UNIVERSITY OF SALERNO

Faculty of Mathematical, Physical and Natural Science

Master Thesis in Computer Science

Automating the Idle Gaze of Virtual Humans in Waiting Situations

Supervisor
Prof. Hannes Högni Vilhjálmsson
Assistant Supervisor
Prof. Vittorio Scarano

Author
Raffaele Gaito

Academic Year 2008-2009

"...there is nothing new under the sun."
Ecclesiastes 1:9
"...non v'è nulla di nuovo sotto il sole."
Ecclesiaste 1:9

Acknowledgments

There would be to many people to thank. Here I just want to concentrate on who made possible the realization of this work.

Thanks to Hannes, he was not just my teacher and my advisor but like a friend helping me every moment during this fantastic experience. I'm very glad to met him and I really hope to keep in touch with him forever.

Thanks to the Reykjavik University and all the CADIA team for the support and the company during my work in Iceland and in Italy. It's the ideal environment to study and work.

Thanks to the Erasmus Project and the University of Salerno for making possible to live and study six month in Iceland, this amazing island.

Thanks to Claudio Pedica for the infinite suggestions and for the great collaboration on this project.

Thanks to Angelo Cafaro for sharing every single moment related to this work, from the first video recorded to the last line of code written.

Thanks to God because He gave me the opportunity to live and enjoy this experience.

Contents

Chapter 1

Introduction

1.1 Scenario

In the world we live, the importance of virtuality is growing exponentially. The large proliferation of more powerful computers, consoles and broad-band Internet, is contributing to a greater variety of virtual environments.

In particular, in MUVEs[1] like the MMORPGs[2], it is important to create believable situations that simulate depth and engagement of real life interaction as well as possible. This realism is obtained by improving different aspects of the virtual environment.

(a) EVE On-line - Copyright by CCP Games

(b) Star Trek On-line - Copyright by CBS Studios

Figure 1.1: Examples of modern MMORPGs.

In multi-user environments there is a lot of interaction between all characters, both avatars (managed by real users) and agents (managed by the system). Their behavior is probably one of the most important aspects to improve in order to increase the users' feeling of being surrounded by socially intelligent beings [Gillies and Dodgson, 2004]. Social human behavior is very complex and strongly related to the social situations and some interrelational and environmental factors. The gaze is one of the most important

[1] Multi-User Virtual Environment

[2] Massive Multiplayer Online Role-Playing Game

behaviors because it conveys awareness and attention as well as playing a role in expressing emotions, feelings and personality among other things. etc ... [Gillies and Dodgson, 2002]. Due to the high number of behaviors and the difficulty for a user of an avatar to manage of all them, it is important to leave some of them to autonomous control [Pedica and Vilhjálmsson, 2008]. For example, every avatar and agent needs to show social intelligence by reacting spontaneously to the social environment and situation.

In this work I want to focus on one specific behavior in a specific situation: gaze behavior of people waiting for a bus. In this situation I chose subjects that are idling alone to minimize the external factors influencing the behavior. Before proceeding is important to explain that with the term "alone" I mean: people surrounded by other persons but not in communications with them.

In order to achieve a good result in automating social awareness in avatars and agents, I used sociological studies to analyze the actual human behaviors and transpose them into the virtual environment.

1.2 State of the Art

State of the art is reviewed here in two categories: virtual environments and automation of gaze.

Regarding the virtual environments, it's immediately noticable that they have started to look very advanced and well realized from the graphical point of view (e.g. Eve OnLine[3], Star Trek OnLine[4], etc ...) and from the social functionality point of view (e.g. Second Life[5], The Sims[6], etc ...) but there is still a lot lacking in social realism and naturalness. Even when users can control the avatar inside the virtual world and perform some social action (like play games, go to a bar, talk with others, etc ...). The nonverbal behavior of those avatars is almost null, they look more robots than humans: the micro-movements seem to be random and repetitive, especially in situations with groups of avatars.
It's well studied that it is not just visual realism or the complexity of interaction that matters for engaging the users, but the naturalness in the behaviors (gaze in particular) is very important for increasing the user's involvement during the simulation [Es et al., 2002]. The paper "Making agents gaze naturally - Does it work?" is a perfect example of the importance of the gaze. They made experiments comparing three version of an agent: one with faze behavior that is typically found to occur in human-human dialogues, one with gaze that is fixed most of the time, and a third version with random

[3]http://www.eveonline.com
[4]http://www.startrekonline.com
[5]http://www.secondlife.com
[6]http://thesims.ea.com

gaze behavior. The experiments were conducted on 48 participants and the results show that the optimal version (with natural gaze) was found to be significantly more efficient than the suboptimal version (with fixed gaze) and the random version.

(a) The sims - Copyright by Electronic Arts

(b) Second Life - Copyright by Linden Research

Figure 1.2: Examples of modern MUVEs.

Regarding previous research into automating gaze, there are several studies that cover different ways to generate it. I will go into more depth in chapter 2 as I review the related work, but lets consider some interesting ways to implement gaze here. With the BodyChat avatar-based chat system [Vilhjálmsson and Cassell, 1998] made an automated visual attention system for the avatars based on rules. The avatars are able to generate behaviors, such as gaze, head movements and eyebrow movement, based on rules extracted from the literature about human communicative behaviors. Some examples of these researches are:

Argyle and Cook (1976) "Gaze and Mutual Gaze"
: This work was used mainly to study the behaviors during turn, listener and speaker exchange. And information management with body movements.

Goffman (1983) "Forms of Talk"
: This work was focused more on the process of the interaction management and there are some important theories about communicative behaviors.

Kendon (1990) "Conducting Interaction"
: Here the focus is on the field of attention and the use of gestures and body orientation during an interaction.

The system developed by Gillies and Dodgson [Gillies and Dodgson, 2004] also implements autonomous behaviors (gaze included) for avatars and agents in virtual environments based on informal observations of human behavior. The main difference to my system is that they focused on non-social situations rather than social ones.

A very complex framework was made by [Chopra-Khullar and Badler, 1999] where they realized a set of parameters to manage attention, a hierarchy of eye behavior and other components. My work is highly related to this work because I went deeper into the "Spontaneous Looking" case, which was one of the cases they modeled.

There are other recent works related to gaze behaviors (i.e. [Pennock et al., 2005] [Shao and Terzopoulos, 2005]) but none of them use observed patterns and statistical data obtained from video studies.

1.3 Contribution

As mentioned in the previous paragraph, the main difference, between this work and the previous literature, is the approach for generating the gaze behaviors. I used the sociological literature in order to study and understand how we behave, but I also observed persons in the real world to confirm the information gained from the literature and to collect some new empirical data.

In my work I design the gaze behavior using those statistical data and general patterns to achieve gaze behavior in all its complexity. As we will see in chapter 3, every moment the humanoid has to make a choice (like real person) and this choice is made using this information.

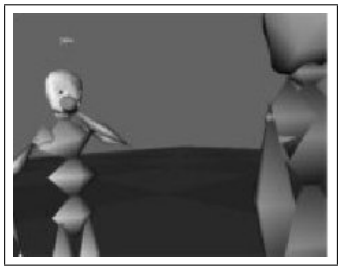
This is important because it prevents the gaze from being fully generalizable but relates it to the social situation (people idling in my case [Argyle et al., 1981]) and to personal factors (cognitive activity in my case [Lee et al., 2007]).

