Design and Experimental Evaluation of a Context-aware Social Gaze Control System for a Humanlike Robot

Thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
in
Automatic, Robotics and Bioengineering

By
Abolfazl Zaraki

February 17, 2014
Pisa, Italy

University of Pisa
Interdepartmental Research Center E. Piaggio
Automatic, Robotics and Bioengineering
PhD Cycle XXVI (2011-2013)

Thesis title:

Design and Experimental Evaluation of a Context-aware Social Gaze Control System for a Humanlike Robot

Author:
Abolfazl Zaraki

Advisor:
Prof. Danilo De Rossi

Tutor:
Dr. Daniele Mazzei

Dedicated to

My Beloved Wife, Maryam

Without Whom None of My Success Would Be Possible.

Abstract

Nowadays, social robots are increasingly being developed for a variety of human-centered scenarios in which they interact with people. For this reason, they should possess the ability to perceive and interpret human non-verbal/verbal communicative cues, in a humanlike way. In addition, they should be able to autonomously identify the most important interactional target at the proper time by exploring the perceptual information, and exhibit a believable behavior accordingly. Employing a social robot with such capabilities has several positive outcomes for human society.

This thesis presents a multilayer context-aware gaze control system that has been implemented as a part of a humanlike social robot. Using this system the robot is able to mimic the human perception, attention, and gaze behavior in a dynamic multiparty social interaction. The system enables the robot to direct appropriately its gaze at the right time to the environmental targets and humans who are interacting with each other and with the robot. For this reason, the attention mechanism of the gaze control system is based on features that have been proven to guide human attention: the verbal and non-verbal cues, proxemics, the effective field of view, the habituation effect, and the low-level visual features.

The gaze control system uses skeleton tracking and speech recognition, facial expression recognition, and salience detection to implement the same features.

As part of a pilot evaluation, the gaze behavior of 11 participants was collected with a professional eye-tracking device, while they were watching a video of two-person interactions. Analyzing the average gaze behavior of participants, the importance of human-relevant features in human attention triggering were determined. Based on this finding, the parameters of the gaze control system were tuned in order to imitate the human behavior in selecting features of environment.

The comparison between the human gaze behavior and the gaze behavior of the developed system running on the same videos shows that the proposed approach is promising as it replicated human gaze behavior 89% of the time.

Riassunto

Al giorno d'oggi, i robot sociali sono sempre più sviluppati per una varietà di scenari antropocentrico, in cui interagiscono con le persone. Per questo motivo, tali robot dovrebbero possedere la capacità di percepire e interpretare segnali verbali e non verbali della comunicazione umana. Inoltre, dovrebbero essere capaci di identificare autonomamente il target più importante al momento opportuno esplorando le informazioni percettive ed esibire di conseguenza un comportamento credibile. Impiegare un robot sociale con tali capacità ha diversi risultati positivi per la società umana.

Questa tesi presenta un sistema basato sul contesto per il controllo dell'attenzione di un robot sociale umanoide. Con questo sistema il robot è in grado di imitare la percezione, l'attenzione e il comportamento dello sguardo dell'uomo durante un'interazione sociale dinamica tra più partecipanti. Il sistema consente al robot di indirizzare lo sguardo in modo appropriato e al momento giusto verso target ambientali o verso persone che stanno interagendo tra loro e con il robot. Per questo, il meccanismo di attenzione del sistema di controllo dello sguardo si basa su caratteristiche che si sono dimostrate guidare l'attenzione umana: segnali non verbali e verbali, prossemica, il campo di vista effettivo, l'effetto dell'adattamento e le caratteristiche di basso livello.

Il sistema di controllo dello sguardo utilizza l'identificazione dello schel- etro, il riconoscimento vocale e delle espressioni facciali e l'individuazione

della salienza.

Come studio pilota, è stato registrato il comportamento dello sguardo di 11 partecipanti con un dispositivo professionale di eye tracking mentre guardavano video relativi all'interazione tra due persone. Analizzando il comportamento medio dello sguardo dei partecipanti, è stata determinata l'importanza delle caratteristiche umane nella cattura dell'attenzione umana. Sulla base di questi risultati, sono stati regolati i parametri del sistema di controllo dello sguardo al fine di imitare il comportamento umano nella selezione di caratteristiche dell'ambiente.

Il confronto tra il comportamento dello sguardo umano e quello del sistema sviluppato applicato allo stesso video dimostra che l'approccio utilizzato è promettente replicando il comportamento dello sguardo umano per l'89% del tempo.

Acknowledgements

Thank you ALLAH, the most Beneficent and the most Merciful, for giving me the strength to keep going. Thank you for granting me health and prosperity to finish this work in a foreign country far from my soil. This work is done because whenever all I really wanted to do was give up, you held me to take two more steps. Your words kept me all through the journey of completing this thesis.

First and foremost I would like to thank my beloved wife Dr. Maryam Banitalebi Dehkordi for her endless love, support, permanent trust of my abilities, giving me confidence, and standing next to me throughout my PhD study. Utmost appreciation also goes to my beloved parents for encouraging and supporting me throughout my life.

I would like to express my sincerest gratitude to my adviser Prof. Danilo De Rossi for his support, supervision, valuable comments, and advice throughout my PhD. I would also like to appreciate Dr. Daniele Mazzei for his valuable advice, which directed my PhD research toward the right point.

I would like to extend my sincerest gratitude to Prof. Alois Knoll and Dr. Manuel Giuliani from Technische Universität München of Germany, who provided me the opportunity to do a part of my research in Germany. I appreciate their supports, knowledge sharing, instructive advice and

valuable guidance on my research.

I would like to appreciate the Azienda Regionale per il Diritto allo Studio Universitario di Pisa (DSU) Toscana, for supporting me during my PhD studying.

I would also like to thank all my colleagues and friends in Centro E. Piaggio, in particular Dr. Nicole Lazzeri for the collaboration and knowledge sharing throughout my PhD.

Last but not the least, I would also like to appreciate Dr. Andreas Haslbeck from the Institute of Ergonomics at the Technische Universität München for providing the facilities and for his collaboration.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

Nowadays, robotic technologies are making their way into human society as powerful devices that possess capability to perform complex tasks. They are increasingly being designed and used to assist people in improving the quality of life. For example, industrial robots as *conventional class* of robots perform complex tasks in companies with highest efficiency where human is not able to do. This class of robots are being designed to accomplish limited tasks in a non-human centered scenario. Although the conventional robots serve for humans but they do not have any direct interaction with human.

In addition to the conventional robots, the new generation of robots -*social robots*- are being developed to be used in tasks and positions alongside human, and unlike the conventional robots, they are required to perform tasks in human-centered scenarios [12]. For example, they can be used as tutor for educational purposes, as toys for kids, as therapeutic aids, as companion for humans, as domestic stuff, and they can serve in

Figure 1.1: Social robots inspired from animals: (a) Paro the therapeutic seal robot developed at AIST [1]. (b) Leonardo developed at MIT Lab [2]. (c) AIBO the robotic dog developed by Sony for entertainment [3].

many other positions in which they interact directly with humans.

Social robot class consists of three main groups that are being developed for various purposes: animal-inspired, humanoid, and human-like (android) robots. Figure 1.1 shows examples of first group -animal-inspired- social robot that have been designed to serve in human-centered scenarios, i.e. therapeutic assistance [1], human-robot interaction and research purposes [2], and entertainment purposes [3]. To function in such positions, they should be able to make an effective interaction with human. In other words, in addition to the task-performing capabilities, they must have the ability to make a meaningful and behaviorally acceptable interaction with humans.

The second group of social robots are humanoid robots that usually have the body shape and size similar to humans. Due to vision and audition capabilities, as well as their physical characteristics, humanoid robots are able to communicate with people through verbal and non-verbal communicative cues and perform task in collaboration with human, in a dynamic environment. Thus, they can replicate the human

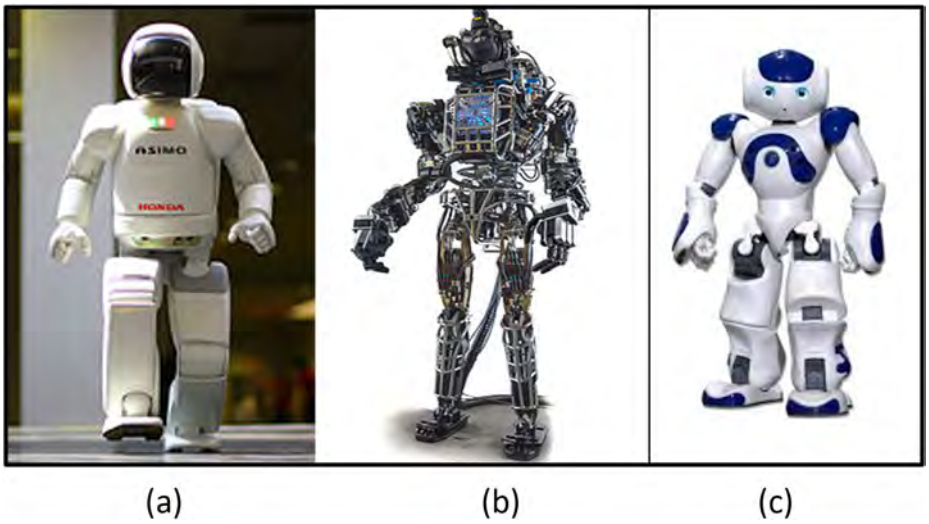


Figure 1.2: Social humanoid robots: (a) ASIMO developed by Honda to be a multi-functional mobile assistant [4]. (b) Atlas is a bipedal humanoid military robot primarily developed by the American robotics company Boston Dynamics [5]. (c) Nao is an autonomous humanoid robot developed by Aldebaran Robotics, a French robotics company [6].

role in society as they can replicate human motions. However, due to non-human face appearance, people perceive and accept them as non-human creatures. Figure 1.2 shows examples of three famous humanoid robots developed for multiple purposes such as education, military, and entertainment [4] [5] [6].

1.1 Social Humanlike Robots

With the rapid advancement of bio-mimetic materials and advances in control and computing techniques, the third group of social robots called *humanlike* was born. A humanlike social robot mostly is being developed

with an appearance similar to human, particularly in facial features and skin. They are being designed to interact with humans and become more integrated into human daily life.

Due to their appearance similarity with human, people perceive and accept them as non-machine and believable creatures. It allows the humanlike robots to be used for positions, as people treat them as human and since they can replicate the human role in society. As a believable creature, they can even potentially make an empathic and emotional relationship with human.

This is the dream of social humanlike robots development that an artificial creature be ultimately a companion for humans. For that aim, the development of several humanlike robots is already under the way that can be engaged in long-term and short-term interactions with human. Figure 1.3 shows several humanlike robots developed to be used in various human-centered positions. As shown, the appearance especially facial features, skin and body shape of these robots are very similar to human. Due to especial humanlike appearance, they can be beneficial for human in various scenarios.

The underlining assumption of designing such robots, as revealed in many scientific and practical researches, is that people unconsciously treat with humanlike robots in the same way that they interact with other people by demonstrating politeness, showing concern of their feeling, etc. It is promising and most likely because of the human brain structure. It treats with a creature with humanlike *appearance* and *behavior* in the same way that it treats with other people. Thus, the main concern about social humanlike robots, after designing a humanlike appearance (i.e., head, face, hair, teeth, skin, body shape and size, etc.), is to enable them to exhibit appropriately humanlike behaviors. Designing appearance together with behavior models similar to the human, social humanlike robots will

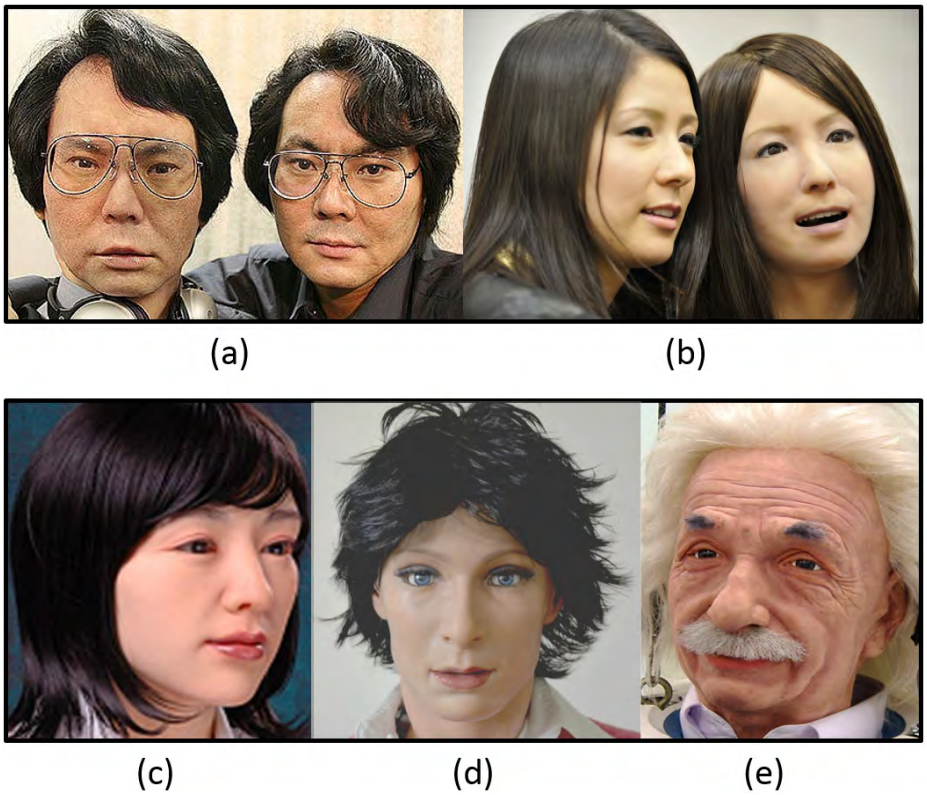


Figure 1.3: Social humanlike robots able to replicate facial features and head-eye movements in a humanlike way: (a) Hiroshi Ishiguro (Right) and his robotic Doppelgänger Geminoid HI-1 (Left) developed in Japan [7]. (b) Geminoid-F (Right) developed in Japan [8]. (c) EveR-2 developed by the Korea Institute of Industrial Technology [9]. (d) FACE developed by Hanson robotic [10]. (e) Robotic android of Albert Einstein developed by Hanson robotic [11].

benefit human life in various positions.

To make a behaviorally acceptable robot, firstly, the robot should be able to respond appropriately the dynamic environment. For that, the robot requires a human-level perceptual model to perceive and interpret

human-relevant verbal and non-verbal communicative cues.

Then, the robot requires a human-level attention model to identify the potential interactional target point of environment.

Finally, the robot requires a behavior model to generate an appropriate behavior (i.e., facial expressions, body gesture, vocal signals, gaze, etc.), according to the target modality. Clearly to generate any behavior that human perceive as natural, a motion model derived from human data is required.

In short, the robot must be equipped with a system that can afford to understand human-relevant features, autonomously identify the target points and generate an appropriate human-level behavior for the robot. Integrating the *perception*, *attention*, and *behavior* models enable a humanlike robot to make a natural, intuitive, and enjoyable interaction with human that has many positive outcome for human society.

1.2 Thesis Motivation

As discussed, an effective human-robot interaction is highly depended on how appropriate the robot responds to human and how natural it is being perceived by the human. To display an acceptable behavior, a human-level *attention* system is fundamental in social robot development. It closes the interaction loop between robot and environment and enables the robot to afford a context-aware behavior (see Figure 1.5).

The system should performs two major tasks for the robot. On one hand, it explores actively, the perceptual information of environment and identifies the most important human/non-human target based on a human-level selection mechanism.

On the other hand, the system controls the robot's behavior according to the target modality, in order to enable the robot to exhibit a natural humanlike and believable behavior. Such a system should be used as

a middle layer that correlates the robot's behavior with current social-context.

Before designing an attention system, implementing two additional components are imperative: perceptual system and behavior control system.

- Perceptual system:

A human-level perceptual system able to support familiar verbal and non-verbal cues of human is fundamental for the robot development. Perceiving and interpreting a social-context, in the same way that human is doing, allows the attention system to evaluate human and environmental features and to select a right target point at the right time for the robot, in a dynamic social interaction with multiple people.

- Behavior control system:

A behavior control system containing human-level behavioral models, allows the attention system to control correctly the robot's behavior. It moves the robot's actuators in a the way that robot shows a humanlike motion. For example, the behavior control system adjusts the dynamic of head and eyes and generates a humanlike gaze for the robot.

Fig. 1.4 shows our social humanlike robot FACE [13, 14], involved in a social scenario where it interacts with a group of people. To display behavior that humans perceive as natural, the robot should direct its attention at the most important person at the right time based on the current social-context. It thus requires a mechanism that is able to control attention and gaze, based on social cues and information extracted by the perceptual system from raw visual-auditory data.

To design attention systems for social robots, it is necessary to consider the psychological, neurological and computational aspects of human

Figure 1.4: The FACE humanoid robot interacts with a group of people. Picture courtesy of Enzo Gargano.

attention [15–21] as well as the social cues and conventions. This information can support the gaze control system to direct the robot's attention at the appropriate target, during a social interaction with multiple people.

1.3 Main Objectives and Thesis Scopes

The main aim of this work as shown in Figure 1.5 is designing, implementing and evaluating a multilayer context-aware social Gaze Control System (GCS) as a part of a social humanlike robot called FACE (Facial Automaton for Conveying Emotion) [10, 22–24]. The system enables the robot to engage autonomously multiple people in a social interaction by

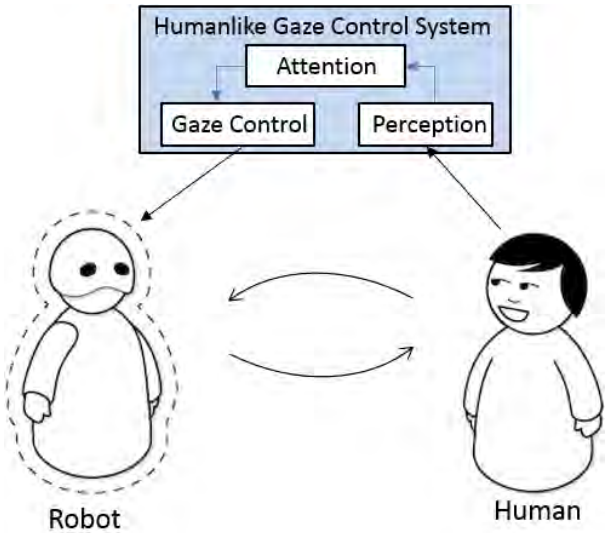


Figure 1.5: The general structure of the proposed gaze control system. Using this system a robot is able to perceive and interpret human-relevant features, direct its attention to the most important human and controls its behavior, accordingly.

generating an acceptable social gaze behavior for the robot.

GCS has three standalone-interconnected layers that simulate human perception, attention, and gaze control system for our humanlike robot. For that, development of GCS requires implementing these three layers and underlying models. Thus, the objectives of this research are as follows:

- Design and implementation of a perceptual layer that perceives and interprets the surrounding environment for the robot. It provides a human-level understanding for the robot by recognizing human-relevant verbal and non-verbal cues of multiple humans through several parallel algorithm.

- Design and implementation of an attention layer that actively identifies the most important human/non-human target exploring recognized perceptual information of humans, based on the human-level attention model.

- Design and implementation of a gaze control model that controls the dynamic of head and eyes of the robot and generates a humanlike gaze for the robot.

- Design and implementation of a communication unit that makes bidirectional channels between layers, components and the robot, in order to send and receive data.

These three interconnected layers together, manage low-level sensory information, and make a high-level human-level interpretation of environment for the robot and enables it to interact autonomously with multiple people in a dynamic social interaction, displaying appropriate and acceptable gaze behavior.

1.4 Organization

The remainder of this thesis is organized as follows.

Chapter 2 extensively reviews the previous works in the area of human attention and gaze control modeling particularly, in social robotics applications. It discusses the major limitations of the current existing models for HRI application and explains why the current class of attention models are inefficient when they be used in a social robot.

The last part of the chapter describes an ideal human-based attention and gaze control models that can be potentially benefits social robots development and it is the focus of this thesis.

Chapter 3 explains the theoretical aspects of several high-level human-like features that have been proven to guide human attention in a social human-human interaction. The high-level features that are the focus of the thesis, and have been considered in attention and gaze modeling are non-verbal and verbal cues, proxemics, the effective visual field of view, the habituation effect, and people intention.

The chapter then introduces the new generation of attention system for social robotic application that simulates human attention dealing with the proposed high-level human-relevant features, and low-level visual features instead of merely low-level visual features.

Chapter 4 presents the proposed context-aware social gaze control system and its layers and components. It details the system structure and describes the theoretical and informatics design and implementation aspects of the system. It then describes the performance of the final integration of the system and evaluation process, which shows the compatibility and data synchronicity between layers and components.

Chapter 5 describes a gaze tracking study that we carried using a professional eye-tracker device, due to adjusting the weight parameter of the attention model of gaze control system. It details the experiments process and data-collecting steps. It then follows by data analysis and explains the way that the parameter of the proposed system are tuned according to the human data.

The last part of the chapter describes the test and the performance of the system in compare to the human.

Finally, this thesis is concluded in chapter 6, and contributions of this work to the state-of-the-art, and the future works are summarized.

Chapter 2

Human Attention and Gaze Modeling

Contents

Several studies in the past demonstrated that, human gaze behavior in a complex scene viewing is highly correlated to the human attention behavior in the same scene. In other words, people most often look at the point of environment that is selected by the attention system. Thus, to simulate human gaze behavior, several efforts investigated two aspects of the problem: the *strategy* that human attention selects a target in a complex-scene viewing (attention mechanism), and the dynamic of head and eye movements when shift from one point to another to look at a target (gaze behavior).

While the term attention and gaze behavior are often used interchangeable, each of them has a more subtle definition, which allows their

delineation. For that, the work described in this chapter draws on research and techniques from two main areas: attention and gaze modeling.

The chapter begins with Section 2.1 that reviews the previous works in the area of human and robot *attention modeling*. It describes in detail, the current existing classes of attention models and their applications and strengths. It then discuss the main limitations and drawbacks of the models in human-robot interaction applications.

Section 2.2 discusses that how the *attention modeling* results have been used to explain human *gaze models*, and to implement robot gaze models.

Considering the limitations of the current attention and gaze models in human-robot interaction application, Section 2.3 proposes a *context-aware attention* model that drives gaze behavior. In the proposed model, number of shortcomings of the previous works are resolved, and thus it is powerful to be employed in HRI applications. The last section reports the summary.

2.1 State-of-the-art on Attention Modeling

Modeling human attention has been an active research field over the last two decades. The main concern in this field resides in identifying which features and phenomena influence human attention selection mechanism, and how they influence human attention.

Investigations on human attention have described two different aspects of how the human mind in selecting its target in a scene viewing [25]: top-down and bottom-up processing.

The first aspect is top-down processing [25, 26] also known as goal-driven attention that is a voluntary process of selecting target of environment that is under the control of the person who is attending the environment. Clearly, the individual-relevant features affect top-down

processing and attention selection mechanism. The individual-relevant features refers to those features of individual that vary in person to person. In other words, due to variety of these features, people show different attention behavior in a same scene viewing.

Although modeling such features is a complex multidisciplinary issue, but some efforts investigated the effect of some of them such as working memory [27], personality and familiarity [28], culture [29, 30], etc.

The second aspect is bottom-up processing [25, 31] also known as the scene-driven features that is an involuntary process of selecting target of environment. The features of environment influence the human attention selection mechanism. They can be either low-level visual features of environment (e.g., color, intensity, etc.) or high-level features of objects (e.g., shape, distance, characteristic, etc.) and humans (body gesture, facial expression, etc.) in environment.

Since the aim of this work is to design and implement an attention system for a social robot, this chapter reviews those previous works that investigated the effect of scene-driven features on human attention.

The following section describes the state-of-the-art in the area of attention modeling and discusses the major limitations of this class of attention system for HRI applications. At the last part of this section we answer this question that why the current existing attention class is inefficient in HRI applications and why we need to emerge new attention class. Then, we introduce new features and characteristics, which are essential in designing a new attention class called *humanlike attention* class.

2.1.1 Salience-based Attention Class

- **Origin**

The extensive psychophysical literature in the field of human attention modeling shows that the basic low-level visual features of a 2D image

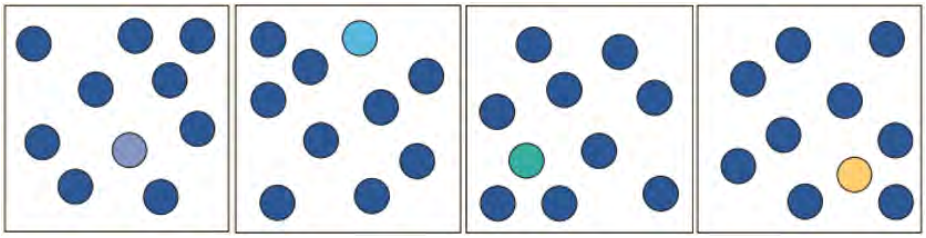


Figure 2.1: The attention targets are different from background in a single low-level feature "color", and they immediately attract human attention.

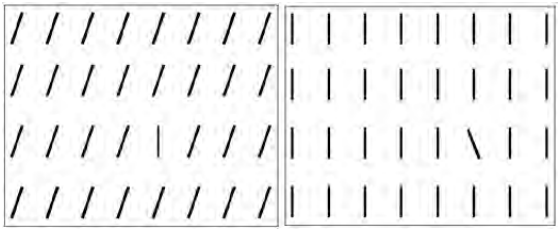


Figure 2.2: The attention targets are different from background in a single low-level feature "orientation", and they immediately attract human attention.

(e.g., color, orientation, etc.) attract and guide attention to specific points (targets) of a visual scene [32, 33]. If a target differs in a single feature (e.g., color, orientation) from its surrounding regions, it can be detected very fast by attention [34] (See Figure 2.1 and Figure 2.2).

In a more complex visual scene such as Figure 2.3, where a target is presented by a conjunction of two or more than two features, the attention selection mechanism can successfully be explained using serial selection driven by image features. In this case, the attention explores perceptual information and after a while it will be attracted by a target that is different from surrounding regions.

Treisman & Gelade [34] in an initiative work called "Feature Integration Theory" stated that which low-level visual features of an image are

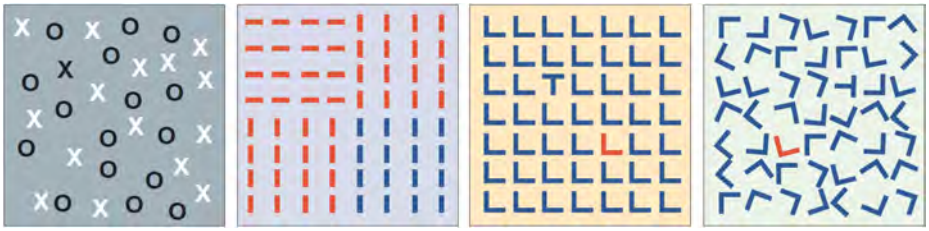


Figure 2.3: In a more complex scene attention target are presented in conjunction of two or more than two low-level features.

important and how these feature should be combined to influence human attention. Then, Koch & Ullman in [35] proposed a model to combine these features and they introduced the concept of salience map, which is a map that is created from original image and shows the salience of the image regions. They also proposed winner-take-all model that selects the most salient point of a salience map as the attention target.

The concept of salience map have been the origin of a large class of attention system called *salience-based attention* system. Following this principle many attention systems have been proposed.

