# iCu: Gaze Communication Towards Multiple Users

Rui Araújo
Instituto Superior Técnico, Taguspark Campus
Avenida Prof. Cavaco Silva,
Taguspark 2780-990 Porto Salvo, Portugal
ruiaraujo@ist.utl.pt

***Abstract*—Gaze is a crucial method to communicate through non verbal signals. Due to our social nature, users usually expect to interact with synthetic characters just like they interact with other people. In our work, we are concerned about the authenticity and believability of synthetic characters during this reciprocal sequence of actions. Our work focuses on gaze as a multi-user communication tool. We present a solution, inspired on psychology theories of visual attention, animation principles, and anticipatory mechanisms, where multiple users interact with a synthetic character without breaking the suspension of belief.**

***Index Terms*—Synthetic character, believability, gaze, anticipation, attention, multi-user interaction.**

## I. Introduction

Most, if not all known species developed any kind of communication system which assured their survival throughout time. Humans communicate via oral and non verbal expressions, using a complex system of symbols. Body language, eye contact, and sign language are some of the means of non verbal communication. These kind of signals, like eye contact and facial expressions, provide important information about the emotional state, and are so important that, sometimes, we may use only non verbal expressions to communicate with others. These symbols not only complement and reinforce spoken messages, but they also can be used per se. For example, if an individual is asked about the whereabouts of the TV remote control, he may look at the person, and then glance at the remote control, saying "there". Nonetheless he could also inform the other person, only glancing at the direction of the remote control, without speaking. The other person would understand that the remote control could be found where the other subject was gazing at, without hearing one word.

Due to our social nature, users usually expect to interact with synthetic characters just like they interact with other people [18]. Yet, actual computer science relies mainly on written communication. But, what if our computers had "eyes"? What if we asked our computers about the TV remote control and they answered with a glance? Would not it be easier for most of us just to briefly glance each other to achieve a certain task, like humans commonly do? What if we could engage in multi-party conversations, with human and non-human participants, in which we depend on our eyes to share relevant information? An engaging interaction between a machine and an human may be seen as a enjoyable communication. In order to generate an interesting communication, both the speaker as the interlocutor must be believable. Since humans are sociable beings, communication is intrinsic to our nature. Thus, we must provide characters a realistic communication capability. The communication with virtual beings can be achieved through several ways, such as tracking cameras, joysticks, keyboard, or plush toys. Nevertheless, the choice of technology should not be underestimated, since it is a major factor to create a affective interaction system. Plush toys are more familiar to the users, than joysticks, just like body gestures are more natural and engaging that complex controls. The chosen platform influences the emotional bond with the virtual character [36]. Synthetic characters are virtual creatures that present life-like behaviours [1]. Quality synthetic characters make the audience believe they are authentic by giving the illusion of having cognitive properties. However, creating believable characters is a demanding task to perform. Computer science, drama, animation, are just some of the fields that are still interested in the pursuit of the recipe for the development of such a creature.

Given this scenario, our work main focus is synthetic character's visual attention and how it can be used to communicate emotions, beliefs, and goals. This paper introduces how to increase believability through attentional and emotional control regarding gaze interaction with multiple users. We present an anticipatory mechanism that supports multiple users, called iCu (*I See You*), which was implemented our synthetic character, Bit (see Figure 1). Bit is an expressive mummy cat designed to interact with two human players in a word puzzle game.

Figure 1. Bit: a synthetic character mummy cat

## II. Related Work

### A. Believable Characters

The Oxford dictionary of English defines 'believable' as "convincing", while 'realism' is defined as "the quality or fact of representing a person or thing in a way that is accurate and true to life". Although both notions present a similar idea, in animation context they are distinct. In order to understand the concept of a believable character we need to look back to Aristotle, who proposed such character should be able of capture, represent and demonstrate believable states. According to Laurel [9], a dramatic believable character must fulfil the following requirements: good, appropriate, "like" reality, and consistent. Hayes-Roth [10] reinforces the idea that a believable character has "to seem" instead of "to be". The author highlights that believable characters must demonstrate communication skills, intelligence, individuality, social skills, empathic capability, and dynamic and coherent behaviour. However neither of these definitions work for all audiences, since believability is subjective to the human being.

Prendinger and Ishizuka [13] pointed out that realistic looking characters could represent a higher risk than cartoon-style ones: "As opposed to cartoon-style characters, users have high expectations of the performance of realistic looking characters, which bears the risk that even small behaviour deficiencies lead to user irritation and dissatisfaction". Years before, Masahiro Mori [12], a robot designer, explained the hypothesis of the uncanny valley which defends that, when a character looks and acts almost but not entirely like humans, it causes an unnerving effect among observers. Bill Tomlinson also reinforces the idea that the design of the characters creates expectations: "We do not make human characters who look like humans because we cannot make human minds" [6].

Frank Thomas and Ollie Johnson, ex-members in Walt Disney Productions, stated that their goal was to "make the audience feel the emotions of the characters, rather than appreciate them intellectually" [3]. If these targets are met "the audience will now care about the character and about what happens to him, and that is audience involvement" [3].

Squash and stretch are mentioned as the two of the most important animation principles, since "it is the change in shape that shows what the character is thinking. It is the thinking that gives the illusion of life. It is the life which gives meaning to the expression" [3]. As such, character's personalities depend more from their movements, than from their appearance [3].

Regarding the eye expressiveness, in the beginning, Disney Studios had the main figures looking up, which confered them an appealing, pure and innocent aspect. But as time went by, the eyes became an important way of expressing emotions and personality, since "the eye is the most eloquent tool of communication" [3]. Walt Disney said "the audience watches the eyes, and this is where the time and money must be spent if the character is to act convincingly" [3].

