

SPECIFICATION FOR MODEL OF AWARENESS

Conveying Affectiveness in Leading-edge
Living Adaptive Systems

CALLAS

Project IST-34800

Deliverable D1.3.2 WP1.3

Programme Name: IST
Project Number: 34800
Project Title:..... CALLAS
Partners:..... Coordinator: ENG (IT)
 Contractors:
 VTT Electronics, BBC, Metaware, Studio
 Azzurro, XIM, Digital Video, Humanware,
 Nexture, University of Augsburg, ICCS/NTUA,
 University of Mons, University of Teesside,
 Helsinki University of Technology, Paris 8,
 Scuola Normale Superiore di Pisa, University of
 Reading, Fondazione Teatro Massimo,
 HITLaboratory New Zealand

Document Number: D1.3.2
Work-Package: WP1.3
Deliverable Type: Report
Contractual Date of Delivery: Month 15
Actual Date of Delivery:
Title of Document: Specification for Model of Awareness
Author(s): Christopher Peters (PAR8)

Approval of this report Executive Committee

Summary of this report: Description of components that form a design
 for a model of agent awareness

History:

Keyword List: Visual attention, shared attention, social
 signals, virtual agents

Availability This report is public

Table of Contents

EXECUTIVE SUMMARY	1
1. INTRODUCTION	2
2. BACKGROUND	4
3. TECHNOLOGICAL FOUNDATIONS	5
3.1 SYNTHETIC VISION	6
3.2 VISUAL PERCEPTION PIPELINE	6
3.3 VISUAL ATTENTION	7
3.4 INTERPRETATION OF VISUAL SIGNALS	8
4. NEW CAPABILITIES	10
4.1 INPUT FROM THE REAL ENVIRONMENT	10
4.1.1 <i>Technologies</i>	10
4.1.2 <i>Posture</i>	10
4.1.3 <i>Eye Gaze and Blinking</i>	11
4.2 CONTEXTUAL ASPECTS	11
4.3 ATTENTION METRICS	13
4.3.1 <i>Mutual Gaze</i>	13
4.3.2 <i>Level of Attention</i>	13
4.3.3 <i>Level of Interest</i>	14
4.4 SHARED ATTENTION MODULE	14
5. OUTLOOK	15
6. REFERENCES	16

Table of Figures and Tables

Figure 3-1: High-level illustration of the framework.....	5
Figure 3-2: Simplified schematic of our version of Theory of Mind.....	9
Figure 4-1 HFSM's representing the context of the interaction from agent perspective.....	11
Table 4-1 Algorithm 1	13

Executive Summary

In this document we present a specification for a model of agent awareness. Essentially, this task is working towards the creation of attentive computer agents that can have various modes of awareness regarding the environment in which they are placed, be it a real location, e.g. an agent placed occupying a computer screen who can interact with passers-by, or a virtual one, e.g. a virtual agent located inside a virtual environment interacting with other virtual agents or with the user through an avatar.

After a description of introductory and background material (see Sections 1 and 2), we present the technological infrastructure, in Section 3, required as the basic foundation into which new capabilities can be assembled. The design of these new capabilities is described in Section 4.

The purpose of these new capabilities is to enhance the basic framework and improve some social awareness capabilities of agents, particularly by allowing them to sense through a visual modality.

1. Introduction

While *awareness* is a broad concept with many possible definitions, not all of which are particularly suitable to computer agents, we use the term here to refer to the idea of endowing agents with the means to perceive, interpret and remember aspects of their environments, as sensed through a visual modality.

A focus of this work is on the detection of social signals, while still accounting, to some degree, for events in the surrounding environment. For example, when we engage in dialogue with another person, although we may try to give them the impression that they alone possess our undivided attention, by orienting our senses and being responsive, this is usually never the case: when we gaze away briefly from the other during social discourse, we may not only be thinking about what they have said or planning what to say next, we may also, unsuspectingly, be appraising other aspects of the environment, or studying in more detail a recent topic of the conversation. Furthermore, given the characteristics of the human eye, we may also be involved in mutual gaze with the other, but paying attention to something in the periphery.

Achieving this level of perception and behaviour in virtual agents encompasses the detection, analysis and storage of socially *relevant* signals from an interactant. It also requires mediation with the continued background processing of signals of potential importance from the environment, while attempting to maintain the interaction, if that is the goal of the agent. This is, of course, an extremely difficult task for a computer agent: it must essentially balance its perceptual input (through sensory orienting) with the sending of social signals. We do not hope to propose a solution to this problem, but rather to describe how progress can be made towards this goal by focusing on a number of additions and integrations to an ongoing computational framework created around the theoretical model proposed in [1].