- Implementation and Applications

Itti & Koch [36] proposed the first and complete implementation with verification of salience-based attention class. The process of attention selection for a given visual scene, in the original computational implementation of salience-based attention class, is illustrated in Figure 2.4. The process is as follows: first, the algorithms extract visual low-level features of the given visual scene, through parallel channels. Then, local competition across image space and feature scales is computed yielding to the so-called feature maps. Finally, individual feature maps are combined by weighted sums creating the salience map. Based on the salience map, the algorithms can then select attention targets, for example by applying

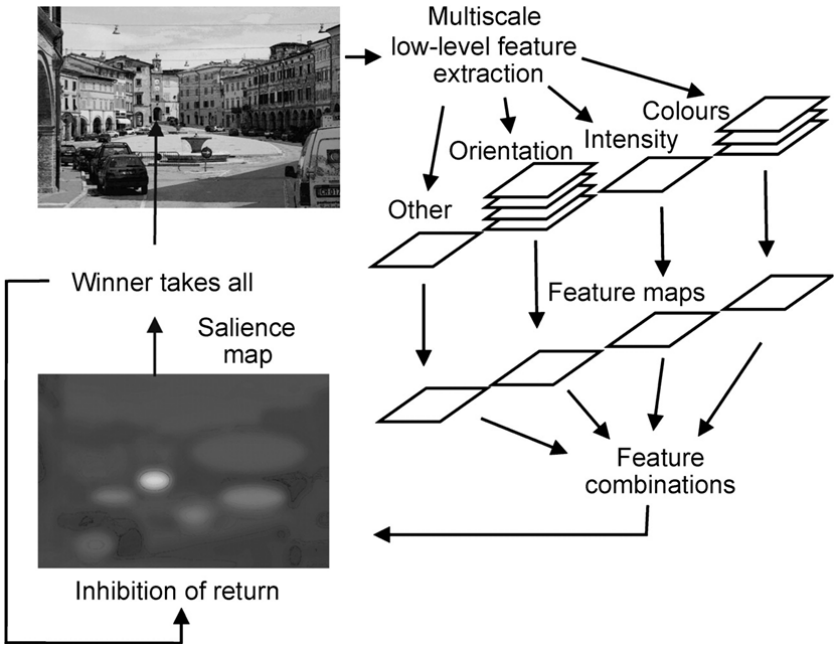


Figure 2.4: The salience-based attention model identifies attention target (salient point) by analyzing the low-level features of a given visual scene through parallel channels.

the "winner-takes-all" principle. The identified target points show the points that have the most chance to attract human attention in a scene viewing.

According to the Treisman & Gelade's work [34] presented in "Feature Integration Theory", the conventional computational attention models were implemented to find the salient point of a visual scene dealing with three main features of a visual scene: intensity, color, and orientation. Intensity is defined as average of three colors' channel; color is defined as red-green and blue-yellow channels; and orientation is implemented as convolution with a oriented Gabor filter [37].

To enable conventional attention models to mimic human attention

Table 2.1: Other low-level features and image processing techniques that have been added by several studies to the conventional attention model, to improve its performance in simulating human attention behavior.

No.	**Features and Image Processing Techniques**	**Ref.**
1	Human face	[38]
2	Skin hue	[39]
3	Motion	[40]
4	Depth information	[41]
5	Texture contrast	[42]
6	Gist of the image	[43]
7	Spatial resolution filter on image	[44]
8	Horizontal line	[43]
9	Wavelet filter on image	[45]
10	Center-bias filter	[46]
11	Optical flow	[47]
12	Above ave. salience of image	[42]
13	Center-surround contrast filter	[48]

with higher accuracy rate, a larger number of features (e.g., motion, depth information, etc.) had to be taken into account. Several studies have been added different low-level visual features and applied several image processing techniques to the original implementation of attention model, which allow the model to consider a larger set of features in computing the salient point. The additional features increase the performance of the original model in simulating human attention. Table 2.1 reports the additional features and image processing techniques together with references who proposed these features.

Salience-based attention class have been successfully applied to the several research areas ranging from computer vision and graphic to the

Table 2.2: The salience-based attention class have been successfully applied to several filed of research

No	Field of Study	Ref.
1	Image segmentation	[50]
2	Image matching	[51]
3	Scene classification	[52]
4	Object detection	[53]
5	Object recognition	[54]
6	Visual tracking	[55]
7	Face segmentation and tracking	[56]
8	Active vision	[57]
9	Robot localization	[58]
10	Robot navigation	[59]
11	HRI applications	[40, 60, 61]

robotics applications. Table 2.2 reports the works that employed this class of attention system to their fields (the information of the table was taken from [49]), however in this work, we are more interested to describe works that are relevant to HRI especially social robot development, in dynamic environment.

As reported in Table 2.1 and Table 2.2, salience-based attention class is a powerful tool especially for the image processing application that are required to analyze low-level features of an image to detect objects and motions areas (Figure 2.5 and Figure 2.6), but the question raises whether these class of attention models can be generalized to the real-world and human-centered situations? Especially, in a social robotic application, that the attention model should imitate the gaze behavior of speakers and listeners involved in a social scenario. Is the salience-based attention model effective in task-oriented circumstances?

The next section answers this question and discusses the major limitation of salience-based attention class. It then, talks about new generation of attention model that can overcome these drawbacks and can be more powerful in HRI, which is aim of this work.

- **Major Limitations**

Imagine you are in a meeting with your colleagues. They enter into the room, one after another. During the meeting, they speak, show different meaningful body gestures and facial expressions, discuss some issues, and at the end, they leave the room one after another. The important question raises here: where did you gaze, as the one participated, during the meeting? Which parameters influenced the selection mechanism of your attention? Do these parameters can be extracted by analyzing low-level features of the visual scene? Is the salience-based attention class is able to simulate your attention behavior?

The answers of these questions clearly show that since the target selection strategy of the model is only based on the low-level feature analysis, it is not aware about high-level communicative features that guide human attention in a social interaction and therefore it is incapable in simulating human gaze behavior in such meeting.

Tatler et al. [62] reviewed the major limitations of salience-based attention model and found this class of attention poor in accounting many important aspects of complex scenes that cannot be explained only through low-level features analysis.

In the design of social robot attention system, thus it is necessary to take into account also the high-level communicative and social features that are fundamental in human attention system. Following section details important issues that demonstrate the inefficiency of the salience-base attention class in social robotic applications.

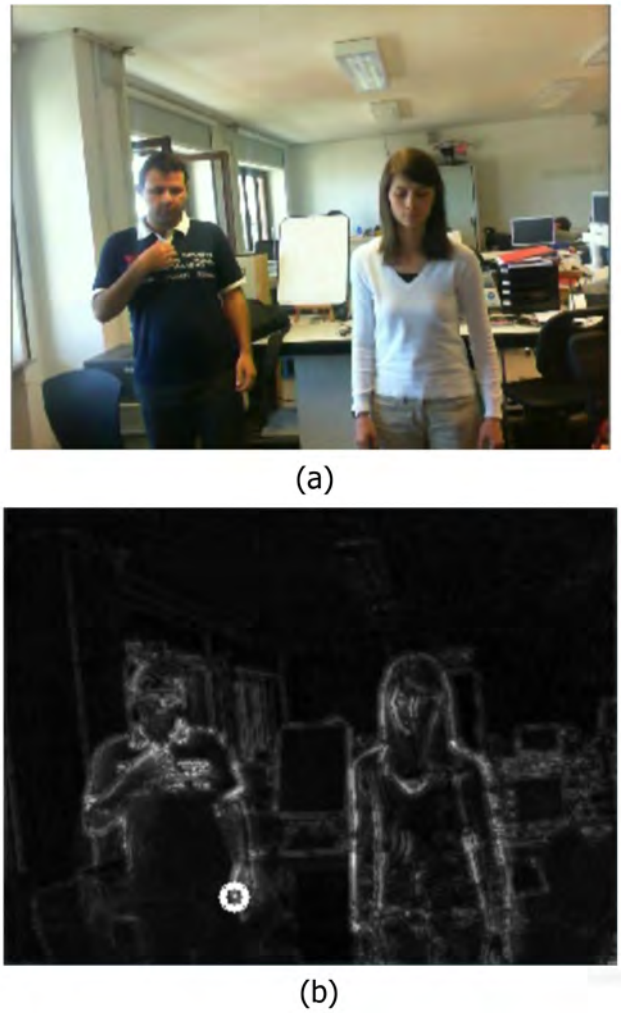


Figure 2.5: The salience-based attention model is a powerful method capable of extracting objects regions by analyzing low-level features of a scene. The image (a) shows the given visual scene and image (b) shows the corresponding salience-map created by the model. The circle on the image (b) shows the salient point (attention target) identified by the model.

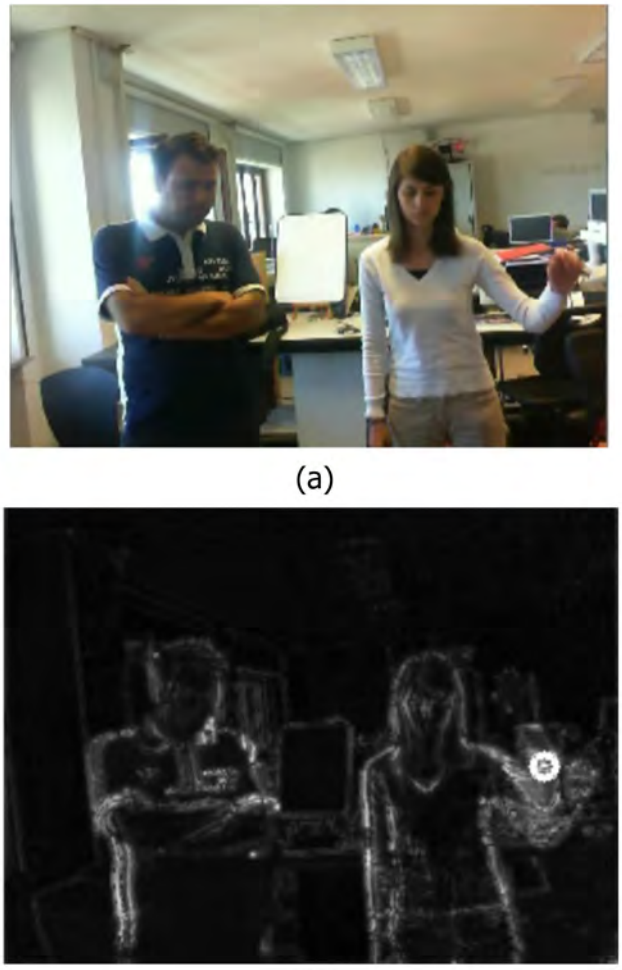


Figure 2.6: The salience-based attention model is a powerful method capable of detecting motion regions analyzing low-level features of a scene. The image (a) shows the given visual scene and image (b) shows the corresponding salience-map created by the model. The circle on the image (b) shows the salient point (attention target) identified by the model.

- **Scene Context Guides Human Attention:**

The main limitation of salience-based attention class is inability to discriminate between different information when analyzing a scene image. In other words, it identifies attention targets by analyzing the image in a pixel-by-pixel way without emphasizing human and objects regions. While, there are some evidences that show, human attention has its priority to choose region of a visual scene. For example with the presence of human and object in a scene, the attention is directed more to them and associated features, rather than the environment. Thus, it is wrong if the same process is being used to identify a target point in the image, with and without presence of humans/objects.

To prove this fact, Rothkopf et al. [63], in a gaze study analyzed participants' gaze behavior to figure out which points of environment was looked by a human during a virtual walking experiment. They also analyzed for the same scene, the identified points that were obtained by a salience-based class. The comparison result between the participants and salience-based model's gaze behaviors showed that the identified target points of them were totally different: humans mainly looked at objects and only 15% of their fixation directed to the background while using salience-based model 70% of fixation directed to the background. These result shows that salience-based class is not capable of simulating human attention in a human/object centered environment.

- **Type of the Task Affects Human Attention:**

The main assumption in designing the salience-based attention model is simulating the attention of a person who watch a visual scene in a "task-free" condition. This assumption gives license to the viewer to look at the visual scene, without any purpose, which is not a reasonable assumption.

There are several evidences [62] show that human fixation in a task-free viewing is very different with human fixation when is engaged in a task. To support this idea, Yarbus [64] in a very famous example showed that human attention/fixation is dependent on the current task. He carried out the following experiment to demonstrate his theory. Several humans were asked to watch the same scene (a room with a family and an unexpected visitor entering the room) under different conditions such as "estimation the material circumstances of the family", "estimation the age of the people of the scene" and "freely viewing". The result was interesting: The human fixation/attention was differed considerably in each of cases. It can be thus concluded that salience-based attention class, which is designed based on the "task-free" assumption of humans, is not reasonable way to simulate human attention, particularly in a social interaction that a person is engaged to different tasks.

- **Auditory Signals Affect Human Attention:**

Auditory signals mostly cause unintentional shift of human attention and must be taken into account in modeling human attention, while salience-base attention class identifies its target points of a scene dealing only with visual features. The effect of auditory signals on human attention will be discussed in detail in the next chapters.

- **Human Social Signals:**

In a social human-human interaction scenario, intention of the interactional partner that is expressed through social signals has been proven as a factor that guides the attention to select its target. For example, the attention behavior is different for a person that approaches to initiate a social interaction and for a person that leave a social interaction. This concept also has been demonstrated in the gaze study that we have done

as a part of this work. The result shows human intention that can be verbally/non-verbally expressed is an important factor that influence the attention of others. Next chapters describe the study in detail.

- **Other Parameters:**

In spite of several well-known factors that influence human attention, modeling some of them is an extremely or even impossible task. For example, human's cognitive process at the time of scene-viewing, memory and internal emotional states are factors that influence human attention, but due the diversity in person to person, it is difficult to model them. These factors are beyond the focus of this thesis and we do not consider them in our implemented attention model. We considers mostly those factors that usually occur in a social interaction between people.

By reviewing literature, it can be concluded that the salience-based attention class is poor in accounting of many aspect of a scene that guide human attention in a social scenario. It can be thus, failed if be used in a human-centered applications.

To overcome the several shortcoming of this class, emerging a new class of attention system called "Visual-Auditory" or "Humanlike" attention class is essential.

2.1.2 Visual-Auditory Attention Class

Unlike the conventional attention class, the visual-auditory attention class deals with both high-level human-relevant features and auditory signals of environment, in order to identify a target point in a human-centered scenario (e.g., social interaction). In spite of investigation of some works in this area, there are still several open challenges in this topic.

Following section reviews works that are close to the aim of our work in attention modeling for HRI applications. It, then discusses the missing

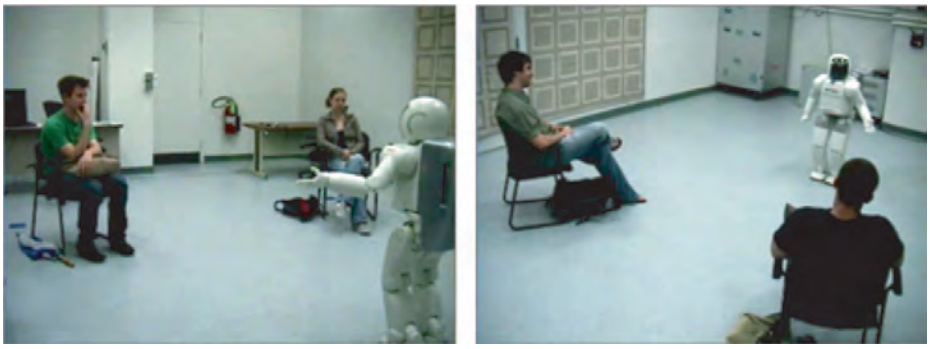


Figure 2.7: ASIMO telling a Japanese fairy tale to two listeners.

components of each work that prevent generating a natural context-aware attention and gaze behaviors.

In an attention/gaze study, Mutlu et al. [65] derived gaze patterns (attention points) of human subject during storytelling, in order to design a model that replicates these patterns on a robot. For that, they first collected the gaze patterns (locations of attention points, target selection frequencies) of a professional storyteller during storytelling. Then, they designed a model that generated the same gaze behavior on the humanoid robot ASIMO, and evaluated the model. The results show that humans can recall the story better when the robot looked at them during storytelling. Another important result of this study was that, the frequency of gaze has an effect on how women and men perceive the robot. They found that women like the robot more, when it looks at them less frequently. Figure 2.7 shows ASIMO during storytelling (taken from [65]).

However, this work used visual-auditory information to model the same human behavior for the robot, but a different process has been followed. They used a pre-programmed model rather than autonomous humanlike attention model for their application that is not aware about the scene-context. It just replicates pre-recorded behavior without under-

Figure 2.8: The robot is able to locate speaker in a group.

standing the current-context. In a HRI application, robot should be able to adjust its gaze in real-time according to the current social-context.

Trafton et al. [66] integrated vision and audition within a cognitive architecture, which enabled a social robot to track conversations and focus its attention to the speaker among multiple humans. As shown in Figure 2.8 (taken from [66]) they evaluated their system on a social mobile robot, which resulted in a natural conversation tracking in a dynamic environment.

Although the proposed architecture correctly guided the robot's attention to the speakers at the right time, but it did not take into account of many other communicative cues that have been proven to guide human attention and are known to be fundamental for social attention modeling. A cognitive architecture should be able to identify the attention target, taking into account of all human-relevant features as well as social conventions between people.

Figure 2.9: A person in social distance to the receptionist.

Holthaus et al. [67] proposed a spatial model for a robot attention system. It drives the attention of a receptionist robot according to the spatial information of humans interacting with the robot. For that, the robot localized and tracked humans in the field of view by monitoring their distances. The Holtaus's robot moved its head and body, when they were getting closer to the robot, in order to initiate or terminate a social interaction with humans. Through a questionnaire-based evaluation, Holthaus et al. found that even if the robot made random movements when someone approached, external observer evaluate the interaction as humanlike. This results show the importance of proxemics and contextual reactions in the modeling of humanlike robot behavior. Figure 2.9 (taken from [67]) shows the receptionist robot of this work in the scenario.

All of the discussed related works partially cover some of the current challenges in the area of attention systems. However, we believe that designing of a comprehensive attention, model able to specify the most important attention target of environment based on low-level and high-level environmental visual/auditory features analysis is essential for the development of a new generation of social robots.

This work proposes an innovative context-aware humanlike attention model able to identify the most important target of environment dealing with *low-level* visual features and *high-level* human-relevant features of 2D images, 3D images, and auditory signals. Moreover, the presented framework provides a high-level image interpretation for social robots similar to the human attention system that significantly improve the behaviors of the social robot, and it is imperative for human and robot social exchange.

2.2 State-of-the-art on Gaze Behavior Modeling

As discussed, an ideal attention system of a social robot should be able to identify in real-time the most important target point by analyzing low-level visual feature and high-level human-relevant features in a social scenario. In humans, when the attention system selects its target, human sight line (gaze) most often moves, in order to attend to the selected target.

Gaze is a coordinated motion of eye and head through which the center of human visual attention moves to a specific point identified by the attention system.

Several researchers investigated different aspects of human gaze behavior over the past years. Through analysis of the gaze behavior of humans and monkeys, Goldring et al. [68] demonstrated that gaze behavior is beyond making/breaking eye contact and smooth tracking of moving subjects. They showed that gaze behavior is regulated by complex dynamics that allows a subject to use this attitude, not only for observation but also for delivering meaningful information and drive the conversation flow. Goldring et al. deeply studied the characteristics of head and eye movement of human subjects to understand if they use the same strategies when they gaze at visual, auditory and visual-auditory

targets. They found that target modalities have an effect on human gaze behavior characteristics identifying also some human gaze dynamic (head and eye velocities, motion amplitudes delays) during gaze shifts between targets.

Based on attention modeling research, various researchers proposed models and implementation of robot/agent gaze control systems. Andrist et al. [69] proposed an effective gaze model for virtual agents on which they considered various gaze characteristics such as amplitude, velocity and latency period in a gaze shift. They evaluated their gaze model on a humanlike virtual gent. Andrist's results show that when the agent maintains its head orientation toward the participant to emphasize the social interaction (affiliative gaze), it induces positive feeling to participants while when the agent maintains its head orientation, more toward visual space to emphasize other information (referential gaze), it improves the subjects learning capabilities. Figure 2.10 (taken from [69]) shows the affiliative and the referential gaze of the virtual agent.

Itti et al. [40] presented a gaze model for target shift and smooth tracking that has been implemented on an avatar. In their model, the amplitudes of head and eye movements were estimated and linked with the initial position of eye in its orbit. Figure 2.11 (taken from [40]) shows the sample of the animation gaze model.

Although several works have addressed important issues in human gaze behavior modeling and implementation but due to the complexity of the gaze behavior, a comprehensive context-aware model that estimates the gaze parameters (e.g., velocity, amplitude, latency, etc.), has not been implemented yet.

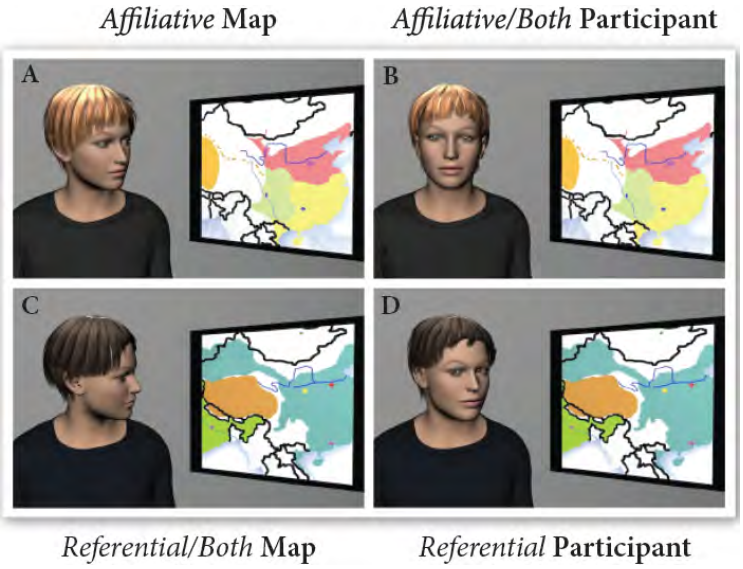


Figure 2.10: The affiliative and referential gaze of the virtual agent during an experiment.

2.3 Proposed Solution for Attention and Gaze Behavior Systems

Human attention and gaze behavior modeling as complex multidisciplinary tasks still have many open issues to solve. Modeling human attention/gaze topic has been the heart of a large research groups in several area such as human robot interaction. In HRI, an effective attention model, on one hand, should be able to simulate *human attention* in selecting the targets. On the other hand, it should be able to control the dynamic of head-eye (gaze) movements in a way that robot displays a *humanlike and acceptable motion.* Due to modeling complexity, a comprehensive model that cover both aspects (humanlike attention and gaze control) is still lacking.

Figure 2.11: Sample of animation gaze.

In this work, we propose a new model that fills the current existing gap in this area. The model has two main components: a humanlike target selection mechanism and humanlike gaze control mechanism.

The target selection mechanism correctly identifies the most important target of environment dealing with low-level visual non-human features and high-level human relevant visual and auditory features (Figure 2.12). While the gaze control mechanism generates a humanlike gaze motion for humanlike robot/3D avatar based on the human-inspired gaze model [40, 68, 70]. Using the proposed model (context-aware gaze control model), a humanlike robot is able to interact with multiple humans and produces humanlike behavior in a dynamic environment, which is essential in HRI.

Next chapter describes all components of the proposed model and its performance compare to human behavior.

2.4 Summary

This chapter extensively reviewed the previous works in the area of human attention and gaze control modeling particularly, in human-robot interaction applications. It discussed the major limitations of the current existing model for HRI application and explained why the current class of attention model is inefficient in simulating human attention in a HRI application. The last part of the chapter proposed potential solution for

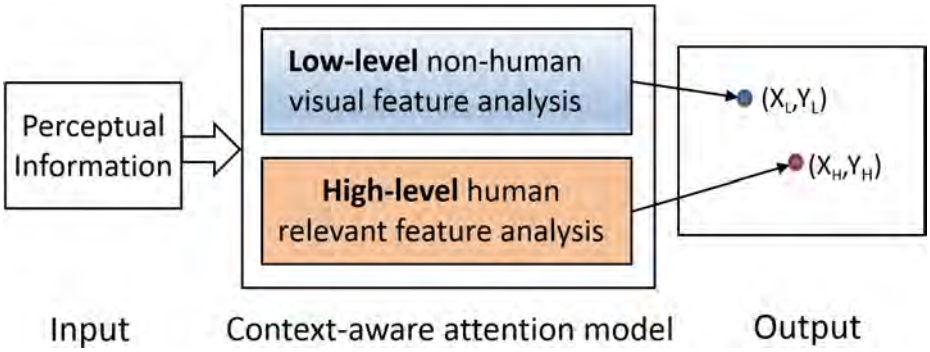


Figure 2.12: The proposed context-aware gaze control model capable of identifying attention targets analyzing both low-level visual features and high-level human-relevant features.

attention and gaze behavior systems, which are the focus of this thesis. We believe that using the proposed model, we can overcome to several shortcomings exist in the state-of-the-art.

Chapter 3

High-Level Features and Phenomena in Attention Elicitation

Contents

The previous chapter extensively reviewed the conventional class of attention system called "salience-based" that identifies attention targets dealing with the low-level features of a visual scene. The low-level features were described as phenomenon that attracts human attention in an environment that is out of human or a specific object. The concept of high-level features then was discussed against low-level features as phenomena that guide human attention in a human-centered situation.

Discussing several aspects of the salience-based attention model and also the requirements for social robots development, we concluded that, since the selection mechanism of the salience-based attention model is only based on the low-level feature analysis, and due to this fact that

human attention in a human-human interaction is affected by high-level human-relevant features, the current class of attention system is inefficient in simulating human attention especially in a human-centered scenario and thus it is not a right option for social robotic applications.

We described characteristics of an attention system that identifies attention targets dealing with both low-level features and *high-level* human-relevant features which is the focus of this work.

In order to have a behaviorally appropriate HRI, robot should focus its dynamic attention correctly toward a right target and at the right time but the concern is how the robot's attention should find its target point. What are the features and phenomena that manipulate the attention?

In spite of numerous unknown factors that manipulate the human attention, the effect of some high-level features on the attention is well explored. The following chapter briefly describes a few phenomenon and features that have been proven to guide human attention in a social scenario. They play a pivot role in the proposed attention system.

3.1 Non-verbal/Verbal Cues

Non-verbal cues comprise a large number of wordless signals that a person mostly uses to deliver a meaningful message to the interactional partner. These cues compose a significant part of the interaction (about two-thirds) between two humans [71].

People use their facial expressions, body gesture, head pose and gaze to attract other people's attention, to express their emotions, and intentions and to manage the flow of interaction while speaking or listening [72]. Figure 3.1 shows a wide variety of non-verbal cues that people usually use as communicative cues in a social interaction.

Therefore, detecting such high-level features is mandatory for human attention modeling and social environment understanding. For example,

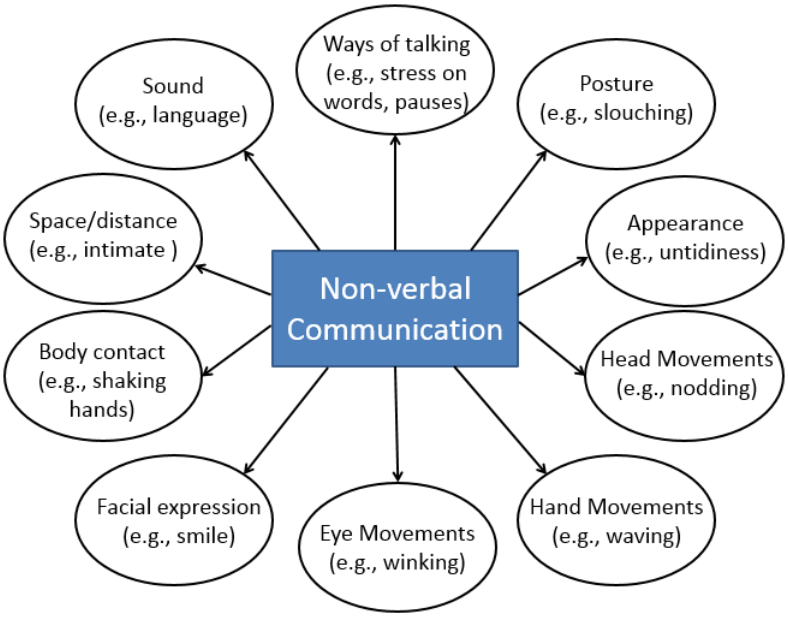


Figure 3.1: Non-verbal cues comprise a large number of wordless signals that a person mostly uses to deliver a meaningful message to the interactional partner.

imagine you are engaging in a social interaction with a group of friends. If one of your friends suddenly raises hand or shows a specific facial expression (e.g., smiling) or body movement (e.g., sitting, rising), your attention immediately will be attracted by that person to get more detail about the purpose of that meaningful motions, then it shifts your gaze to look at that person. We also experimentally proved this fact that high-level humanlike features affect the human attention selection mechanism.