In [3], we find some considerations when drawing eyes: (1) if there is too little of the pupil showing, it is difficult to make a strong statement of either the expression or the direction which the character is gazing, (2) if the pupil is close to the rim of the eye, the direction is well defined, (3) if there is a great part of white surrounding the pupil, the eye conveys excitement and intensity, (4) white all around the pupil increases the vagueness to the expression and makes the direction look uncertain. Blink, which may be first seem like a mechanical move, is used to keep a character alive, and not only increases the amount of move, but also emphasizes it. Blinks are good on any shift of eye direction, since they call the audience's attention to the change. Blink is an useful device to show surprise.

According to Joseph Bates, an emotionless character has no life, which can be merely presented as a machine. Thus, if the character does not have an emotional posture, neither will we [14]. Animators use empathy to suspend disbelief by creating an emotional bond with the viewer. It is used to make the audience care about the characters' emotions and problems, which will distract them from the fact that they are interacting with a lifeless being [11]. Therefore, empathy is often applied to engage the audience emotionally, enriching the scene. For example, Blumberg experienced a new level of engagement and involvement when using Sympathetic Interfaces (plush toy) to interact with the public [5].

Each movement of the character must be coherent. Every action should demonstrate the result of the character's cognitive process. Otherwise the audience will no longer believe in such character. Disney uses an animation trick to transmit the idea of a thinking character, which is known by anticipation. According to Disney Studios principles of animation, anticipation is essential to suspend disbelief towards animated characters, since the audience will not understand the events unless there is a planned sequence of actions that leads them from one point to the next [3]. It is used to capture the audience's attention, and prepare them for the next action. It is also used to guide the audience's eye along the screen [2]. Usually animators simulate anticipation by moving the eyes of the character a few frames before the head.

Providing an anticipation model to a synthetic character means its behaviour system contains a predictive model. It will make use of its predictions to modify its current behaviour. "If we further suppose that the model can approximate by its predictions the future events with a high degree of accuracy, then this system will behave as if it was a 'true' anticipatory system, i.e. a system the behaviour of which depends on future states" [1]. Duncan, the Highland Terrier, developed by Damian Isla and Bruce Blumberg, focused on different kinds of expectation in synthetic characters. They defined anticipation as the "ability to make decisions and react to aspects of the world state that, for one reason or the another, cannot be directly perceived" [8]. Duncan was enhanced with the notion of object persistence by using a probabilistic occupancy map, that allows it to anticipate the position of moving objects that are out of sight.

According to the American philosopher Daniel Dennett, the growing complexity of machines leads people to treat them as intentional beings [15]. His intentional stance theory presents the idea that we predict and explain the actions of animate

things by treating these objects as rational beings, which aim to a fulfil their 'desires', given their 'beliefs' [16]. A character who meets these requirements will appear to have intentions and a distinct personality. Furthermore, Reeves and Nass [18] showed that human-computer interaction is fundamentally social, which reinforces the importance of believability in synthetic beings. Believability also plays an important role in collaborative scenarios involving both humans and agents, like resolution of tasks. According to Prada and Paiva, group believability is achieved by looking at group dynamics and human social psychological sciences [19]. Blumberg, Tomlinson and Rhodes [17] also demonstrate how social status could benefit multi-agent systems, just like it does in the natural world. They present three areas that could benefit: streamline negotiations, alliance formation, and human interface.

### *B. Interaction Methodology*

Piccard [35] suggested that computers should be designed to recognize and express human emotions, which is currently known as 'affective computing'. More importantly than a computer capable of showing an emotion to a user, is to design an interactive system capable of provoking emotions within the user. There are multiple ways to interact with virtual characters and, regardless of the method used, we must not neglect the fact that participants should always be able to read the desires, beliefs and actions of the characters.

The interaction via simple objects of our quotidian is by itself familiar to the participants, and therefore more engaging to work with. Object manipulation can be seen in *sand:stone* [4] installation, which allowed participants to "leave traces of their interactions through the patterns of sand and arrangements of stone". This creative project was inspired in history, when sand and stones were the communication methods. DigiWall [37] is another example of integrating quotidian objects in interactive computer systems. DigiWall is a standard, artificial rock climbing wall and a computer game. The interface has no computer screen, however music and sound are the main drivers of its interaction models.

A sympathetic interface is a concept introduced by Blumberg in *Swamped!* [5] , which represents a physical interface that senses the environment. The major objective of *Swamped!* is the interaction enhancement between the user and the character, in the most immersing and compelling way possible. In 2005 another MIT researcher, Stefan Marti, has developed a physical embodied pet cell phone [34]. This pet was developed to manage a person's phone calls by taking messages, answering calls and alerting the user when it decides it is important for the owner to answer the phone. The assumption is that having a cuddly pet to signal calls is less intrusive than the regular ringing. A study that compared people's response to both, showed that they rated the squirrel as being more friendly, fun, humorous and cute. SenToy is another example of a sympathetic interface that allows users to influence the emotions of a character in a 3-D system [20]. The underlying assumption was that, by performing a set of actions in the avatar that allow users to influence the emotions of the character, would "pull them into the game". Sentoy captures certain patterns of movements from the users, which are associated with emotional responses.

The "magic mirror" metaphor was presented by Maes in ALIVE [7], an "Artificial Life Interactive Video Environment". The ALIVE system allows the user to interact with his full-body, without requiring the use of any special kind of devices. The ALIVE system captures the person's movements with the help of a single video camera, and projects his image onto a large screen presenting a rich graphical world inhabited by autonomous agents. The "magic mirror" idea arises from the necessity to create a non-intrusive interface, that allows rich an intuitive gestures to navigate and control a virtual world. Instead of using tethered goggles-and-gloves interfaces, it is preferable to use a wireless sensor, like vision, that provides a safer solution, since the user can still see his whereabouts. Thus the users can avoid bumping into things or tripping over wires. Also, the users enjoy greater behavioural and expressive freedom. "Users have less difficulty interacting with the agents when they can use simple gestures that are natural and appropriate given the domain" and "interaction is improved when the user receives feedback from the agents, either in terms of movement and/or facial and body expression".