Chapter 2

Related Works

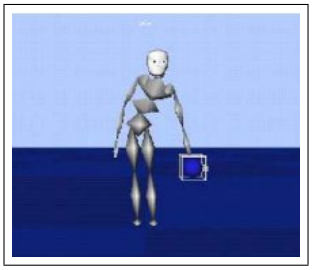
The previous literature about human behaviors, gaze and visual attention is very thick. I selected some of the most recent and important work related to mine. The overview of them follows a chronological order.

(a) [Vilhjálmsson and Cassell, 1998]



(b) [Gillies et al., 2002]

(c) [Lee et al., 2007]



(d) [Gillies and Dodgson, 2004]

Figure 2.1: Screenshot of some related systems.

In 1998 Vilhjálmsson created a visual attention system for avatars based on communicative rules from the social literature [Vilhjálmsson and Cassell, 1998]. This work is important because was one of the first attempts to represent human knowledge in the avatar's behaviors. The BodyChat system allows users to communicate via text while the avatars show natural behaviors in order to increase the credibility and the naturalism of the interaction.

The great social background is an important feature of the Vilhjálmsson and Cassell's work: the deep studies about conversational rules, human behaviors and communicative facial expression made a solid base for the design and realization of the BodyChat system. This is the same kind of study and research that I conducted during my work. I build on the same theory, but add my own empirical results to fill into gaps of their system, namely the automation of believable idling behavior.

Another important work is the framework made by Chopra-Khullar and Badler [Chopra-Khullar and Badler, 1999]. In their complex system, they built an architecture able to generate eye and head motions based on some factors like: visual search, eye limits, input parameters, type of agents, etc. That system generates different kinds of gaze related to the execution of certain tasks. My work is related to this one because I went deeper into the "spontaneous looking" case, which they covered, and studied it in a defined social situation. In my approach I used some of their concepts but with different implementations. The attention, for example, is realized in their work using psychologists theories mixed with image processing (to discover significant objects). In my work the attention system is more "social" because is based on the proxemics areas introduced by anthropologist Edward T. Hall in 1966. Moreover, they start a sponteneous looking only when there are not event important for attention. In my model, the subject is continuously under a "Spontaneous Looking mode". It means that, even if he is engaged in a conversation, everything appends around him could attract his attention. I also used a primitive short-memory system in order to remember the last events (objects or persons seen) that happened.
The implementation of these factors draws on empirical and qualitative observations known from human factors and psychology literature. This background was fundamental for designing concepts like *distractability*, *intentional shifting*, *stimuli*, etc.

A more generic attention behavior system is the one made by Gillies and Dodgson [Gillies et al., 2002] where they simulate the secondary behaviors (a behavior that is generated autonomously) for an avatar in order to produce a more complex behavior and a truly expressive avatar. The work is very interesting because every secondary behavior is related to the actions that the user is controlling (the primary behavior). Every primary behavior sends some informations (tags) to the secondary behavior system that manages them and generates the secondary behavior.

Peters and Sullivan also did interesting work in [Peters and Sullivan, 2003] where they realized a bottom-up visual attention system for virtual humans. This framework is made up of different components:

- a synthetic vision system for perceiving the virtual world, this system is based on a network of neurons to continuously calculate a saliency map (the most salient locations in the scene);

- a model of bottom-up attention (very similar to the model used in my work) for processing the data received from the vision system and from other stimuli;
- a memory system for the storage of previously sensed data (in my work there is a very simple version of it);
- a gaze controller for the generation of the final gaze behavior.

This work was one of the main reference models for me.

In 2004 Gillies and Dodgson made another important step in the work "Behaviourally rich actions for user-controlled characters" [Gillies and Dodgson, 2004]. They build an autonomous behavior system based on their previous work [Gillies et al., 2002]. They introduced, moreover, different types of required attention (immediate and monitor) and peripheral vision. The main difference between my work and this one is that while they focused on non-social situations using informal observations, I concentrated on specific social situations and I based the gaze behavior generation on video studies in order to achieve results as realistic as possible.

Another important work to point out is the Rickel Gaze Model [Lee et al., 2007]. The main reason I include this work here, is its thorough analysis of cognitive operations in face-to-face human interaction. They argue that it is not enough to merely imitate a person's eye movements. The gaze behaviors should reflect the internal states of the virtual human and onlookers should be able to derive those states from observing the visual behaviors.
The gaze model of their system is based on the one developed by Jeff Rickel and extended by the authors. In my work I also took into account the cognitive activity during the modeling phase in order to relate the gaze to the personal factors as we will see in chapter 3. Another important common point with my work is the definition of some gaze properties that can be specified: type, speed, target, reason, priority, etc.

Finally and more recent is the SBGC[1] system made by [Thiebaux et al., 2009]. This system is based on input parameters to generate various types of gaze, but the main point of this work is character animation through parameterizable joints. In fact, the framework, provides a highly flexible and reusable interface to the gaze behavior, applying motion calculations to a selected set of skeletal joints.
The main common parts between this work and mine is the classifications of gaze movement produced by the model and the gaze warping transformations. As we will see in chapter 5 I also describe different gaze movements (eye-only shift, eye-head shift, eye-head-torso shift) and some types of warping transformations (head posture, torso posture, movement velocity).

[1] SmartBody Gaze Controller

Chapter 3

Methodology

3.1 Overview

It is well established that social situations have a great impact on all aspects of behavior [Argyle et al., 1981]. A situation could be described as the sum of the features of a social occasion that influence an individual person. However this description does not allow for the fact that a person contributes to the situation himself, so a better description might be the sum of the features of the behavior system, for the duration of a social encounter [Argyle et al., 1981]. The main idea behind my methodology is the following: people behave in different ways (in particular they express different kinds of gaze behavior) according to the social situation in which they are. Moreover, fixing one social situation, people still have different gaze behaviors because there are some factors that directly impact and modify their behavior. The combination of all these factors, in a defined social situation, influences and causes the appropriate gaze behavior.
With this picture in mind, now, we can provide the model reading key:

*"A combination of a **Social Situation** and a number of **Factors** influence people's **Gaze Behavior**, consisting of eye, head and torso movements. The amount of time in which these factors are present, in the given social situation, may vary from one factor to another."*

Due to the high number of combinations between social situations and impact factors I focused on a particular configuration that is very common:

Idle Gaze Waiting
Social Situation: idling "alone" (waiting in a bus station);
Principal Factor: cognitive activity.

The combination of the factors in the defined social situation produces body movements that involve eyes, head and torso [Thiebaux et al., 2009].
In the next three sections I will give a little explanation of what I mean by

"Social Situations", "Personal State Factors" and "Human Behaviors" with an overview and some examples of each one.

3.2 Social Situations

The main part of the study is the choice of a social situation to analyze. Let's describe what I mean by Social Situations using a citation from Goffman:

> *"By the term social situation I shall refer to the full spacial environment anywhere within which an entering person becomes a member of the gathering that is (or does then become) present."* [Goffman, 1967, p. 144]
>
> Erving Goffman.

The everyday human life is full of social situations, some examples are:

- Idling alone in a relatively "non" social situation;
- Having a conversation with one or more people;
- Walking through the environment, filled with people;
- Being visually attended to by another person;
- Watching some public display that requires attention, with or without other people;
- Engaging in greeting and farewell rituals;
- Talking with someone while performing a task (cooperative or competitive interaction);
- Being involved in an emergency situation;
- Etc.