Due to this fact, before designing an attention system, a perceptual system is designed to collect and recognize various high-level human-relevant features. It allows attention model to select its target point based on the dedicated selection mechanism, according to the human features.

Table 3.1 reports the high-level features that are considered in human attention modeling in this work.

Verbal cues consist of a variety of cues that directly affect the human attention and cause involuntary gaze shift. Literature [29, 73] indicates that, when someone speaks in a group, human attention immediately locates the speaker among others, and tries to understand the content. Vertegaal et al. [74], through a very accurate study using a gaze tracker in a group of four people, found that listeners looked at the person who was speaking 88% of the time. It means that, once a person speaks, attracts other people attentions. We have also experimentally found the importance of auditory signals and other features in triggering attention in a social interaction.

In a careful human gaze study, with a professional gaze-tracking device, we found that the auditory signals have a profound effect in attracting human attention in a social interaction, thus it should be considered in the attention modeling.

Therefore, in the proposed attention model of this work, we take into account the 3D position of speaker respect to the robot, and pronounced words are considered as strong cues that manipulate the robot's attention.

3.2 Proxemics

In addition to the gestural behavior, physical distance between people influences implicit and explicit interaction between them. Hall [75] investigated the effect of the physical space as the important non-verbal cue on the interpersonal communications. According to the anthropologist Edward T. Hall's theory, there are four invisible bubbles in the certain distances around our body that influence the social communication level between us. Figure 3.2 shows these spaces as *Intimate* (too close), *Personal*, *Social* and *Public* (too far). Subject's social cues thus,

elicit different level of attention depending on their spatial locations. For example, if a person shows a specific cue (i.e., hand motion, rising eyebrow, smiling); the attention of the surrounding people will be attracted based on their distances; It means that, the other people attention will be drawn quickly if the person is in the personal space, while it takes longer time if the person is in the social space and finally, there is a risk of losing attention if the person is in the public space. Besides, in a multiparty social interaction, humans prefer to interact with others that are in a closer physical distance.

On the other hand, proxemics phenomenon controls human behaviors when respond to the other people. For example, when we talk to our colleague in a social interaction, depending on the distance between us, we tune the loudness of our voice.

Due to importance of the proxemics, the proposed attention model of this work is able to classify people in trigging attention, based on their distances (spaces), in four groups. The importance of people for the attention model varies from high to low, according to their spaces. The spaces and the associated distances are illustrated in Figure 3.2.

Although, the proxemics factor can be influenced by other environmental features (e.g., lighting [76], setting [77], location in setting and crowding [78], size [79], and permanence [75]) and individual-relevant factors (e.g., involvement [80], sex [81], age [82], ethnicity [83], and personality [84]) but, due to limitation of sensing technologies in detection and recognition of such features, we use a standard style (four spaces) for proxemics, and we do not consider the effect of other factors (To see more information about proxemics and its implementation please see [85]).

In addition to the Hall's theory, there are several other behavioral studies for example [86,87] argued that, people look at close targets more frequently than distant target.

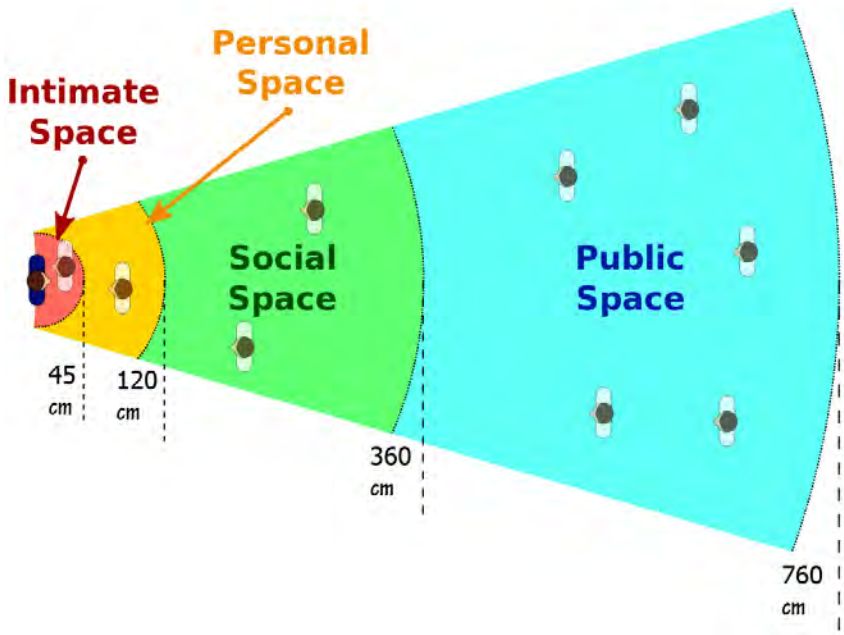


Figure 3.2: According to the Hall's theory, there are four invisible bubbles (spaces) in the certain distances around the human body that influence implicit and explicit interaction between people. The spaces from close to far are Intimate, Personal, Social and Public, respectively

According to discussed phenomenon, we use, the concept of *proxemics* (use of distance) as a sort of non-verbal cue that influences the total attention elicited by a person, in the attention modeling.

3.3 Effective Visual Field of View

Unlike the conventional vision devices that uniformly sample the environment in the field of view, the human eye collects the visual information at high resolution from a small central area called the fovea while, the

peripheral area is sampled at lower resolution [70]. In other words, visual stimulus angle, with respect to human sight line, affects human attention eliciting. Thus, human perception is more attracted by affective and social cues (e.g., facial expressions), in a central small area, known as effective field of view (eFOV). Clearly, the social cues in this area elicit higher level of human attention. Furthermore, several human behavioral studies indicates that, there is a strong eye tendency to look at the center of the image, regardless of the entire image contents [88, 89].

Based on these facts, we considered the angle between a detected feature and the center of eFOV, as a sort of non-verbal cue that influences human attention. As can be seen in Figure 3.3, we discretized angle related non-verbal cue in three levels: High, Medium, and Low.

The physical distances as well as human orientation are considered only for visual features in the FOV and not for auditory signals. Clearly, the auditory signals affect human attention regardless of the spatial position even outside field of view.

3.4 Habituation Effect

To make an effective and believable human-robot interaction, the habituation should be implemented on the robot [90]. Infants responds strongly to the new detected feature of environment, but respond less when they get familiar with that [91]. The habituation effect inhibit robot attention to be continually fascinated by only one target and allows it to see new features and target of environment and select new target point. This capability endows robot to display a dynamic attention behavior, when interact with multiple targets.

The habituation effect is a process that makes human attention adapted to the continuous existence of a new stimulus presented in environment. It is an adaptive behavior that causes a decreasing of the interest to a

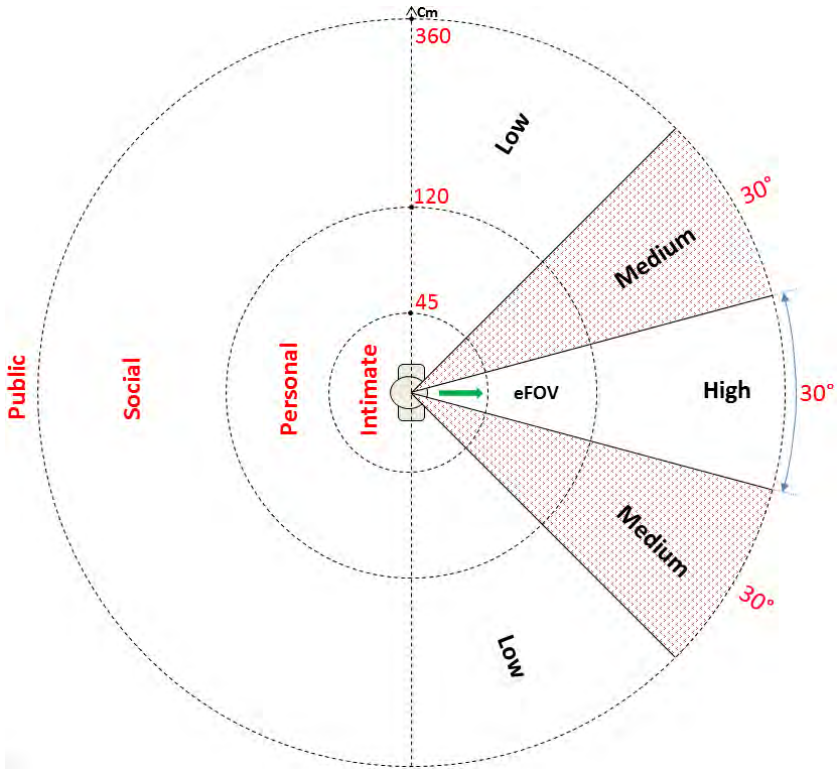


Figure 3.3: Right side semicircle: The orientation of people respect to each other influence the level of engagement in a social interaction. The importance of people, based on their orientation is grouped to High, Medium and Low.

new stimulus. This effect can be considered both for long-term and short-term issues. For example for when a new feature such as a sound signal is detected in environment, it highly attracts the attention of people at the beginning, but after a while, despite the existence of the feature the attention is adapted to that and the feature attract the lower level of attention.

The concept of habituation is implemented in our robot's attention

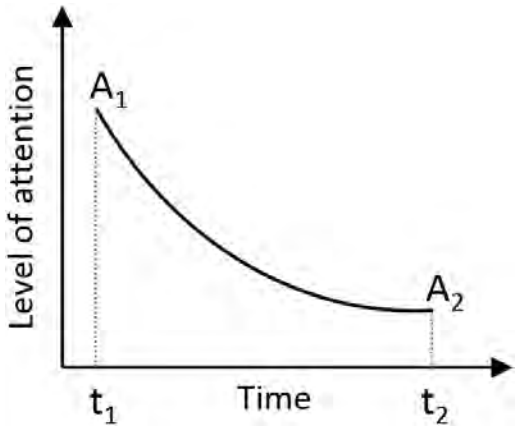


Figure 3.4: The elicited human attention level by a new feature decreases within the time from A1 to A2. Human attention is being adapted to a new detected feature within the time.

module, as a time-variant function (Figure 3.4) that adjusts the level of attention elicited by the selected target [60]. A detected target elicits highest level of attention, but due to habituation function, it lost its attractiveness linearly/exponentially within a time-constant.

3.5 Human Social Signals

Human social signals are known as important factor that influences the level of human elicited attention. A person that expresses a social signal to initiate a social interaction with us attracts our attention quickly. In addition, according to our experimental gaze study, we found that a person who enter and leave the social interaction attracts human attention; however we showed that the levels of elicited attention are different when a person recognizes different social signals.

Thus, in this thesis we consider human signals as the important factor

Table 3.1: High-level human-relevant features of a scene that are essential in human attention modeling.

No.	Features
1	Body gesture and head pose (roll, yaw, pitch angles)
2	Facial expressions (happiness, sadness, surprise, anger)
3	Facial features (age, gender: male/female)
4	Entering/leaving times (sec)
5	Proxemics (spaces: public, social, personal, intimate)
6	Effective field of view (spaces: low, medium, high)
7	Habituation effect

that has impact on our attention model. For example, we consider entering time and leaving time for a person that enter into the room and wants to initiate a social interaction and for a person that wants to leave the social interaction as high-level features that represent the human intention and influence people attention.

3.6 Summary

This chapter reviewed several high-level humanlike features (reported in Table 3.1) and phenomena that affect human attention in a social interaction and cause voluntary/involuntary gaze shift toward people of environment. The chapter also reviewed and discussed the associated theoretical background of the high-level human-relevant features in attention triggering.

Some of the high-level features/phenomena that have been proven to guide human attention in a human-human interaction are non-verbal and verbal cues, proxemics, the effective visual field of view, the habituation effect, and people intention. We consider all these features in designing

the attention model, presented in this work.

We believed that the new generation of attention system for social robotic application should be able to identify target points for the robot, dealing with the proposed high-level humanlike features as well as, the low-level visual features, instead of solely low-level features.

Chapter 4

Design and Implementation of Perception, Attention, and Gaze Control Layers

Contents

This chapter presents a system that enables social humanlike robots to autonomously identify the most important target (human/ non-human) of environment and adjust its gaze to look at the selected target, in a dynamic multiparty social interaction. The system receives as input, the constructed data of the Kinect and sends as output, the control signals to the robot's actuators.

The core of the system is an attention model called *humanlike attention*, which performs two major tasks. On one hand, it actively explores the acquired perceptual information and identifies the most important human/non-human *interactional target* of environment through a context-aware humanlike attention selection mechanism. On the other hand, it controls the robot's head and eyes -*gaze*- movement such that the robot displays a behaviorally acceptable and believable gaze shift toward selected targets, during a multiparty social interaction.

Two other important layer that are linked with the attention layers are perception and gaze control. The perception layer collects the raw data of environment and interprets the sensory information that are required by the attention layer. The gaze control layer manipulates the dynamic of the robot's gaze, such that robot appropriately makes an eye contact to the selected target by attention layer.

As shown in Figure 4.1 the proposed system functions as a middle component between robot and environment and closes the interaction loop between human and the robot. It enables the robot to monitor environment in real time and adjusts its behavior according to the scene content.

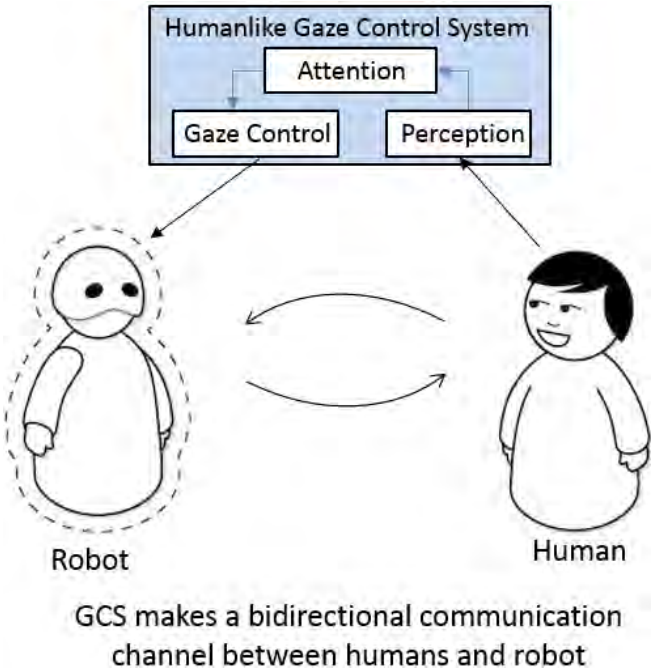


Figure 4.1: The overview of the gaze control system.

As discussed in the chapter 2, the conventional attention class called salience-based attention, identifies the most important target of a visual scene, analyzing only low-level visual features such as color, intensity, and etc., however as discussed in chapter 3, there are several important human-relevant features and phenomena that cannot be expressed through low-level features. Thus, we found that due to many limitations and drawbacks of the selection mechanism of this class of attention model, it is inefficient in HRI applications. Thus, we should move away from that approach toward designing a humanlike attention model. Such a model should be aware of both low-level visual features and high-level human-relevant features when identifies the most important target of a

visual-auditory scene. Hence, this work proposes a new class of attention model, which identifies targets by analyzing both low-level and high-level features of the environment. It explores the perceptual information of a given visual-auditory scene and selects the most important point of environment based on a humanlike selection mechanism. Then, the attention model controls the robot gaze movement based on a gaze model that is derived from real-data of humans.

4.1 System Overview

Figure 4.2 shows the modular structure of the proposed system that we call it context-aware humanlike *gaze control system (GCS)*. It consists of three distinct layers: *perception*, *attention* and *gaze control*.

Employing several parallel algorithm, the perception layer collects in real time the visual-auditory information of the environment, detects and analyzes a variety of low-level visual features and high-level human-relevant social cues, and provides a high-level interpretation of the environment for the robot.

The attention layer actively explores the perceptual information acquired by the perception layer and using a selection strategy, identifies the most important region of environment on which the attention of the robot has to be focused.

Using the humanlike gaze model, which is developed based on the state-of-the-art [40, 68, 70], the gaze control layer updates the robot's head and eye positions and generates a believable gaze movement for the robot.

Using the GCS, the robot has a humanlike understanding of the environment while it has a humanlike behavior in target selection and gaze movements, in a multiparty social interaction. The GCS makes a bidirectional channels between the robot and humans of environment that allows

robot to adjust its behavior according to the current social context.

The following sections introduce the humanlike robot and its capability in generating humanlike motion. Besides, it details the GCS layers and components and explains how the layers contribute the GCS to generate a humanlike context-aware gaze behavior.

4.2 FACE Robot

The proposed GCS has been designed and implemented on the humanlike social robot called FACE (Facial Automation for Conveying Emotions) that is created by Hanson Robotics [10, 22–24] (Figure 4.3). The robot has a female appearance and its artificial skull is covered by a porous elastomer material called FrubberTM which requires less force to be stretched by servo motors than other solid materials. FACE has 32 servo motors that allow it to replicate high-quality facial expressions (Figure 4.4) and humanlike head and eye motions [13, 14, 92]. The movements of head and eyes are in 4-DOF and 2-DOF, respectively. The kinematic structure of the actuation system enables the robot to generate realistic facial expressions and gaze behavior [93, 94].

4.3 Perception Layer

As shown in Figure 4.2, the perception layer is the first layer of the gaze control system that is directly connected to the Kinect. It is designed to simulate human perception by analyzing and interpreting both environmental (non-human) and human-relevant features. It provides a human-level understanding of environment for the robot.

The perception layer receives as input, the visual and auditory data of environment constructed by the Kinect, deeply analyzes the scene by several parallel algorithms and finally, it creates as output, the meta-scene

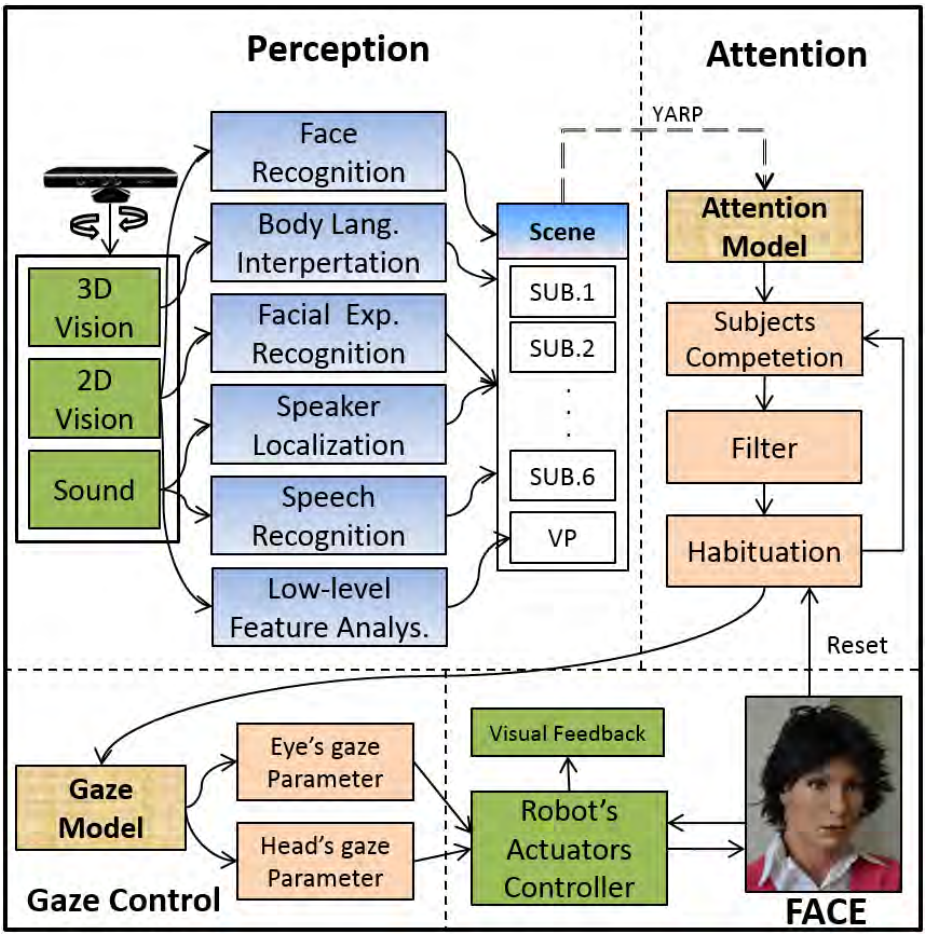


Figure 4.2: Modular structure of our gaze control system: the perception layer receives visual-auditory information from Kinect, extracts low-level and high-level features. Based on these features, the attention layer computes the most prominent target points. The gaze control layer drives the robot's actuators according to target positions using a gaze model.

object that contains two important information: the high-level features of humans presented in the scene, and the salient point that represent the

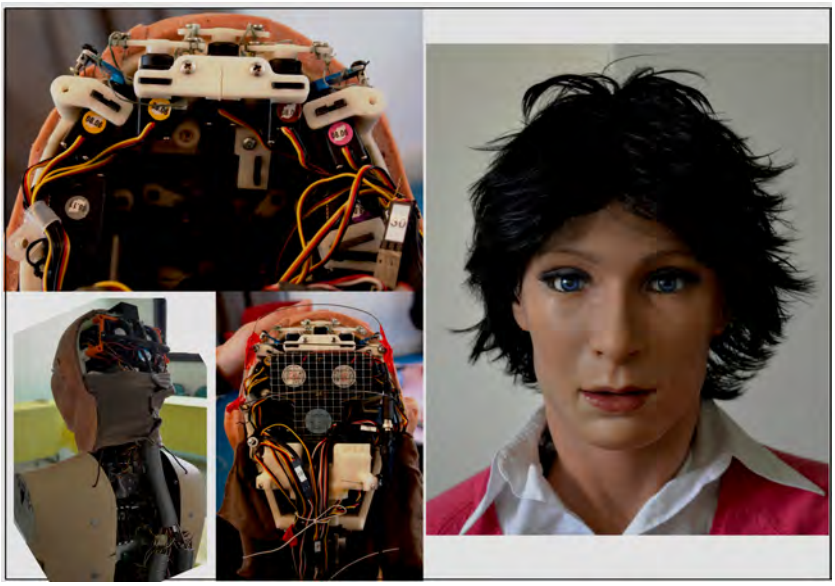


Figure 4.3: FACE's android actuator system consists of 32 servo motors together with artificial skin, allows FACE to reproduce high-quality facial expressions and humanlike gaze movements.

most important point of environment.

When a person enters to the Kinect field of view (FOV), the perception layer creates a sub-object in the meta-scene object, and recognizes and stores several high-level features associated with that person. In addition, it analyzes the low-level visual features of the scene and identifies the salient point in pixel (X,Y). Figure 4.5 shows the structure of constructed meta-scene object and it reports high-level human-relevant features as well as the salient point, recognized by perception layer.

As illustrated in Figure 4.2 the perception layer contains two parts: data acquisition and feature extraction. These parts are deputed to prune data and extract low-level and high-level features from visual-auditory information of a social scene.

Figure 4.4: FACE is capable of reproducing a wide variety of human facial expressions and humanlike motions.

The perception layer acquires raw data through a Microsoft Kinect device as RGB-D images running the Kinect for Windows SDK by Microsoft[1]. Kinect RGB-D camera records 2D video and depth images with a resolution of 640x480 pixels at 30fps, and it has a built-in four-element microphone array for audio beam acquisition.

Kinect acquired raw data are analyzed extracting a variety of low-level visual features and high-level human-relevant features (verbal and non-

[1] `http://www.microsoft.com/en-us/kinectforwindows/`

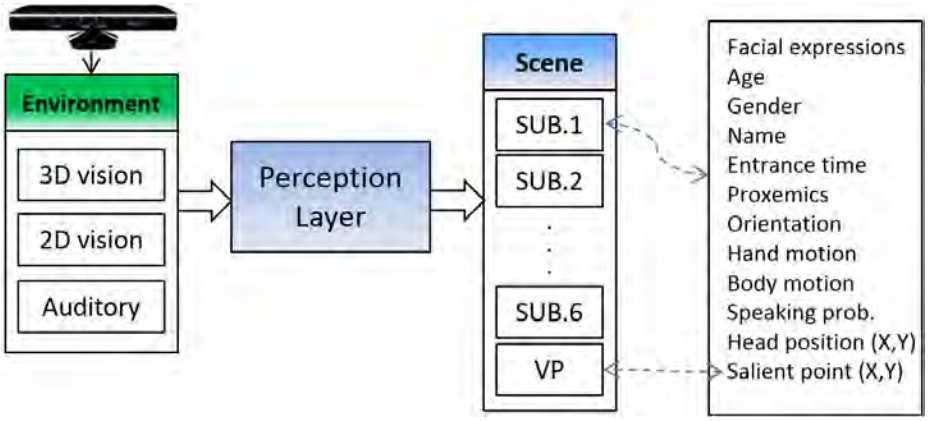


Figure 4.5: The perception layer receives visual-auditory scene constructed by sensor as input and creates a meta-scene object that contains several high-level features of humans presented in field of view. The meta-scene object contains humanlike information that provides a high-level humanlike understanding of the environment for the robot.

verbal cues). The perception layer classifies all the extracted features in different taxonomies and stores them in the *meta-scene object* and streams it to the attention layer through a YARP [95] gateway.

GCS implementation aims to extract social relevant visual features (i.e., human proxemics, orientation, facial properties, gestures, and entry time) and auditory features (i.e., sound source angle and pronounced words), through various parallel algorithms and or dedicated libraries.

In addition to the high-level human-relevant features, the perception layer identifies the most important environmental target, analyzing low-level visual features of 2D image by using a feature analysis engine called SUN.

Table 4.1 summarizes the algorithms and libraries, which are used in perception layer together with corresponding high-level features that are extracted by these libraries. The following section describes in detail each layer and components of the perception layer.

Table 4.1: Human-relevant features and salient point extracted by the perception layer

Used Library	**Extracted Features**
Kinect SDK	Human 3D position of up to six people Twenty body joints coordinates for 2 humans Sound direction and beam angle
SHORE	Positions of face, eyes, nose, and mouth Eyes and mouth state (open/close) Gender classification (male/female) Age estimation (years) Facial expressions Face rotation (up to ±60 image plane)
Face recognition	Name of human (according to pre-trained data set)
Body Gesture and Head Pose Recognition	Gestures and body motion Head pose (roll, yaw, pitch angles)
fastSUN	Virtual point (X,Y)

4.3.1 Face Detection and Facial features Analysis - facial expressions, age, gender

Observation of human visual attention revealed that face-like shapes attract human attention [96]. In addition, various features such as human's age and facial expressions (e.g., happiness, sadness, surprise, anger), directly regulate the social interactions between people [97]. In a social context, it is imperative to know the age and gender of interactional partners and to continuously receive feedback of facial expressions and

mimics. Analogous to humans, robots should have the same ability in localizing faces and understanding facial expressions and related social features.

For face detection and facial features (i.e., facial expressions, age, and gender) analysis, the perception layer uses the Sophisticated High-speed Object Recognition Engine called SHORE [98, 99].

SHORE is a robust detection engine that works based on the illumination invariant approach, and detects multiple faces in a single visual frame and tracks them in real time within a video frame. SHORE engine receives the 2D frame constructed by Kinect, detects faces in real time, and assigns consistent ID to each face. It estimates various facial features such as four universally agreed facial expressions (i.e., happiness, sadness, anger, surprise) in percentage, age, gender (male/female), eyes and mouth state (open/close), positions of face region, eyes, nose, and mouth in pixel.

When the SHORE detects a new face, an internal timer will be generated that shows the entrance time of the user (see Table 4.1). Figure 4.6 and Figure 4.7 illustrate the SHORE capability in recognizing various features of human face in a 2D scene, for a single and multiple faces.

SHORE as a reliable C++ library enables the perception layer to sense the presence of human from distant and in any lighting conditions, and track multiple frontal/rotated faces with the high degree of robustness against background characteristics. It estimates correctly the human facial expressions (i.e., happiness, sadness, anger, and surprise) which allows the perception layer to monitor in real time the apparent emotion of each human during a multiparty social interaction. SHORE precisely estimates the humans gender (as male or female) with almost 100% degree of precision and estimates the age of each person in the FOV.