This model focuses on the use of attention signals from the social entities in the environment for making inferences about their mental states. In our case, we are interested in using it as a general outline for a computational model to make simple inferences regarding the quality of the social engagement and for making judgments about the degree to which one thinks the other is involved in an engagement.

Judging the level of interest that the other has in the conversation is a basic and important issue for conversational agents and other human computer interfaces: one should not continue to talk to the other or expect them to listen if it is clear the other has directed their attention elsewhere. While this may seem obvious, linking external signals in order to infer internal processes is non-trivial problem: many contemporary systems are not sensitive to these types of human signals in the first place and therefore cannot account for them during conversational scenarios. Mutual eye contact, associated with increased psychological arousal, establishes a special connection between speaker and listener where each is the object of the others attention.

As pointed out in [25], the more people share looking behaviours, the more they are involved and coordinate in the conversation. However, it may be more difficult than it seems to infer directly from eye-gaze whether to the other is paying attention to the conversation. The absence of mutual gaze at a certain instant, for example, cannot be a good indicator of the absence or presence of interest in the conversation. First of all, it is usual under normal circumstances for the eyes to disconnect often from those of the other so that continual contact is not maintained, something that can cause social discomfort.

Furthermore, in many cases, the fact that the other is looking away can actually be a signal of their interest, for example, they may look upwards if they are thinking about what is being said [22], or may look at an object being referred to as part of the conversation. In these cases, the absence of a break or even total reorienting of gaze can actually signal that the

other is not really paying attention to what is being said, but is merely pretending to. On the other hand, it is possible that even though the other is looking continuously at the speaker, they are actually giving a 'blank stare', that is, they are not really paying any attention to what is being said. In order to investigate these concepts in more detail, we consider engagement and level of interest of the other as important central metrics in allowing the construction of systems that are socially aware, at least to a minimal degree, when interacting with others.

2. Background

The work presented here covers a range of domains, including visual attention modelling, theory of mind and engagement. A number of models of visual attention and perception have been proposed for virtual agents.

They are based either on bottom-up ([8], [20]) and top-down approaches ([6]), on movement observation and cognitive modelling [3]. Peters et al [19] propose a model of gaze behaviour for the synthetic agent based on the level of interest of the user.

In the field of social robotics, [24] is constructing a humanoid robot as a test bed for the evaluation of models of human social development. The robot, Cog, has been endowed with social abilities using models of social development in both normal and autistic children.

Scassellati has proposed a merger of two models of theory of mind, including Baron-Cohen's model. The model first considers the movement of environmental stimuli in terms of the physical laws in order to distinguish between *animate* and *inanimate* objects. Self-animating stimuli are then further processed by Baron-Cohen's model, which acts as a social perception.

Sidner et al. [25] have studied rules of looking behaviour in order to allow robots to maintain engagement with humans in a collaborative environment. They found that users engaged in mutual gaze with the robots, directed their gaze to them during turns in conversation and responded to changes in head and gaze direction by changing their own gaze or head direction.

Unlike robotics systems, the approach we have been pursuing so far has been easier to implement since we have been dealing solely with a virtual environment and virtual sensors: using the synthetic vision module, difficult and time-consuming issues such as segmentation and recognition are avoided. These modules are discussed in the next Section.

Adapting the modules and framework for a specific range of socially relevant real-world inputs is an important goal in this work, as described in Section 4.

3. Technological Foundations

In this Section, key areas of previous and ongoing research are described as they form a necessary foundation for the new development designs detailed in Section 4. These foundations are necessary for allowing basic input both from the environment, for processing in a homogenous manner and for allowing interpretation to take place in a principled manner.

A high-level overview of the framework, shown in Figure 3-1, illustrates the stages from input to behaviour generation.

We next detail some of the components of this framework that have been the focus of previous research:

- *synthetic vision* for taking input from the virtual environment (Section 3.1)
- a *visual perception pipeline* for providing a principled approach to successive storing and filtering of input (Section 3.2)
- *visual attention* for attending to certain parts of input (Section 3.3), and
- computational methods for interpreting conversation intention based on visual signals (Section 3.4).

Work within CALLAS seeks to elaborate additional parts of this framework as described in Section 4, particularly input from the real environment and a shared attention capability based on attention metrics.

