section 3.2). Based on the interpreted metrics, an assessment of the state of the interaction is made in order to support shared-attention behaviour (Section 3.3).

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 [1], and follows a variant of this method for Head Pose and Eye Gaze estimation. In particular, head pose is estimated by calculating the displacement of the point in the middle of the inter-ocular line, with regards to its position at a frame where the user faces the avatar frontally. This displacement produces the head pose vector which is a good index of where the user's head is turned towards (see Figure 1). Normalisation with the inter-ocular distance (in pixels) guarantees that the head pose vector is scale-independent. In order to distinguish between displacements caused by head rotations and by translations, the fraction between the inter-ocular distance and the vertical distance between the eyes and the mouth is always monitored and, if it is found to be always kept within certain limits, no rotation is decided. Also, occlusion and rapid rotations make it difficult for the visible features to be tracked. In such cases, when the user comes back to his frontal position, the vector corresponding to pose estimation reduces in length and stays fixed for as long as the user is looking at the monitor. In these cases, the algorithm can re-initialise by re-detecting the face and the facial features. For eye gaze estimation, relative displacements of the iris center with regards to the points around the eye give a good indication of the directionality of the eyes with regards to a frame where the user faces the agent frontally. These displacements correspond to the eye gaze vectors (see Figure 2). Again, these displacements are normalised by the inter-ocular distance and, thus, are scale independent. For relative distance changes of the user position with regards to the web-camera, the inter-ocular distance (in pixels) is calculated at each frame, and compared to the inter-ocular distance calculated at the phase when the user is looking frontally. The computational complexity of the method permits real time applications and requires only a simple web-camera to operate. Tracking the features takes 13msec per frame on average for a resolution 288×352 pixels of the input video, using a Pentium 4 CPU, running at 2.80GHz, while re-initialisations, whenever occurring, require 330ms.

3.1.1 Conversion to Scene Coordinates

The raw information about the users head and eye directions are converted into 2D coordinates allowing them to reference the virtual scene. There are two basic possibilities: the user is either looking inside or outside the screen area containing the 3D scene. In this work, we are not only concerned about where the users gaze lands inside of the screen area, but also where it lands outside, as it can be an indicator of lack of attention. In order to facilitate both of these possibilities, at the beginning of each interaction scenario with the user, a calibration process is invoked in order to find the corresponding maximum and minimum extents of the screen boundary in terms of raw head and eye direction values. After conversion, the final coordinates data structure consists of a flag signalling gaze inside or outside the screen, accompanied by a 2D coordinate. If the flag indicates gaze within the scene boundary, the 2D coordinates correspond to the (x,y) screen position with respect to these boundaries. Otherwise, the 2D coordinate signals the screen boundary edge or corner that gaze fell outside.

3.2 Attention and Interest

Initially, the screen coordinates obtained from the users gaze direction are used to compute the nearest virtual object falling under that gaze position (see Section 3.2.2). The attention that the user may have in particular objects, in the scene as a whole and/or in the interaction, is an important issue in this work. While the gaze module (Section 3.1) detects the users gaze direction (i.e. the eye/head direction