SHORE as an important component of the perception layer, empowers

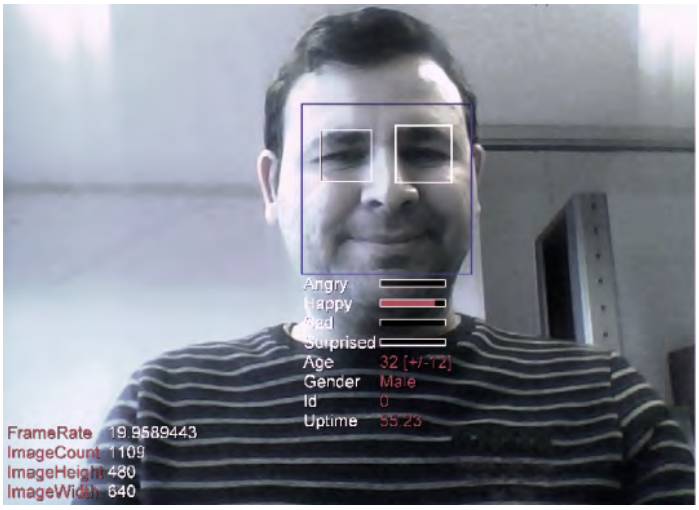


Figure 4.6: SHORE (Sophisticated High-speed Object Recognition Engine) detects multiple faces in a 2D image and tracks them in a real time video. It estimates several facial features such as facial expression (i.e., happiness, sadness, anger, and surprise), age (year), gender (male/female) and entering time. It assigns a consistent identification number to each of the recognized faces and tracks it in real time.

the GCS to simulate more and more the human perception especially in sensing the presence and the localization of multiple people as well as in monitoring their apparent emotions and other facial features. Using SHORE, the perception layer is able to assign a consistent identification number to each of humans and re-assign the same ID in the case of lost tracking, if the human does not change his/her position, which is essential for a successful HRI.

SHORE provides the pixel address (X,Y) of each detected face as well as the pixel address of the eyes, nose, and mouth in 2D visual frame. It enables the perception layer to guide the robot to make an eye contact with humans. Due to internal timer of SHORE, it stores the entrance

Figure 4.7: SHORE is capable of analyzing facial features of multiple humans in an image and track them in real time.

time for each human. This capability enable the perception layer to recognize the new entry human to control the behavior of the robot (e.g., welcoming).

4.3.2 Multiple Face Recognition Using PCA

Facial features and expressions recognition are very important for social context analysis but require the integration of subjects' identity information to allow the robot to adjust its behavior in a context-aware manner. The GCS perception layer integrates a facial recognition engine based on Principal Components Analysis (PCA) [100].

In the face recognition component, face images in a 2D frame are detected and projected into a face space (feature space) that best encodes

the variation among known face images. The face space is defined by the Eigenfaces, which are the Eigen-vector of the set of faces. Therefore, each face is represented by a set of features that require less computation for recognition compare to the whole face image.

The facial recognition module uses a pre-trained data set to assign an identity to recognized faces and stores the extracted features in the faces data set. Figure 4.8 shows an example of perception layer merged information on which recognized the subject's name, the ID, and the facial information (i.e., estimated happiness ratio, age, gender and entry time) from SHORE, are merged.

Using the SHORE ID along with the PCA result, the perception layer is able to recognize faces even for non-fully-frontal faces (see Figure 4.9).

4.3.3 Body Gesture and Head Pose Estimation

People use body gestures and head poses as social signals when they interact with each other [72, 101]. These social signals are strong non-verbal cues that elicit human attention. For example, in a multiparty interaction, if one of the humans raises his/her hand or waves the arm, others will direct their attention to him/her. Robots thus need to be able to react to these social cues.

The perception layer uses the skeleton tracking of the Kinect SDK to recognize a person's movements. The Kinect SDK locates up to six humans by merging information from RGB and depth images and recognizes body joint coordinates for the two closest persons (Figure 4.10).

In order to estimate the head pose, the perception layer computes Euler angles (pitch, roll, and yaw) in real time, using SDK's head data. In addition, we implemented a dynamic body gesture and head pose recognition engine, which continually monitors the body's motion and head pose in the absolute coordinate through extracted skeleton information,

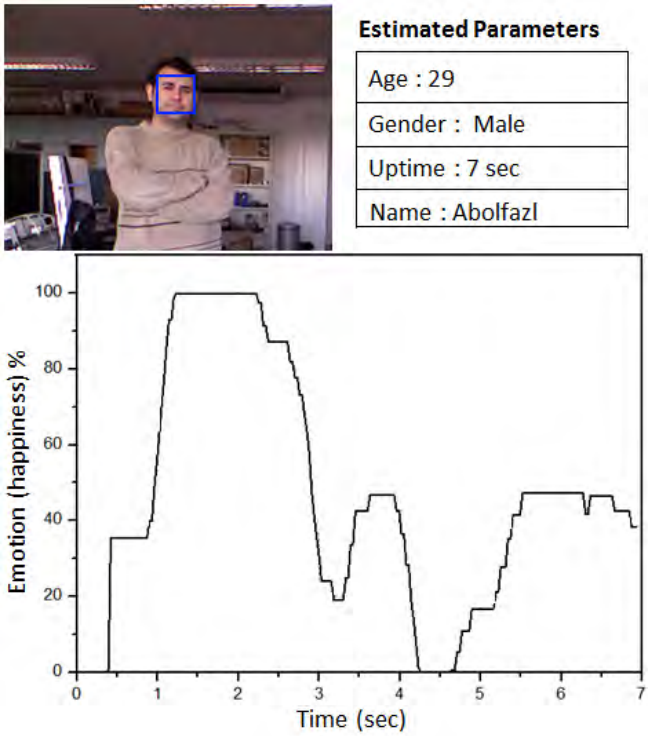


Figure 4.8: Example for final integration of face detection, facial expression/features estimation and face recognition in the perception layer. The module detects a face and extracts estimated happiness ratio, age, gender and entry time. It also compares the detected face with the database in order to identify the name of the person.

and identifies meaningful motions.

4.3.4 Speaker Localization

The auditory streams cause an unintentional shift of attention that most often shifts the gaze of the person towards the sound source. Hence, it is essential for robots to locate the speaker in a multiparty interaction.

Figure 4.9: Face recognition engine compares each of the detected face to the subjects' database and assigned IDs to the detected faces.

The perception layer uses the Kinect SDK to calculate the sound source direction with a triangulation algorithm. It computes the 3D position and beam angle of sound signals received by the microphone array. The algorithm considers only auditory signals that can be associated with humans in the scene by comparing the direction of the sound to the 3D positions of the detected humans.

In a real situation, human attention is also attracted by auditory signals outside the visual field of view, which are not relevant to the visual stimuli. However, due to the limitation of sensor detection range, the system is designed to ignore sound signals, not related with multiparty interaction, as environmental noise. This limitation of the system is one of the issues that prevent natural gaze behavior from being generated.

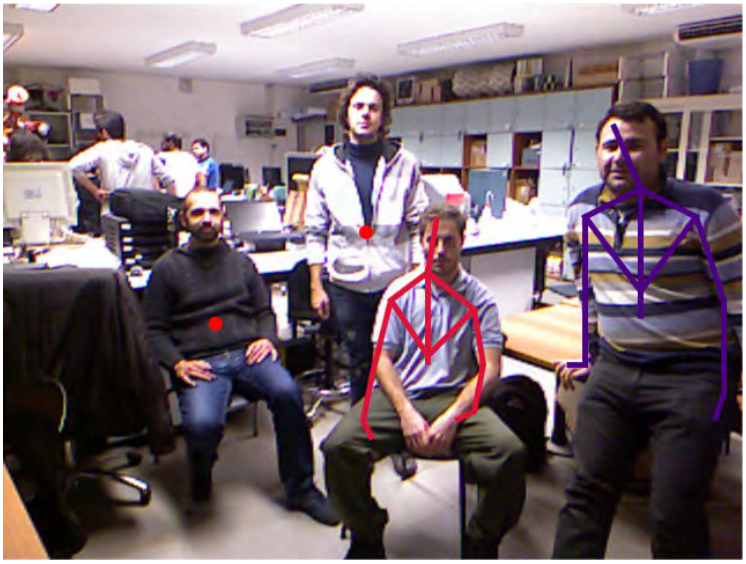


Figure 4.10: The perception layer uses the skeleton tracking of the Kinect SDK to recognize a person's movements. The Kinect SDK locates up to six humans by merging information from RGB and depth images and recognizes body joint coordinates for the two closest persons

Once a sound source is associated with a person, a dedicated engine is used to recognize speech and convert it to text if possible. The human's recognized word are stored in the *meta-scene objects* along with a *speaking probability* parameter calculated based on a comparison between the sound angle and the human's position.

4.3.5 Visual Salient Point as a Virtual Subject

In addition to the high-level human-relevant features extracted by dedicated methods, the perception layer analyzes low-level visual features of the 2D scene and identifies the most important no-human target (salient)

point, using a SUN (Salience Using Natural statistics) based engine.

SUN is a Bayesian framework for identifying salience regions of a visual scene, using natural statistics [102]. It is designed to identify potential targets in a 2D image that attract human attention in a complex scene viewing. To achieve this, the model estimates the probability of a target at every location given the observed visual features. Through a competition between feature space or salience map, the model identifies the salient point that is the most important point of the given scene as no-human target or virtual point (VP), with a very little computational cost while leaves plenty of CPU cycles for other tasks. Figure 4.11 shows a visual scene and the associated salience map reconstructed by the SUN-based engine. As shown, the features regions of the visual scene are illustrated in the salience map, as bright regions and the salient point or VP is identified as the brightest point of the salience map.

To enable the perception layer to identify salient target in real time, we integrated the fastSUN [103] component to the layer, which is an efficient implementation of the SUN algorithm for real time application.

The fastSUN receives the constructed 2D scene through Kinect, creates corresponding salience map according to the features regions of the scene, and identifies in real time the potential target point in pixel (X,Y) analyzing low-level visual feature of scene and without considering high-level features. It identifies in real time salient point that is called in this work, as Virtual Point (VP). In fact, the VP as a point of visual scene shows a salient point of environment in term of low-level visual features. The position of VP actively change, according to the scene-context. Presence of VP in the perception layer allows the robot attention to be attracted by the environment when the interactional partners do not show any specific feature or social cues. In other words, when humans are not interesting enough, the robot's attention prefers to switch on envi-

Figure 4.11: The perception layer analyzes low-level visual features of the 2D scene and identifies the most important no-human target (salient) point on using a method that is based on the attention model called SUN (Salience using Natural Statistics). The upper part of image is the original scene image while the lower part shows the reconstructed salience map. The region of features are illustrated as bright areas.

ronment instead of humans. Such capability significantly improves the believability of the robot gaze behavior.

The perception layer stores the address of identified salient point in pixel (X,Y) into the meta-scene object and updates its address in real time.

4.3.6 Subjects Database

Once the perception layer has identified the virtual point and recognized high-level features, a database of all people seen by the perceptual layer (Figure 4.12) is stored as the *meta-scene object*. The *meta-scene object* has a hierarchical structure through which an arbitrary number of people can be inserted. Each person's object includes the person's unique ID and the associated high-level features.

Once a new person has been identified by PCA identification algorithm and by SHORE, a new person instance is created in the *meta-scene object*, which is populated with the features, extracted by the perceptual layer. Since PCA engine recognizes frontal faces, new pictures are continuously taken by RGB image and stored in the PCA training set. PCA unrecognized humans are inserted into the *meta-scene object* using a temporary ID (taken from SHORE) which is re-assigned, once the operator has inserted the new person name through the GCS interface.

The meta-scene object, in fact is a high-level interpretation of the environment and it provides a humanlike understanding for the robot. It enables the robot to understand the sensory information in the way that human being does. Using the meta-scene object, the robot is able to explore perceptual information and selects its target as well as adjust its behavior according to the target status. For example, if a person intent to initiate a social interaction, robot can greet that person.

Due to importance of the meta-scene object, it should be accessible for

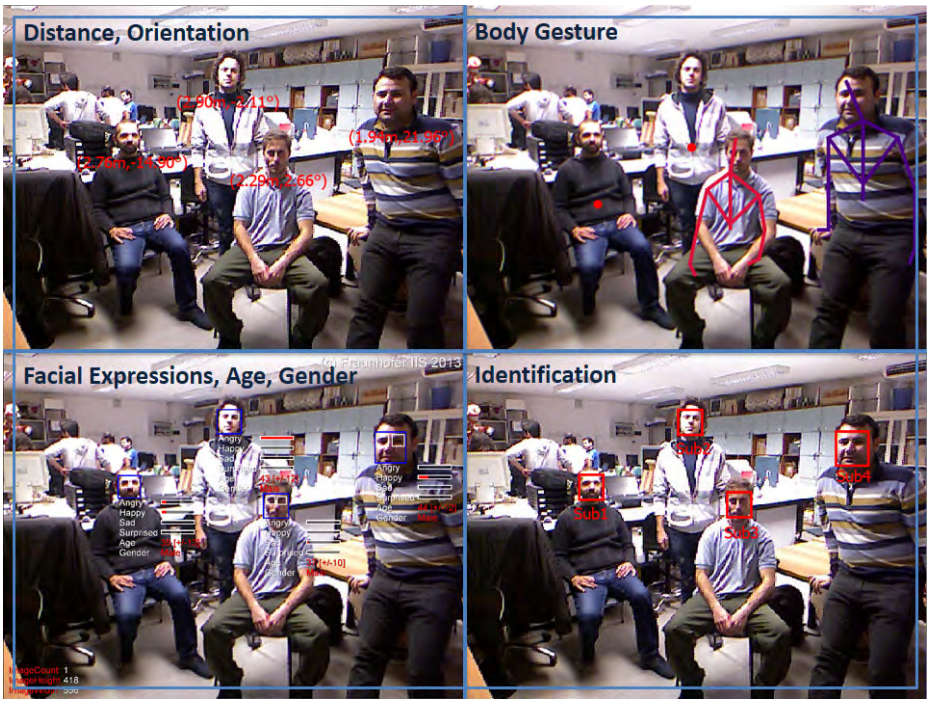


Figure 4.12: The extracted high-level human-relevant features of multiple humans, by the perception layer.

other layers and components especially attention layer. For that reason, through the .NET object serialization, the *meta-scene object* is converted into an XML structure, which is streamed through a dedicated YARP port between the GCS layers and modules.

4.3.7 Communication Channel Through YARP

As discussed, the GCS is responsible for providing a human-level understanding of environment as well as controlling the robot's actuators in real time. Since on one hand, data acquisition and low-level/high-level feature extraction, and on the other hand, controlling the robot in real

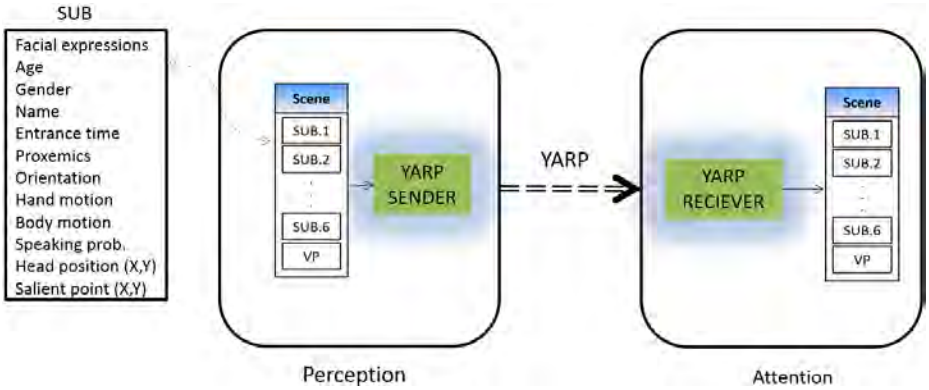


Figure 4.13: YARP creates a bidirectional wireless communication channel between the perception and the attention layers. YARP sender delivers a created meta-scene object from the perception layer to the attention layer. YARP receiver, receives the meta-scene object and converts it in a manageable object.

time take a lot of processing cycle, GCS layers are not able to operate at same machine. To solve this problem, we used two different machines for perception and attention/gaze control layer. The YARP [95] is created as a wireless data communication channel between the perception and attention layers.

YARP (Yet Another Robot Platform) is an open-source software package, written in C++ that makes a reliable communication channel between two machines, sensors and actuators in order to send and receive data. YARP communication gateway is established in order to make a bidirectional connection between modules and layers. The perception layer acquires data from Kinect and creates meta-scene object containing features of subjects as well as the virtual point, identified in the FOV. It then sends the created object to the attention layer by *yarp sender*. The attention layer that is connected to the gaze control layer receives the meta-scene object and through a selection mechanism, specifies target points. It sends the pixel address of the identified target to the gaze

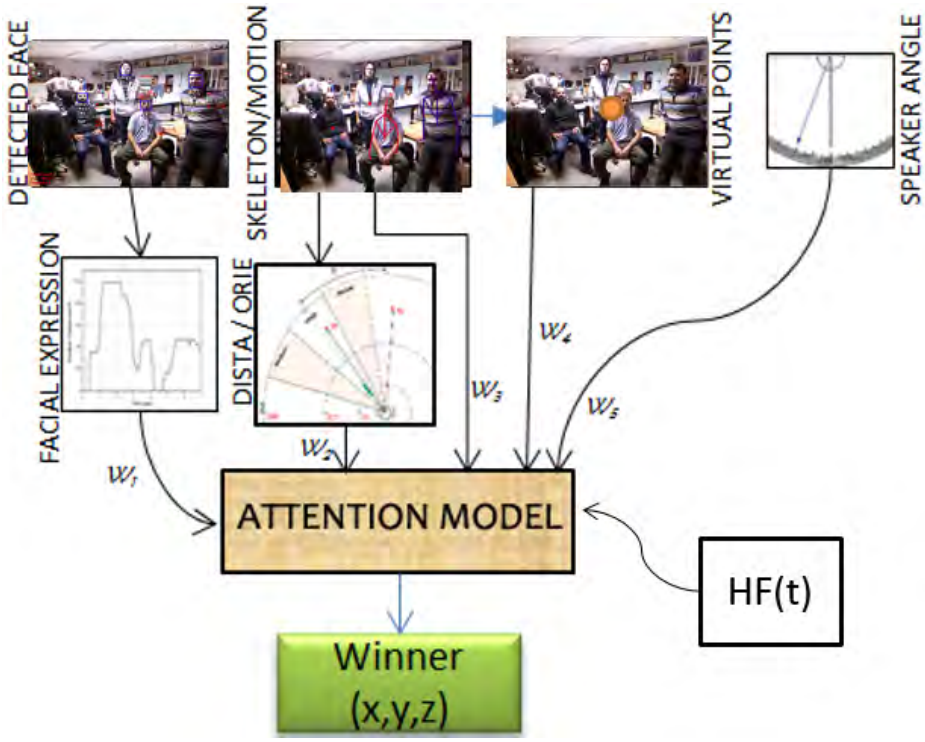


Figure 4.14: The attention layer receives the meta-scene object as XML streamed through yarp receiver, and then deserializes it back in a manageable object.

control layer, in order to drive robot's actuators to shift its gaze.

4.4 Attention Layer

As Figure 4.14 shows, the attention layer receives the *meta-scene object* as XML streamed through *yarp receiver*, and then deserializes it back in a manageable object to be useable for the attention layer. The aim of designing the attention layer is endowing the GCS to find the most prominent region (*target*) of the scene that the robot should focus on.

There are two types of target that the attention layer should select: human subject and virtual subject. The attention layer selects a human subject, based on a target selection strategy, in order to engage people to the social interaction or perform a specific task. While, it selects the virtual subject (salient point) that has been identified by perception layer, when the humans are not enough interesting for the robot. The virtual subject enables the robot to have a dynamic believable motion.

4.4.1 Target Selection Strategy

The core of the attention layer is a model that calculates the elicited attention (EA) level as a score, for each human presented in the scene. The score of each human is calculated according to his/her high-level features. It means that, if a person shows more features, gets higher score with respect to other humans presented in the scene. Evaluating subjects, the attention layer selects a winner in real time as the one with highest score, who is the most important (interesting) for the robot.

Since the numerical values quantifying the features are not within the same range, they are normalized (X_n) to the range $[0, 1]$ by considering the maximum values that features can have according to the sensor properties and features ranges. The overall EA/score of each human in the scene is calculated based on four main components: social features (F), proxemics (P), orientation (O), and a memory component (EAM)

$$EA_{S_j}(t) = F_{S_j} + P(r) + O(\theta) + EAM_{S_j} \tag{4.1}$$

where EAM_{S_j} is a parameter that refers to the memory of the robot not yet included in the database and consequently set to zero.

The social feature elicitation contribution F_{S_j} is calculated as a weighted summation of social normalized features X_n, which can be written as

$$F_{S_j} = (\sum_{i=1}^{n} W_i.X_n) \tag{4.2}$$

where weight W_i is set according to the the feature's importance. The weight parameter adjustment is the most important issue that results a humanlike gaze behavior for the robot. It have to be determined based on the priority that humans have in selection features in a social interaction. For example in a group, in a simple case, if a person speaks, others shift gaze to make eye contact with speaker. In a more complex social scenario, the gaze behavior of people should be determined. For that reason, we conducted a gaze tracking study with a gaze tracker device in a social human-human interaction and determined the average priorities that people shows in selecting features of environment. The feature's importance and priority are explained in chapter 5. Figure 4.15 shows the graphical user interface designed in order to adjust the weights of the attention layer by operator.

The values of $P(r)$ and $O(\theta)$ in Equation 5.1 reflect the proxemics and orientation contributions in the model (described in sections 3.2 and 3.3). When humans have almost the same distances and orientations with respect to the sensor, these parameters reflect the same values for them. Due to the unavailability of sensory data in nearby and distant areas, the attention layer reflects the proxemics effect only for personal and social spaces and the orientation effect only for high and medium spaces. These effects can be expressed as

$$P(r) = Fp_r.(1 - \frac{|r|}{r_{max}}) \quad O(\theta) = FO_\theta.(1 - \frac{|\theta|}{\theta_{max}}) \tag{4.3}$$

where $|r|$ and $|\theta|$ denote the current distance and orientation of humans with respect to the robot. Fp_r and FO_θ convert continuous distance and orientation into discrete values, respectively. These discrete values

represent four proxemics spaces (intimate, personal, social, and public) for Fp_r, and the three zones of the eFOV (high, medium and low) for FO_θ (see Figure 4.16). r_{max} is the maximum distance and θ_{max} is the maximum angle detectable by the sensor. Clearly, the levels of $P(r)$ and $O(\theta)$ are at their maximum when a human is in the *intimate* space and the center of the eFOV of the robot.

Since the human's orientation is calculated with respect to the robot's current head position, the Kinect sensor should simultaneously turn with the robot's head to capture the same scene as the robot. For this reason, a servomotor is used to horizontally rotate the Kinect at the same angle as the robot's head ($\pm\beta$) (see Figure 4.16).

The attention layer shows a strong tendency to move to the VP when the human subjects are not enough interesting for the robot and their scores are lower than a threshold. Hence, a virtual point (VP) is positioned according to the low-level features of the scene, to attract the robot's attention like a virtual human.

The EA is simultaneously calculated for six humans in the robot's field of view. The attention layer selects the winner (i.e., the human with the highest EA level) through a competition among humans and the VP

$$Max(EA_{S_1}, EA_{S_2}, ..., EA_{S_6}, VP) \rightarrow K_{winner} \rightarrow (X, Y) \tag{4.4}$$

where K_{winner} is the winner's ID.

The attention layer then extracts the winner's head position (X, Y) or virtual subject pixel address (X,Y) from the meta-scene object and sends it to the gaze layer.

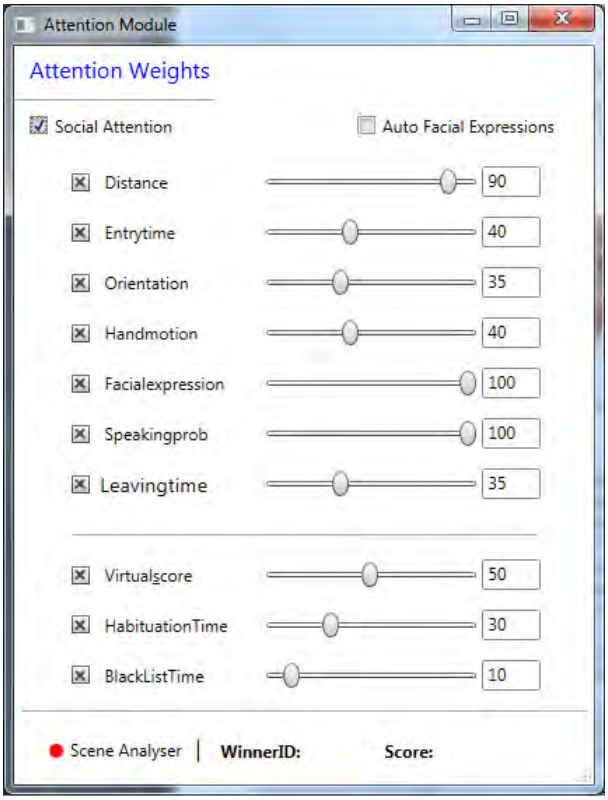


Figure 4.15: Graphical user interface designed in order to adjust the weights of the attention layer by operator.

4.4.2 Habituation Function

The habituation effect is activated, once the robot makes eye contact with the selected human (winner). The attention layer multiplies the habituation function (HF) by the winner's score (EA_{S_k}), in order to make a time-variant decreasing score ($EA_{S_{winner}}(t)$) for the winner, as

$$EA_{S_{winner}}(t) = EA_{S_k} \; . \; HF(t) \tag{4.5}$$

where

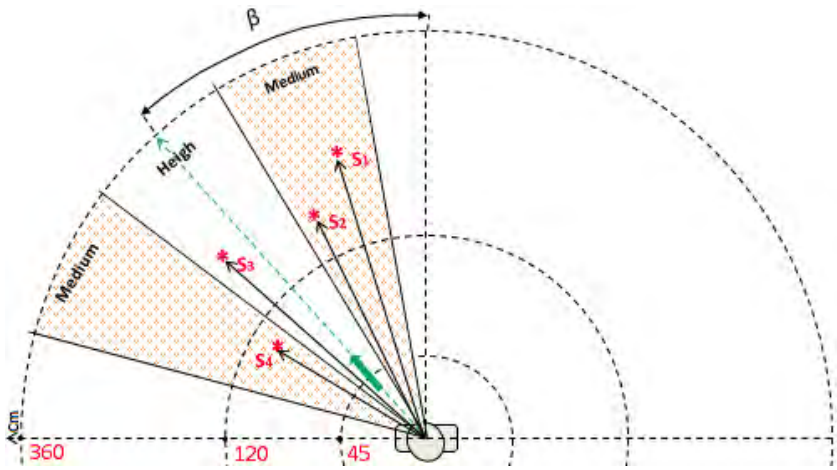


Figure 4.16: The Kinect sensor horizontally rotates at the same angle as the robot's head in order to capture the same scene.

$$HF(t) = Peak \,.\, Max(0, (1 - \frac{\Delta t}{\tau})) \tag{4.6}$$

and τ is a time constant and $Peak$ is the maximum amplitude of HF. Following [60] we set the time constant and peak parameters to 10 and 30 seconds, respectively. The HF value linearly decreases to zero within the time constant τ. When the robot's gaze reaches the new winner, Δt is reset to zero and HF will be maximized. The model searches for a new winner in real time while decays the score of the last winner to zero. Employing this system, the winner's attractiveness for the robot decreases gradually over time thus allowing other people to attract the robot's attention. It empowers the robot to show a more natural and dynamic behavior.

This capability allows the robot to display different emotions through gaze behavior. For example, according to evidences, a happy person makes longer eye contact or an anxious person makes shorter eye contact

with the higher frequency, thus the amount of time constant τ enables robot to shows these emotions.

4.4.3 Time-based Filter

Due to the mechanical limitations of the robot's head and eye actuators, the robot's gaze is not capable of synchronizing with the rapid changes in target positions. To solve this problem, a time-based filter is used that ignores rapid changes of attention point.

The attention layer sends the winner's position to the gaze control (GC) layer in real time, which is entrusted with generating gaze parameters according to the target position. The GC layer continually receives updates from the attention layer and decides how to direct the robot‘s gaze to the selected human.

4.5 Gaze Control Layer

The gaze control layer receives the target address and remap the address (X,Y) to the robot's actuator control coordinates. It then generates the robot's control signals in order to move the eye and head actuators in the way that robot shows a humanlike gaze motion.