Considering that accounting for all pf the situations is impossible, the best way to proceed was to choose a single social situation and to go into more depth with real-world studies. The social situation chosen in my work is "Idling alone" and, in particular, "Waiting alone" for a bus.
The bus station is just a common place that is full of people involved in this situation. But the general case study is useful for every kind of "waiting situation" with people alone (i.e. waiting in a shop, waiting for the doctor, etc.). Moreover, this situation is highly related to the personal state factor chosen and the human behavior being studied.

3.3 Personal State Factors

We mentioned above that in the same social situation people may vary their gaze behavior in many different ways that are influenced or caused by a high number of *factors*. Some examples of personal state factors are:

- Personality;
- Mood;
- Emotion;
- Social Role;
- Target of Attention;
- Cognitive Activity;
- Etc.

One of these factors is **personality**:

> *"The Personality can be defined as a dynamic and organized set of characteristics possessed by a person that uniquely influences his or her cognitions, motivations, and behaviors in various situations."*
>
> Richard Ryckman.

Argyle et al. in [Argyle et al., 1981] affirm that personality-situation interaction concerns with situations and, for example, behavior is at least as much determined by the interaction between situation and personality as by general traits of personality.

The **mood** is well explained by the psychologist Robert Thayer in the study "The biopsychology of mood and arousal" where he describe it as:

> *"A mood is a relatively long lasting emotional state. Moods differ from simple emotions in that they are less specific, less intense, and less likely to be triggered by a particular stimulus or event."*
>
> Robert Thayer

Emotion also plays an important role in this context:

> *"Complex reactions that engage both our minds and our bodies. These reactions include: a subjective mental state, such as the feeling of anger, anxiety, or love; an impulse to act, such as fleeing or attacking, whether or not it is expressed overtly; and profound changes in the body, such as increased heart rate or blood pressure."*
>
> R. Lazarus and B. Lazarus.

With the term **Social Role** I mean a set of connected behaviors, rights and obligations as conceptualized by actors in a social situation. It is an expected behavior in a given individual social status and social position.

The **Target of Attention** is the object or the person looked at by the subject during the studies. Important studies about this psychology field have been made by Michael I. Posner, editor of numerous cognitive and neuroscience compilations, especially in the works "Attention and the detection of signals" [Posner, 1980a] and "Orienting of attention" [Posner, 1980b].

In my studies the control of all these factors was impossible (due to the unawareness of the camera by the subjects) and therefore it was impossible to understand exactly which factors were influencing the gaze behavior of the subjects. The only factor present for sure in every human action is the cognitive activity.
As explained in [Lee et al., 2007], people never stop to think. They think during conversations with others and when alone. At every moment, the brain is involved in some kind of background activity. This is perhaps one of the most important base-line contributors to gaze behavior. These are the main reasons why I choose to consider the cognitive activity as the Personal State Factors for my studies.

3.4 Human Behaviors

In chapter 2 I already pointed out some important research work about human behavior
Of course, it is already obvious that the human behavior analyzed in this work is the gaze. Let's see first a small overview of human behaviors and some example of it.

"Human behavior is the population of behaviors exhibited by human beings and influenced by culture, attitudes, emotions, values, ethics, authority, rapport, hypnosis, persuasion, coercion and/or genetics.
The behavior of people (and other organisms or even mechanisms) falls within a range with some behavior being common, some unusual, some acceptable, and some outside acceptable limits. In sociology, behavior is considered as having no meaning, being not directed at other people and thus is the most basic human action. Behavior should not be mistaken with social behavior, which is more advanced action, as social behavior is behavior specifically directed at other people. The acceptability of behavior is evaluated relative to social norms and regulated by various means of social control."

Wikipedia

There are a lot of communicative/social behaviors that spontaneously occur during social interaction. The literature is also full of example of

them:

Facial Expression
This is, probably, one of the most recurrent behavior during social interaction, the work of Argyle and Cook [Argyle and Cook, 1976] explain well how the people give feedback using facial expression while receiving informations over the speech channel.

Mutual Glances
As described in [Cary, 1978], mutual glances are foundamental during an initiation of of a conversation in order to set the type of conversation (formal, informal, etc ...), set the distance, the first speaker and the next turns.

Sounds and Pitch
Chovil in the paper "Facial Displays in Conversation" [Chovil, 1992] remarks the importance of the facial expression and, analyz them in combination with the voice pitch. People give and request feedback unconsciously and continuously with these two behaviors.

Hands Movements
Another important and common communicative behavior is the use of the hands. In [Kendon, 1990] there are some studies and example of how people move the hands in combinations with gaze movements as well as speech. It is also observed how people often shake the other's hands before to start a conversation.

It's easy to see that, similar to the personal factors, human behaviors are many and impossible to consider them all. In my study, due to my interest in studying the human gaze, the obvious choice are the three behaviors that relate to it:

- Eye movement;
- Head movement;
- Torso movement.

Erroneously, you might think that gaze is related only to eye movement, but, as we will see in chapter 4, for each session. I took into account all movements of those three body parts, which are necessary to realize the gaze behavior [Thiebaux et al., 2009].

Chapter 4

Video Studies

One of the main goals of this work was to realize gaze movements based on statistical/empirical data gained from video studies of naturally occurring behavior in public places. In order to obtain these data, I performed the following steps:

1. I video recorded people in a bus station (matching the chosen study situation);
2. I analyzed all the video frame-by-frame with a video annotation software called Anvil[1] (see fig. 4.4 [Kipp, 2001]);
3. I analyzed the obtained annotations in order to extract statistical data and common patterns.

4.1 Description

In this section I will refer to the study subjects using the abbreviation Ss.

For my studies, I first chose an S and then I recorded him or her for a duration between 60 sec and 120 sec in a public area. The public area choosen for this study was the Hlemmur bus station (the main bus terminal in Reykjavik). The camera was placed 30 meters away from the bus station as depicted in fig. 4.1 and 4.2. Each subject was recorded with two cameras: one with fixed zoom on the eyes and another with a large field of view to record the head and body movements. The two cameras used were:

- A Sony DSR-PD170P DV-VCR video camera (placed on a tripod) focused on the S's eyes;
- A Canon IXUS 900ti photo camera (capable to record video) for the S's surroundings.

[1] http://www.anvil-software.de/

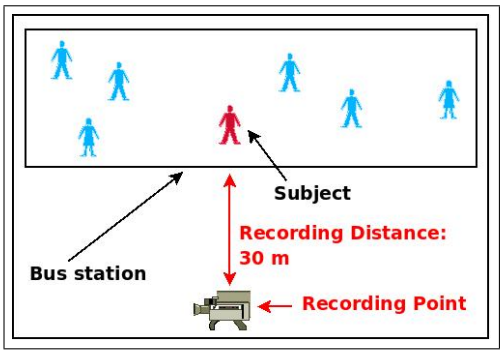


Figure 4.1: The setup of the study

Figure 4.2: Hlemmur bus station.

In my study setting there are some environmental factors influencing the results, and it's important to take them into account. These factors are:

- Videos were recorded in Reykjavik, Iceland:
  - The data represents the culture of this particular country, social norms and habits may differ in other countries;
  - The season was winter (November) so there wasn't too much light[2] and the temperature was low (around zero Centigrades), so peoples' posture was also affected by this factor;
- The Ss were for the most part unaware of the studies. Camera was recording them in a public setting to obtain the most natural behaviour possible. Utmost care was taken to keep the identities of Ss hidden and

[2]Also impacting the video quality.

privacy protected. No-one other than the experimenter had access to the video data and identity has been removed from public snap-shots.