### *C. Eye Gaze*

The Oxford dictionary of English defines 'gaze' as the act of "look steadily and intently, especially in admiration, surprise, or thought". In other words, gaze is about the way our eyes are fixed on something that get our attention. When our attention is caught, we reveal our emotions with gaze expressions, which are fundamental to communicate via nonverbal messages with other beings.

The human vision is limited by two aspects: the capacity of perception and the selectivity of information perceived [23]. Due to these limitations, attention is consequently affected, since there are multiple stimuli competing in the visual cortex. The selectivity is accomplished by saccades, which are quick eye movements that search for new points of interest. According to Thiebaux's study "a gaze could be used to attend to task performance, regulate conversational interaction, express intimacy or emotion, or exercise social control" [25]. Gaze is a versatile way of communication and expression. Poggi and Pelachaud's work showed that different gaze expressions may have different functions in a conversation. They analysed gaze items and related them to specific gaze states and behaviours [29].

Visual attention is currently known for operating in three different phases [21]: pre-attentive stage, attentive stage, and afterwards there is an habituation stage to the surrounding environment. In the first phase occurs a crude analysis of the external visual scene, and all the information retrieval is performed in parallel. Features are extracted from the objects and the entire field of view is analysed. During this process, the distinction between objects is not voluntarily. According to Treisman's theory, this stage is essential to execute perception [24]. In the attentive stage, the mind map

constructed in the previous phase is analysed with greater detail. This phase deals with limited-capacity problems, therefore it treats only one item (or at best a few items) at a time [22].

In 1980, Posner distinguished attention towards a stimulus in two ways: an endogenous system (covert), which is controlled by the subject's intentions, and an exogenous system (overt), which automatically shifts attention towards external stimuli, and therefore cannot be ignored [26].

According to original filter theory [26], there is only one processing channel, and task combination is achieved by rapidly switching filters and multiplexing, or time-sharing tasks."Attention is like a bottleneck where selection has to take place owing to the fact that parallel processing must change to serial processing to protect the limited capacity component of the processing system" [26].

Visual joint attention is known as the process by which one individual alerts another to a stimulus via gazing. Imai [27] and Yonezawa [28] developed a robot and showed that eye contact is essential to make the subjects become aware of some desired stimulus. Brooks [33] work on Leonardo demonstrated that joint attention can be used to perform a task, like getting attention.

Khullar and Badler elaborated a research [30] where they proposed a computational framework to generate visual attending behaviour, such as eye control, head motion and other locomotion actions. Their work on AVA (Automated Visual Attending) was based on both psychology and human anatomy. The framework captured important notions of psychology theories: endogenous factors, exogenous factors, scanpaths, saccades, free viewing, divided attention, overt and covert attention.

In [31], Vertegaal et al. dealt with speech recognition problems, by trying to understand when the system is being addressed or expected to talk. In multiparty conversations, it is not obvious who will be the next speaker. The eye contact, as we seen before, plays an important role by determining who is going to be the next speaker. "Multiparty conversational structure is much more complicated than its dyadic equivalent. As soon as a third speaker is introduced, the next turn is no longer guaranteed to be a specific non-speaker" [32].

## III. iCu Arquitecture

Our architecture will group several principles and ideas applied in similar works. We developed an architecture resemblant to AVA, the one used by Kullar and Badler, and we included an anticipation model, simple eye expressions, principles of animation, and extended it to a multi-user paradigm. The architecture is divided in three main systems: the Perception System, the Selective Attention System, and the Expression System (see Figure 2). There are also two another important structures: a Self Initiative structure and a Event Evaluator. The information flow is performed by arrays containing the information about events. Events are composed by the following information: (1) the type of event, (2) the object that caused such event, (3) the force of that event, (4) the eye expression related to that event, and finally (5) the time the agent will take to gaze at that object. The type of event is a string that reveals its nature. There are three different types of events: endogenous, exogenous and spontaneous. Spontaneous events refers to the spontaneous looking. Events are all treated the same way, hence it is essential to distinguish them by type. The object that caused an event is, for example, a moving box, or a waving arm. The force of the event is the strength associated to that event. The force is represented by a float that varies between the value 0 and 1. The eye expression may vary between neutral, affirmative, and negative expressions. Finally, the time information refers to how long the agent will gaze at the object.

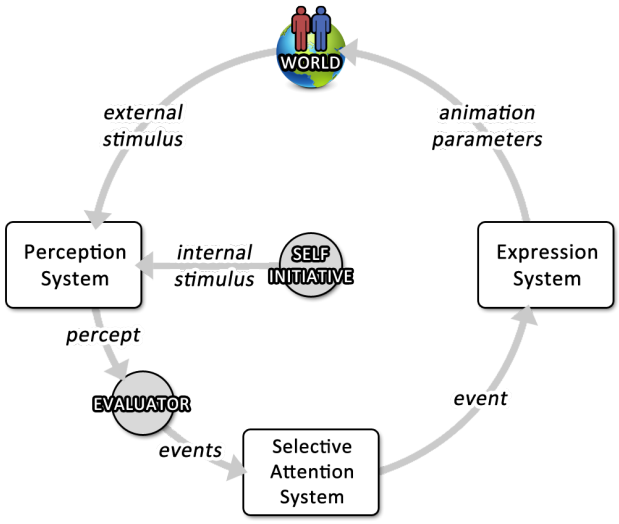


Figure 2. iCu architecture

### A. Virtual and Real Environments

The agent is vulnerable to two different kinds of external stimuli. One of them is entirely virtual, i.e. it occurs only inside the virtual environment and its origin is not caused by real environment events. The other kind of external stimulus is a representation of the events occurring in the real world caused by the users. Both worlds are susceptible to the agent output. The users in the real world observe the agent behaviour, and the virtual world is also sensitive to its response.

### B. Perception System

External stimuli caused both by virtual and real environments are redirected to the Perception System. This system is responsible to receive each input and, possibly, turn it into a percept. Not all stimuli will be promoted to percepts, but only those whose object remains inside the agent's field of view. The Perception System also retains information about something we call Spontaneous Looking, which can be seen as a blank stare around the environment. Then, the Perception System redirects relevant input information – percepts – to the Event Evaluator, which will later transform them into events.