4.5.1 Head and Eye Movements

A gaze is composed of two components: eye movement and head movement. The summation of these components (gaze) is relatively constant [68] (Figure 4.17). The amplitude of the gaze can be written as:

$$\theta_g = \theta_e + \theta_h \tag{4.7}$$

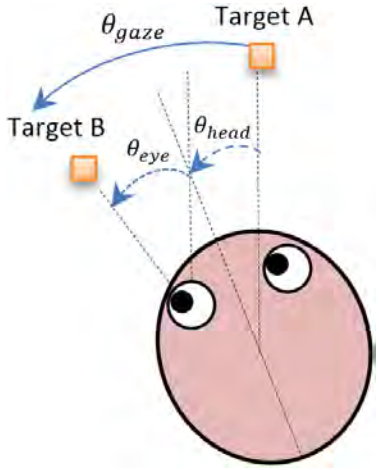


Figure 4.17: A gaze is composed of two components: eye movement and head movement. The summation of these components (gaze) is relatively constant.

where θ_e is the eye angle in its orbit with respect to the head (internal coordinates), θ_h is the head angle with respect to the environment (global coordinates) and θ_g is the gaze angle in the global coordinate system. Since the gaze angle is constant, any combination of head and eye is possible; where the angle of the eye increases, the angle of the head decreases and vice versa.

Assuming that the eyes are at the center of their orbit before gaze shift, $\theta_e(t = 0) = \theta_0$ is equal to zero. In order to accomplish a gaze, the eye moves until it reaches the threshold θ_{thr} and the head movement starts to compensate for the eye movement. If the eye's current position is not at the center of its orbit, θ_{thr} varies. In fact, the initial angle of the eye (θ_0) and the position of the selected human determine whether the gaze needs to be accomplished by eye movement alone or together with head movement.

In order to ensure a humanlike gaze shift, we use a humanlike gaze

model [68,70], which is derived from a motion capture of human subjects, using high-speed video-based eye and head tracking. The equations for θ_{thr} and θ_h were estimated, using empirical data. In this model, θ_{thr} varies depending on the initial position of the eye in its orbit (θ_0), and were obtained as

$$\theta_{thr} = -0.28\theta_0 + 11.2 \tag{4.8}$$

where θ_0 is positive if the initial eye deviation has the same direction as the subsequent movement. This equation is obtained from [70]. The constant numbers express head and eye dynamics in vertical and horizontal movements.

Following this notation, to accomplish a given gaze (θ_g), θ_h can be obtained as

$$\theta_h = \begin{cases} 0 & \text{if } -\theta_{thr} < \theta_g < \theta_{thr} \\ \\ \theta_{thr} + k(\theta_g - \theta_{thr}) & Otherwise \end{cases} \tag{4.9}$$

where

$$k = 0.0185\theta_0 + 0.715 \tag{4.10}$$

and k is a parameter that controls eye and head movement, in order to generate a humanlike gaze shift. This equation is derived from [70] based on empirical data (Figure 4.18).

4.5.2 Head and Eye Velocities

Investigation on humans and monkey's gaze behavior revealed that the gaze is not simply shift head and eyes to a target point. In a movement especially social gaze, the velocity of head-eye is an important factor

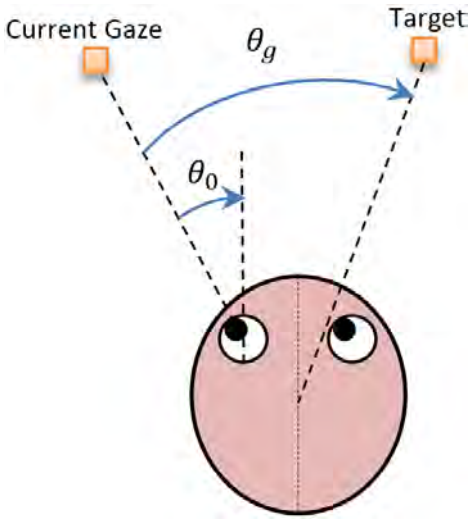


Figure 4.18: Depending on the initial eye position, gaze is accomplished by either only eye movement or head-eye movements.

that show meaningful information to the interactional partner. For that, we should apply a velocity model to the gaze control layer such that it performs a humanlike movement.

The head and eye velocities vary according to target eccentricity and modality [68]. However, the auditory and visual targets influence the velocities of the head and eyes in different ways. In this work, it is assumed that visual and auditory stimuli have the same effect on the robot's gaze. When the attention layer selects the target's coordinate in a pixel (X,Y), the GCS gaze control layer calculates the amount of target eccentricity with respect to the current sight line of the robot.

The authors of [68] showed that there is a relatively linear relationship between target eccentricity and head and eye velocities. Thus, due to the limitation of the robot's mechanical structure, three levels of velocities are defined as *high*, *medium* and *low* for the robot's actuators. The GCS

gaze control layer calculates the level of the head and eye velocities as a function of head and eye amplitudes, assuming that the eye always moves faster than the head.

The concept of velocity is implemented in the GCS by the amount of gaze angle (in degrees) over time needed to reach the target point (in seconds). Velocity can be expressed for the head as

$$[Vh_{high}, Vh_{medium}, Vh_{low}] = [75, 45, 22]\,^{\circ}/sec \tag{4.11}$$

and for the eye as

$$[Ve_{high}, Ve_{medium}, Ve_{low}] = [450, 150, 90]\,^{\circ}/sec \tag{4.12}$$

4.5.3 Head and Eye Latencies

Latency is the delay in reaction time when people shift their gaze to a target. It is influenced by target eccentricity and modality. Head latency is longer than eye latency [104], and varies approximately in the range of 50 ms to 300 ms.

Auditory stimuli have the longest reaction latencies for central targets, $\theta_g < 20$, and the visual targets elicit the longest reaction latencies in $\theta_g > 40$ (see [68]). In order to reach the target points, the model generates rapid saccadic eye movements with a 50 ms delay, then after a 200 ms delay, it generates head movements for the robot. Two constant values (l_e , l_h) denote eye and head latencies in the model, respectively.

The GCS gaze control layer estimates the gaze parameters for eyes and head, based on the proposed gaze model and target eccentricity as (θ_{thr} , θ_e , V_e) and (θ_{thr} , θ_h , V_h) for the robot actuators. It also generates reaction latency. All of the derived information is sent to the Robot Control (RC) layer, which is directly connected to the robot actuators.

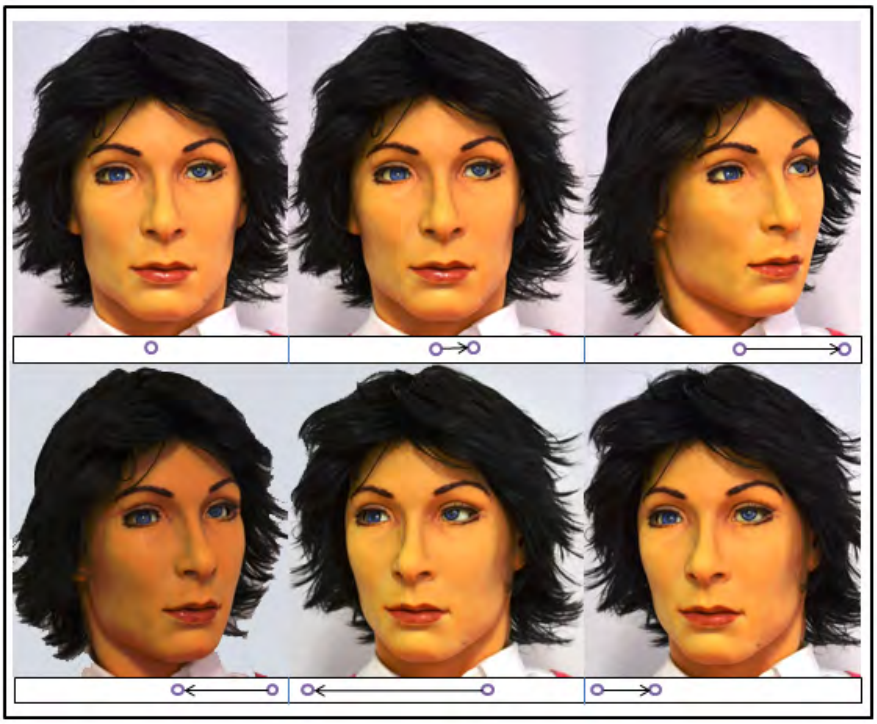


Figure 4.19: Sample gaze of FACE when it gazes at a target point. For a small movement only the eye actuator is driven while for the large movements head-eye actuators are driven.

Figure 4.19 shows how the robot control layer drive the robot's actuators in order to reach the robot's gaze to a target point.

4.6 Summary

This chapter described a context-aware gaze control system (GCS) that empowered a humanoid social robot to interact autonomously with a group of people in a social scenario. The proposed system consisted of three layers for perception, attention, and gaze control, which mediated

the robot and the environment.

The perception layer that was connected to the input sensor was responsible to create a humanlike understanding for the robot. It created and sent a meta-scene object that contained all features of human presented in the scene to the attention layer.

The attention layer was deputed to select the most important target point of environment, through an attention mechanism. It sent the pixel address of target to the gaze control layer.

The gaze control layer, which was connected directly to the robot control, moved the robot actuators based on the humanlike gaze control model to generate a humanlike gaze movement for the robot.

The YARP as a communication channel was established, in order to make an interconnection among layers and modules. Overall integration of the system with the robot showed that the system was able to select properly the target point of environment and generated a humanlike gaze motion for the robot.

Chapter 5

Proof of Concept Evaluation

Contents

5.1 Overview

Chapter 4 described a context-aware gaze control system (GCS) that enabled a humanlike social robot to select autonomously its target point exploring low-level and high-level features of the environment. As discussed, the GCS actively identifies the most important target point based on a scoring strategy. The system calculates a score for each human in the FOV and selects a human with highest score as winner. The overall score of each human in the scene is calculated based on four main components: social features (F), proxemics (P), orientation (O), and a memory component (EAM)

$$EA_{S_j}(t) = F_{S_j} + P(r) + O(\theta) + EAM_{S_j} \tag{5.1}$$

where EAM_{S_j} is a parameter that refers to the memory of the robot not yet included in the database and consequently set to zero.

The social feature elicitation contribution F_{S_j} is calculated as a weighted summation of social normalized features X_n, which can be written as

$$F_{S_j} = (\sum_{i=1}^{n} W_i.X_n) \tag{5.2}$$

where *weight* W_i is set based on the feature's importance. The weight parameter adjustment is the most important issue that results a humanlike gaze behavior for the robot and it must be determined based on the priority that humans have in selection features in a social interaction.

To identify the importance of features in attracting human attention, a gaze tracking study was performed with a professional eye tracker device. In this study, the gaze behaviors of 11 participants were collected in a social interaction, similar to our scenario. The purpose of the study was to tune the parameters of the GCS such that the system displays a humanlike gaze behavior in a social interaction. Thus, one aim was to determine which social cues have the most prominent effect on the attention of the study participants. The second aim was to compare how well the GCS was able to replicate human gaze behavior on the same context (input videos).

This chapter describes in detail, the humans gaze study, the evaluation process of the GCS and discusses the obtained results.

5.2 Participants

A total of 11 participants (9 males and 2 females), from the Department of Mechanical Engineering at the Technical University of Munich took part in this experiment. The mean age of the participants was 27.3 (range 22–35). Eight of the participants were native German speakers, the three other participants spoke English, but were not native English speakers. The participants received a chocolate for taking part in the experiment.

5.3 Experiment Procedure

The participants were asked to watch a video showing two humans discussing different research topics. In this video, two humans enter the room separately, sit down on two chairs with the same distance from the camera, and then leave the room separately. During the discussion, both humans talk to the video camera from time to time, as if they were interacting with a third person (the robot/the experiment participants), in

Table 5.1: Attention of participants towards person A and B while speaking, while performing non-verbal cues, and the average from the entire video (Avg. att.)

Person	**Speaking (%)**	**Non-verbal (%)**	**Avg. att. (%)**
Person A	41.8	20.3	54.4
Person B	32.5	23.1	43.6

order to help with experiment participant engagement.

The video was taken in parallel with an HD video camera and a Kinect RGB-D camera placed side by side. The scene captured by HD video camera was shown to the participants for human gaze analysis, while the Kinect-acquired RGB-D data was used as input for the GCS for the system behavior performance analysis.

The video lasted 7:20 minutes and consisted of three sub-scenes. In the first and third sub-scene, the people in the videos talked in English, in the second sub-scene, they spoke in German. In each sub-scene, only one person spoke at a time, while the non-speaking participant executed diverse gestural and postural acts, in order to attract the attention of the viewer. Gestures and movements included: stretching while being seated, raising an arm, getting up from the chair to get a drink, and retrieving a smart phone from their pocket. Table 5.1 shows the average features that person A and person B have shown during the video. Table 5.4 and Table 5.5 detail the behavior of person A and person B in the first recorded scene while Figure 5.11 and Figure 5.12 illustrate the average gaze behavior of participants during watching the scene.

While the participants watched the video, they wore a DIKABLIS eye tracking system to record gaze behavior (Figure 5.1). The eye tracker included a field camera in order to capture the scene and an infrared camera to capture a video of the left eye.

The participants sat roughly 75 cm away from a 23 inch display 5.2.

Figure 5.1: The DIKABLIS eye tracking system has two separate cameras: the field camera looks to the front in order to capture the scene the participants are looking at, and an infrared camera captures a video of their left eyes.

Before starting the experiment the DIKABLIS eye tracker was calibrated, both for improving its pupil detection and gaze estimation capabilities. Experiments were carried out in a room with controlled lighting to prevent any external light sources interfering with the eye tracking system.

5.3.1 Eye-tracker Calibration

The DIKABLIS eye-tracker device should be calibrated for each participant, in order to estimate correctly the user gaze on the environment in real-time. The device calibration can be done in two steps: user's pupil detection and gaze estimation.

The user's pupil detection have to be done in order to ensure the full detection of participant's pupil, when it moves in its orbit, in order to

Figure 5.2: The participants sat roughly 75 cm away from a 23-inch display while they wore the eye-tracker device.

gaze at a target. For that, the infrared camera of the eye tracker device has to be fixed at the right point. Besides, the lighting condition must be adjusted such that reflects the minimum light to the participant's pupil. Figure 5.3 shows a sample of well-calibrated device where the user's pupil is detected in its whole its movement range.

The next step of calibration is due to estimating the user's gaze point correctly on the environment. For that, we asked user to look at a few fixed points of the environment, then we calibrated the device by adding offset values to the vertical, and horizontal axises of the gaze point. The test was done, when the user gazed at the distant as well as close target such as screen. At the final step, we asked user to gaze at the corners of

Figure 5.3: Pupil detection step: The user's pupil detection have to be done in order to ensure the detection of pupil in its whole movement orbit when it shifts to gaze at the target.

the display to ensure the device precision. Figure 5.4 shows a sample of a calibrated system.

After the calibration steps, the gaze tracker is ready to collect, the participants gaze behavior.

5.4 Data Collection and Analysis

5.4.1 Data Collection

Using the eye tracking device together with the DIKABLIS recorder software, two separated video files are generated from pupil (through infrared camera) and from field (through front camera). The DIKABLIS analysis software produces a single video file of the field camera with an overlaid cross-hair showing where the participants look.

We showed to the participants the three sub-scene of a social interaction between two people while they assumed to be a member of the social interaction. Then, we recorded their gaze behaviors during watching the scenes. At the end of the experiments, we collected 11 video of all participants where their gaze point were indicated on the screen by a red-cross (see Figure 5.5).

Figure 5.4: Gaze detection step: Users gazed at the specific points in order to ensure the precision of the eye-tracker.

5.4.2 Data Analysis

ELAN [105] is used, as a powerful annotation software, to annotate the participants' recorded videos on a frame-by-frame basis: timing looking at either *person A* (the person on the left in the scene), *person B* (person on the right), or at the *environment* (other regions). We also annotated when and how often the person not speaking provided a non-verbal social cue. Table 5.1 summarizes how often the experiment participants looked at each person and how often the person was either speaking or providing a non-verbal cue.

After annotation, log files containing time duration (in ms) and position (i.e., person A, person B, environment) of the participants' gaze fixations were exported. The average attention of the participants was calculated using Matlab.

Analyzing the average attention/gaze behavior of the participants, we identified two types of gaze shift: saccadic and non-saccadic. Saccadic gaze was the eye movement where it quickly gazes at the target and then returns to the previous target. While non-saccadic gaze shift was the eye movement that takes a longer time.

In order to identify the verbal/non-verbal cues that cause non-saccadic and saccadic gaze shift to person A, person B and the environment, we an-

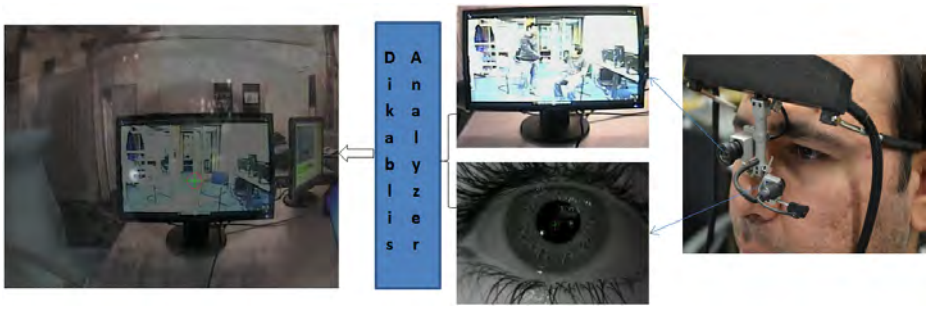


Figure 5.5: Dikublis Analyzer generate a single video in which the participants gaze point are shown as a cross.

alyzed the average participants' gaze behavior obtained by using Matlab. The average gaze pattern of participants was divided into 15 segments (A–N) identifying regions where the observers' attention were on an individual person (A or B). The various peak points of the average gaze pattern were also selected by identifying verbal and non-verbal cues that attracted participants' attention thus triggering the gaze shift.

The GCS parameters were extracted according to the target selection priorities of participants on the basis of the method described in Section 5.6.

After the GCS parameters had been extracted through human gaze analysis and interpretation, the GCS generated gazes were compared to the average gaze pattern of participants. The Kinect-acquired RGB-D data was used as input to the GCS module, which generated a new video similar to the one, obtained through the DIKABLIS eye tracker analysis software. A red circle identifying the FACE robot gaze point was streamed through YARP to the robot control library. The GCS generated video was annotated using ELAN with the same modalities used for the participant's video annotations. The error between the two gaze paths was calculated as an average of the absolute difference between the human gaze and GCS

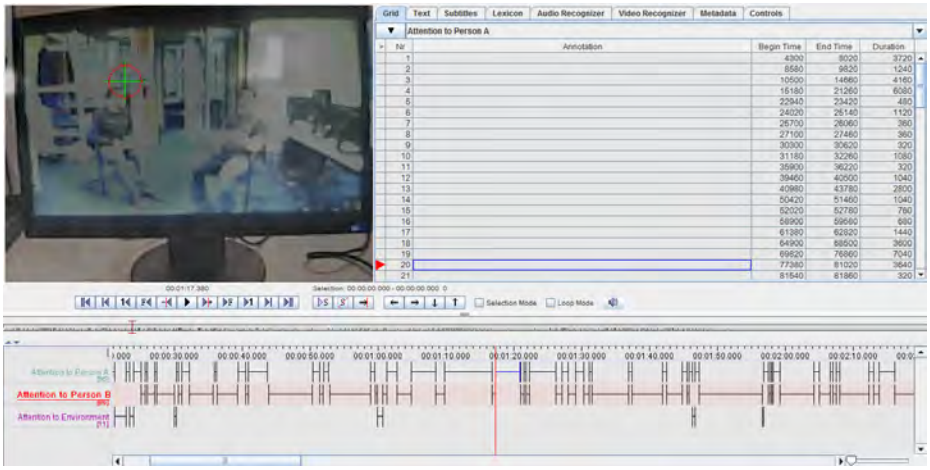


Figure 5.6: ELAN Environment.

pattern functions (error function).

5.5 Gaze Behavior Results

5.5.1 Participants Gaze Behavior

We collected the participants gaze behavior using the eye-tracking device, and then generated a single video that the participants' gaze were indicated on the environment as a red-cross. Since the movement of the red-cross on the generated video shows the participants' gaze behavior, we annotated the video using ELAN to figure out the correlation between the gaze behaviors and the actors' features of the scene at the corresponding time. Annotating the participants gaze behavior, 11 log files were generated in which the time and the positions that participant gazed were obtained. Figure 5.6 shows the ELAN software environment.

Analyzing the generated log files using Matlab, we obtained the average gaze behavior graphs of all participants in which the average gaze

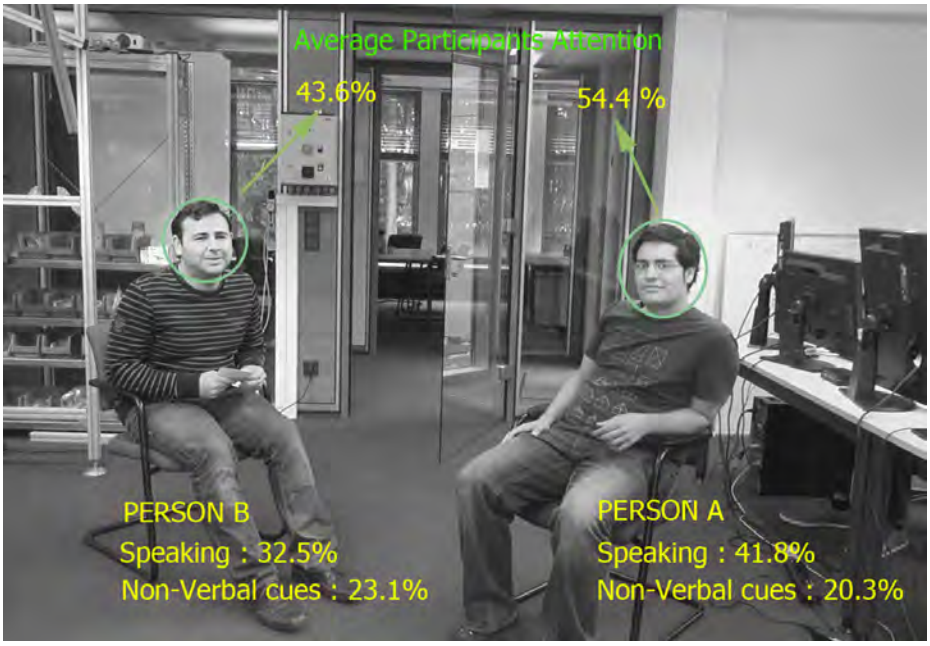
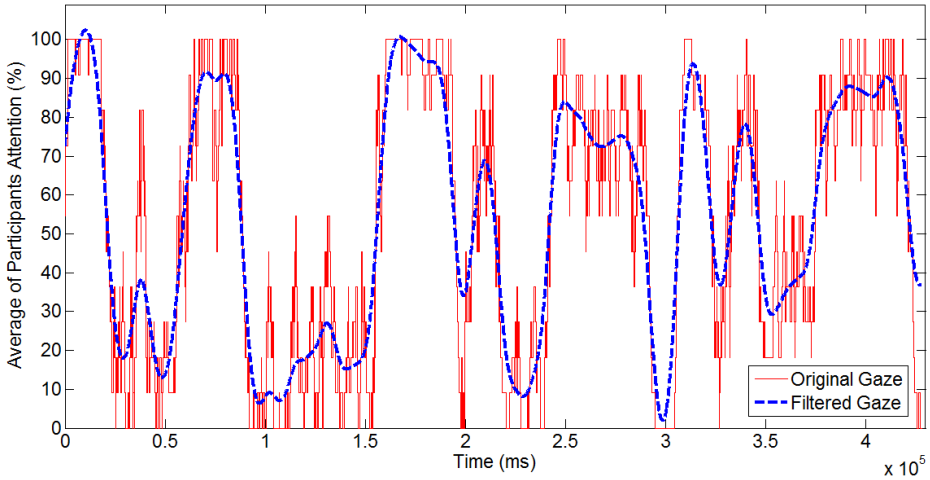


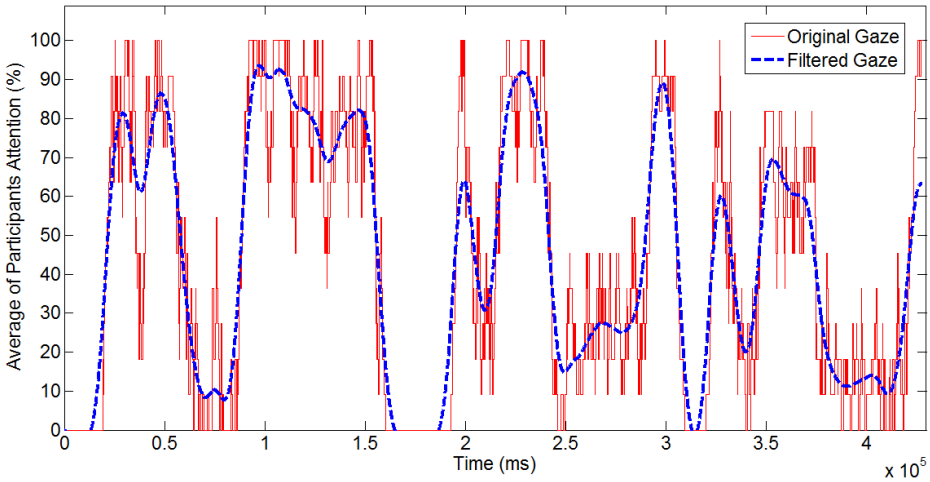
Figure 5.7: At the top of the figure, the Average participant attention on person A and person B are shown. At the bottom of the image, the average verbal/non verbal behavior of person A and person B are shown.

behavior to the person A and B and environment were illustrated. Figure 5.8(a) shows the average participants attention/gaze on the person A, and Figure 5.8(b) the average participants attention/gaze on the person B, respectively. Table 5.1 reports average gaze behavior of participant to the person A and person B, together with the high-level that two actors showed during the three sub-scenes (see Figure 5.7).

In order to obtain the overall gaze behavior, high-frequency gaze shifts were filtered through a second order low-pass filter, that resulted the graph with dotted line in Figure 5.8.



(a) Average attention of 11 participants on person A.



(b) Average attention of 11 participants on person B.

Figure 5.8: Average attention on person A and person B in the recorded video.

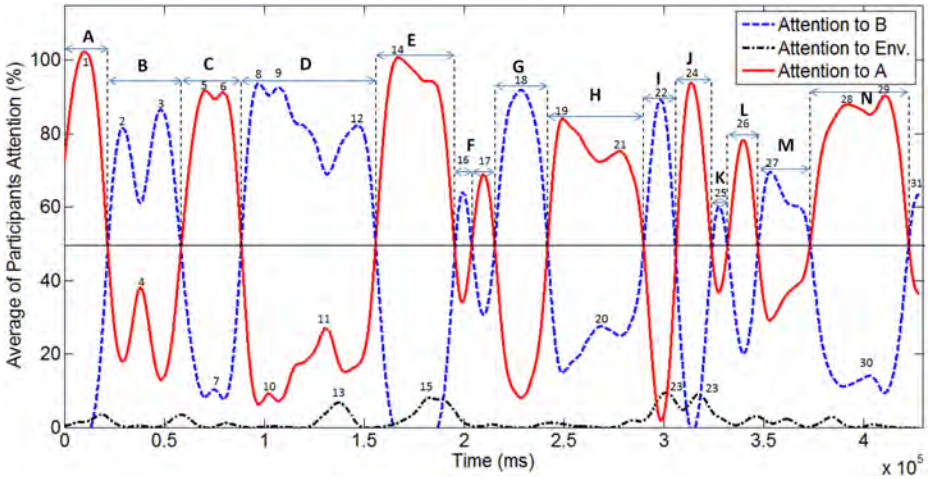


Figure 5.9: Average participant attention on person A, person B, and the environment. The segments identify regions when the gaze is kept on a person (A or B). The peaks identify specific events that triggered the participant's attention.