I recorded 9 subjects for a total duration of 14,30 minutes, during which 142 gaze shifts were observed. All the study was recorded in the same day (from 09.00 AM to 16.00 PM). The bus station was not really full of persons, but for each study there were other people on the scene with the subject. There was a normal bus traffic and also cars passing for almost all the time.

More details about each video study are available in appendix A with snap-shots, accurate descriptions and summarized tables.

4.2 Analysis and Annotations

After the video recording sessions I started to analyze frame-by-frame all the collected video and I annotated, for each case, the following information[3]:

- Eyes direction;
- Head orientation;
- Torso facing;
- Target of attention.

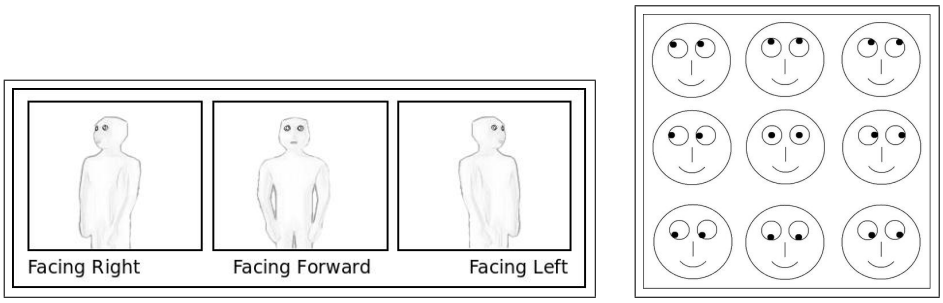
In order to describe the direction faced by the subject's torso I used three different values: facing left, facing forward and facing right as is shown in figure 4.3. To establish one of these values I took into account the direction of the torso, during the studies, relative to the camera from the S point of view. For the head orientation I used nine different values: up left, up center, up right, center left, center center, center right, down left, down center and down right. In each pair the first value represent the direction on the vertical axis and the second one the direction on the horizontal axis. In order to choose the correct value I made the head orientation relative to the torso facing. For the eyes directions I used the same set of values, but the direction is relative to the head orientation.

I also have information about the duration. Moreover, only for the eyes, head and torso I annotated:

- Time interval in which movement happens;
- Speed of that movement[4];
- Potential gaze targets for the Ss (as described below, categorized as objects or persons);

[3]Referring to the focus subject of each recording.

[4]I used three approximate values based on empirical observation: slow, neutral and fast



(a) The three directions used for the torso

(b) The nine directions used for the head orientation and eyes directions

Figure 4.3: Directions used for torso, head and eyes.

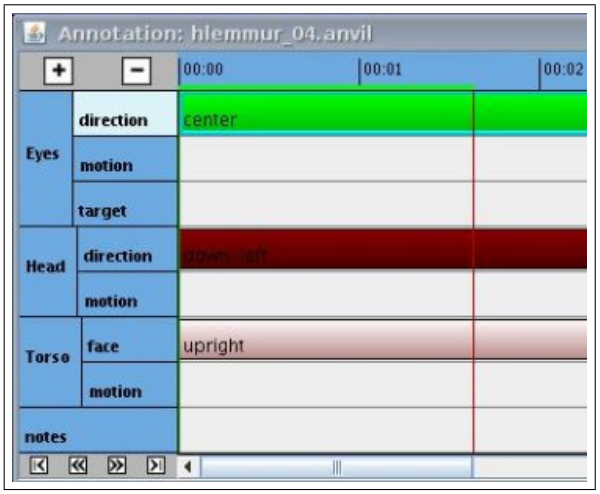
- Proxemics area in witch the potential target appears [Hall, 1966];
- Whether the potential target in the given area produced actual gaze movement towards that target.

With the term potential targets I mean all the objects and persons present in the scene in that moment, that reflect some characteristics in order to be a potential gaze target. Some of those characteristics are fixed whereas others are contextualized to the scenario. As we will see in chapter 5, the use of visual perception and proximity was the base of the targets detection. For example, a person close to the subject but positioned behind him was not considered as potential target because it was out of the visual field (both central and peripheral). In general, it's important to know that a target was considered potential only if the subject was able to notice it. Another important characteristic used to identify these potential targets was movement with a priority for the targets moving towards the subject.

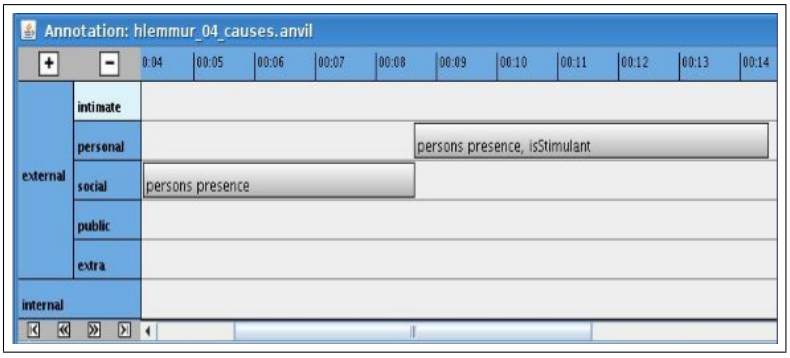
The potential targets were categorized into persons and objects because I expected different behavior from the subjects towards these two kinds of targets. At the end of this chapter we will see some statistical data that seems to confirm this expectation. When the subjects looked towards the camera, I considered that case an "object target" due to the fact that the camera was far away from the subject and typically was hidden, and the subjects seemed more or less unaware that someone was recording them.

Once collected, these potential targets, as we will see in chapter 5, help determine what typically gets looked at in a scene, giving rise to an automated decision process.

I was interested the proxemics areas because different kinds of social interactions seem to occur at different ranges according to [Hall, 1966]. Proxemics is the study of set measurable distances between people as they interact. He said:



(a) Directions analysis



(b) Causes analysis

Figure 4.4: Screenshots of Anvil

"Like gravity, the influence of two bodies on each other is inversely proportional not only to the square of their distance but possibly even the cube of the distance between them."

Edward T. Hall.

Hall lists 4 different areas at different distances[5]:

1. Intimate distance (15 to 46 cm)
2. Personal distance (46 to 120 cm)
3. Social distance (120 to 370 cm)
4. Public distance (370 to 760 cm or more)

Taking these distances into account is important in order to simulate realistically how people react to the social environment around them.

[5]This distance may change with different cultures.

4.3 Study Results

At the beginning, the focus of the video annotations was to understand where Ss generally look (which direction) and what Ss look at (target of attention) and for how long (gaze duration). After analysing the videos I also discovered some more general patterns of gaze behavior, some corresponding to the literature, which I incorporated into our implementation as well. These observed patterns include:

1. Ss that produce short glances were observed to keep their glances short throughout the study;
2. Conversely, Ss producing longer glances or gazes were observed to keep them long throughout the study;
3. The shorter glance Ss were observed to pick many different targets around them;
4. The longer glance Ss were picking targets from a narrower set;

In figure 4.5 we can see for each S which common patterns were observed. Due to the long duration of every video, it's common to see that in some sessions there is more than one pattern.

		Video Sessions								
		1	2	3	4	5	6	7	8	9
Pattern	1			x	x		x	x		x
	2	x	x		x	x	x		x	
	3			x						x
	4	x	x		x			x		

Figure 4.5: An "x" denotes that the given pattern (vertical axis) was observed in the given session (horizontal axis).