### C. Event Evaluator

This is where the anticipation mechanism lays. The Event Evaluator receives the input information from the Perception System and analyses it. Depending on the nature of the event, for instance, if the event fulfils any of the agent's intentions, the evaluator will rearrange the information and create an endogenous event, otherwise, it creates an exogenous event. The exogenous events force depends on the velocity parameter of the object. The greater the velocity, the greater the force of that exogenous event. Afterwards, the system merges the exogenous events occurring at that time into a list of exogenous events. The main difference from the exogenous events is on the force of the event and the eye expression. Rather than simply show surprise when an exogenous event occurs, endogenous events are more complex since they result of cognitive processes. Similarly, the force associated to an endogenous event cannot depend only on visual noise, like its velocity. Once the endogenous event is created, its force is calculated. The force calculator, it takes into account three components: the task progress, the distance between the agent and the object of interest, and the user model. The main cause of believable behaviour in our agent comes from the correct attribution of forces to the events. Thus, it is critical that the provided forces do not contradict the agent's goals. The evaluator is also responsible to make an appreciation of the input, given the agent's current state. It evaluates the proper emotional response and fills both the eye expression and time fields of the array. For example, if the event is considered positive, regarding the goals of the agent, the eye expression will be also positive. In the end, the Evaluator merges all the endogenous events inside an endogenous list and redirects it to the Selective Attention System.

### D. Self Initiative

Although the agent's behaviour depends greatly on the external environment, the lack of input also triggers responses. If the agent goals are not being met, it takes measures to try to ensure they will be fulfilled. The Self Initiative mechanism is responsible for promoting the success of the agent goals. When the agent is performing a spontaneous look for a long time, the mechanism triggers the self initiative event. The time that takes to trigger the mechanism is variable, but it stays between a range of specified values. Therefore, the Self Intention structure will redirect an internal stimulus to the Selective Attention System.

### E. Selective Attention System

The Selective Attention System is, like the name indicates, the system responsible for choosing an event from a range of events. This system receives three different types of events: an exogenous list, an endogenous list, and spontaneous look, all coming from the Evaluator. The Selective Attention System is not always active. It only chooses an event when it finishes dealing with the last one. If that is the case, once the Selective Attention System receives the events it evaluates them. First it determines what was the most recent endogenous event sent by each user, if a user sent more than one event during the time the Selective Attention System was occupied. Next, it puts each of those last endogenous events inside a list, along with the self initiative event, and chooses what is the strongest event inside that list, in other words, which event has the highest force. Then, it chooses what is the strongest event inside the exogenous list. Finally, the system proceeds to a selection process.

The selection process arbitrates which event should be chosen, depending on the type of event and its force. The selection process privileges the endogenous events. If there is a maximum endogenous event, the selection process activates that event. If not, it chooses the strongest exogenous event, only if that event's force is higher than the spontaneous look force, and if the agent is not in the habituation stage. We impose a restriction to the exogenous events, otherwise the agent would be constantly paying attention to them. We applied two different strategies: the spontaneous look has a low force that inhibits the agent to pay attention to exogenous events with minimal forces; the habituation stage inhibits the agent to pay attention to exogenous events that keep occurring. If those conditions do not apply, the spontaneous look event will be activated.

### F. Expression System

The Expression System is responsible to manage the physical representation of the agent. This system gets the chosen event from the Selective Attention System and extracts the relevant information, such as the target and the eye expression. Lastly, the agent gazes at the chosen target with the associated eye expression, playing the adequate animation.

### G. Attention

Our model takes into account the three visual attention stages: pre-attentive stage, attentive stage, and habituation stage. The Perception System, and the Self Initiative mechanism is where the pre-attentive stage occurs. The attentive stage occurs inside the Event Evaluator, and the Perception System. In the habituation stage, the system becomes insensible, and ignores external events. This stage can be related to the anticipation of certain events.

Our habituation mechanism holds an habituation variable that represents the threshold in which the system enters in the habituation stage – threshold$_{hab}$. The system retrieves the strongest exogenous event at each time cycle, which is the most significant one from the sample – exog$_{max}$. If the habituation threshold is greater than the force coming from the strongest exogenous event, the system enters in the habituation stage. Consequently, during a variable amount of time the system will ignore exogenous events. The habituation threshold updates at each time cycle according to the following formula:

$$threshold_{hab} = threshold_{hab} + \frac{exog_{max} - threshold_{hab}}{attenuation}$$

The *attenuation* factor is a defined constant. The greater the attenuation, the longest takes to the system to enter in the

habituation stage.
The joint attention functionality in our agent results from the feedback animation given by the Expression System. There are times the agent fixes its eyes in a certain object of interest. Other times, the agent shifts glances between the user and the object it wants the user to look at, as it was saying: "user, look over there".
Our model is capable of performing different gaze functions: perform a task, regulate conversational interaction, and express emotion. The agent is capable of giving feedback to the users, in order to perform a task, while expressing emotion. When self initiative mechanisms are active, it regulates conversational interaction. The agent also follows patterns of spontaneous looking when there is no event to be attended.
Our model takes into account the limited visual capacity and selectivity of information perceived. Although they might not be attended, the competing events increase the perceptual load. The Selective Attention System works like a bottleneck that selects the event the agent will attend to.

### H. Anticipation Model

The existence of an anticipation model in our architecture enables the occurrence of believable behaviours. Our anticipation model is present in the Event Evaluator, which builds endogenous events. The anticipation model attributes forces to the endogenous events, taking into account the distance between the object and the agent, the Progress Model, and the Player Model. The force that results from the Anticipation Model is given according to the following formula:

$$force = x * distance_f + y * |progress_f| + z * user_f$$

The $distance_f$ represents the force factor given by the distance between the agent and the object related to the event. The $progress_f$ represents the force given by the progress model. The $user_f$ represents the force given by the user model. Each of these forces, as well as the resultant force, vary between 0 and 1. The variables $x$, $y$, and $z$ are percentages that represent the weight of each component to the force of the endogenous event. The sum of these variables must be equal to 1. Our model assigns the following values to the variables: $x = 0.1$, $y = 0.7$, and $z = 0.2$.