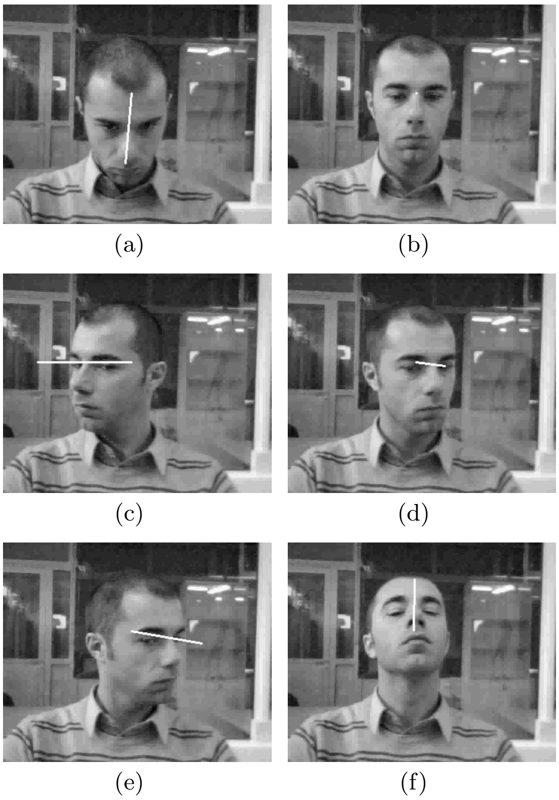


Figure 1: Examples of Head Pose Vectors

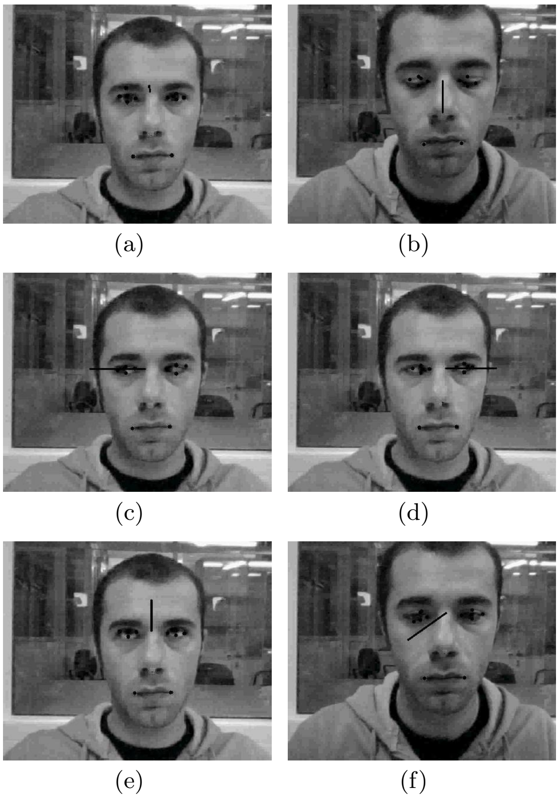


Figure 2: Examples of Eye Gaze Vectors

of the user mapped into scene coordinates), this information must be converted into a knowledge of *what* they are looking at for use in interaction understanding. Temporal integration is an important concept here: if at one time instant, the system detects that the user is looking outside of the screen, this does not necessarily imply that they are uninterested in what is happening - they may simply be glancing momentarily towards a distraction in their environment or staring upwards to think about what the ECA is saying. To aid in assessing the detection of the users attentive behaviours over different time-frames, we define a number of metrics.

3.2.1 Directness and Level of Attention

We use a *directedness* metric to refer to the momentary orientating of the users sensory body parts (in this case, the eyes and head) with respect to an area on the screen and record the ratio between them. For example, the user may have their head rotated directly towards an object in order to look at it (e.g. Figure 1(b)) - this would be considered a high degree of directedness. On the other hand, the user may have their head turned to the side, but be looking back at the object with their eyes (e.g. Figure 1(c),(d),(e)) - this would be considered a lower degree of directedness. Since metrics based on user gaze configuration during a single frame are highly unreliable indicators of attention, we define a *level of attention* metric. It refers to a clustering of a users focus of interest in a single region over multiple frames.

3.2.2 Virtual Attention Objects

In order to simplify the analysis of what is being looked at in the scene, in a methodology similar to [13], we define *virtual attention objects*, or *VAO's*. A single VAO is attached to each object for which we wish to accumulate attention information - for example, 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. If the screen-coordinate of the gaze fixation is located inside a VAO, then its corresponding level of attention is updated to reflect this. Thus, as the users 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 users gaze around the scene and attention to specific objects.