In addition, using more detailed analysis I extracted numerical data describing: the gaze attraction of potential targets that were either objects or persons, in relation to the proximity of the target. This is based on how much of the total time an object or a person stayed at a given proximity was spent being looked at by the subject. In addition the duration of each gaze was recorded. This is shown in Table 4.1.

Table 4.1 shows for each proximic area and for each type of target three values:

1. Time on Target: indicates the time spent by the subject looking at a potential target out of the total time that target was in that area;
2. %: indicates the percentage value of the Time on Target;

	Objects			Persons		
Prox. Area	Time on Target	%	Avg. Dur.	Time on Target	%	Avg. Dur.
Intimate	84.11 out of 84.11	100%	6	0 out of 6.43	0%	-
Personal	10.26 out of 10.26	100%	5.13	7.3 out of 61.16	12%	2.43
Social	14.36 out of 22.56	64%	4.80	45.17 out of 96.68	47%	22.60
Public	1.33 out of 1.33	100%	1.33	6.90 out of 29.46	23%	2.30
Extra	5.90 out of 8.66	68%	5.90	9.00 out of 35.56	25%	3.00

Table 4.1: Observations about the relationship between gaze targets and proxemics. All durations are in seconds.

3. Avg. Duration: indicates the average duration (in seconds) of each gaze.

From this table and figure 4.6 we can draw some interesting conclusions:

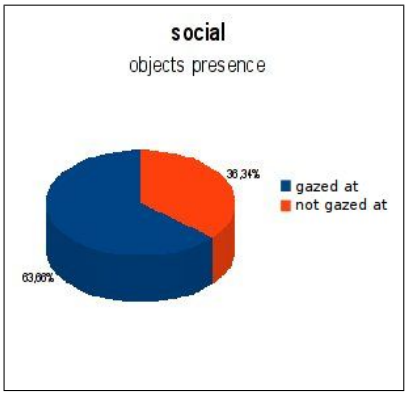
1. The objects are more looked at than the persons (both in social and public areas, as we can see also from figure 4.6). This most likely happens because when we look to other people there are factors (shyness, shame, etc ...) influencing the gaze while with objects we don't have any of these.

2. We have the impression that the initial hypothesis described in section 4.2 (that objects and persons were looked at in different ways) is good and, in fact, we can clearly see, for each area, a very different percentage value.

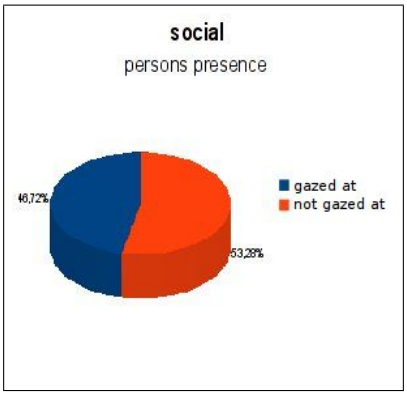
3. When the target is an object and it's really close to the subject (intimate or personal area) the percentage value is 100 %. This is an interesting value because, probably, this is the case when the subjects are holding something in the hands (looking at the phone, reading a book, etc ...) or are really interested in a particular object (reading the bus time table, etc ...).

4. When a person is in the intimate area of a subject the subject never looks at him or her. This could happen because, as Hall said in [Hall, 1966], it's unbecoming to keep this distance in a public place. Maybe people feel unconfortable and they try not to look at the other person.

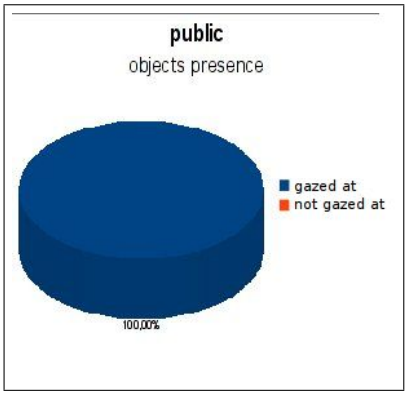
5. On the other hand, we can see that people in the social area receive a long gaze duration. This could happen for the same reason as before: this is the distance that normally people keep for the occasional encounter.



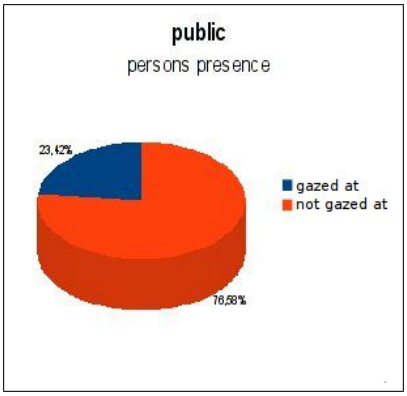
(a) Objects in social area



(b) Persons in social area



(c) Objects in public area



(d) Persons in public area

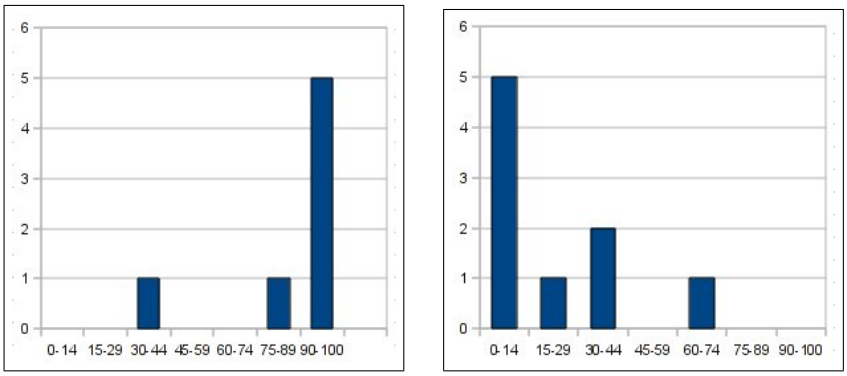
Figure 4.6: Different percentage values of received gaze between objects and persons.

4.4 Statistical Data Analysis

In order to establish whether there truly is a difference between the way people gaze at objects and at persons, I performed some additional statistical analysis. First I created two data sets from my data, summarizing the amount of gaze that objects and persons received across the different video sessions. these data sets are shown in Table 4.2 and shown visually in two histograms in Figure 4.7. I executed two different kinds of tests (t-test and

Session	Objects	Persons
1	-	0.00%
2	85.60%	9.20%
3	100.00%	67.40%
4	100.00%	41.80%
5	100.00%	0.00%
6	23.60%	9.70%
7	100.00%	41.00%
8	-	3.70%
9	100.00%	25.30%

Table 4.2: Average values of Percentage of Time on Target for each session.



(a) Dataset 1 - Objects (b) Dataset 2 - Persons

Figure 4.7: Histograms representing the distribution of the datasets.

z-power) in order to establish the probability that the apparent difference is accidental and the power of my data since the number of samples is relatively low (N=9). Before showing the result of these tests, I first provide some intermediate values essential for the calculation:

Average Values

Objects: 87.03%

Persons: 22.01%

Standard Deviations

Objects: 28.48%
Persons: 23.56%

With these two intermediate values I can show now the results of the two tests:

Paired t-test: t=5.45, $p < 0.002$ (probability of no difference)

Z-power: 0.99

Both the results are good. The t-test with a p-value so low[6] indicates the low probability that the difference that I observed was accidental. The z-power being so high[7] indicates enough statistical power in the datasets used, in spite of a low N. However, even if the results are very good, these kinds of studies should be supported with higher number of sessions to provide a more detailed analysis, for example to study the effect of distance. In fact, as we will see in the last chapter, one of my main goals in future works is to record more videos in order to achieve additional data and execute a more complex and deeper statistical analysis.