*1) Object Distance:* The object distance influences the resultant force by the following formula:

$$distance_f = \frac{1}{distance_{obj} + 1}$$

The $distance_{obj}$ represents the distance between the agent and the object. This way, the farther the object, the lower the resultant force. This formula ensures that nearer objects have preference over farther objects.

*2) Progress Model:* The Progress Model is the source of our agent's anticipation. The Perception System input allows the agent to make assumptions about the users' next action. If the user is near a certain object, the agent assumes he will probably grab it. These assumptions allow the system to evaluate the progress of the task, taking into account the most likely next action of the user. The force associated to each action anticipated by the agent varies between 0 and 1. Our model divides the overall progress into short-term goals. Dividing the overall progress into smaller sub-objective progress allow us to predict with more precision and with lower uncertainty what the next action will be. However, sometimes the users' actions not only decrease their own sub-objective progress, but the others as well, which results in a negative force. A negative force could mean that its event was not important, which is not true. A negative force means the overall progress will decrease, which makes its event very important. Therefore, we use the absolute value to define the anticipation force, $|progress_f|$.

*3) User Model:* This model is responsible to differentiate the users due to their previous actions. A user that contributes to the task progression has privilege over one that does not help to fulfil the agent's goals. The system memorizes the events caused by each user and the individual progress of each one. Depending on the agent's primary goals, the User Model may have different approaches. We applied the following scheme:

$$user_f = x * correct_f + y * incorrect_f + z * punish_f$$

The $correct_f$ represents the force related to the correct actions of the user. The $incorrect_f$ represents the force associated to the incorrect actions of the user. The $punish_f$ represents force related to the punishable actions of the user. Our model assigns the following values to the variables: $x = 0.4$, $y = 0.4$, and $z = 0.2$. We tried to make a fair system that does not punish the user too much, when he makes a mistake. These weights guarantee that a correct action will nullify an incorrect action, and vice versa.

## IV. Case Study: Bit

Bit is the name of our synthetic character, a synthetic ancient mummy cat, stuck inside an Egyptian tomb. Bit's body is totally rigid, except for its head. Its emotions are conveyed by the head motion. Its body is similar to a cat's body, and is composed by a torso, two upper paws, two lower paws, and a tail. The interactive elements in its head are the eyeballs, the eyebrows, the whiskers, and the eyelids. While Bit's body is fixed, the head has three degrees of freedom (DOF). The eyeballs have two DOF, the whiskers, the eyebrows, and the eyelids have only one DOF.
The expression of attention can be expressed by two distinct but complementary mechanisms. The first mechanism is obtained by changing the direction of the character's eyes and head, in such a way that it points toward the object which is the focus of attention. The eyes always move faster than the head, and the speed is dictated by the nature of the event. The delay existing between the eyes and the head motion increases if the character is free viewing. If the character is confronted with an exogenous event, its head will move faster than when it is following patterns of spontaneous looking. The second mechanism happens when the character takes initiative, using joint attention. It helps the users by shifting glances between

the user and the objective. When the character shifts the object of attention, it blinks to emphasize that change. The absence of focus of attention leads to spontaneous looking, which makes the character gaze around the environment.
The expression of emotion is directly related to the feedback upon user input. There are three sets of emotional expressions: positive, negative and neutral. The greater the task progress anticipation, the higher the emotional intensity. For example, if a user correctly finishes a task, the feedback coming from the character is strong. However, if the user is about to finish the task, its feedback is weaker. If the users do not advance their tasks the cat starts to despair. The character also emphasizes surprise every time the object of attention changes. It blinks and the eyebrows raise.
Our case study follows the 'magic mirror' paradigm shown in [7]. We use Microsoft's Kinect camera to capture the users movements, so that they are able to interact with the synthetic character with their hand gestures. Kinect's sensors capture two images: a color map and a depth map. According to the users position in the depth map, the system determines where users joints are located, such as the head, the arms, the knees, etc. In order to grab the objects in the environment, the user must get near to the object with his right hand, and reach his forehead with the left hand – making use of the 'telekinesis' metaphor – as he was thinking hard to move that object with his own mind.

## V. Experiment

Believability is a difficult concept to evaluate. Usually it is measured by direct questions to assess the satisfaction of the audience and by relating it to the suspension of disbelief. Each subject is presented a task in six distinct scenarios. The first three scenarios are meant to validate the character's expressiveness – training scenarios. The remaining three scenarios are meant to evaluate our model – word puzzle scenarios. Although the synthetic characters differ throughout these scenarios, they look all alike. The forth scenario verifies the impossibility of the task without consistent help from the character. The remaining two scenarios tested our model. Our model verifies the hypothesis that having a gaze architecture that supports multiple users will improve synthetic characters believability. Afterwards, each user is asked to answer a questionnaire evaluating the character's believability.

We chose a game as a means of developing our experiment to create a motivating experience for users. We decided our game would take place in a fully three-dimensional virtual environment in Egypt (see Figure 3). The users must escape an ancient Egyptian tomb before their oxygen runs out. They must fight for their lives by extracting all passwords from Bit, a mysterious mummy cat. But there is a catch: he is not be able to tell the users a single word. The game is a two-player experience where the players must form an alliance to escape the tomb alive. However, there is also an interesting competition between the players, since only the player that


Figure 3. Word puzzle environment

gathered the greater amount of treasures will escape alive from the tomb. Each treasure is won individually, and Bit reproves the players that choose to steal treasures, instead of contributing to the main task, which is to escape alive. Each player has two words: one of them is shared by both players, and the other is individual. Once the shared word is correctly completed, the gate of the tomb opens and only the richer player will escape alive, and consequently, win the game. For each letter correctly inserted in the individual word, the player wins a treasure. One player cannot access the other player's individual word, but he can modify the letters present in the shared word, even if the letters were chosen by the other player. Bit always plays as the non-guessing player. It is the only one that knows the three words. The users always play as the guessing players. To construct the word, the players must use a set of letter-boxes, each one with an alphabet letter inscribed onto it. Each word is constructed by placing the box onto sockets that represent word letters and their relative position in the word. Users do not interact directly with the character. Instead, they interact with Bit through their hands' movement, and the letters they choose. The character returns them feedback via gazing. Bit is the only help the users can rely on.