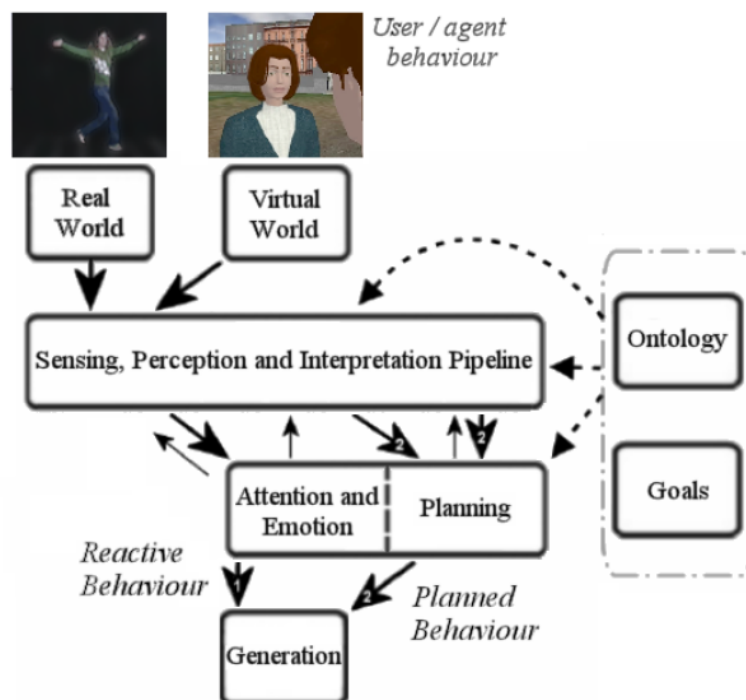


Figure 3-1: High-level illustration of the framework

With reference to Figure 3-1, input is taken from the real or virtual environment, processed in a homogenous manner through the perceptual pipeline where it is progressively filtered and interpreted, before being used to generate either reactive or planned behaviours. A key contribution within the CALLAS is the integration of real-world input with virtual, and to add an extra layer of interpretation relating to level of interest based on eye movement for exploring shared attention behaviours.

3.1 Synthetic Vision

The synthetic vision module processes stimuli from a *virtual* environment in a snapshot manner by means of an orientable synthetic vision sensor that is locked to the gaze direction of the agent.

This sensor renders the scene from the viewpoint of the agent in two modes, allowing for two different basic types of scene representation to exist: spatial and object.

The first of these, a full-coloured rendering, contains the final colour of each element corresponding to the agent's visual field, including the contributions of textures, lighting, special effects and so on.

The second, false-coloured rendering, consist of objects in the visual field rendered according to their uniquely preassigned colours.

By scanning this image, elements of the scene can be queried for their associated object ID in the scene database: this could be thought of as a fast method of scene segmentation. By combining the full-coloured and false-coloured representations, the agent has access to a view-dependant representation of the scene, which can be sensitive to both object and spatial information.

These two maps form the initial input into the visual perception pipeline, described in Section 3.2, which can modulate these inputs to allow attention allocation in an object or spatial manner, according to the specific requirements.

The synthetic vision module could be thought of as analogous to a highly simplified human eye and visual processing in the brain: the full-coloured rendering could be thought of as the variant retinal image that is spatial in nature, while the result of the false-coloured rendering could be thought of as links to invariant object representations and their properties.

3.2 Visual Perception Pipeline

A visual perception pipeline (see Figure 3-1) is an important step in providing a homogenous, robust design for input, filtering, storage and interpretation of input taken from the environment, regardless of whether that input has been derived from a real or a virtual environment.

The visual perception pipeline is a branched pipeline structure consisting of a number of different stages and partitions, some of which may be hierarchical in nature. Every stage in the early stages of the pipeline contains a number of maps, each one the result of a processing operation on either the input maps or a map from a previous stage in the pipeline.

These maps are referred to as Synthetic Perceptual Maps (SPM) and they have been introduced [16] as a robust method for allowing information obtained through synthetic senses to be represented and operated on using homogenous representations.

For the visual modality, synthetic perceptual maps are a virtual analogy of topographic retinotopic maps that represent the visual world as seen through the eyes of a viewer. They are rectangular, 2D grey-scale maps corresponding to the agent's field-of-view, where the

value of a location in the map represents the strength of some particular feature or resultant operation based on the corresponding spatial location. Bottom-up saliency maps [8] and task relevance maps [12], for example, be viewed as instantiations of SPM's in our model.

SPM's can be combined together to form master maps for driving behaviour, or data can be extracted for hierarchical representation, although this is still a work-in-progress. For example, a master attention map is created from a visual attention algorithm, as described next, for judging salient parts of the environment.

3.3 Visual Attention

A simplified categorisation of visual attention, and one of great utility for computational modelling, is the distinction between top-down and bottom-up processing. This categorisation distinguishes between

- (a) bottom-up resource allocation in the brain solely due to basic characteristics of the sensory input and
- (b) top-down resource allocation wilfully allocated due to, among other factors, the current goals of the entity.