3.2.3 Level of Interest

Over a larger time-frame, and for a specific set of VAOs, the *users level of interest*, or *LIU*, for that set can be computed based on the stored attention levels for each member of the specified set (see Table 1 for possible member types). By defining a set of VAOs that contains only those objects currently relevant to the interaction, such as a recently pointed to or discussed object, and comparing the attention paid to these objects with the rest of the scene, we can obtain a measurement of how interested or engaged the user is with respect to the interaction itself, rather than superficial scene details (see Section 3.3).

3.2.4 Parameters for Agent Behaviour Generation

In addition to the metrics used for interpretation of the users attention and interest, a level of interest is defined for the

Case	User Gaze Direction	General Level
(a)	Outside of screen area	Low
(b)	Background	Medium
(c)	Scene object	Medium/High
(d)	Agent	High

Table 1: Possibilities for user gaze direction and general associated interest levels. Scene objects are ranked more highly when they are *relevant* to the interaction.

agent, *LIA*. Unlike the *LIU*, which is based on the users detected behaviour, the LIA helps define how the agent should generate its behaviour. The *LIA* is determined by the agents motivation in interacting (a preset selection in the current scenario: see Section 4). The agent should attempt to generate behaviours in order to convey the desired level of interest to the user, for example, if interested, maintaining gaze with the user, or being active in informing them about an object they are looking at. Other behaviours relate to the cues and cue strengths when referring to an object. For example, the agent may add vocal emphasis when mentioning the object, may make a pointing gesture towards it and momentarily direct its gaze towards it. In future, we hope to implement such behaviours, and do so in a manner that is graded according to the level of interest. By linking the level of interest of the ECA to dynamic internal variables related to motivations, emotions and personality traits, we hope to construct an ECA with more believable behaviour, for example, that does not put as much behavioural emphasis on objects when they are not so important to the discourse and in turn does not rate the detected users shared attention behaviours towards them as highly.

3.3 Shared Attention and Engagement

Although we are attempting to construct a shared attention model, we view engagement as being a complementary related topic underlying this aim. For example, 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, the case where a mother establishes relatively prolonged mutual eye contact with her infant before providing the gaze cue to the cuddly toy to be attended to.

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” [18] 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” [5] [12]. We regard engagement as being facilitated by both attentive and emotional processes between the interactants, and regard the presence of some degree of engagement as being a necessary prerequisite for shared attention behaviours to take place. We also view it as a process that may possibly be modulated by the presence of ensuing shared attention behaviour. For example, engagement may be diminished due to not engaging in shared attention behaviour.

Case	Quality	Description
(a)	High	Interested in *interaction* with agent: attends to relevant objects
(b)	Medium	Superficial interest in agent/objects not relevant to the interaction
(c)	Low	Uninterested in the scene

Table 2: Different levels of engagement quality

3.3.1 Engagement Metrics and Dynamics

The *level of engagement* is related to how much the user has been looking at the *relevant* objects in the scene at the *appropriate times.* When we discuss relevance here, we refer to gaze synchrony with the contents of the ongoing discussion or interaction. For example, if the agent makes reference to an object, by pointing to it or mentioning it in conversation, a high level of engagement would be signalled by the user attending to the referenced object or to the agent shortly afterwards. On the other hand, it is less serious if the user is not paying attention when the agent is not doing anything important. Therefore, the behaviour of the user is not considered in isolation, but rather in the context of what the ECA is doing. This is a key factor for us in making a distinction between engagement and attention and interest.

The *quality of engagement* relates to an assessment of the type of engagement that the user has entered into with the scene and the agent. This is an important metric, as not all attention paid to the scene necessarily indicates engagement in the interaction with the agent. For example, the user may be looking a lot at the graphical properties of the agent (or following some other self-defined agenda), but they may not be attending to where the agent is looking or what it is saying. We define three quality levels (see Table 2). The purpose of the engagement quality metric is to differentiate between when the user is superficially paying attention to aspects of the scene and when they are actually engaged in interaction with the agent, by synchronising attention with relevant aspects of the interaction. Thus, we regard the highest quality of engagement to involve the user attending to the agent and parts of the scene that are *relevant* to the interaction (i.e. what is being looked at and what is being said).