### A. Experimental Settings

Our experiments are divided in two main categories: the interaction-based experiment, and video-based experiment. Both experiments are similar: in the first, the users interact with the character directly playing the word puzzle; in the second, the users are spectators of video samples recorded during the previous phase.

*1) Training Scenarios:* The first scenario aims to verify whether the user is able to understand Bit's gaze behaviour. In this scenario, the character randomly chooses two of the five available letter-boxes that represent the five vowels: a, e, i, o, u. Bit focus its attention on one letter at a time, throughout the scenario. Then, the subject is asked to identify which letters Bit was looking at.
The second scenario introduces Bit's emotional reactions to the subject. Figure 4 shows the training environment. It verifies whether the subject is able to differentiate between positive and negative emotions. In the beginning of this scenario, the

subject is told that Bit is sad because it has lost its favourite letter. There are the five letter-boxes available that represent the five vowels: a, e, i, o, u. The avatar interacts with the letter-boxes and the user is asked to find out which is Bit's favourite letter.

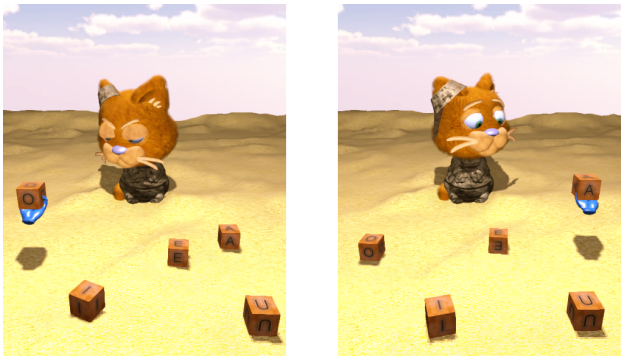

Figure 4. Emotion training

The last training scenario verifies whether the subject is able to understand Bit's intentions. The third scenario introduces the notion of inserting letter-boxes onto the sockets, an essential task in the word puzzle scenarios. Figure 5 shows the training environment. The avatar interacts with the letter-boxes triggering reactions on Bit. Once the avatar grabs a letter, the character shifts attention between him and the place where the letter should be inserted.

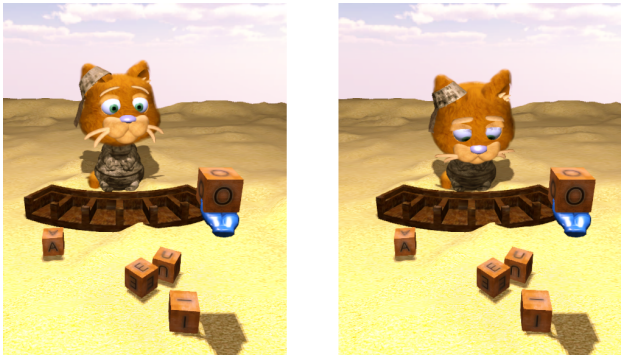

Figure 5. Intention training

*2) Word Puzzle Scenarios:* In the idle behaviour scenario, the character never reacts to any kind of subject's action, unless if it is an exogenous event. Bit does not process endogenous events. The cat presents an idle behaviour, moving its head around the environment, exhibiting a blank gaze.

In the player-n scenario, the character will interact only with one avatar. Firstly, the character chooses one avatar to interact with, while the other one will be ignored throughout the whole scenario. Bit handles both exogenous and endogenous events, ignoring every possible events associated to the non chosen player. This scenario asserts if the character gaze behaviour becomes more believable without a support to multiple players.

The iCu behaviour uses our model to control the attention and basic sensations, towards multiple users.

## VI. Results

During the interaction-based experiment, in the iCu cat scenario, we observed that the subjects where deeply immersed in the experience, since they were interacting with the environment with their own gestures. During the idle gazing cat scenario, when the users realized the cat was not going to help them, the suspension of disbelief was broken and they started to play with the letter-boxes instead of focusing on the task. The reaction observed during the player-n cat scenario was similar, since the cat was ignoring one of the players. Even the players that were being helped by the character were not fully immersed in the experience, since they did not understood why they were chosen by the cat. When the players did not found the character believable, their focus of attention usually turned to the interaction methodology. Usually, they started to interact with the scenario with their gestures. However, during the iCu cat scenario, the players' state of mind showed that they were completely focused on the cat, and on the task. In this case, the interaction methodology worked as a catalyst to the suspension of disbelief. The almost immediate suspension of disbelief suggests that the chosen interaction methodology does not make a character more believable, but it contributes to improve the users experience.

### A. Training Analysis

Regarding the training analysis, the gaze training results show that 83.1% of the users was capable of identifying the letters the cat was gazing at. The remaining users identified only part of those letters. The emotion training results show that 93.5% of the subjects were able to identify the letter. As such, the majority of the subjects undoubtedly recognized the character's emotional expressions. The intention training results show that 49.4% of the subjects were able to fully understand the cat's intentions, and 90.9% were able to identify, at least, part of the character's intentions. Interestingly, from the remaining 9.1% of subjects that could not identify correctly the character's intentions, only 2.5% of the population could not identify any possible intention. As such, about 97.5% of the subjects recognized one possible intention of the character, even if it is not correct. This value strongly suggests the intentional stance proposed by Dennett[15].