5.5.2 Non-saccadic Gaze Shift

As shown in Figure 5.9 and Figure 5.10 the average human gaze behavior can be divided into saccadic (high frequency) and non-saccadic (low frequency) movements. For example, peak 1 of segment A shows that 100% of the participants looked at person A. Analyzing the video at corresponding peak time shows that peak 1 corresponds to the instant when person A entered the room and initiated the conversation with the observer. Similarly, at Peak 2 of segment B, person B entered the room while person A was still there. At this point 82% of participants looked at person B and the rest of the participants kept their focus on person A. With the same methodology, we analyzed the entire average gaze pattern (Figure 5.9) by identifying various social verbal and non-verbal cues that attracted participants' attention. Social cues identified in the videos and associated peak numbers are reported in Table 5.2.

Table 5.2: Social cues identified in the average gaze pattern and their associated peak numbers

Social Cue	Peak Number
Entering	1,2,14,16,17,25
Speaking	3,8,9,14,16,18,19,22,24,26,27,28,29
Leaving	12,15,21,22,29,31

Table 5.3: Analysis of saccadic gaze behavior

Social Cue	Section
Facial Expression	A,B,E,F,K
Body Gesture	C,D,G,H,M,N

5.5.3 Saccadic Gaze Shift

To identify social cues that cause a saccadic gaze shift, the non-filtered data were analyzed. Figure 5.10 details the saccadic gaze shift during the experiment. As illustrated, while almost all participants were attracted by person B (sections B1-B6), some of them had several very quick and short gaze shifts to person A (A1-A4). With the same methodology, beside the cues that caused non-saccadic gaze shifts (i.e., entering, speaking, and leaving), analyzing the non-filtered average gaze behavior, the cues that caused saccadic gaze shift were identified: Facial expressions, body gestures and hand motions. Saccadic triggering cues are reported together with the associated log segment in Table 5.3.

5.5.4 Detail Analysis of the Participants Gaze Behavior

The following section describes in detail, the first scene (2:40 min) of the eye-tracking experiment that watched by participants and lists the cues

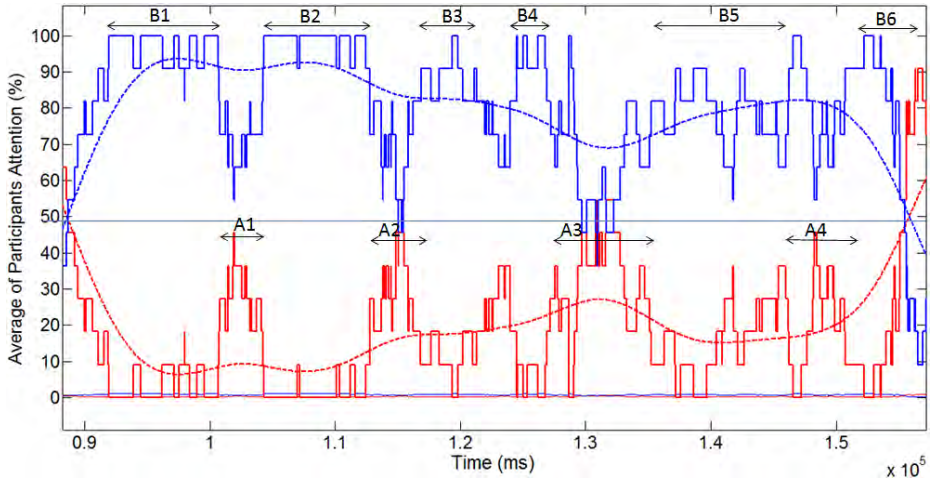


Figure 5.10: Participants gaze shift between person A (in section A1-A4) and person B (in section B1-B6) in the video. Analyzing the corresponding videos demonstrates that peak points are associated with the verbal/non-verbal cues that person A and person B performed.

that caused saccadic/non-saccadic gaze shift considering time that these cues have been shown on the video. In addition, it shows the corresponding participants gaze behavior during watching the video. The above information helps us to get more details about the effect of social cues (signals) on the human attention and shows how some cues in a social interaction cause non-saccadic gaze shift while some other cause saccadic gaze shift.

To see a more clear image of the gaze movements, the average participants gaze behavior was illustrated in two figures where Figure 5.11 and 5.12 show the first and second part of the scene respectively. Moreover, Table 5.4 and Table 5.5 describe the activities and cues that actors of the scene (person A and B) showed during the scene. Thus, analyzing the participants gaze behavior graphs (Figure 5.11 and 5.12) and considering

Table 5.4: Detail of the activities and cues that person A and person B showed in the first recorded scene - part 1

From (sec)	**To** (sec)	**Person A**	**Person B**
0	10	enter	–
10	20	sit, speak	–
20	30	sit	enter
30	40	speak,mov.	mov.,sit,facialexp.
40	50	speak	speak
50	60	speak	speak
60	70	hand motion	speak
70	80	body mov., facial exp.	speak

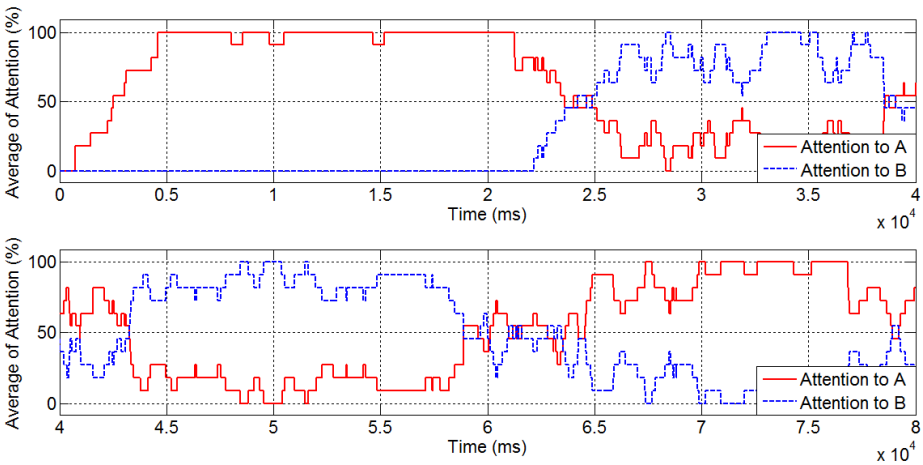


Figure 5.11: Detail of the average attention of 11 participants during watching the first scene - part 1.

the content of the scene at the corresponding time period (Table 5.4 and Table 5.5) allow us to track the effect of the different social cues on the human attention.

Table 5.5: Detail of the activities and cues that person A and person B showed in the first recorded scene - part 2

From (sec)	**To** (sec)	**Person A**	**Person B**
80	90	speak	sit
90	100	sit	speak
100	110	hand/body mov.	speak
110	120	hand/body mov.	speak
120	130	hand/body mov.	speak
130	140	speak, hand/body mov.	sit, speak
140	150	speak,sit	speak
150	160	sit	leave

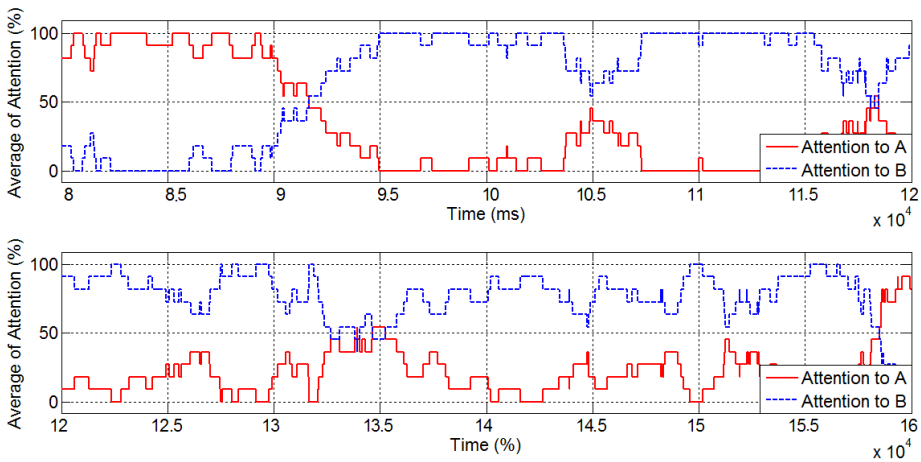


Figure 5.12: Detail of the average attention of 11 participants during watching the first scene - part 2.

As illustrated in Figures 5.11 and 5.12, the vertical axis shows the percentage of participants that gaze at person A and person B. As expected, at some moments, almost all participants gazed at person A or B

in the video however, in some other moments, only half or less than half of them gazed at the targets. Besides, the participants gazed at the same target with different durations in different times. This fact shows that the cues of environment have not had the same effect on the participants' attention.

As some examples of data analysis, the gaze behavior and associated cues in the video are detailed as follows. For example, within the period (0-22 sec) person A entered into the room while person B was outside the room. As illustrated, person A was attracted the attention of all participants. However, within the period (25-30 sec) that person B entered into the room, most of the participants shifted gaze to person B. Or for example, at the time like (t = 60) in which both person A and B showed cues, participants showed different gaze behaviors.

Analyzing the recorded scene and the average participants gaze behavior in the frame-by-frame basis, we identified the features and their importance in attraction human attention. With the same methodology, we identified which features make saccadic and non-saccadic gaze shifts.

5.6 GCS Parameter Estimation and Priorities Features

As discussed, the most important factor that enables the GCS to generate a humanlike behavior is weight parameter of the attention model however, gaze data analysis results showed that there is no generic gaze behavior that can be used to implement a unique model as a standard for *humanlike* gaze patterns. Especially in cases where speaking and hand or body motions occurred at the same time, participants demonstrated different gaze behaviors. However, through the analysis the maximum peaks in saccadic and non-saccadic gaze shifts of participants, the priorities of

verbal and non-verbal cues in attracting human attention were estimated.

The strongest cue attracting the attention of all participants was the new-entry (person A/person B) who joined the interaction. Even when person A was speaking with a participant and at the same time person B arrived, all participants were attracted by the new-entry (person B). Thus, the highest priority should be given to the new-entry.

The second priority is given to the auditory signals (speaker). Once person A/B started speaking, all participants were attracted by the speaker. It should be noted that, if one person showed body gesture while another one was speaking, most participants were distracted by the body gesture for a very short time and gazed back to the speaker quickly, which shows the higher importance of auditory signals.

In addition, a few of participants were attracted by the speaker all the time and ignored the body gesture/hand motion of the other person. Therefore, the third priority goes to body gesture/hand motion.

The last two priorities are given to the person leaving and facial expressions, respectively, which attracted less attention compare to other cues.

The identification of a set of parameters that enables the GCS to generate in the FACE robot, a similar gaze to that observed in humans, both saccadic and non-saccadic movements triggering cues, were considered. The weight parameter W_i of GCS introduced in Equation (5.2) was calculated considering a maximum value of 100, based on the identified empirical priority order extracted, by analyzing the maximum peaks for each cue (Figure 5.10) during the video. The priority order and the assigned GCS weight are reported in Table 5.6. In addition to Table 5.6, we set a distinction factor for those features that cause a saccadic gaze shift. This enables GCS to have both saccadic and non-saccadic gaze behavior.

Table 5.6: Verbal and non verbal cues identified as attention triggers and their associated GCS weight calculated on the basis of human observed priorities

Priority	**Social Cue**	**GCS Weight**
1 (highest priority)	Entering	100
2	Speaking	100
3	Hand motion/body gesture	65
4	Leaving	55
5 (lowest priority)	Facial expressions	45

5.7 Gaze Control System Behavior

After the weight parameter of the attention model had been tuned based on the empirical data of humans, it is expected that the GCS shows the humanlike behavior in the same social interaction.

To test the performance of the GCS, the Kinect-acquired RGB-D streams were used as input to the GCS. With the same method, a new video was captured from the screen, similar to the one obtained through the DIKABLIS eye tracker analysis software. A red circle identifying the FACE robot gaze point was streamed through YARP to the robot control library (Figure 5.13).

The GCS generated video was annotated using ELAN with the same modalities used for the participant's video annotations and the system behavior plotted using Matlab (Figure 5.14).

The error between the two gaze behaviors was calculated as a mean of the absolute difference between the human gaze and GCS behaviors (error function).

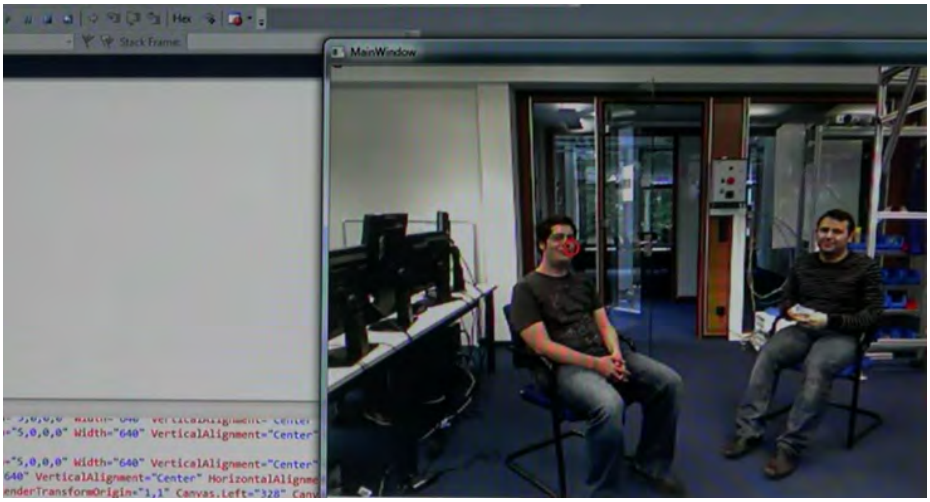
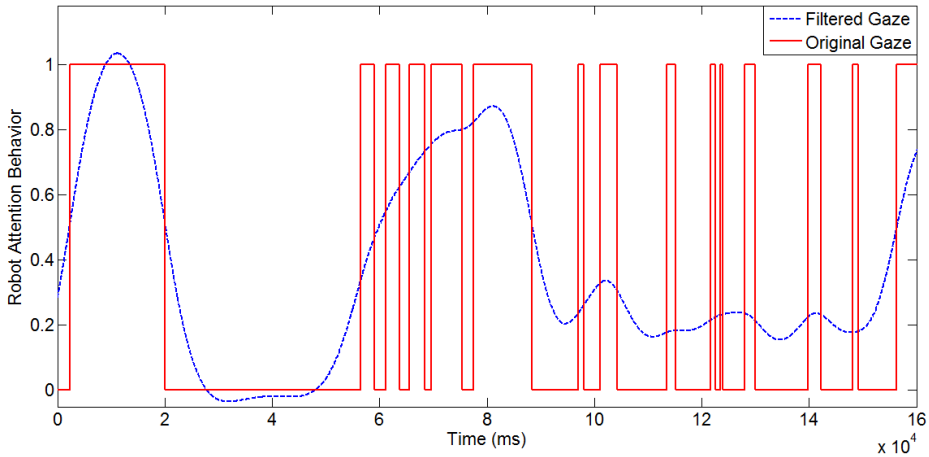


Figure 5.13: The gaze control system's behavior was recorded in which the identified targets by the system was indicated with a red-cross.

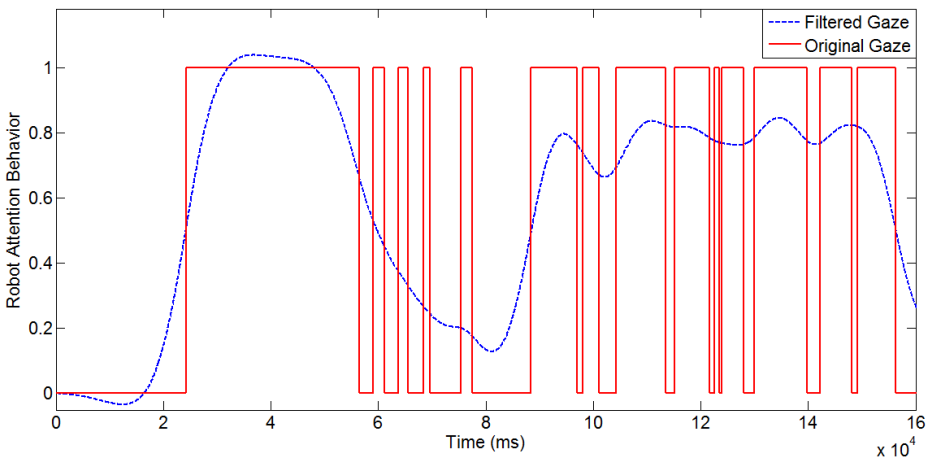
5.8 Human and GCS Generated Gaze Behavior Comparison

As discussed, through an in-depth gaze study, we obtained the average gaze behavior of humans in the social interaction and the GCS behavior in the same scene of social interaction.

In order to obtain the performance of the GCS, its results should be compared to the human data. Since we identified two gaze movements (i.e., saccadic and non-saccadic), the performance of the system in replicating these behaviors should be evaluated.



(a) Gaze control system attention on person A.



(b) Gaze control system attention on person B.

Figure 5.14: Gaze control system attention on person A and person B in the same recorded video.

5.8.1 GCS Performance in Replicating Non - saccadic Gaze Behavior

To evaluate the ability of the system in replicating non-saccadic gaze shifts, the filtered gaze data of humans and the system are compared. Figure 5.15 shows the comparison of the average gaze of participants to the GCS generated pattern for the same scenes. The upper image shows the attention on person A; the lower image shows the attention on person B. The graph shows that the system follows the human gaze behavior for the entire duration of the video.

We derived the error function as difference between human average gaze behavior and the system behavior at person A and person B as follows

$$error(t) = F_{human}(t) - F_{gcs}(t) \tag{5.3}$$

where $F_{human}(t)$ shows the average gaze behavior of humans and $F_{gcs}(t)$ shows the system gaze behavior for the same scene. Figure 5.16 illustrates the error function between humans behavior and the system behavior during the same scene. The mean of the absolute error function shows that the system is able to replicate the average human gaze behavior with a replication factor of 89.4% throughout the video.

5.8.2 GCS Performance in Replicating Saccadic Gaze Behavior

To evaluate the capability of the system in replicating saccadic gaze shifts, the non-filtered gaze data of humans and the system are compared. Figure 5.17 compares the average gaze of participants to the GCS generated pattern for the same scenes. The upper image shows the attention on person A; the lower image shows the attention on person B. The graph

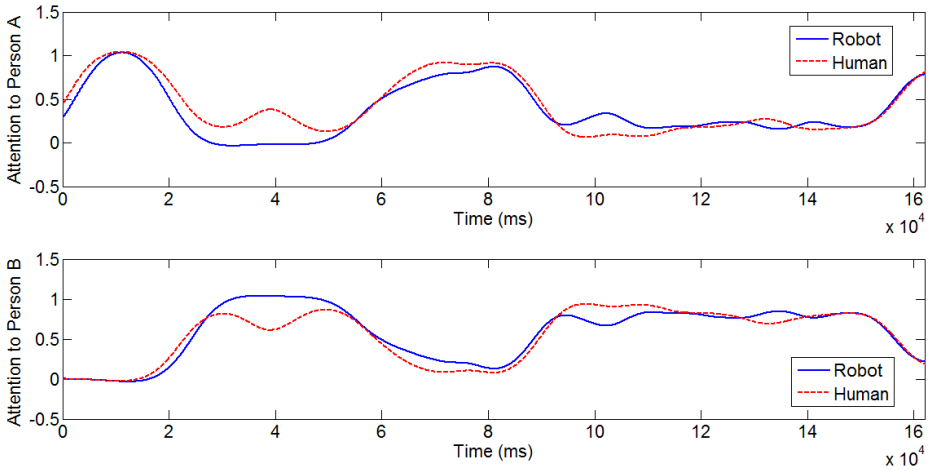


Figure 5.15: Comparison of human and robot gaze behavior.

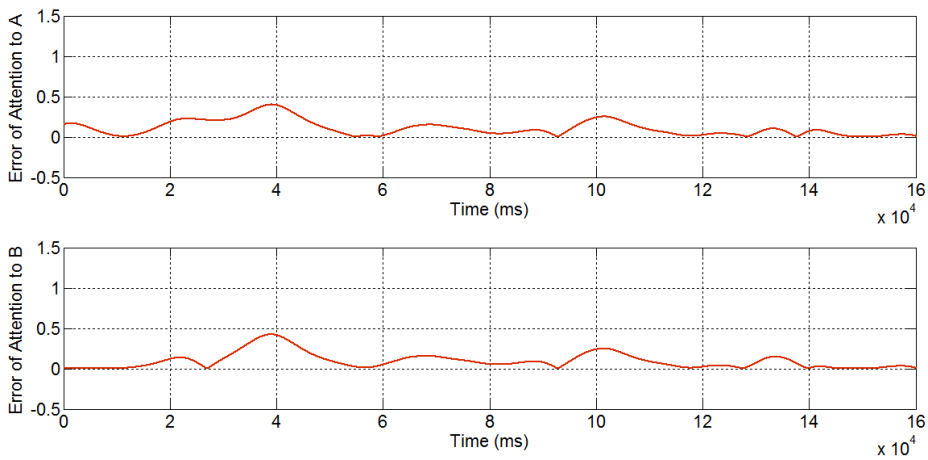


Figure 5.16: error between human and robot gaze behavior.

shows that the system follows the human gaze behavior for the entire duration of the video.

Considering saccadic eyes movements (red continuous line in Figures

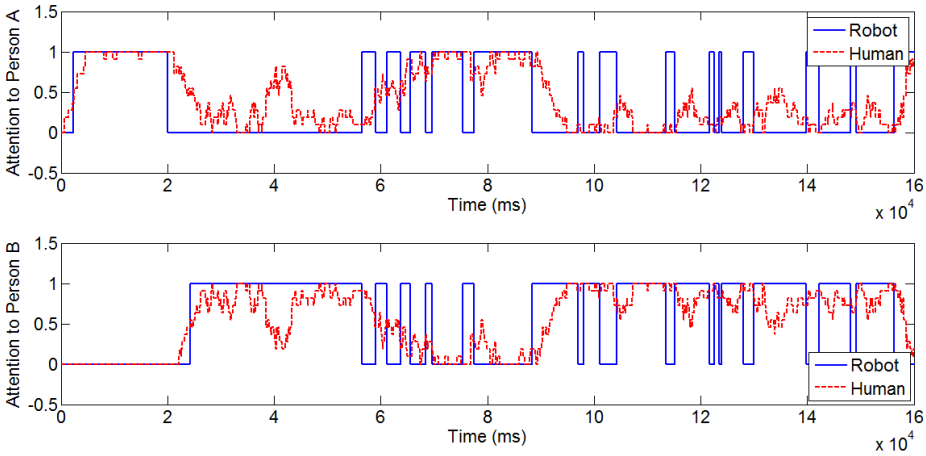


Figure 5.17: Non-filtered data comparison of human and robot gaze behavior.

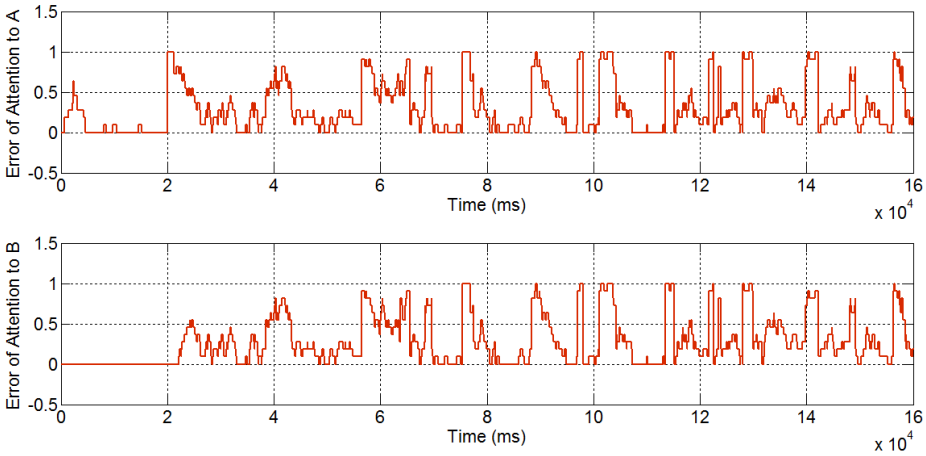


Figure 5.18: Non-filtered error between human and robot gaze behavior.

5.8(a) and 5.8(a)), the accuracy rate of the GCS decreases to 75.2%, which is likely due to limitations in sensor detection range and speed in comparison to the human eye.

5.9 Discussion

A context-aware social gaze control system, which enables the social humanlike robot FACE to display humanlike gaze behavior has been presented chapter 4. The underlying attention mechanism of the implemented gaze control system used high-level social features, such as nonverbal and verbal cues, proxemics, an effective visual field of view, and the habituation effect, to determine where the robot should direct its attention. In order to enable the proposed system in generating a humanlike behavior, we tuned the GCS's parameters based on the human attention priority in selecting features in a dynamic human-human interaction. To identify the feature priorities, we performed a gaze tracking study with an eye tracker.

Although the GCS with the tuned parameters should perfectly display gaze behavior similar to human, experimental results showed that the GCS is able to replicate average human behavior for both non-saccadic (89.4% accuracy) and saccadic (75.2% accuracy) movements with some errors. The lower accuracy especially in the case of saccadic movement replication may be due to several points.

Diversity in human gaze behavior: individual human gaze behavior is correlated especially in saccadic movements to the factors such as personality, age, and gender [29]. Thus, gaze behavior is different from person to person. Our model replicates the average gaze behavior of the participants of our experiment. These personal differences are not replicated in the GCS due to the extraction of average based parameters. However, these differences are common in humans and consequently not perceived as being strange but more as personal and unique peculiarities.

Limitations of the input sensor used: compared to the human eye, the Kinect sensor has a narrower field of view and a much lower resolution, which affects the attention computation of the GCS. The most influencing

system limitation was probably the sensor range of the Kinect, which is between 800 mm and 4000 mm while humans are able to see much further. Thus, the experiment participants were able to detect people who entered the room shown in the video, when they were for example in their public space (see Figure 3.2). The maximum sensor range of Kinect is similar to that of a human's social space.

Unmodeled human attention features: although our human attention model already considers many features that guide human attention, there are still other unknown factors that we did not use in our model. For example, Figure 5.9 shows the participants looking at the environment over time.

In addition, there are further external features that guide the attention selection mechanism, which we did not include in the current implementation. For example, taking into account the auditory information that comes from outside visual FOV and considering the intentions of people during a social interaction could help the attention mechanism to generate a more natural humanlike social gaze behavior.

However, as shown in Figure 5.15 and Figure 5.17, the proposed implementation of the GCS is able to select the appropriate gaze target points at the right time, which is essential for the development of believable social robots.

5.10 Summary

This chapter described a gaze tracking study using a professional eye-tracker device that is carried due to adjusting the weight parameter of the attention model of the GCS. Analyzing the average gaze behavior of 11 participants, the priorities that humans have in selecting features in a social interaction were identified. According to the obtained gaze analysis results, the system attention parameters were tuned.

To test the performance of the system, the same scene were used that we had showed to the participants as input for the GCS, and analyzed the behavior of the system. The comparison between average human gaze behavior and the system behavior showed that the GCS is able to replicate human gaze behavior with high degree of precision, in a social interaction. The result showed that the GCS, as an essential component of the humanlike robot was able to select right target point and at the right time which is important for an effective HRI.

Chapter 6

Conclusion and Future Work

Contents

An innovative context-aware social gaze control system (GCS) has been implemented as a part of a humanlike robot called FACE. Employing GCS, the robot, possesses the perceptual capabilities similar to humans, that is fundamental in regulating the robot's behavior to perform a meaningful human-robot interaction. The robot, is able to perceive and interpret its surrounding environment in the same way that human being does. Using GCS, the robot and humans have the same attention mechanism to identify the same types of stimuli salient at the same time. Integrating the system, the humanlike robot is able to autonomously analyze the environment, identify its target regarding the current social scene

and generate a context-aware gaze shift to the selected target such that the dynamic of the head and eyes movements are similar to humans.

In short, robot is now able to autonomously spend an hour with multiple people in a dynamic environment, while during the interaction exhibits a context-aware social humanlike gaze behavior in a such a way that it is being perceived by human as natural, that is a fundamental step toward developing a believable creature which yields several positive outcome for human society.

This thesis discussed in detail the human-robot interaction problem from two perspectives: theoretical and technical. The first part of the thesis discussed the psychological parameters and phenomena that affect human attention and gaze behavior in a social interaction, and should be considered when developing humanlike model for social robotic application. The second part presented the proposed system that is developed to fill the existing gap in the state-of-the-art. Finally, the last part of this thesis discussed the results and efficiency of the system as well as the evaluation process that was done with real data of humans.