Top-down systems deal with the allocation of attention based on the goals of the entity, often related to object properties when modelled for character control [3][6].

On the other hand, bottom-up systems base the allocation of attention on the contrast between different low-level image features. Bottom-up models are also important controllers for the animation of characters interacting with humans [9][21] and within virtual environments [20][4].

Bottom-up attention is behaviourally significant, as it constitutes a fast, powerful alerting mechanism that allows primates to instantly become aware of unexpected predators or dangers. In terms of characters, this type of attention is useful for generating spontaneous looking behaviours and for interrupting task-level attention with potentially important events.

The bottom-up model of attention that we use is based on a model by Itti [8], that traces its origins to a biologically plausible architecture proposed by Itti, Koch, Ullman and Nieber [10][11].

The model attempts to mimic the low-level, automatic mechanisms responsible for attracting our attention to the salient locations in our environment and closely follows the neuronal architecture of the earliest hierarchical levels of visual processing. It has been demonstrated to be effective for processing natural [8] and rendered [26] scenes.

As it is discussed in the following (see Section 4.1), this makes the visual attention module especially useful, as it allows for scenes from both the real and the virtual environment to be processed in a bottom-up manner.

The model itself takes as input an RGB image and successively calculates local contrast over multiple scales for intensity, orientation and colour features respectively.

Although these are the only features handled in the basic model, it is easily upgraded to also incorporate motion, depth and other important features.

Feature contrast is computed in a biologically plausible centre-surround fashion, so it is sensitive to the local relative spatial contrast rather than global amplitude in a given feature.

In practice, this means, for example, that a very bright area of the image will not be highlighted by the algorithm as being salient, whenever the rest of the image is also very bright. However, the algorithm will highlight a dark area of the image whenever it is surrounded by a bright area and vice-versa. Thus, there is sensitivity (at multiple different scales) to the surrounding context of the area that is being processed. At the end of the

process, the resulting saliency map encodes a ranking of spatial coordinates of the image according to their saliency value.

A number of strategies can then be employed to put the saliency map to use for generating behaviours (see for example [20]).

In the context of this work, the visual attention component described here is particularly useful as it works not only with input from a virtual input, but is also robust for operation with real-world input. In order to facilitate real-time operation, we have created a GPU version of the visual attention algorithm [17].

3.4 Interpretation of Visual Signals

As mentioned in Section 3.3, attention is a vital, if not fundamental, aspect of engagement.

Indeed, it is doubtful that one could be considered engaged to any great extent in the absence of the deployment of attention. There are many facets of attention that are of relevance to engagement.

Attention primarily acts as the control process for orienting the senses towards stimuli of relevance to the engagement, such as the speaker or an object of discussion, in order to allow enhanced perceptual processing to take place. In social terms, the volitional deployment of attention, manifested as overt behaviours such as gaze and eye contact, may also be used for signalling one's desires, such as to become or remain engaged [22]. Therefore, the perception and interpretation of the attentive behaviours of others is an important factor for managing agent engagements in a manner consistent with human social behaviour.

In relation to the interpretation of visual signals, there are many ways in which visual signals may be interpreted depending on the context, goals and so on.

In this work, we focus on the analysis of attentive behaviours of the other, in particular relating to the eyes.

In this respect, the theory of mind model proposed by Baron-Cohen [1] is suitable. It suggests that the ability to read the behaviour of others in terms of their mental states is advantageous for the survival and reproduction of an organism and that this may have strong links to the interpretation of another's gaze [2]. Baron-Cohen suggests that the brain contains a series of specialised modules that enable humans to attribute mental states to others (see Figure 2).

These modules are thought to be present and functioning in most humans by approximately four years of age. The modules are:

- Eye-direction Detector (EDD): the EDD is a social cognition module exclusively based on vision. It functions by detecting the presence of eyes or eye-like stimuli in the environment and computing the direction of gaze (e.g. directed or averted).
- Intentionality detector (ID): the ID module attributes the possibility of an object having goals and desires based on self-propulsion, i.e. notions of animacy and intention. One should not, for example, attribute volitional behaviour to a brick, even if it is moving in the environment.
- Theory of Mind Mechanism (ToMM): this module stores the attribution of mental states to the other agent and is based on the results of interactions between the other modules. It contains working theories that may not necessarily be correct, but are nonetheless vital for forming an internal representation of the possible motives behind the actions of other living entities.