### B. Word Puzzle Analysis

*1) Help and Believability:* For each model, we asked the users to categorize each cat according to its help, and believability. Looking at the results of the help given by the character, we conclude that: 83.1% of the users considered the idle model below 'medium', 80.6% found the player-n model above 'medium', and 96.1% found the iCu model above 'medium'. The iCu model clearly shows the better results, followed by the player-n model. The idle model, as expected, shows a negative response. As for the character's

believability, we observe similar results: 53.3% considered the idle model below medium believable, 68.9% considered the player-n model above medium believable, and 93.5% of the users thought the iCu model was above medium believable.

*2) Gaze and Help:* For each model, we asked the users to categorize each cat according to the gaze, and help given to the players. We cannot confirm the correlation between gaze and help, since the idle model does not show a strong correlation – which makes sense, since the idle behaviour cat does not help any of the players. As expected, the results show a stronger correlation in the player-n model, than in the iCu model. Since the player-n cat gazes and helps only one player, the correlation must be stronger than the one observed in the iCu model. In the iCu model, the cat starts the video liking both players the same way, therefore gazing and helping both players the same way. Throughout the interaction, given both players actions, the cat starts to pay more attention and helping more the cooperative player. The results clearly show that the users are well aware of the character's gaze and help.

*3) Intentional Stance:* For each model, we asked the users to explain why the cat was acting in such way. The results show that: 81.8% of the users was able to identify an intention in the iCu behaviour cat, 57.1% of the users associated an intention to the player-n behaviour cat, and 61% of the users gave the idle behaviour cat one possible intention. Surprisingly unexpected, the results show that it was easier to attribute a possible intention to the character in the idle model, than in the player-n model, even if it was not the correct one. Even though the users were provided a very limited context and 1 minute length video, 49.4% of them was able to understand the cat's true intentions in the iCu model. Only 13% of the subjects was able to understand the cat's true intentions in player-n model.

*4) Believability:* In order to evaluate the character's believability, we analyse the the character's gaze, the character's emotions, and how easy it was to complete the task with each one of the tested models. First we present the results of how hard it was to understand where the cat was looking at. The results show that the idle model is evenly distributed, 78% of the users considered the player-n behaviour cat above 'medium', and 91% of the users considered the iCu behaviour cat above 'medium'. The results of how easy it was to perceive what the cat was feeling show that: 66.3% of the users found the idle model below 'medium', 67.6% considered the player-n model above 'medium', and 87% of the users found the iCu model above 'medium'. The results of how easy it was to understand if a letter belonged to the hidden word show that: 75.3% of the users considered the idle model below 'medium', 62.4% found the player-n model above 'medium', and 81.9% considered the iCu model above 'medium'. Finally, the results of how easy it was to understand if a letter was in the right position of the hidden word, similarly to the previous evaluation, show that: 85.7% of the users considered the idle model below 'medium', 76.6% found the player-n model above 'medium', and 93.5% considered the iCu model above 'medium'.

*5) Satisfaction:* Finally, we asked the users to choose which cat they found more believable, and which one would they choose to play. As for the most believable cat, the results show that: 2.6% of the users considered the idle model the most believable, 18.2% found the player-n model the most believable, and 79.2% considered the iCu model the most believable. Similarly, 2.6% of the users would choose the idle model to play with, 15.6% would choose player-n model, and 81.9% would choose the iCu model. Closely studying the results we observe that most of the subjects that chose the n-player model to play with, justified it with two reasons: its expressions were easier to understand, and its preference for one of the players makes the game intellectually more challenging. The subjects that chose the idle model justified it for considering it more believable and because none of the players got advantage over the other.

## VII. Conclusion

The correct use of audience expectations, coherent behaviour, awareness, anticipation, and emotion will hopefully build a character with believable behaviour. The audience will percept the character's emotions, and develop a mental model to coherently explain the character's intentions. The viewers will attribute a certain personality to the character, and create a set of expectations regarding it. If these expectations are not unrealistic broken, the character will gain life of its own. Our work manages multiparty eye gaze conversations between a synthetic character and human participants. Our results showed that a character that supports an multiparty gaze architecture, while interacting with different users, is more believable than one that does not.

## VIII. Future Work

Our anticipation mechanism depends on the task progression, on a simple user model, and on the distance between the user. Our user model is based on the users' previous actions to predict what will be the next most probable action of the player. So, the character internally poses the following question: given the player's history, will his next action lead to progress? But we can reformulate the character's problem: given a set of correct and incorrect behaviours of the player, how does the character feel towards the player? Our work could be extended in such way that the character's anticipatory mechanism would also depend on personality models. Not only the character would be provided with different personality characteristics (e.g. patience, sympathy, optimistic), but it would also attribute different personalities to the players, during the interaction.

Our model's pre-attentive visual stage only extracts the object's velocity to evaluate the objects relevance in the pre-attentive stage. This stage could be improved by extracting another interesting visual features, such as the color, the brightness, etc. Furthermore, the character's perception of the virtual world is a simulation of the character's vision, i.e. the

character's visual sensor is based on environmental triggers. Then, in order to reproduce a realistic visual system, one could insert a camera in the character's eyes. The system would analyse each frame that was captured by the camera, to realistic retrieve information about the environment.
Our model could be considerably improved regarding the character's animations. Our Expression System could be extended to different variations of emotions.

## REFERENCES


[1] C. Martinho. *Emotivector: Mecanismo Afectivo e Anticipatório para Personagens Sintéticas*. Ph.D. Thesis. September 2007.

[2] J. Lasseter. Principles of traditional animation applied to 3d computer animation. In M. Stone, editor, *Proceedings of the 14th annual conference on Computer Graphics and Interactive Techniques*, number 4 in 21, pages 35-44, New York, NY, USA, 1987. ACM Press.

[3] F. Thomas, and O. Johnson. *The Illusion of Life*. Hyperion Press, New York, NY, USA, 1994.

[4] Synthetic Characters Group. M. Downie, B. Tomlinson, A. Benbasat, J. Wahl, D. Stiehl, and B. Blumberg. *sand:stone*. Leonardo Vol 32, N.5, pp. 462-463. 1999.