The following sections summarizes the proposed system, contributions of the work to the state-of-the-art, and the future work.

6.1 Conclusion

In this thesis different methods and algorithms were proposed, to take a step forward in the field of human attention and gaze behavior modeling for social humanlike applications. The proposed systems are concluded as follows.

6.1.1 Human-level Perceptual System

As a fundamental component of social humanlike robots, a human-level perceptual system that simulates human perception was developed. The system receives as the input RGB-D images and sound signals constructed by Kinect and creates as the output a human-level interpretation of environment for the robot. The system consists of several well-integrated real time perceptual components that work together and in parallel, to actively recognize several high-level human-relevant features and to identify non-human environmental salient points. The main aims of designing a human-level perceptual system is essentially twofold: on one hand, to recognize and estimate the high-level and low-level features that are required by the attention in order to identify its target point, and on the other hand, to give the robot the ability to make a social exchange with the selected target according to the detected perceptual information of that target. The perceptual system consists of the following components.

- Using a high speed object recognition engine called SHORE, the perceptual system detects multiple faces (frontal/rotated) in a single frame and tracks them in a real-time video while it assigns a consistent identification number to each of detected faces. In addition, SHORE recognizes wide variety of facial features including four universally agreed facial expressions (i.e., happiness, sadness, surprise, and anger) in percentage, gender (male/female), estimates age (years) the eyes and mouth state (open/close), and locates face, eyes, nose, and mouth in pixel (X,Y). SHORE has an internal timer that stores the entry time in FOV of humans.

- Using Microsoft Software Development Kit (SDK), the perceptual system locates the 3D positions (X,Y,Z) of 6 people while it extracts full skeleton information (the 3D positions of 20 body joints) for 2

closer human. A dynamic gesture recognition system monitors the extracted body joints position in order to detect specifics motions and meaningful gestures.

- Using PCA-based face recognition engine, the perception layer compares the detected faces with the subject data-based in order to identify the subjects name. It enables the robot to adjust its long-term and short term interaction with people in specific social ways.

- Using a salience-based engine called fastSUN, the perceptual system analyzes low-level features of an image and identifies the most important environmental point as salient point in pixel (X,Y).

The perceptual system was developed to create a human-level interpretation that simulates human attention for the robot. As discussed, the system collects several high-level communicative features of human and environment and creates a meta-scene object that is a human-level interpretation of environment.

6.1.2 Human-level Attention System

As a fundamental component of social humanlike robots, a human-level attention system that simulates human attention was developed. It receives in real-time as input the meta-scene object constructed by the perceptual system and actively identifies the most important human/non-human point that the robot should gaze at, based on a feature-based human-level attention model. The model evaluates all features of the humans presented in the scene and computes a score for each of people by summation of weighted features. Clearly, it selects one with highest score as winner, among people and environmental point identified by the perceptual system. Thus, if the people are not interesting enough for the

robot, the attention model shifts to the environment otherwise it selects a human for interaction.

The attention, system computes scores based on the evaluation of those features that have been proven to guide human attention in a social interaction. In addition it weights the features according to the priorities that human attention has in selecting features. This priority is identified, through an in-depth gaze study, using a professional eye-tracking device.

6.1.3 Gaze Control System

As a fundamental component of social humanlike robots, a gaze control system that controls the robot's gaze movement, was developed. It receives in real-time as input, the identified target point's address in pixel from the attention system and controls the eyes and head of the robot in such a way that robot displays a humanlike, meaningful and believable gaze movements. System moves the robot's head and eyes actuators, adjusting their amplitudes, speed, latencies, and priorities. As result, the robot gaze moves from one point to another in a humanlike way.

6.1.4 Data Communication Unit

Due to required computational processing cycle, the perceptual and attention systems were implemented in two different machines. YARP as a reliable wireless communication channel was developed for data communication between different systems and components. For example, it transfers the created meta-scene object by perceptual system to the attention system, which is implemented in different machine with different IP.

6.1.5 System Evaluation

As part of a pilot evaluation, the gaze behavior of 11 participants were collected using a professional eye-tracking device. Participants were shown videos of two-person interactions and tracked their gaze behavior. Analyzing human gaze data, we identified two types of gaze behavior as saccadic and non-saccadic. A comparison of the human gaze behavior to the behavior of the gaze control system running on the same videos showed that the system replicated human gaze behavior with an accuracy of 89.42% for non-saccadic movements and 75.23% for high-speed saccadic movements. The system allows the control of the robot in performing the social attentive tasks in which believable behaviors are mandatory.

6.2 Main Contributions to the State-of-the-art

Many efforts have been conducted to design attention systems which guide robot gaze fixation based on the salience of low-level features presented in the visual scene (colors, intensity, orientation, and etc.). However, due to several shortcoming, the salience-based attention model dramatically failed in replicating human attention and gaze behavior in a human-centered scenario. For that this thesis presented a social context-aware gaze control system for humanlike robot applications that considers both low-level visual features and high-level human-relevant social features in the robots attention. The attention mechanism of the system identifies targets based on the low-level visual feature analysis and high-level human-relevant feature analysis. Parameters of the system were tuned based on the gaze study according to the human attention/gaze behavior in selecting features of environment, in a social interaction.

An innovative gaze model was developed based on the previous studies. It controls not only the amplitudes of head and eye movements

but also their velocities, latencies and movements' priorities. Using this model, robots shows the humanlike motion when it shifts its gaze in a social interaction.

The overall integrated GCS system has been implemented as a part of a humanlike social robot called FACE, and drive its dynamic attention and gaze in real-time in a multiparty social interaction.

6.3 Future Work

System evaluation showed that the major limitation of the GCS is due to the narrow field of view of the used Kinect camera. As a future plan to circumvent this problem, by either using more than one Kinect or by replacing the Kinect camera with other sensors.

Moreover, the possibility to continuously adapt the GCS weight and parameters according to the robot emotional mood will be investigated. This mood based GCS tune will allow the FACE android to adapt its behavior not only based on the social scenario on which it is trying to be involved, but also in accord to its internal emotional state.

Bibliography

[1] Kazuyoshi Wada, Takanori Shibata, Tomoko Saito, and Kazuo Tanie. Effects of robot-assisted activity for elderly people and nurses at a day service center. *Proceedings of the IEEE*, 92(11):1780–1788, 2004.

[2] Cynthia Breazeal, Andrew Brooks, Jesse Gray, Guy Hoffman, Cory Kidd, Hans Lee, Jeff Lieberman, Andrea Lockerd, and David Chilongo. Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robotics*, 1(02):315–348, 2004.

[3] Masahiro Fujita. On activating human communications with pet-type robot aibo. *Proceedings of the IEEE*, 92(11):1804–1813, 2004.

[4] Kenji Kaneko, Kensuke Harada, Fumio Kanehiro, Gou Miyamori, and Kazuhiko Akachi. Humanoid robot hrp-3. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 2471–2478. IEEE, 2008.

[5] ROBOTS IN COMBAT. Ethical issues relating to engineering and humanoid robots.

[6] David Gouaillier, Vincent Hugel, Pierre Blazevic, Chris Kilner, Jérôme Monceaux, Pascal Lafourcade, Brice Marnier, Julien Serre,

and Bruno Maisonnier. Mechatronic design of nao humanoid. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 769–774. IEEE, 2009.

[7] Christoph Bartneck, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. My robotic doppelgänger-a critical look at the uncanny valley. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, pages 269–276. IEEE, 2009.

[8] Christian Becker-Asano and Hiroshi Ishiguro. Evaluating facial displays of emotion for the android robot geminoid f. In *Affective Computational Intelligence (WACI), 2011 IEEE Workshop on*, pages 1–8. IEEE, 2011.

[9] Dong-Wook Lee, Tae-Geun Lee, B So, Moosung Choi, Eun-Cheol Shin, KwangWoong Yang, MH Back, Hong-Seok Kim, and Ho-Gil Lee. Development of an android for emotional expression and human interaction. In *Seventeenth world congress the international federation of automatic control, Seoul*, 2008.

[10] Daniele Mazzei, Lucia Billeci, Antonio Armato, Nicole Lazzeri, Antonio Cisternino, Giovanni Pioggia, Roberta Igliozzi, Filippo Muratori, Arti Ahluwalia, and Danilo De Rossi. The face of autism. In *RO-MAN, 2010 IEEE*, pages 791–796. IEEE, 2010.

[11] David Hanson. Hanson robotics inc @ONLINE, 2014.

[12] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166, 2003.

[13] D. Mazzei, N. Lazzeri, L. Billeci, R. Igliozzi, A. Mancini, A. Ahluwalia, F. Muratori, and D. De Rossi. Development and evaluation of a social robot platform for therapy in autism. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 4515–4518. IEEE, 2011.

[14] Daniele Mazzei, Nicole Lazzeri, David Hanson, and Danilo De Rossi. Hefes: An hybrid engine for facial expressions synthesis to control human-like androids and avatars. In *Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on*, pages 195–200. IEEE, 2012.

[15] Albert L Rothenstein and John K Tsotsos. Attention links sensing to recognition. *Image and Vision Computing*, 26(1):114–126, 2008.

[16] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[17] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

[18] Koen Lamberts and Rob Goldstone. *Handbook of cognition*. Sage, 2004.

[19] Roxanne L Canosa. Real-world vision: Selective perception and task. *ACM Transactions on Applied Perception (TAP)*, 6(2):11, 2009.

[20] Farhan Baluch and Laurent Itti. Mechanisms of top-down attention. *Trends in neurosciences*, 34(4):210–224, 2011.

[21] Marisa Carrasco. Visual attention: The past 25 years. *Vision research*, 51(13):1484–1525, 2011.

[22] David Hanson, Andrew Olney, Steve Prilliman, Eric Mathews, Marge Zielke, Derek Hammons, Raul Fernandez, and Harry Stephanou. Upending the uncanny valley. In *Proceedings of the national conference on artificial intelligence*, volume 20, page 1728. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[23] David Hanson. Expanding the aesthetic possibilities for humanoid robots. In *IEEE-RAS international conference on humanoid robots*, 2005.

[24] David F Hanson, Giovanni Pioggia, Yoseph Bar-Cohen, and Danilo De Rossi. Androids: application of eap as artificial muscles to entertainment industry. In *SPIE's 8th Annual International Symposium on Smart Structures and Materials*, pages 375–379. International Society for Optics and Photonics, 2001.

[25] Michael I Posner and Steven E Petersen. The attention system of the human brain. Technical report, DTIC Document, 1989.

[26] Michael I Posner and Mary K Rothbart. Attention, self–regulation and consciousness. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1377):1915–1927, 1998.

[27] Duncan E Astle and Gaia Scerif. Using developmental cognitive neuroscience to study behavioral and attentional control. *Developmental psychobiology*, 51(2):107–118, 2009.

[28] Frank Broz, Hagen Lehmann, Chrystopher L Nehaniv, and Kerstin Dautenhahn. Mutual gaze, personality, and familiarity: Dual eye-tracking during conversation. In *RO-MAN, 2012 IEEE*, pages 858–864. IEEE, 2012.

[29] Michael Argyle and Mark Cook. Gaze and mutual gaze. 1976.

[30] Nikolaus Bee and Elisabeth André. Cultural gaze behavior to improve the appearance of virtual agents. In *IUI Workshop on Enculturating Interfaces (ECI)*, 2008.

[31] Jan Theeuwes. Exogenous and endogenous control of attention: The effect of visual onsets and offsets. *Perception & psychophysics*, 49(1):83–90, 1991.

[32] J.M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.

[33] Jeremy M Wolfe and Todd S Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6):495–501, 2004.

[34] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[35] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. Springer, 1987.

[36] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10):1489–1506, 2000.

[37] Judson P Jones and Larry A Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987.

[38] Moran Cerf, Jonathan Harel, Wolfgang Einhäuser, and Christof Koch. Predicting human gaze using low-level saliency combined

with face detection. *Advances in neural information processing systems*, 20, 2008.

[39] R. Rae. Gestikbasierte mensch-maschine-kommunikation auf der grundlage visueller aufmerksamkeit und adaptivitat. *PhD thesis, Universitat Bielefeld*, 1, 2000.

[40] Laurent Itti, Nitin Dhavale, and Frederic Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Optical Science and Technology, SPIE's 48th Annual Meeting*, pages 64–78. International Society for Optics and Photonics, 2004.

[41] Atsuto Maki, Peter Nordlund, and Jan-Olof Eklundh. Attentional scene segmentation: integrating depth and motion. *Computer Vision and Image Understanding*, 78(3):351–373, 2000.

[42] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123, 2002.

[43] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[44] F.H. Hamker. The emergence of attention by poulation-based inference and its role in distributed processing and cognitive control of vision. *Computer Vision Image Understanding*, 100(3):64–106, 2005.

[45] Jia Li, Yonghong Tian, Tiejun Huang, and Wen Gao. Probabilistic multi-task learning for visual saliency estimation in video. *International journal of computer vision*, 90(2):150–165, 2010.

[46]]B.W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor bases and distributions. *J. Vision*, 14(1):1–17, 2007.

[47] Sethu Vijayakumar, Jörg Conradt, Tomohiro Shibata, and Stefan Schaal. Overt visual attention for a humanoid robot. In *Intelligent Robots and Systems, 2001. Proceedings. 2001 IEEE/RSJ International Conference on*, volume 4, pages 2332–2337. IEEE, 2001.

[48] Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. Measuring visual clutter. *Journal of Vision*, 7(2), 2007.

[49] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[50] Ajay Mishra, Yiannis Aloimonos, and Cheong Loong Fah. Active segmentation with fixation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 468–475. IEEE, 2009.

[51] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

[52] Christian Siagian and Laurent Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(2):300–312, 2007.

[53] Sara Mitri, Simone Frintrop, Kai Pervolz, Hartmut Surmann, and Andreas Nuchter. Robust object detection at regions of interest with an application in ball recognition. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 125–130. IEEE, 2005.

[54] Albert Ali Salah, Ethem Alpaydin, and Lale Akarun. A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):420–425, 2002.

[55] Vijay Mahadevan and Nuno Vasconcelos. Saliency-based discriminant tracking. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1007–1013. IEEE, 2009.

[56] Hongliang Li and King N Ngan. Saliency model-based face segmentation and tracking in head-and-shoulder video sequences. *Journal of Visual Communication and Image Representation*, 19(5):320–333, 2008.

[57] Ali Borji, Majid Nili Ahmadabadi, Babak Nadjar Araabi, and Mandana Hamidi. Online learning of task-driven object-based visual attention control. *Image and Vision Computing*, 28(7):1130–1145, 2010.

[58] Christian Siagian and Laurent Itti. Biologically inspired mobile robot vision localization. *Robotics, IEEE Transactions on*, 25(4):861–873, 2009.

[59] Shumeet Baluja and Dean A Pomerleau. Expectation-based selective attention for visual monitoring and control of a robot vehicle. *Robotics and Autonomous Systems*, 22(3):329–344, 1997.

[60] Cynthia Breazeal, Aaron Edsinger, Paul Fitzpatrick, and Brian Scassellati. Active vision for sociable robots. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 31(5):443–453, 2001.

[61] Guido Schillaci, Sasa Bodiroza, and Verena Vanessa Hafner. Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *International Journal of Social Robotics*, 5(1):139–152, 2013.

[62] B.W. Tatler, M.M. Hayhoe, M.F. Land, and D.H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5), 2011.

[63] Constantin A Rothkopf, Dana H Ballard, and Mary M Hayhoe. Task and context determine where you look. *Journal of Vision*, 7(14), 2007.

[64] Alfred L Yarbus, Basil Haigh, and Lorrin A Rigss. *Eye movements and vision*, volume 2. Plenum press New York, 1967.

[65] Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, pages 518–523. IEEE, 2006.

[66] J Gregory Trafton, Magda D Bugajska, Benjamin R Fransen, and Raj M Ratwani. Integrating vision and audition within a cognitive architecture to track conversations. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 201–208. ACM, 2008.

[67] Patrick Holthaus, Karola Pitsch, and Sven Wachsmuth. How can i help? *International Journal of Social Robotics*, 3(4):383–393, 2011.

[68] J.E. Goldring, M.C. Dorris, B.D. Corneil, P.A. Ballantyne, and D.R. Munoz. Combined eye-head gaze shifts to visual and auditory targets in humans. *Experimental brain research*, 111(1):68–78, 1996.

[69] Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher. Designing effective gaze mechanisms for virtual agents. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 705–714. ACM, 2012.

[70] L. Itti, N. Dhavale, and F. Pighin. Photorealistic attention-based gaze animation. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 521–524. IEEE, 2006.

[71] Jürgen Ruesch and Weldon Kees. *Nonverbal communication: notes on the visual perception of human relationships.* University of California Press, 1956.

[72] Michael Argyle. *Bodily communication 2nd edition.* Routledge, 1988.

[73] Michael Argyle and Janet Dean. Eye-contact, distance and affiliation. *Sociometry*, pages 289–304, 1965.

[74] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 301–308. ACM, 2001.

[75] Edward T. Hall. *The Hidden Dimension.* Anchor Books New York, 1969.

[76] Leslie Adams and David Zuckerman. The effect of lighting conditions on personal space requirements. *The journal of general psychology*, 118(4):335–340, 1991.

[77] Elizabeth A Geden and Ann V Begeman. Personal space preferences of hospitalized adults. *Research in Nursing & Health*, 4(2):237–241, 1981.

[78] Gary W Evans and Richard E Wener. Crowding and personal space invasion on the train: Please dont make me sit in the middle. *Journal of Environmental Psychology*, 27(1):90–94, 2007.

[79] John R Aiello, Donna E Thompson, and David M Brodzinsky. How funny is crowding anyway? effects of room size, group size and the introduction of humor. *Basic and Applied Social Psychology*, 4(2):193–207, 1983.

[80] Emanuel A Schegloff. Body torque. *Social Research*, pages 535–596, 1998.

[81] Gay H Price and James M Dabbs. Sex, setting, and personal space: Changes as children grow older. *Personality and Social Psychology Bulletin*, 1974.

[82] John R Aiello and Tyra De Carlo Aiello. The development of personal space: Proxemic behavior of children 6 through 16. *Human Ecology*, 2(3):177–189, 1974.

[83] Stanley E Jones and John R Aiello. Proxemic behavior of black and white first-, third-, and fifth-grade children. *Journal of Personality and Social Psychology*, 25(1):21, 1973.

[84] Mark Baldassare. Human spatial behavior. *Annual Review of Sociology*, 4:29–56, 1978.

[85] Ross Mead, Amin Atrash, and Maja J Matarić. Automated proxemic feature extraction and behavior recognition: Applications in

human-robot interaction. *International Journal of Social Robotics*, pages 1–12, 2013.

[86] Benjamin W Tatler, Roland J Baddeley, Benjamin T Vincent, et al. The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision research*, 46(12):1857–1862, 2006.

[87] A Terry Bahill, Deborah Adler, and Lawrence Stark. Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology & Visual Science*, 14:468–469, 1975.

[88] Benjamin W Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 2007.

[89] Benjamin T Vincent, Roland Baddeley, Alessia Correani, Tom Troscianko, and Ute Leonards. Do we look at lights? using mixture modelling to distinguish between low-and high-level factors in natural image viewing. *Visual Cognition*, 17(6-7):856–879, 2009.

[90] Cynthia L Breazeal. *Designing Sociable Robots with CDROM*. MIT press, 2004.

[91] Susan Carey and Rochel Gelman. *The epigenesis of mind: Essays on biology and cognition*. Psychology Press, 1991.

[92] Abolfazl Zaraki, Daniele Mazzei, Manuel Giuliani, and Danilo De Rossi. Designing and evaluating a social gaze-control system for a humanoid robot. *IEEE Transactions on Human-Machine Systems*, PP(99):1–12, February 2014.

[93] Yoseph Bar-Cohen, David Hanson, and Adi Marom. The coming robot revolution, 2009.

[94] D Hanson and Victor White. Converging the capabilities of eap artifical muscles and the requirements of bio-inspired robotics. *Smart Structures and Materials 2004: Electroactive Polymer Actuators and Devices*, 5385:29–40, 2004.

[95] Giorgio Metta, Paul Fitzpatrick, and Lorenzo Natale. Yarp: yet another robot platform. *International Journal on Advanced Robotics Systems*, 3(1):43–48, 2006.

[96] C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. *rn*, 255:3, 1999.

[97] Robert A Hinde. *Non-verbal communication*. Cambridge University Press, 1975.

[98] C. Küblbeck and A. Ernst. Face detection and tracking in video sequences using the modifiedcensus transformation. *Image and Vision Computing*, 24(6):564–572, 2006.

[99] Fraunhofer IIS. Cognitive systems @ONLINE, June 2009.

[100] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991.

[101] Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[102] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.

[103] Nicholas J Butko, Lingyun Zhang, Garrison W Cottrell, and Javier R Movellan. Visual saliency model for robot cameras. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 2398–2403. IEEE, 2008.

[104] Wolfgang H Zangemeister and Lawrence Stark. Gaze latency: variable interactions of head and eye latency. *Experimental neurology*, 75(2):389–406, 1982.

[105] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. Elan: a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006, 2006.

Appendix A

Appendix A - Publication: Designing and Evaluating a Social Gaze Control System for a Humanoid Robot

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Designing and Evaluating a Social Gaze-Control System for a Humanoid Robot

Abolfazl Zaraki, Daniele Mazzei, Manuel Giuliani, and Danilo De Rossi

***Abstract*—This paper describes a context-dependent social gaze-control system implemented as part of a humanoid social robot. The system enables the robot to direct its gaze at multiple humans who are interacting with each other and with the robot. The attention mechanism of the gaze-control system is based on features that have been proven to guide human attention: nonverbal and verbal cues, proxemics, the visual field of view, and the habituation effect. Our gaze-control system uses Kinect skeleton tracking together with speech recognition and SHORE-based facial expression recognition to implement the same features. As part of a pilot evaluation, we collected the gaze behavior of 11 participants in an eye-tracking study. We showed participants videos of two-person interactions and tracked their gaze behavior. A comparison of the human gaze behavior with the behavior of our gaze-control system running on the same videos shows that it replicated human gaze behavior 89% of the time.**

***Index Terms*—Active vision, context-dependent social gaze behavior, human–robot interaction, scene analysis, social attention.**

Fig. 1. FACE humanoid robot interacts with a group of people. *Picture courtesy*: E. Gargano.

I. Introduction

With the rapid advancement of humanlike robots and of related computing methods in robotics, social robots that interact with humans are becoming more integrated into daily life [1]. Social robots are designed for tasks and scenarios that require close interaction and collaboration with humans. Thus, in addition to task-performing capabilities, social robots must be able to display socially acceptable behavior. For example, Fig. 1 shows the facial automaton for conveying emotion (FACE) humanoid robot [2], [3], involved in a social scenario where it interacts with a group of people. To display behavior that humans perceive as natural, the robot should direct its attention at the most important person at the right time based on the current social context. Social robots thus require a mechanism that is able to control attention and gaze on the basis of social cues and information that are extracted from raw visual-auditory data.

To design attention systems for social robots, it is necessary to consider the psychological, neurological, and computational aspects of human attention [4]–[10] as well as the social cues and conventions. This information can support a robot gaze-control system (GCS) to direct attention at the appropriate target during interactions with humans.

This paper presents a modular context-dependent social GCS which has been implemented as part of the Hanson humanoid robot FACE [11]–[14]. The GCS enables the robot to analyze high-level features and cues of a complex social scene in order to direct the gaze at the most important social target. The selection of these points is based on high-level visual and auditory features, which are extracted from two-dimensional (2-D) videos, depth data, and auditory signals. The GCS captures and analyzes incoming sensory inputs, identifies humans in the robot's environment, and extracts their high-level social features (i.e., facial expressions, age, gender, body gestures, head pose, distances, orientation, and speaking probability), using parallel algorithms. Using the extracted high-level social features, an attention model selects the most prominent attention target. The GCS continuously adjusts the robot gaze parameters using an algorithm that is based on an implementation of Itti *et al.*'s model [15]. As a proof of concept, we evaluated a prototype in a social context comparing GCS-generated attention points and gaze trajectories with human gaze data that are acquired using an eye tracking device.

Manuscript received March 17, 2013; revised October 17, 2013 and December 18, 2013; accepted January 19, 2014. This work was partially funded by the European Commission under the 7th Framework Program projects EASEL, "Expressive Agents for Symbiotic Education and Learning," under Grant 611971-FP7-ICT-2013-10, and JAMES, "Joint Action for Multimodal Embodied Social Systems," under Grant 270435-FP7-STREP. This paper was recommended by Associate Editor B. F. Mettler.

A. Zaraki and D. Mazzei are with the Research Center "E. Piaggio," University of Pisa, 56122 Pisa, Italy (e-mail: ab.zaraki@gmail.com; mazzei@di.unipi.it).

M. Giuliani is with the Department of Cyberphysical Systems, Fortiss GmbH, 80805 Munich, Germany (e-mail: giuliani@fortiss.org).

D. De Rossi is with the Department of Information Engineering, University of Pisa, 56122 Pisa, Italy and also with the Research Center "E. Piaggio," University of Pisa, Italy (e-mail: d.derossi@centropiaggio.unipi.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2014.2303083

II. Background and Related Work

Section II-A discusses human and robot *attention modeling* and Section II-B reviews how *attention modeling* can support human *gaze models* and the implementation of robot gaze models.

A. Attention Modeling

The main aim when modeling human attention is to identify which features guide human attention in a complex scene and how these features influence human attention. Researchers have described two different aspects of the human attention target selection process: *top-down* and *bottom-up* processing [16]. Predicting human attention is a complex issue, which involves both aspects of attention processing and needs models that simulate the working of the mind [17] which is beyond the scope of this paper. The attention modeling in this study focuses on the bottom-up cues (features) without a top-down cognitive effect.

The low-level visual features of an image (i.e., color, intensity, and orientation) guide human attention to specific target points of a visual scene [18]–[21]. In the computational implementation of salience-based attention models of [18], attention selection for a given visual scene is as follows. First, the algorithms extract visual low-level features of the given scene. Next, local competition across image space and feature scales is computed yielding feature maps. Finally, individual feature maps are combined by weighted sums creating the salience map. Based on the salience map, the algorithm can then select attention targets, for example by applying the "winner-takes-all" principle. Salience-based attention models have been used in several robotics applications [15], [21]–[23]. Some researchers have also extended salience-based attention models, by adding low-level features and using image-processing techniques, including human face region [24], depth and motion information [25], and spatial resolution of an image [26].

For human–robot interaction, attention models must be capable of mimicking the gazes of speakers and listeners. Tatler *et al.* [27]'s review of the major limitations of salience-based attention modeling showed that such models do not account for many important aspects of complex scenes that cannot be explained only through low-level features analysis. Therefore, when designing a social robot attention system, high-level communicative and social features (e.g., verbal/nonverbal cues) must be accounted for, which are fundamental to the human attention system.

Mutlu *et al.* [28] derived a gaze model (attention points) of a human during story telling by first collecting gaze patterns (locations of attention points, target selection frequencies, and fixation durations) of a professional storyteller. They designed and evaluated a model that reproduced human natural gaze behavior on the humanoid robot ASIMO. They assessed the efficacy of their gaze model by manipulating the frequency and fixation duration of the robot's gaze between two participants. Participants recalled the story better when the robot looked at them during storytelling, although women liked the robot more when it looked at them less frequently. Although Mutlu's gaze model generated a natural gaze, it was not aware of the social context of the dynamic scene, and displayed only a predefined gaze pattern.

Trafton *et al.* [29] integrated vision and audition within a cognitive architecture, which enabled a social robot to track conversations and focus its attention on the speaker. They evaluated their system on a social mobile robot. The proposed architecture correctly guided the robot's attention to the correct speakers, but it did not account for many of the human communicative cues (e.g., gesture, motions, proxemics) known to be fundamental for social attention calculation.

Holthaus *et al.* [30] proposed a spatial model for a robot attention system. The system drives the attention of a receptionist robot according to the spatial information of humans interacting with the robot. The robot located and tracked humans in its field of view (FOV) by monitoring their distance. The robot moves its head and body in order to initiate or terminate a social interaction with humans when they are getting closer to the robot. Through a questionnaire-based evaluation, Holthaus *et al.* found that even if the robot made random movements when someone was approaching, external observers evaluated the interaction as humanlike. Although the results show the importance of proxemics and contextual reactions when modeling humanlike robots to enable robots to have a natural social gaze behavior, their system lacks other factors (e.g., gesture analysis, auditory signal analysis) that have been proven to guide attention in human attention modeling.