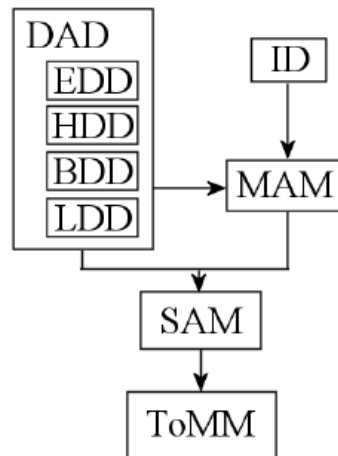


Figure 3-2: Simplified schematic of our version of Theory of Mind

Our version of Theory of Mind is based on description by Baron-Cohen and elaborated by Perrett and Emery. A key component that has not yet been investigated in the agent literature is the SAM, or Shared Attention Module.

Perrett and Emery [13] have advanced this work to propose further module classifications:

- Direction of attention detector (DAD): this is a more general form of the EDD above, that combines information from separate detectors that analyse not only gaze, but also body and direction of locomotion.
- Mutual attention mechanism (MAM): this is a special case of shared attention where the relationship is dyadic, involving mutual gaze and eye contact. In this situation, the goal of the participants attention is each other.

These models, provided by Baron-Cohen and Perrett and Emery, have been inspirational to us for creating a direction of attention and theory of mind model applicable to autonomous human-like agents that can initiate interactions with each other within virtual environments [14]. Furthermore, our studies involving human participants have shown that at a basic level, humans can perceive attentive behaviours from computer character based on the orientation of their body parts e.g. eyes, head, torso and locomotion direction [15].

The next step in our research, as outlined in the Section 4, is the addition of new capabilities in our design to allow agents to detect visual signals from the real world as well as the virtual (within the same general framework), and to adapt the level of interest metrics described here to work with this new input. This will allow more sophisticated high-levels inferences for driving interaction and supporting shared attention.

4. New Capabilities

Here we identify new additional capabilities necessary for creating more elaborate agents that are more sensitive and aware of their surroundings. Of course, these capabilities do not exist in isolation, but are designed to work beside or be integrated within the foundations already described in Section 3.

4.1 Input from the Real Environment

Thus far, our agent models have been contained solely within the virtual environment i.e. aimed towards agents that interact with each other. Important new work involves the input of data from the real environment in order to generate agent behaviour.

We will not be concentrating on creating these input modules ourselves, but rather, will be assembling currently available technologies from within CALLAS in order to 'plug' them into the early stages of the existing framework for real-world input (Figure 3-1).

4.1.1 Technologies

Our target platforms are laptop machines with small, mounted web-cameras.

OpenCV, an open source library of computer vision routines, is an example of a good solution we (and other partners) are using, as it can conduct low-level operations, and is also capable of higher-level feature detection, such as face detection. When driving agent behaviours, such seemingly trivial inputs are very important: for example, if a face is not detected in the field of view or the camera, the agent can assume the *No Interaction* state (see Section 4.2) and adopt idle behaviours. Such reasoning is not fool-proof (i.e. the user's face might actually be occluded or light conditions may not allow detection), but it provides a useful grounding for basic interaction decisions.

Other libraries are also available, that build on *OpenCV* to provide higher-level analysis and interpretation, often for different types of user behaviour. The following describes two important types of user behaviour for our design, for which practical real-time solutions are already available.

4.1.2 Posture

In computer games, a number of light-weight real-time approaches have been used to process the movements of user in order to transfer their side-to-side movements to an avatar in real-time during play, for example, in order to look around corners (see [23] for one such example using *OpenCV*).

We are investigating these approaches for detecting the posture of the user, particularly in terms of their movements towards or away from the computer screen, as such movements are likely an important consideration for calculating the user's attention metrics (see Section 4.3).

4.1.3 Eye Gaze and Blinking

A crucial input component in the system is the detection of eye and gaze direction from users.

An eye gaze direction detector essentially fulfils the role of the EDD and HDD in the models proposed by Baron-Cohen and Perrett and Emery, which are an integral part of our design (see Section 3.4). For this work, we will be using the detectors being provided by ICCS/NTUA partners in CALLAS from activities in WP1.2.

We will be endeavouring to use these to obtain fast estimations of the users gaze direction as an early, low-level input into the perception and interpretation pipeline (see Section 3.4).. From there, it will be interpreted into a number of attention-related metrics (see Section 4.3), over different time-scales, for helping to determine the agents behaviours and how it perceives the state of the interaction (see Section 4.2).

Blinking is a further input of high importance that needs to be considered, since mutual gaze does not necessarily infer mutual attention: one may be staring *blindly* at the other while thinking about something else. Blink detection will help to detect such situations and reduce uncertainty about the user's actual interest.

4.2 Contextual Aspects

Another important consideration is that of context. There are many different types of context, each of which may have important effects on how an interaction should be perceived and judged. For example, the actions of a user who looks away while apparently engaged in an interaction with the agent should be perceived differently or invoke different responses to one who looks away while not in an interaction.