[5] M. Johnson, A. Wilson, B. Blumberg, C. Kline, and A. Bobick. *Sympathetic Interfaces: Using a Plush Toy to Direct Synthetic Characters*. In Proceedings of the CHI 99 Conference on Human Factors in Computing Systems, pp. 152-158. 1999.

[6] B. Tomlinson, M. Downie, and B. Blumberg. *Multiple Conceptions of Character-Based Interactive Installations*. Submitted to CHI2001. 2001.

[7] P. Maes, T. Darrell, B. Blumberg, and A. Pentland. *The ALIVE System: Wireless, Full-Body Interaction with Autonomous Agents*. In the ACM Special Issue on Multimedia and Multisensory Virtual Worlds. 1996.

[8] D. Isla, adn B. Blumberg. Object persistence for synthetic characters. In C. Castelfranchi and W. Lewis Johnson, editors, *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems*. Part 1, pages 1356-1363, New York, NY, USA, 2002. ACM Press.

[9] B. Laurel. *Computer as Theatre*. Addison-Wesley Longman Inc., Reading, MA, USA, 1991.

[10] B. Hayes Roth. What makes characters seem life-like In H. Prendinger and M. Ishizuka, editors, *Life-like characters: tools, affective functions, and applications*, pages 447–462. Springer, Berlin, Germany, 2003.

[11] H. Maldonado, and B. Hayes-Roth. *Towards cross-cultural believability in character design*. In R. Trappl and S. Payr, editors, Agent Cultures, pages 143–175. Lawrence Erlbaum and Associates, Mahwah, NJ, USA, 2003.

[12] M. Mori. Bukimi no tani (the uncanny valley). *Energy*, 7:33–35 1970.

[13] H. Prendinger, and M. Ishizuka, editors. *Introducing the Cast for Social Computing: Life-Like Characters*, Springer-Verlag. Berlin, Germany, 2004.

[14] J. Bates. The nature of characters in interactive worlds of Oz project. Technical Report CMU-CS-92-200, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1992.

[15] D. Dennett. *True Believers: The Intentional Strategy and Why It Works*. Cambridge, MA, USA MIT Press. 1987.

[16] D. Dennett. *Brainchildren: Essays on Designing Minds*. Cambridge, MA: MIT Press. 1998.

[17] B. Tomlinson, B. Blumberg, and B. Rhodes. *How Is an Agent Like a Wolf*? In International ICSC Symposium on Multi-Agents and Mobile Agents in Virtual organizations and E-Commerce(MAMA), Wollongong, Australia. 2000.

[18] B. Reeves, and C. Nass. *The Media Equation. How People Treat Computers,Television, and New Media Like Real People and Places*. Cambridge University Press. 1998.

[19] R. Prada, and A. Paiva. *A Believable Group in the Synthetic Mind*. AISB'2005 - Joint Symposium on Virtual Social Agents: Mind-Minding Agents. 2005.

[20] A. Paiva, G. Andersson, K. Hook, D. Mourao, M. Costa, and C. Martinho. Sentoy in FantasyA: Designing an Affective Sympathetic Interface to a Computer Game. In *Personal and Ubiquitous Computing* vol. 6. pages 378-389, 2002.

[21] J. Jonides. *Further towards a model of the mind's eye's movement*. Bulletin of the Psychonomic Society, 1983.

[22] J. Theeuwes. *Visual selective attention: A theoretical analysis*. Acta Psychologica 83, 1993.

[23] R. Desimone, and J. Duncan. *Neural Mechanisms of Selective Visual Attention*. Cambridge CB 2 2EF, England, 1995.

[24] A. Treisman, and G. Gelade. *A feature-integration theory of attention* . Cognitive Psychology, 1980.

[25] M. Thiebaux, B. Lance, and S. Marsella. *Real-Time Expressive Gaze Animation for Virtual Humans*. In Proc. of 8th Int. Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS 2009). May 2009.

[26] E. Styles. *The Psychology of Attention*. Psychology Press, 1995.

[27] M. Imai, T. Ono, and H. Ishigum. *Physical Relation and Expression: Joint Attention for Human-Robot Interaction*. IEEE, 2003.

[28] T. Yonezawa, H. Yamazoe, A. Utsumi, and S. Abe. *Gaze-communicative Behavior of Stuffed-toy Robot with Joint Attention and Eye Contact based on Ambient Gaze-tracking*. Proc. of ICMI07, 2007.

[29] I. Poggi, and C. Pelachaud. *Signals and Meanings of Gaze in Animated Faces*. In 8th International Workshop on Cognitive Science of Natural Language Processing. Galway, 1999.

[30] S. Khullar, and N. Badler. *Where to look? Automating Attending Behaviors of Virtual Human Characters*. Proc. 3rd Annual Conf. Autonomous Agents, Seattle, WA, USA, May, 1999.

[31] R. Vertegaal, R. Slagter, G. Veer, and A. Nijholt. *Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes*. In Conference on Human Factors in Computer Systems, United States, 2001.

[32] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung. *GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Direction*. In Proceedings of the SIGCHI Conference of Human Factors in Computing Systems (CHI'03), ACM Press, 2003.

[33] A. Brooks, J. Gray, and G. Hoffman. *Robot's Play: Interactive Games with Sociable Machines*. MIT Media Laboratory, 2004.

[34] S. Marti, and C. Schmandt. Physical embodiments for mobile communication agents. In *Proceedings of UIST'2005*, Seattle, Washington, October, pp. 23-26, 2005.

[35] R. Picard. *Affective Computing*. MIT Press, Cambrige, MA, 1998.

[36] H. Sharp, Y. Rogers, and J. Preece. Interaction Design: beyond human-computer interaction. 2rd edn. John Wiley & Sons, Ltd. 2007.

[37] M. Liljedahl, and S. Lindberg. DigiWall - an audio mostly game. *Proceedings of the 12th International Conference on Auditory Display*, London, UK. 2006.