[6] For this kind of test it is considered good when $p < 0.05$

[7] 1 is the maximum value.

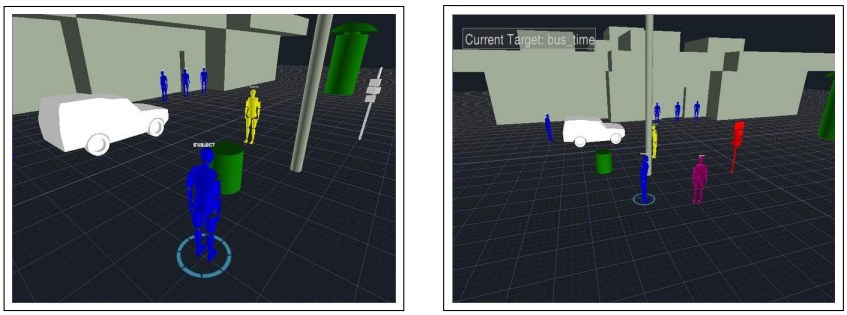
Chapter 5

Autonomous Generation of Idle Gaze Behavior

5.1 Approach

I implemented the idle gaze generation within a tool for implementing social behaviors for avatars and agents in game environments: CADIA Populus [Pedica and Vilhjálmsson, 2008]. The process was simply a matter of plugging in the new behavior along with conditions to activate it during the particular situation that I was modeling. The new behavior fit nicely into the steering behavior framework of CADIA Populus, adding a new set of motivations for turning the head and eyes when appropriate.

CADIA Populus is a tool that combines full on-line multi-player capability with clear visual annotation of the social environment. Game developers can drive avatars around and simulate social situations (conversations, lines, etc ...) between them using the set of built-in tools (i.e. text chat, perception system, etc ...). Constructing an environment and manipulating the social situation is made very easy for the developer. Screenshots from the CADIA Populus environments are shown in figure 5.1. CADIA Popu-



(a) (b)

Figure 5.1: The "Hlemmur" environment inside CADIA Populus.

lus supports starting a behavior with given pre-conditions and updating it with a fixed frequency. In my case the only necessary pre-condition for the activation of the behavior was the idle waiting situation, which I simplified to an avatar standing still and not engaged in a conversation. The update frequency was 20Hz.

5.2 General Process

The general process behind the implementation of my study is showed in figure 5.2. This process supports having at each moment a target (object,

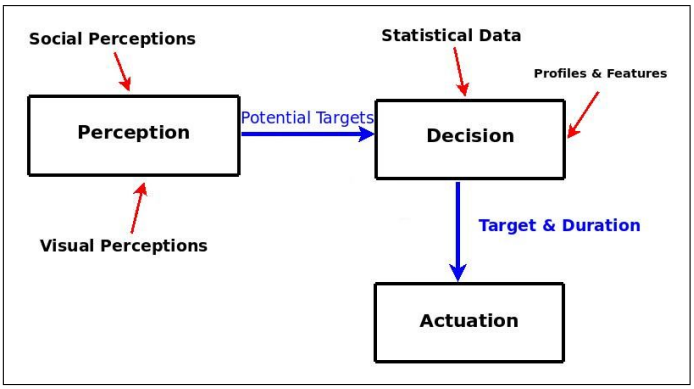
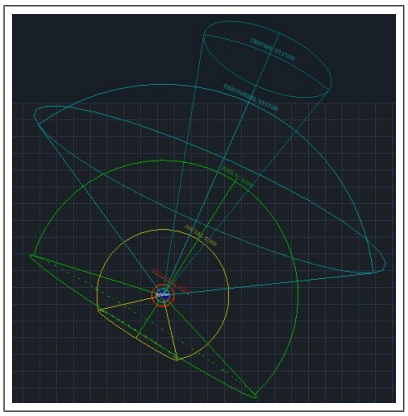


Figure 5.2: The general process diagram.

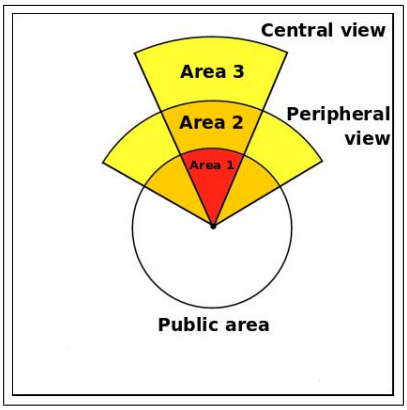
person, direction) and a duration. This process is split in three main parts:

Perception

The first step uses the perception system provided by CADIA Populus in order to obtain all the potential targets in the scene. The perception system is composed of the visual perception (using central view and peripheral view) and social perception (using the proxemics areas). Notice that the perception of social and public space has a blind cone behind the avatar, of respectively 90 and 150 degrees [Pedica and Vilhjálmsson, 2008]. The perception system is depicted in figure 5.3 (a). I used the combinations of these three different zones (central view, peripheral view and public area) in order to create an attention model (figure 5.3 (b)). This model is basically a priority list of potential targets. It is realized using the intersection between the three areas. In order to have the list of the potential targets the system adds into this list all the available entities in the highest priority area. This scan starts from Area 1 and ends as soon as the available entities are found. At this point, only the entities moving toward the avatar present in the remaining areas, are added as they represent high potential for upcoming interaction. A detailed attention model



(a) Perception system in CADIA Populus



(b) Priority areas for the attention model

Figure 5.3: Perception system.

implementation is beyond the purpose of this paper, in fact ours is primarily a way to obtain a reasonable list of targets to work with.

Decision

The decision process is described in detail in the next section. Here I will only say that into this step arrives a list of potential targets. This list is analyzed and, using the video study results (the statistical data, general observations and common patterns) and a minimal profile/features framework. The system chooses a target and a duration in order to actuate, finally, the gaze.

Actuation

The actuation step is the easiest among the three. It is just a system call to CADIA Populus with the two parameters described above (target and duration) to activate all the necessary movements (eyes, head and torso) for the gaze.

5.3 Decision Process

The decision process is the main step of the implementation phase. As mentioned above, this step is responsible for deciding a target and a duration in order to generate the gaze. This decision is made based on four values: *Choice Probability*, *Look Probability*, *Minimum Duration* and *Maximum Duration*. The meaning of these values is the following:

1. Choice Probability: is the probability of a target to been choosed from among all the potential targets;

2. Look Probability: for this single target, we may or may not decide to actually look at it, this decision is left to a Look Probability;

3. Minimum Duration: is the minimum duration of a gaze obtained from the data;

4. Maximum Duration: is the maximum duration of a gaze obtained from the data.

For each target, these four values are obtained combining the type of target (person or object) and the proxemies area of it. Once a target has been chosen and it's going to be looked at, a random value between *Minimum Duration* and *Maximum Duration* is chosen for the effective duration of the gaze.
For example if there are two persons and one object in the Social Area of the subject, the three potential targets get these normalized choice probabilities: 47/158, 47/158 and 64/158 respectively. Suppose person A gets chosen. Now there is the 47% of probability that it actually gets looked at (Look Probability). If this happens, the duration of the gaze will be a random value between 2,80 sec. and 6,80 sec.