An important aspect of engagement for us is the continuous evolution of the interaction between agent and user. The *engagement state space* (see Figure 3) represents the history and current state of the engagement from the agents perceptive. It is a plot of the LIA and LIU for the duration of the interaction so far. Figure 3 also depicts a typical example of the continuous evolution of interaction in the engagement state space. In this example, at the start of interaction, both the ECA and the user are interested in engaging (1). They begin to interact, with the detected gaze behaviours of the user signalling they are interested (2). The motivation of the ECA here is to show interest in interacting, so it maintains the interaction by providing appropriate interaction maintenance behaviours, such as paying attention to the user and providing feedback when listening (3). After a while, the user becomes less interested; the ECA interprets a drop in interest, theorising that they are moving towards closing the interaction (4) and to a state where no interaction exists (5).

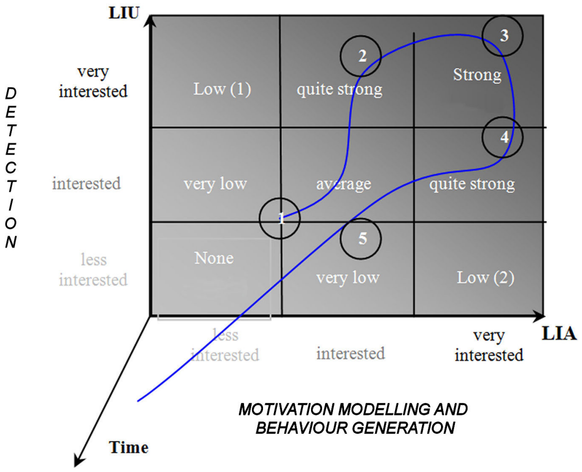


Figure 3: The engagement state space, according to the interest level of the user (LIU) and the agents own level of interest (LIA). At any time, the developing state of engagement (an example illustrated here by the continuous line) can be categorised according to its zone membership.

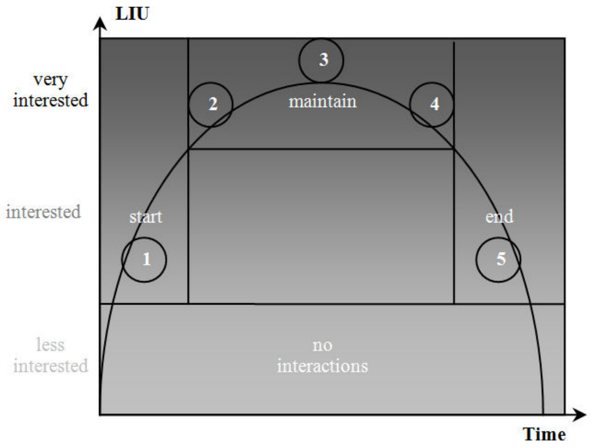


Figure 4: The evolution of an interaction between an ECA and a user according to the estimation of the interest level of the user (LIU) over the time. The interest level of the ECA (LIA) is preset so that it acts as if it is very interested in interaction with the user.