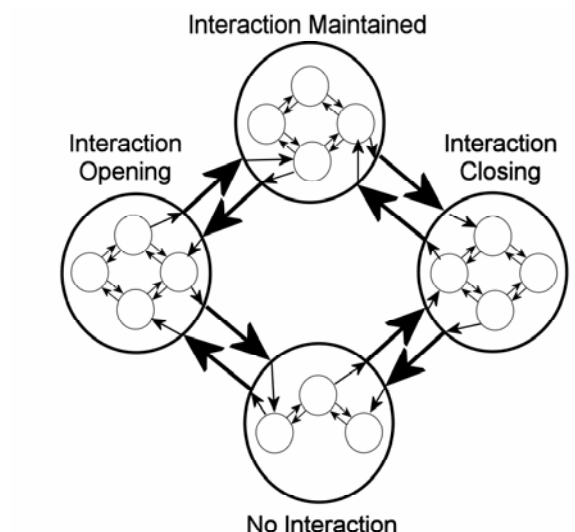


Figure 4-1 Hierarchical Finite State Machine (HFSM) representing the context of the interaction from agent perspective

In Figure 4-1 each state is associated with a subset of the behaviour repertoire and transitions are triggered according to a comparison between the user's attentive behaviours over a time-course with respect to thresholds for the respective state.

In our design, the agent keeps a track of the context in terms of where in the interaction it deems itself to be i.e. in high-level terms, whether it thinks the interaction is just beginning, is being maintained or is ending (or if it is not involved in an interaction at all). This is of key importance for helping to quickly interpret users' actions and determine thresholds for behavioural responses.

We will be investigating the modelling of this narrow type of contextual awareness using hierarchical finite state machine (HFSM). Each state in the HFSM is representative of an interactive context that the agent can be involved in with the user. The current interactive context that the agent thinks it is in defines the way in which it will treat incoming signals and react to the user's behaviour. It is important to note that the HFSM does not represent the *actual* context of the interaction with the user, but rather its *theory* about what the current context of the interaction is.

The hierarchical nature of HFSM allows multiple refinements of different states: for example, although the *interaction maintained* state would represent the situation where the user has been looking at the agent regularly, during special situations, such as those involving shared attention, they should still be judged to be in the *interaction maintained* state (even though the user might be directing their gaze elsewhere, towards an object). Both of these situations and transitions are accounted for within the same high-level *interaction maintained* node.

State transitions in the HFSM are determined by comparison of state values with a number of attention metrics, themselves derived from the visual input, which are described next (see Table 4-1 Algorithm 1 for example).

Input :

Sensory memory *STSS*

Short-term Memory *STM*

UPDATEVISUALPERCEPTION(*STSS*, *STM*)

```
//capture visual snapshot into sensory memory for image processing
VisualSnapshot(STSS)
```

```
//Detect faces in input
STSS.ExtractFacePercepts(facePerceptList)
```

```
for each face in facePerceptList do
```

```
    //calculate Direction of Attention
    eyeDir <- Direction(eye,me)
    headDir <- Direction(head,me)
```

```
    //calculate Attention level
    CalculateAttentionLevel(eyeDir, headDir)
```

```
    //detect Mutual Attention
    mutualGaze <- Direction(eyeDir,myEyeDir)
```

```
    //add information to short term Memory
    STM.AddEntry(facePercept, AL, mutualGaze)
```

CALCULATEINTERESTLEVEL(*agent*, *timeInterval*)

```
//get attention level over a time interval
IL <- STM.Integrate(agent, all AL's over timeInterval)
```

CHECKINTERACTION(*face*, *timeInterval*, *interestThreshold*)

```
//assume we are in the No Interaction state to begin with
if (facePerceptList > 0) and
```



```
(CalculateInterestLevel(face, timeInterval) > interestThreshold) and
(STM[face].AttentionProfile(timeInterval) == RISING) and
(mutualGaze == TRUE) then
    HFSM Transition to Interaction Opening State
```

Table 4-1 Algorithm 1

With reference to Table 4-1, updates perceived gaze information, calculates interest level over a time interval, and decides when to move from the *No Interaction* to *Interaction Opening* state.

4.3 Attention Metrics

Here, we seek to extend our previous level of interest metric suited to agents in virtual environments [19] to operate with the real-world input (see Section 4.1). This consists of a number of metrics relating to different time-scales over the course of the interaction.

Although a potentially huge amount of information will be arriving through the visual input of the agents, an important factor in the perceptual pipeline is that this information be stored at higher and higher levels of representation as it progresses through the perceptual pipeline, accounting for longer periods of time: for example, the fact that the user is looking at the agent at a single time instant does not provide as much useful information as knowing over what time period the user has been looking.