With this system, the avatar has a gaze behaviour all the time. There are some cases when the decision process could not select a target for two possible reasons: there aren't potential targets surrounding the avatar or the decision process chose not to produce gaze towards the most likely target (due to the *Look Probability*). In these cases the decisional process generates a *default* gaze behavior with a relative direction based on the discovered common patterns mentioned in chapter 4.

In order to avoid a repeated choice of the same target and to have a more realistic gaze behavior, I also used a minimal short-term memory system of recent targets. The decisional process after choosing a target, checks if it is in this list and, if so, the target is cancelled and the process restarted. This addition avoids the situation where some targets around the avatars seem to be looked at for an unusually long time. The targets are deleted from this list after around 20 seconds[1].

Another important implementation detail, is the introduction of a simple management of avatar profiles and preferences in order to bring in more factors such as the avatar's gender (using the profile) or possible preference bias[2]. While profiles and preferences are typically set up ahead of time, in my system they can also change at run-time, and this is very common and useful to reflect a change in interest based on activity or needs. In this way, it is possible to repeat the same scenario with different preferences or profiles in order to analyze the different results in the gaze behavior.

[1]This is the duration of the Short-term memory in a human.

[2]For example how much he/she likes one kind of object present in the environment, I used a scale of values in the range 0.0 - 1.0 to express that preference.

With this profile system I also simulated some characteristic like the extroversion level of a person to realize more complex aspects of the gaze behavior like the gaze aversion between two persons[3]. Of course, my implementation was a very basic one but the objective was just to show that a profile/preferences system is useful and expandable to suit differents needs.

[3]The gaze aversion is the typical consequence when a mutual gaze happens.

Chapter 6

Results and Conclusion

6.1 Results

As we can see in Fig. 5.1, I have constructed a virtual model of the place where we recorded our video in order to test our steering behavior. In order to better see the target of the currently selected avatar, in the simulations, I highlighted it with a different color (red) for the whole gaze duration.

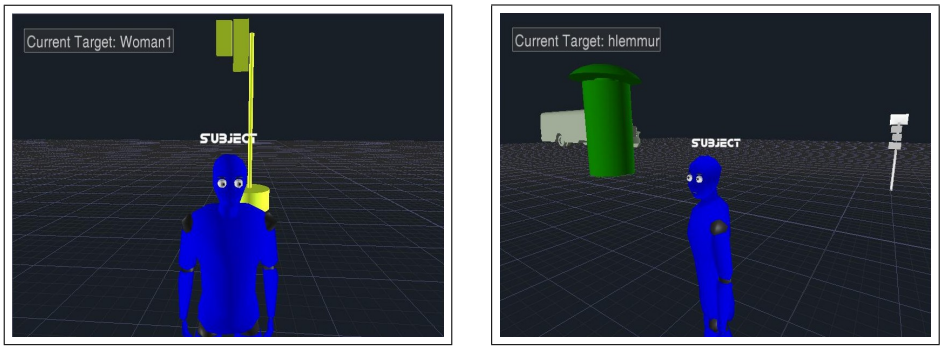


Figure 6.1: Eye gaze results.

In my opinion the new steering behavior inside *CADIA Populus* produced avatar behavior that is relatively natural and realistic in terms of gaze. The common stares that pervade game environments are gone, replaced by much more plausible gaze patterns that bring the social environment to life. I attribute the life-likeness of the results to having both taken into account the existing theory on gaze behavior and to having gathered data on the particular situation that I wanted to model.

Moreover my implementation supports bringing more factors into the gaze decision process, for example through avatar profiles or through features. This can even be modified at runtime, which provides the maximum amount of flexibility and upgradeability.

6.2 Limitations

Some limitations should be considered. First of all, the data was gathered in a single location, which may or may not generalize to other locations or cultures.

Secondly, the videos were gathered in a natural setting so it was impossible to obtain detailed data on many of the personal state factors that we know can influence gaze, for example how the subjects were feeling. I simply assumed the same basic underlying cognitive factor associated with the activity of waiting.

Thirdly, given that only certain kinds of potential targets were present in my study environments, it is not clear how the gaze generation algorithms will handle completely different scenes.

6.3 Future Work

There are many things I would like to do to expand and refine this work and its results.

For example, to achieve greater naturalness in the gaze behavior, I would like to incorporate the *speed* of *head* and *torso movements*. I would also like to add *eyelids* to the avatar models to include the eyes-closing movement.

With regards to the video analysis, the next steps, as indicated in section 4.4, is to record more video in order to acquire more data and execute an exhaustive statistical analysis of some of the more intricate patterns that Table 4.1 may be hinting at. It will be also important to conduct a formal evaluation with a human comparing my system with a random-gaze system and a fixed-gaze system. This will give me much needed feedback from a real user to better understand what level of realism is being obtained and how it compares to the state of the art.

Finally, I think, it is important to start with new studies with different configurations, to cover more *factors* and *social situations* as reviewed in chapter 3.

Bibliography

[Argyle and Cook, 1976] Argyle, M. and Cook, M. (1976). *Gaze & Mutual Gaze*, chapter 5, pages 98–124. Cambridge University Press.

[Argyle et al., 1981] Argyle, M., Furnham, A., and Graham, J. A. (1981). *Social Situations*. Cambridge UniversityPress.

[Argyle et al., 1973] Argyle, M., Ingham, R., Alkema, F., and McCallin, M. (1973). The different functions of gaze. *Semiotica*, pages 19–32.

[Ashida et al., 2001] Ashida, K., Lee, S., Allbeck, J. M., Sun, H., Badler, N. I., and Metaxas, D. (2001). Pedestrians: Creating agent behaviors through statistical analysis of observation data. In *Proceedings of Computer Animation*, pages 84–92. IEEE Computer Society.

[Cary, 1978] Cary, M. S. (1978). The roole of gaze in the initiation of conversation. *Social Psychology*, 41(3):269–271.

[Chopra-Khullar and Badler, 1999] Chopra-Khullar, S. and Badler, N. I. (1999). Where to look? automating attending behaviors of virtual human characters. In *Proceedings of the third annual conference on Autonomous Agents (Agents '99)*, pages 16–23, New York, NY, USA. ACM.

[Chovil, 1992] Chovil (1992). Facial displays in conversation. *Research on Language and Social Interaction*.

[Es et al., 2002] Es, I. V., Heylen, D., Dijk, B. V., and Nijholt, A. (2002). Making agents gaze naturally - does it work? In *Proceedings of Advanced Visual Interfaces (AVI '02)*, pages 357–358.

[Gillies and Dodgson, 2002] Gillies, M. F. P. and Dodgson, N. A. (2002). Eye movements and attention for behavioural animation. *The Journal of Visualization and Computer Animation*, 13:287–300.

[Gillies and Dodgson, 2004] Gillies, M. F. P. and Dodgson, N. A. (2004). Behaviourally rich actions for user controlled characters. *Journal of Computers and Graphics*, 28(6):945–954.

[Gillies et al., 2002] Gillies, M. F. P., Dodgson, N. A., and Ballin, D. (2002). Autonomous secondary gaze behaviours. In *Proceedings of the AISB Workshop on Animating Expressive Characters for Social Interactions*, pages 37–42.

[Goffman, 1967] Goffman, E. (1967). *Interaction Ritual: Essays on face-to-face behavior*. Anchor Books.

[Hall, 1966] Hall, E. T. (1966). *The Hidden Dimension*. Doubleday, Garden City, 1st edition.

[Kendon, 1990] Kendon, A. (1990). *Conducting Interaction: Patterns of Behavior in Focused Encounters (Studies in Interactional Sociolinguistics)*. Cambridge University Press.