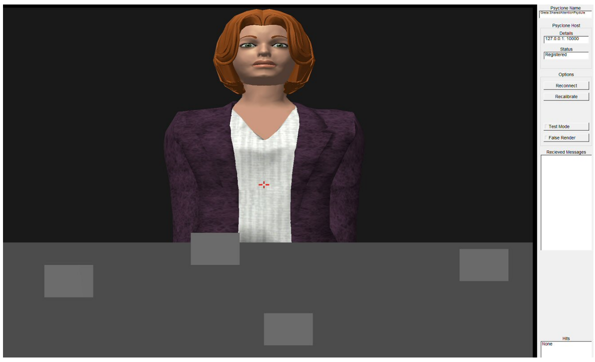
4. SHARED-ATTENTION SCENARIO

In practice, our system is comprised of two key modules: a gaze detector module and a shared attention player module. These modules communicate via a *Psyclone* cnnnection - a blackboard system for use in creating large, multi-modal A.I. systems. The gaze detector module comprises the capabilities described in Section 3.1, employing facial feature analysis of images captured from a standard web-camera in order to determine the direction of the users gaze. The shared attention player contains the graphical representation of the agent and the scene, and receives updates of the users gaze from the gaze module. It implements the agents interpretative capacities (metrics described in Sections 3.2 and 3.3) and behaviour generation (see Section 3.2.4). Figure 5 is a screenshot of the shared attention scenario between the agent and a user. The agent stands behind a table containing a number of simple objects, represented in this case by the gray rectangles. As the scenario progresses, depending on the set-up, the agent may refer to objects either by (1) making a deictic (i.e. pointing) gesture, (2) looking at the object momentarily, (3) making a short predefined verbal description of the object, or (4) a combination of the former. We generally expect interactions to assume a wave-like form as the conversation progresses, with the interaction progressing through the stages of opening, maintenance and closing (see Figure 4). In our primary scenario, the agent assumes a *proactive mode* of operation, and is tasked with reading out a predefined story that relates to some of the object shown on the table graphically. The user is expected to fulfill the role of passive listener in this case, and their interest in the scenario is determined by their gaze towards the agent and objects described in the story (see Section 3.3). If the user interest in the story is detected to be falling off, the agent first of all becomes more explicit in how it cues the relevant story objects by conducting more expression pointing gestures and gazes motions. If user interest falls below threshold, the agent will interrupt the story and go into a *reactive mode* of operation. In this mode, it will describe objects that the user looks at, or not interact at all if the user is not paying attention to the scene. If the user again pays attention to the agent, it will return to proactive mode and continue telling the story.

We intend to use this basic scenario to study user impressions of their interaction with the ECA under different circumstances. For example, in one trial, the agent will not conduct any shared attention behaviours with the user during story telling, in another we may provide limited cues, and in a final trial we may include very explicit cueing. It will be interesting to assess user reporting of their interaction experience, in terms of the level and quality of engagement, during these varying trials.

5. CONCLUSIONS AND FUTURE WORK

We have presented preliminary work outlining a model of shared attention that encompasses low-level detection from a user and seeks to integrate information into high-level metrics that the agent can use to determine the state of engagement with the user. We hope these metrics can provide the basis for an agent that can make improved inferences relating to the interaction goal of the user during shared attention scenarios. We are constructing a prototype scenario to test the metrics. An important issue for us is to improve the naturalness of the interaction - we use an image processing approach to negate the need for head mounted eye-tracking devices. As future work, in terms of gaze detection, we are continuing to improve the robustness of the detector, particularly for eye direction detection. As expected, careful attention must be paid to lighting conditions in order to obtain reliable measurements. We intend integrating detection of facial expressions into the system: the metrics described here are based heavily on attention aspects relating to gaze, and a facial expression detector would allow an affective dimension to also be considered, for example, for the detection of empathic and imitation behaviours. This would be



Psyclone Name
Greta.SharedAttentionPsydule
Psyclone Host
Details
127.0.0.1: 10000
Status
Registered
Recieved Messages
Hits
None

Figure 5: The shared attention scenario: a user is presented with a depiction of the agent and a number of objects. As the user moves his head, the gaze path and current fixations (signalled by red crosshair) are detected and update interest metrics related to objects.

complemented by a model for recognising users in order to support long-term interactions. We also plan to integrate a motivational system into the agent to allow it to manage its level of interest and show appropriate behaviours dynamically. Motivations would arise from internal variables, evolve continuously during the simulation and be influenced by parameters relating to perceptions, personality and emotions, which may make the ECA more interesting and engaging.

6. ACKNOWLEDGMENTS

This work has been funded by the STREP SEMAINE project IST-211486 (http://www.semaine-project.eu) and the IP-CALLAS project IST-034800 (http://www.callas-newmedia.eu).