The purpose of the interpretation stages in the pipeline is to create a number of straightforward, high-level metrics that are representative of a vast amount of complicated incoming sensory information over differing time scales. Furthermore, it is not practical to store this amount of information over a large time scale, so interpretation also serves the purpose of compressing the information so it can be stored in memory.

Each of these proposed metrics represents progressively longer and longer time frames although these time frames are not specific and are only passed when making queries from memory. Furthermore, each metric may include the addition of further modalities of input: for example, mutual gaze does not include blinking, whereas level of attention does.

We use the terms *mutual gaze*, *level of attention* and *level of interest* to refer to the different time-scales. All relate to engagement at some level.

4.3.1 Mutual Gaze

Mutual Gaze is a variable set in sensory memory, when the agent deems the user to be looking at it directly in the eyes: this is different from *mutual attention*, as gaze does not necessarily infer attention.

Our studies have also shown that direct gaze in the absence of any other movements can result in uncertainty in human viewers, who may regard the agent either as paying a lot of attention to them, or merely staring at them blankly and paying no attention [15].

4.3.2 Level of Attention

Level of attention refers to the agent's interpretation of the amount of attention, in terms of looking behaviours, that the user has been paying to it over a variable time period, according to gaze at, gaze away and mutual gaze behaviours (in this case, *gaze at* refers to the user

staring at a part of the agent other than its eyes).

This time period is specified by a decision-making module (part of the HFSM described in the previous Section 4.2), which is making the query, and depends on what the level of attention is to be used for: for example, a decision may need to be made regarding how much attention has been paid by the user since the start of the last state transition in the HFSM, or merely starting at the beginning of the last behaviour that it has been made by the agent, in the case where the agent is trying to attract attention and wants to see if it has succeeded.

4.3.3 *Level of Interest*

Level of interest is similar to level of attention, except for the fact that it attempts to establish not only that the user is paying attention to the agent, according to its gaze direction, but also that the user is interested in engaging with the agent. This involves consideration not only of aspects of gaze, but also integrates any blinking, posture-related behaviour or feedback behaviours that are capable of being detected: for example head-nods. It is also measured over a variable time period, as specified in a query from the decision-making module.

The inputs to the metrics described above, especially the *level of interest*, are dependant to a large extent on what inputs are available for processing from the real environment. These metrics are necessary for the design of the shared attention module, described in Section 4.4.

4.4 Shared Attention Module

Shared attention refers to the manner in which two or more entities may simultaneously focus their attention on a single object in the environment and is a cornerstone in human social intelligence [5] [7].

The Shared Attention Module (SAM), is a centrepiece in Baron-Cohen's model [1], allowing triadic relationships to be formed and integrating information from other lower-level modules. This module is concerned with ones tendency to follow the line of sight of a person staring intensively at a particular object or location. It is an important stepping-stone for forming relationships and more complicated theories about the intentions of others. To begin approaching the design of a shared attention module, one must first have in place the metrics and modules described in Section 4.

Although shared attention may seem like a straightforward concept, this is not the case: shared attention is more than just *gaze following*, although detecting important attentive behaviours of other entities in the environment appears to be an important alerting mechanism for possible threats or opportunities. However, one need not be involved in an interaction in order to follow, automatically or volitionally, the gaze of other entities: this sort of gaze following can be passive and does not necessarily require interaction or engagement to exist between the two entities before or remain after the act, whereas the notion of shared attention described here does. Thus, a vital component in shared attention is the ability to know, during the act, that one is still engaged with the other and is expected to look back and regain engagement afterwards, even in the absence of the usual mutual gaze behaviours.

Our study of concrete metrics relating to the level of engagement (as presented in Section 4.3) is the first step towards investigating the design of such a module.

5. Outlook

We have described both a basic framework and additional capabilities for creating agents that can obtain certain inputs from both the real and the virtual environment, and to explore the creation of metrics for supporting a shared attention module as part of Baron-Cohen's theory of mind model.

These engagement and interest metrics are in no way meant to account for every range of behaviour possible from the user.

Rather, they investigate a generality solely from visual signalling: that the user's gaze behaviour can be harnessed for helping to infer where their attention might or might not be directed and this can be used to relate it to their interaction intention.

As such, this work could be viewed as just one component in a larger, more robust theoretical system. An understanding of the dialogue would, for example, help to greatly reduce the uncertainty of inferences made relating to the interaction.

The investigation of how to try to integrate these diverse components at an engineering and functional level, and implementation of the prototype capabilities described here as part of a real-time framework, will nonetheless provide an important step towards the creation of agents that can interact in a more natural way with users and, above all, provide agent's with the potential to be more sensitive towards them.