[Kipp, 2001] Kipp, M. (2001). A generic annotation tool for multimodal dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology*.

[Kipp, 2003] Kipp, M. (2003). *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. PhD thesis, Saarland University.

[Lance and Marsella, 2008] Lance, B. and Marsella, S. (2008). The relation between gaze behavior and the attribution of emotion: An empirical study. In *Proceedings of Intelligent Virtual Agents (IVA '08)*, pages 1–14.

[Lee et al., 2007] Lee, J., Marsella, S., Traum, D., Gratch, J., and Lance, B. (2007). The rickel gaze model: A window on the mind of a virtual human. In *Proceedings of the 7th international conference on Intelligent Virtual Agents (IVA '07)*, pages 296–303, Berlin, Heidelberg. Springer-Verlag.

[Pedica and Vilhjálmsson, 2008] Pedica, C. and Vilhjálmsson, H. H. (2008). Social perception and steering for online avatars. In *Proceedings of Intelligent Virtual Agents (IVA '08)*, pages 104–116.

[Pennock et al., 2005] Pennock, C., Shao, W., and Terzopoulus, D. (2005). Gaze control for autonomous pedestrians.

[Peters and Sullivan, 2003] Peters, C. and Sullivan, C. O. (2003). Bottom-up visual attention for virtual human animation. In *Proceedings of the 16th International Conference on Computer Animation and Social Agents (CASA '03)*, page 111, Washington, DC, USA. IEEE Computer Society.

[Posner, 1980a] Posner, M. I. (1980a). Attention and the detection of signals. *Journal of Experimental Psychology*.

[Posner, 1980b] Posner, M. I. (1980b). Orienting of attention. *he Quarterly journal of experimental psychology*.

[Reynolds, 2006] Reynolds, C. (2006). Big fast crowds on ps3. In *Proceedings of the 2006 ACM SIGGRAPH symposium on Videogames (Sandbox '06)*, pages 113–121, New York, NY, USA. ACM.

[Shao and Terzopoulos, 2005] Shao, W. and Terzopoulos, D. (2005). Autonomous pedestrians. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation (SCA '05)*, pages 19–28, New York, NY, USA. ACM.

[Siegman and Feldstein, 1978] Siegman, A. W. and Feldstein, S. (1978). *Nonverbal Behavior and Communication*. John Wiley & Sons Inc.

[Thiebaux et al., 2009] Thiebaux, M., Lance, B., and Marcella, S. (2009). Real-time expressive gaze animation for virtual humans. In *Proceedings of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS '09)*.

[Vilhjálmsson, 2003] Vilhjálmsson, H. H. (2003). *Avatar Augmented Online Conversation*. PhD thesis, Massachusetts Institute of Technology.

[Vilhjálmsson and Cassell, 1998] Vilhjálmsson, H. H. and Cassell, J. (1998). Bodychat: Autonomous communicative behaviors in avatars. In *Proceedings of the 2nd International Conference on Autonomous Agents (Agents '98)*, pages 269–276, New York. ACM Press.

Appendix A

Video Sessions Descriptions

In this appendix we will see more detailed data from each session.

Session 1
Recording Time: 01 min 58 sec

Additional Information
The subject is listening music and this could influence the behaviour (for example the head movements).

Behaviour
The subject looks the camera 6 times in two minutes for an interval between 1 and 2 seconds. It could be possible that at the beginning the subject looks the camera more frequently (after 2 seconds) because he wants to be sure about this "external object" and later he looks again in intervals between 13 and 32 seconds just for check if the camera is still there.

Snap-shots

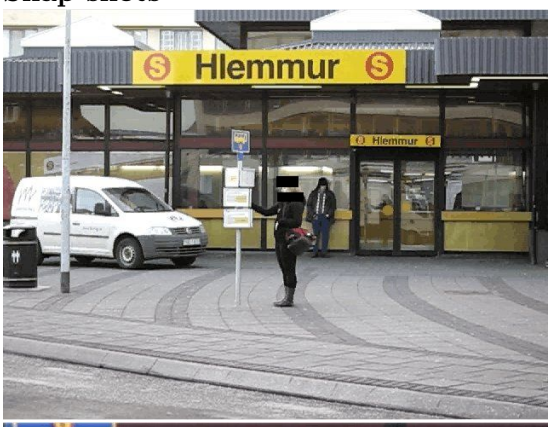

P2
GJALD

Session 2
Recording Time: 02 min 00 sec

Additional Information
After the middle of the video the subject's gaze is covered by the bus.

Behaviour
The subject walks around a lot of the time.
The subject during the video looks at various objects (the bus time table, the cellphone, a paper, etc.)

Snap-shots



Session 3
Recording Time: 01 min 01 sec

Additional Information
None

Behaviour
The subject moves the head and the torso for the whole duration of the session. He looks around but evidently without a specific target. Two times he looks in the direction of the camera but maybe he was still unaware of it.

Snap-shots

Session 4
Recording Time: 01 min 00 sec

Additional Information
None

Behaviour
The subject looks for a long time in the center direction. He walks around two times and he looks towards the camera two times. He also speaks with another person.

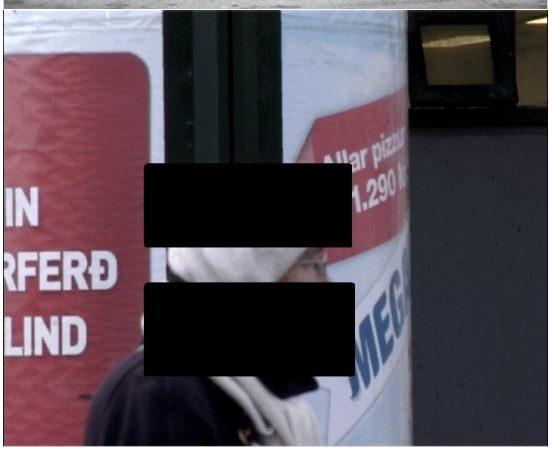
Snap-shots

Session 5
Recording Time: 02 min 01 sec

Additional Information
The subject is chewing a chewing-gum.

Behaviour
The subject looks for the whole study duration to the left, which is the direction a bus is expected from. She looks at the camera once.

Snap-shots



Session 6
Recording Time: 01 min 22 sec

Additional Information
The subject is chewing a chewing-gum

Behaviour
The subject for almost all the session is looking to the center or to the left (in the direction of arriving buses). She sometimes looks to the camera. After every look to the camera she moves the gaze to the left or down.

Snap-shots

Session 7
Recording Time: 02 min 00 sec

Additional Information
None

Behaviour
The subject recognizes that the camera was recording her and she looks more times to the camera than the others. She's in movement during the whole video to perform some action: throwing away something, picking up a notebook, write something, taking a bag, etc ... In the last minutes the subject turns the back to the camera for writing something.

Snap-shots

Session 8
Recording Time: 02 min 00 sec

Additional Information
This is an example of gaze nothing.

Behaviour
The subject is still for almost the whole session. He seems to gaze at nothing. For a long part of the video it is impossible to recognize the gaze because of the subject's position.

Snap-shots

Session 9
Recording Time: 01 min 13 sec

Additional Information
The subject has sun glasses not completely dark. They don't make harder recognizing the gaze.

Behaviour
The subject is moving all the time. The movements are related to some action: taking the bus ticket, closing the jacket, etc ... For a long time it is impossible to recognize exactly what the gaze it because she's looking down or back.

Snap-shots