7. REFERENCES

[1] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias. Non-verbal feedback on user interest based on gaze direction and head pose. 2nd International Workshop on Semantic Media Adaptation and Personalization (SMAP 2007), London, United Kingdom, December, 2007.

[2] S. Baron-Cohen. How to build a baby that can read minds: cognitive mechanisms in mind reading. *Cahiers de Psychologie Cognitive*, 13:513–552, 1994.

[3] E. Bevacqua, M. Mancini, and C. Pelachaud. A listening agent exhibiting variable behaviour. In *Intelligent Virtual Agent (IVA'08)*, Tokyo, 2008.

[4] D. Beymer and M. Flickner. Eye gaze tracking using an active stereo head. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 451–458, Madison, WI, USA, 2003. IEEE Computer Society.

[5] R. Conte and C. Castelfranchi. *Cognitive and Social Action*. University College London, 1995.

[6] A. Gee and R. Cipolla. Non-intrusive gaze tracking for human-computer interaction. In *Int Conference on Mechatronics and Machine Vision in Pract.*, pages 112–117, Toowoomba, Australia, 1994.

[7] N. Gourier, D. Hall, and J. Crowley. Estimating face orientation from robust detection of salient facial features. In *International Workshop on Visual Observation of Deictic Gestures (ICPR)*, Cambridge, UK, 2004.

[8] C. Hennessey, B. Noureddin, and P. D. Lawrence. A single camera eye-gaze tracking system with free head motion. In *Proceedings of the Eye Tracking Research & Application Symposium, (ETRA)*, pages 87–94, San Diego, California, USA, 2006. ACM.

[9] Y. Mao, C. Y. Suen, C. Sun, and C. Feng. Pose estimation based on two images from different views. In *Eighth IEEE Workshop on Applications of Computer Vision (WACV)*, page 9, Washington, DC, USA, 2007. IEEE Computer Society.

[10] A. Meyer, M. Böhme, T. Martinetz, and E. Barth. A single-camera remote eye tracker. In *Lecture Notes in Artificial Intelligence*, pages 208–211. Springer, 2006.

[11] C. Peters. 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*, pages 321–340, 2006.

[12] I. Poggi. *Mind, hands, face and body. A goal and belief view of multimodal communication.* Weidler, Berlin, 2007.

[13] H. Prendinger, T. Eichner, E. André, and M. Ishizuka. Gaze-based infotainment agents. *Advances in Computer Entertainment Technology*, pages 87–90, 2007.

[14] P. Qvarfordt and S. Zhai. Conversing with the user based on eye-gaze patterns. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–230, New York, NY, USA, 2005. ACM.

[15] B. Scassellati. Mechanisms of shared attention for a humanoid robot. In *Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium, AAAI*, 1996.

[16] K. Seo, I. Cohen, S. You, and U. Neumann. Face pose estimation system by combining hybrid ica-svm learning and re-registration. In *5th Asian Conference on Computer Vision*, Jeju, Korea, 2004.

[17] E. d. Sevin. *An Action Selection Architecture for Autonomous Virtual Humans in Persistent Worlds.* PhD thesis, VRLab EPFL, 2006.

[18] C. Sidner, C. Kidd, C. Lee, and N. Lesh. Where to look: A study of human-robot interaction. In *Intelligent User Interfaces Conference*, pages 78–84. ACM Press, 2004.

[19] R. Stiefelhagen. Estimating Head Pose with Neural Networks - Results on the Pointing04 ICPR Workshop Evaluation Data. In *Pointing 04 Workshop (ICPR)*, Cambridge, UK, August 2004.

[20] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung. Gaze-2: conveying eye contact in group video conferencing using eye-controlled camera direction. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 521–528, 2003.

[21] M. Voit, K. Nickel, and R. Stiefelhagen. Multi-view head pose estimation using neural networks. In *Second Canadian Conference on Computer and Robot Vision (CRV)*, pages 347–352, Victoria, BC, Canada, 2005. IEEE Computer Society.