6. References

1. 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.
2. S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, and I. Plumb. The “reading the mind in the eye” test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42(2):241 – 251, 2001.
3. S. Chopra and N. Badler. Where to look? automating attending behaviors of virtual human characters. *Autonomous Agents and Multi-Agent Systems*, 4(1/2):9–23, 2001.
4. N. Courty, E. Marchand, and B. Arnaldi. A new application for saliency maps: Synthetic vision of autonomous actors. *IEEE Int. Conf. On Image Processing (ICIP03)*, 3:1065–1068, 2003.
5. G.O. Deak, I. Fasel, and J. Movellan. The emergence of shared attention: Using robots to test developmental theories. In *Proceedings of the First International Workshop on Epigenetic Robotics*, pages 95–104, Lund University Cognitive Studies, 2001.
6. M. Gillies. Practical behavioural animation based on vision and attention. PhD dissertation, University of Cambridge Computer Laboratory, 2001.
7. M. W. Hoffman, D. B. Grimes, A. P. Shon, and R. P. N. Rao. A probabilistic model of gaze imitation and shared attention. *Neural Netw.*, 19(3):299–310, 2006.
8. L. Itti. Models of Bottom-Up and Top-Down Visual Attention. PhD thesis, Pasadena, California, Jan 2000.
9. L. Itti, N. Dhavale, and F. Pighin. Photorealistic attention-based gaze animation. In *Proc. IEEE International Conference on Multimedia and Expo*, pages 1–4, Jul 2006.
10. L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 20(11):1254–1259, November 1998.
11. C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.
12. V. Navalpakkam and L. Itti. A goal oriented attention guidance model. In *Lecture Notes in Computer Science*, volume 2525, pages 453–461, Nov 2002.
13. D.I. Perrett and N.J. Emery. Understanding the intentions of others from visual signals: neurophysiological evidence. *Current Psychology of Cognition*, 13:683–694, 1994.
14. C. Peters. Direction of attention perception for conversation initiation in virtual environments. In *International Working Conference on Intelligent Virtual Agents*, pages 215–228, Kos, Greece, September 2005.
15. C. Peters. Evaluating perception of interaction initiation in virtual environments using humanoid agents. In *Proceedings of the 17th European Conference on Artificial Intelligence*, pages 46–50, Riva Del Garda, Italy, August 2006.
16. C. Peters. Designing an emotional and attentive virtual infant. In A. Paiva, R. Prada, and R. W. Picard, editors, *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*, volume 4738 of LNCS, pages 386–397. Berlin: Springer-Verlag, 2007.
17. C. Peters. A GPU-accelerated visual attention for animating virtual characters. In J. Ryan, editor, *Eurographics Ireland Workshop 2007, Proceedings*, 2007.

18. C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi. Engagement capabilities for ECAs. In workshop Creating bonds with ECAs, Fourth International Joint Conf. on Autonomous Agents and Multi-Agent Systems, Utrecht, July 2005.
19. C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi. A model of attention and interest using gaze behaviour. In International Working Conference on Intelligent Virtual Agents, Kos, Greece, September 2005.
20. C. Peters and C. O' Sullivan. Bottom-up visual attention for virtual human animation. Proceedings of Computer Animation for Social Agents (CASA) 2003, 2003.
21. Picot, G. Bailly, F. Elisei, and S. Raidt. Scrutinizing natural scenes: controlling the gaze of an embodied conversational agent. In J-C Martin, E. Andre, G. Chollet, K. Karpouzis, and D. Pele, editors, Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA'07), volume 4722, pages 538–539. Springer, September 2007.
22. Poggi, C. Pelachaud, and F. de Rosis. Eye communication in a conversational 3d synthetic agent. *AI Communications*, 13(3):169–182, 2000.
23. Ramisa, E. Vergara, and E. Marti. Game Programming Gems 6, chapter Computer Vision in Games Using the OpenCV Library, pages 25–38. Game Development Series. Charles River Media, Rockland, MA, USA, 2006.
24. Scassellati. Investigating models of social development using a humanoid robot. In Barbara Webb and Thomas Consi, editors, *Biorobotics*. M.I.T. Press, 2000.
25. C.L. Sidner, C.D. Kidd, C. Lee, and N. Lesh. Where to look: a study of human-robot engagement. In *IUI '04: Proceedings of the 9th international conference on Intelligent user interface*, pages 78–84, New York, NY, USA, 2004. ACM Press.
26. H. Yee, S. Pattanaik, and D. P. Greenberg. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. In *ACM Transactions on Graphics*, pages 39–65. ACM Press, 2001.