With regard to the current challenges in attention systems for social robots, we hypothesize that a comprehensive attention model should specify the most prominent target points on the basis of high-level environmental visual and auditory features analysis. Here, we propose a features-based attention model based on empirical data and not on a neurological model. We show that the proposed attention model can emulate human social gaze behavior based on high-level human-relevant features of 2-D images, 3-D images, and auditory signals. Our framework also provides a similar high-level image interpretation for social robots to the human attention system. Thus, the GCS enables a social robot to naturally interact with multiple people in a dynamic environment and take into account the social context.

B. Gaze Modeling

A gaze is a coordinated motion of eye and head movements through which the center of human visual attention is moved to a specific point that is identified by the human attention system on the basis of various attractive cues.

Through analysis of the gaze behavior of humans and monkeys, Goldring *et al.* [31] demonstrated that gaze behavior is regulated by complex dynamics that support observation and deliver meaningful information, thus driving the conversation flow. They studied the characteristics of the human head and eye movement to understand the strategies when people gazed at visual, auditory, and visual-auditory targets. They found that target modalities have an effect on human gaze characteristics, some of which they identified (head and eye velocities, motion amplitude delays) during gaze shifts between targets.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Several models and implementations of robot/agent GCSs have been proposed. Andrist *et al.* [32] proposed an effective gaze model for virtual agents with various gaze characteristics such as amplitude, velocity, and latency period in a gaze shift. They evaluated their gaze model on a humanlike virtual agent. Andrist's results show that when the agent maintains its head orientation toward the participant to emphasize the social interaction (affiliative gaze), it induces positive feelings. In addition, when the agent maintains its head orientation more toward visual space to emphasize other information (referential gaze), it improves the participants learning capabilities.

Itti *et al.* [15] presented a gaze model for target shift and smooth tracking which was implemented using an avatar. In their model, the amplitudes of head and eye movements were estimated and linked with the initial position of the eye in its orbit.

However, due to the complexity of human gaze behavior, a comprehensive context-dependent model that estimates gaze parameters (e.g., velocity, amplitude, latency), has not yet been implemented for robots and avatars.

In this study, an innovative GCS, which is based on a combination of work presented in [15], [31], and [33], was implemented and tested in order to control the head and eye movements (gaze) of the FACE robot.

III. Social Cues for Attention Elicitation

The GCS calculates the robot's attention on the basis of various social cues that are extracted as the high-level interpretation of raw images, sounds, and depth data acquired through the robot's hardware. In this section, we describe nonverbal/verbal cues, proxemics-derived features, an effective visual FOV, and habituation effects in humans and their implementation in the GCS.

A. Nonverbal/Verbal Cues

Nonverbal cues are wordless signals that are used to deliver a meaningful message and consist of approximately two-thirds of human–human interaction [34]. People use facial expressions, body gestures, head poses, and gazes to attract other people's attention, to express their emotions and intentions and to manage the flow of interaction [35].

Verbal cues such as vocalization, prosody, and speech, in particular, directly affect human attention. Argyle and Dean [36], [37] found that humans are able to immediately locate single speakers in a group. Using a gaze tracker, Vertegaal *et al.* [38] showed that in a group of four people, listeners looked at the person who was speaking 88% of the time.

A gesture recognition system was therefore integrated in the GCS to recognize the motions and gestures of humans. Verbal cues are analyzed in the GCS combining the human-tracked information with a speech angle detection device, thus identifying the speaker in the social context in which the robot is involved.

B. Proxemics

Proxemics, i.e., the physical distance between two humans, influences implicit and explicit interaction between people. Hall [39] investigated the effect of physical space as an important nonverbal cue in the interpersonal communication. He defined four "reaction bubbles," which are circles (*intimate*, *personal*, *social*, and *public* distances) located around the human body at varying distances (see the left semicircle of Fig. 2). Human social cues elicit different levels of attention depending on their spatial location. For example, if a human raises his/her hand or eyebrow, the attention of the surrounding people will be attracted to different extents, based on their distances. Their attention will be drawn more often toward the human giving the social cue, if the human is in their personal space, while they look less often at this person if he/she is located in their social space. Tatler *et al.* [40] and Bahill *et al.* [41] showed that people look at close targets more frequently than at distant targets. Thus, in the GCS, distance is considered as a nonverbal cue that influences the attention.

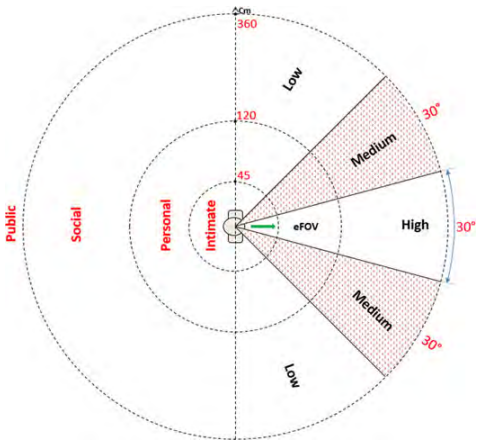


Fig. 2. *Left semicircle*: according to Hall's theory, there are four reaction bubbles at certain distances around the human body that influence implicit and explicit interaction between people. The *intimate distance* is used to embrace or touch a person, the *personal distance* for interactions between family and close friends, the *social distance* for interactions between acquaintances, and the *public distance* for speaking in public. *Right semicircle*: human sight is centered on the eFOV. We regard social signals shown in the eFOV and in the areas left and right of the eFOV as having a high, medium, or low relevance, depending on the distance of the eFOV.

C. Effective Visual Field of View

The human eye collects visual information at high resolution from a small central area called the fovea, while the peripheral FOV is sampled at lower resolutions [33]. Human attention is more attracted by affective and social visual features in a small central area known as the *effective field of view* (eFOV). Social visual features collected in this area elicit higher levels of human attention. Behavioral studies indicate that humans have a strong tendency to look at the center of an image, regardless of the content of the whole image [42], [43]. Thus, we consider the angle between a human position in the scene and the center of

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

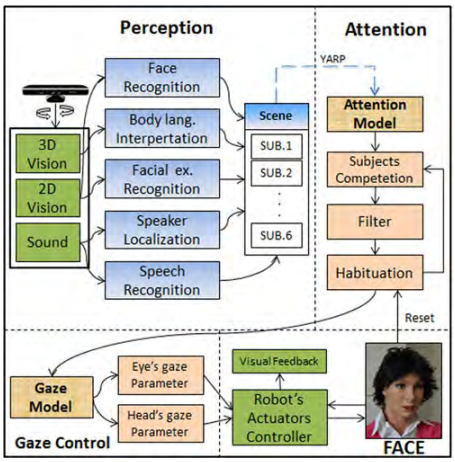


Fig. 3. Modular structure of the GCS: the perception layer receives audiovisual information and extracts human social cues. Based on these cues, the attention layer computes the most prominent target points. Using a gaze model, the GC layer drives the robot's actuators according to target positions.

eFOV as a nonverbal cue that influences the robot's attention toward the human. The right semicircle in Fig. 2 shows the relevance levels implemented in the GCS: a human located in the eFOV is considered as highly relevant; signals viewed in a 30° radius left or right of the eFOV with medium relevance, and all other human social signals as low relevance. In this study, we used the concept of the eFOV only for visual features of a human presented in FOV.

D. Habituation Effect

The habituation effect is a decrease in response to a stimulus after repeated presentations [44]. In the GCS, habituation is implemented as a time-variant function that adjusts the level of attention elicited by the selected target, similar to the work of Breazeal *et al.* [22].

IV. Gaze-Control System

Our GCS consists of three distinct layers: *perception*, *attention*, and *gaze control* (GC) (see Fig. 3). The GCS collects visual-auditory information from the environment, detects and analyzes a wide range of human social cues. It then selects the most important region to focus attention on. In order to ensure humanlike head and eye movements, attention selected points are passed to a gaze dynamic control layer implemented on the basis of [15], [31], and [33].

In this section, we describe the three GCS layers together with details of the FACE robot hardware and control software.

TABLE I
Parameters Extracted by the Perception Layer

Used Library	Extracted Features
Kinect SDK	Human 3D position of up to six people Twenty body joints coordinates for two humans Sound direction and beam angle
SHORE	Positions of face, eyes, nose, and mouth Eyes and mouth state (open/close) Gender classification (male/female) Age estimation (years) Facial expressions Face rotation (up to ±60 image planes)
Face recognition	Name of human (according to pre-trained data set)
Body gesture Head pose recognition	Gestures and body motion Head pose (roll, yaw, pitch angles)

A. Perception Layer

The perception layer contains two parts: data acquisition and feature extraction. These parts prune data and extract high-level features from the visual-auditory information of a social scene. The perception layer acquires raw data through a Microsoft Kinect device running the Kinect for Windows SDK.[1] A Kinect RGB-D camera records 2-D video and depth images with a resolution of 640 × 480 pixels at 30 fps and has a built-in four-element microphone array for audio beam acquisition.

Kinect-acquired raw data are analyzed by extracting a variety of verbal and nonverbal cues that are classified using different taxonomies and stored in a *meta-scene object* which is streamed to the attention layer through a YARP [45] gateway.

The GCS implementation aims to extract socially relevant visual features (i.e., human proxemics, orientation, facial properties, gestures, and entry time) and auditory features (i.e., sound source angle and pronounced words), through various parallel algorithms and/or dedicated libraries. The algorithms/libraries and extracted features are summarized in Table I.

1) Face Detection, Facial Expression Analysis, and Face Recognition: Observation of human visual attention revealed that face-like shapes attract human attention [46]. In addition, various features such as a human's age and facial expressions (i.e., happiness, sadness, surprise, anger), directly regulate the social interactions [47]. In a social context, it is imperative to know the age and gender of the interactional partners and to continuously receive feedback of facial expressions and mimics. Like humans, robots should have the same ability to locate faces and understand facial expressions and related social features.

For facial expressions analysis, the perception layer uses the sophisticated high-speed object recognition engine (SHORE) [48], [49]. SHORE is a robust detection engine that is based on the illumination invariant approach that detects multiple faces in a single frame and tracks them in real time. The SHORE engine receives the 2-D frame that is acquired from Kinect, detects faces, assigns consistent IDs to each face, and estimates various facial features which are reported in Table I.

[1] http://www.microsoft.com/en-us/kinectforwindows/

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

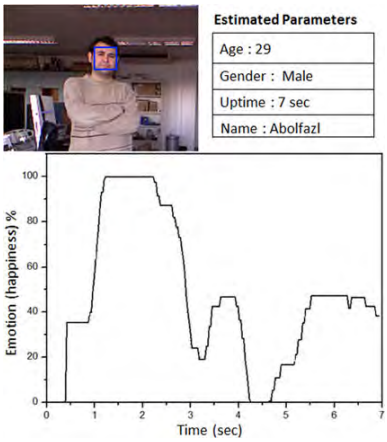


Fig. 4. Example for face-related feature extraction: the module detects a face and extracts the estimated happiness ratio, age, gender, and entry time.

Recognizing facial features and expressions is very important for social context analysis but requires information on the identity of the humans to be integrated to enable the robot to adjust its behavior in a context-dependent manner.

The GCS perception layer integrates a facial recognition engine that is based on principal component analysis (PCA) [50]. The facial recognition module uses a pretrained dataset to assign an identity to the recognized faces and stores the extracted features in the face dataset. Fig. 4 shows an example of the perception layer merged information where the person's name and facial information recognized from the SHORE engine (i.e., estimated happiness ratio, age, gender, and entry time) are merged.

2) Body Gesture and Head Pose: People use body gestures and head poses as social signals when they interact with each other [35], [51] and these signals are one of the strongest nonverbal cues that elicit human attention. For example, in a multiparty interaction, if one of the humans raises his/her hand or waves the arm, others will direct their attention to him/her. Robots thus need to be able to react to these social cues.

The perception layer uses the skeleton tracking of the Kinect SDK to recognize a person's movements. The Kinect SDK locates up to six humans by merging information from RGB and depth images and recognizes body joint coordinates for the two closest persons. In order to estimate the head pose, the perception layer computes Euler angles (pitch, roll, and yaw angles), using SDK's head data in real time. In addition, we implemented a dynamic body gesture and head pose recognition engine which continually monitors the body's motion and head pose through extracted skeleton information, and identifies meaningful motions.

3) Speaker Location: The auditory streams cause an unintentional shift of attention that usually shifts the gaze toward the sound source. Hence, it is essential for robots to localize the speaker in a multiparty interaction. The perception layer uses the Kinect SDK to calculate the sound source direction with a triangulation algorithm. It computes the 3-D position and beam angle of sound signals received by the microphone array. The algorithm considers only auditory signals that can be associated with humans in the scene by comparing the direction of the sound to the 3-D positions of the detected humans. In a real situation, human attention is also attracted by auditory signals outside the visual FOV, which are not relevant to the visual stimuli. However, because of the limitation of the sensor detection range, the system is designed to ignore sound signals, not related to multiparty interaction, as environmental noise. This limitation of the system is one of the issues that prevent natural gaze behavior from being generated.

Once a sound source is associated with a person, a dedicated engine is used to recognize speech and convert it to text if possible. A human's recognized words are stored in the *meta-scene objects* along with a *speaking probability* parameter that is calculated on the basis of a comparison between the sound angle and the human's position.

4) Database: A database of all people seen by the perceptual layer is stored as the *meta-scene object*. The *meta-scene object* has a hierarchical structure through which an arbitrary number of people can be inserted. Each person object includes the person's unique ID and the associated high-level features.

Once a new person has been identified by the PCA identification algorithm and by SHORE, a new person instance is created in the *meta-scene object* which is populated with the features, extracted by the perceptual layer. Since the PCA engine recognizes frontal faces, new pictures are continuously taken by the RGB image and stored in the PCA training set. PCA unrecognized humans are inserted into the *meta-scene object* using a temporary ID which is reassigned once the new person name has been inserted by the operator through the GCS interface.

Through the NET object serialization, the *meta-scene object* is converted into an XML structure which is streamed through a dedicated YARP port between the GCS layers and modules.

B. Attention Layer

The attention layer receives the *meta-scene object* as XML streamed through YARP, and then deserializes it back in a manageable object. The aim of the attention layer is to find the most prominent region of the scene that the robot should focus on.

1) Target Selection Strategy: The core of the attention layer calculates the elicited attention (EA) level of each human present in the scene, on the basis of various features. Since the numerical values quantifying the features are not within the same range, they are normalized (X_n) to the range $[0, 1]$ by considering the maximum values that features can have according to the sensor properties and features ranges. The overall EA of each human in the scene is calculated on the basis of four main components: social features (F), proxemics (P), orientation (O), and a memory component (EAM):

$$\mathrm{EA}_{S_j}(t) = F_{S_j} + P(r) + O(\theta) + \mathrm{EAM}_{S_j} \tag{1}$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

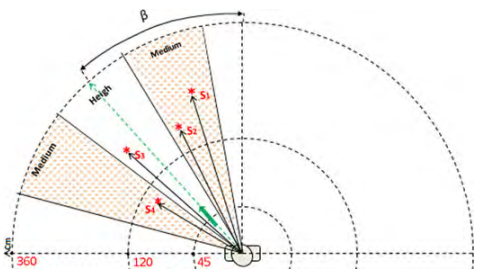


Fig. 5. Green line shows the sight line of FACE, while S_1 to S_4 represent the human's 2-D position according to distances and orientations.

where EAM_{S_j} is a parameter that refers to the memory of the robot not yet included in the database and consequently set to zero.

The social feature elicitation contribution F_{S_j} is calculated as a weighted summation of social normalized features X_n, which can be written as

$$F_{S_j} = \left(\sum_{i=1}^{n} W_i . X_n \right) \tag{2}$$

where weight W_i is set on the basis of the feature's importance. We explain the feature's importance and priority in Section V-F.

The values of $P(r)$ and $O(\theta)$ reflect the proxemics and orientation contribution in the model (described in Sections III-B and III-C). Because of the unavailability of sensory data in nearby and distant areas, the attention layer reflects the proxemics effect only for personal and social spaces and the orientation effect only for high and medium spaces. These effects can be expressed as

$$P(r) = Fp_r . \left(1 - \frac{|r|}{r_{\max}}\right) \qquad O(\theta) = FO_\theta . \left(1 - \frac{|\theta|}{\theta_{\max}}\right) \tag{3}$$

where $|r|$ and $|\theta|$ denote the current distance and orientation of humans with respect to the robot. Fp_r and FO_θ convert continuous distance and orientation into discrete values, respectively. These discrete values represent four proxemics spaces (intimate, personal, social, and public) for Fp_r, and the three zones of the eFOV (high, medium, and low) for FO_θ (see Fig. 5). $r_{\max}$ is the maximum distance and $\theta_{\max}$ is the maximum angle detectable by the sensor. Clearly, the levels of $P(r)$ and $O(\theta)$ are at their maximum when a human is in the *intimate* space and the center of the eFOV of the robot.

Since the human's orientation is calculated with respect to the robot's current head position, the Kinect sensor should simultaneously turn with the robot's head to capture the same scene as the robot. For this reason, a servomotor is used to horizontally rotate the Kinect at the same angle as the robot's head ($\pm\beta$).

The attention layer shows a strong tendency to move to the center of image [42], [43]. Hence, a virtual point (VP) is positioned at the center of the image, to attract the robot's attention like a virtual human. The EA is simultaneously calculated for six humans in the robot's FOV. The attention layer selects the winner (i.e., the human with the highest EA level) through a competition among humans and the VP

$$\mathrm{Max}(\ \mathrm{EA}_{S_1}, \mathrm{EA}_{S_2}, \ldots, \mathrm{EA}_{S_6}, \mathrm{VP}) \rightarrow K_{\mathrm{winner}} \rightarrow (X, Y) \tag{4}$$

where K_{winner} is the winner's ID.

It finally extracts the winner's head position (X, Y) from the meta-scene object and sends it to the gaze layer.

2) Habituation Function: The habituation effect is activated, once the robot makes eye contact with the selected human (winner). The attention layer multiplies the habituation function (HF) by the winner's score (EA_{S_k}), in order to make a time-variant decreasing score ($\mathrm{EA}_{S_{\mathrm{winner}}}(t)$) for the winner, as

$$\mathrm{EA}_{S_{\mathrm{winner}}}(t) = \mathrm{EA}_{S_k} \ \cdot \ \mathrm{HF}(t) \tag{5}$$

where

$$\mathrm{HF}(t) = \mathrm{Peak} \ \cdot \ \mathrm{Max}\left(0, \left(1 - \frac{\Delta t}{\tau}\right)\right) \tag{6}$$

and τ is a time constant and peak is the maximum amplitude of the HF. Following [22], we set the time constant and peak parameters to 10 and 30 s, respectively. The HF value linearly decreases to zero within the time constant τ. When the robot's gaze reaches the new winner, Δt is reset to zero and HF will be maximized. The model searches for a new winner in real time while decreasing the score of the last winner to zero. Employing this system, the winner's attractiveness for the robot decreases gradually over time thus allowing other people to attract the robot's attention. It empowers the robot to show a more natural and dynamic behavior.

3) Time-Based Filter: Because of the mechanical limitations of the robot's head and eye actuators, the robot's gaze is not capable of synchronizing with the rapid changes in target positions. To solve this problem, a time-based filter is used. The attention layer sends the winner's position to the gaze control (GC) layer in real time, which is entrusted with generating gaze parameters according to the target position. The GC layer continually receives updates from the attention layer and decides how to direct the robot's gaze to the selected human.

C. Gaze Control Layer

1) Head and Eye Movements: A gaze is composed of two components: eye movement and head movement. The summation of these components (gaze) is relatively constant [31]. The amplitude of the gaze can be written as

$$\theta_g = \theta_e + \theta_h \tag{7}$$

where θ_e is the eye angle in its orbit with respect to the head (internal coordinates), θ_h is the head angle with respect to the environment (global coordinates), and θ_g is the gaze angle in the global coordinate system. Since the gaze angle is constant and any combination of head and eye is possible, where the angle of the eye increases, the angle of the head decreases and vice versa. Assuming that the eyes are at the center of their orbit before gaze shift, $\theta_e(t = 0) = \theta_0$ is equal to zero. In order to accomplish a gaze, the eye moves until it reaches the threshold

$\theta_{\rm thr}$ and the head movement starts to compensate for the eye movement. If the eye's current position is not at the center of its orbit, then $\theta_{\rm thr}$ is changed. In fact, the initial angle of the eye (θ_0) and the position of the selected human determine whether the gaze needs to be accomplished by the eye movement alone or together with the head movement.

In order to ensure a humanlike gaze shift, we use a humanlike gaze model [31], [33], which is derived from a motion capture of human subjects, using high-speed video-based eye and head tracking. The equations for $\theta_{\rm thr}$ and θ_h were estimated, using empirical data. In this model, $\theta_{\rm thr}$ varies depending on the initial position of the eye in its orbit (θ_0), and can be obtained as

$$\theta_{\rm thr} = -0.28\theta_0 + 11.2 \tag{8}$$

where θ_0 is positive if the initial eye deviation has the same direction as the subsequent movement. This equation is obtained from [33]. The constant numbers express head and eye dynamics in vertical and horizontal movements.

Following this notation, to accomplish a given gaze (θ_g), θ_h can be obtained as

$$\theta_h = \begin{cases} 0, & \text{if } -\theta_{\rm thr} < \theta_g < \theta_{\rm thr} \\ \theta_{\rm thr} + k(\theta_g - \theta_{\rm thr}), & \text{otherwise} \end{cases} \tag{9}$$

where

$$k = 0.0185\theta_0 + 0.715 \tag{10}$$

and *k* is a parameter that controls the eye and head movement, in order to generate a humanlike gaze shift. This equation is derived from [33] on the basis of empirical data.

2) Head and Eye Velocities: The head and eye velocities vary according to target eccentricity and modality [31]. However, the auditory and visual targets influence the velocities of the head and eyes in different ways. In this study, it is assumed that visual and auditory stimuli have the same effect on the robot's gaze. When the attention layer selects the target's coordinate in a pixel (X,Y), the GCS GC layer calculates the amount of target eccentricity with respect to the current sight line of the robot.

In [31], a relatively linear relationship between target eccentricity and head and eye velocities has been shown. Thus, because of the physical limitation of the mechanical structure of the robot used, we define three levels of velocities as *high*, *medium*, and *low* for the robot's actuators. The GCS GC layer calculates the level of the head and eye velocities as a function of head and eye amplitudes, assuming that the eye always moves faster than the head. The concept of velocity is implemented in the GCS by the amount of gaze angle (in degrees) over time needed to reach the target point (in seconds). Velocity can be expressed for the head as

$$[Vh_{\rm high}, Vh_{\rm medium}, Vh_{\rm low}] = [75, 45, 22]\,^\circ/\text{sec} \tag{11}$$

and for the eye as

$$[Ve_{\rm high}, Ve_{\rm medium}, Ve_{\rm low}] = [450, 150, 90]\,^\circ/\text{sec}. \tag{12}$$

3) Head and Eye Latencies: Latency is the delay in reaction time when people shift their gaze to a target. It is influenced by target eccentricity and modality. Head latency is longer than eye latency [52], and varies approximately in the range of 50 to 300 ms. Auditory stimuli have the longest reaction latencies for central targets $\theta_g < 20$, and the visual targets elicit the longest reaction latencies in $\theta_g > 40$ (see [31]). In order to reach the target points, the model generates rapid saccadic eye movements with a 50 ms delay, then after a 200 ms delay, it generates head movements for the robot. Two constant values (l_e, l_h) denote eye and head latencies in the model.

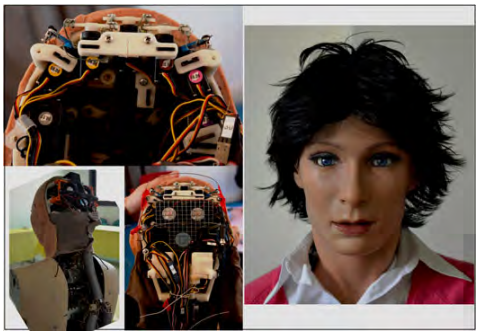


Fig. 6. FACE's android actuator system consists of 32 servo motors together with artificial skin, allowing FACE to reproduce high-quality facial expressions and humanlike head and eye movements.

The GCS GC layer estimates the gaze parameters for eyes and head, based on the proposed gaze model and target eccentricity as ($\theta_{\rm thr}$, θ_e, V_e) and ($\theta_{\rm thr}$, θ_h, V_h) for the robot actuators. It also generates reaction latency. All the derived information is sent to the robot control (RC) layer which is directly connected to the robot actuators.

V. Proof of Concept Evaluation

In order to assess the performance of the GCS and the underlying model, a gaze tracking study was performed. The purpose of the proof of concept evaluation was to tune the parameters of the GCS. Thus, one aim was to determine which social cues have more of a prominent effect on the attention of the study participants. A second aim was to compare how well the GCS was able to replicate human gaze behavior on the same context (input videos).

A. Facial Automaton for Conveying Emotion Robot

We implemented our GCS on the humanoid social robot FACE created by Hanson Robotics [11]–[14] (see Fig. 6). The robot has a female appearance and its artificial skull is covered by a porous elastomer material called FrubberTM which requires less force to be stretched by servo motors than other solid materials. FACE has 32 servo motors to replicate high-quality facial expressions and humanlike head and eye motions [2], [3]. The movements of head and eyes are in 4 degrees of freedom (DOF) and 2 DOF. The kinematic structure of the actuation

system enables the robot to generate realistic facial expressions and gaze behavior [53], [54].

B. Participants

A total of 11 participants (nine males and two females), from the Department of Mechanical Engineering at the Technical University of Munich took part in this experiment. The mean age of the participants was 27.3 (range 22–35). Eight of the participants were native German speakers, the three other participants spoke English, but were not native English speakers. The participants received a chocolate for taking part in the experiment.

C. Experiment Procedure

The participants were asked to watch a video showing two humans discussing different research topics. In the video, the two humans enter the room separately, sit down on two chairs, and then leave the room separately. During the discussion, both humans talk to the video camera from time to time, as if they were interacting with a third person (the robot/the experiment participants), in order to help with experiment participant engagement.

The video was taken in parallel with an HD video camera and a Kinect RGB-D camera placed side by side. The scene captured by HD video camera was shown to the participants for human gaze analysis, while the Kinect-acquired RGB-D data were used as input for the GCS.

The video lasted 7:20 min and consisted of three subscenes. In the first and third subscene, the people in the videos talked in English, in the second subscene, they spoke in German. In each subscene, only one person spoke at a time, while the nonspeaking participant executed diverse gestural and postural acts in order to attract the attention of the viewer. Gestures and movements included: stretching while being seated, raising an arm, getting up from the chair to get a drink, and retrieving a smart phone from their pocket.

While the participants watched the video, they wore a DIKABLIS eye tracking system to record gaze behavior (see Fig. 7). The eye tracker included a field camera in order to capture the scene and an infrared camera to capture a video of the left eye. The participants sat roughly 75 cm away from a 23-inch display. Before starting the experiment, the DIKABLIS eye tracker was calibrated to enable it to detect the whole pupil. Experiments were carried out in a room with controlled lighting to prevent any external light sources interfering with the eye tracking system.

D. Data Collection and Analysis

The DIKABLIS eye tracker analysis software produces a video of the field camera with an overlaid cross-hair showing where the participants look. We used ELAN [55] to annotate these videos on a frame-by-frame basis: timing looking at either *person A* (the person on the left in the scene), *person B* (person on the right), or at the *environment* (other regions). We also annotated when and how often the person not speaking provided a nonverbal social cue. Table II summarizes how often the experiment participants looked at each person and how often the person was either speaking or providing a nonverbal cue. After annotation, log files containing time duration (in milliseconds) and position (i.e., person A, person B, environment) of the participants' gaze fixations were exported. The average attention of the participants was calculated using MATLAB. In order to identify the verbal/nonverbal cues that cause nonsaccadic gaze shift to person A, person B, and the environment, we analyzed the human gaze behavior obtained by averaging the participants' logs. The average gaze pattern of participants was divided into 15 segments (A–N), identifying regions where the observers' attention was on an individual person (A or B). The various peak points of the average gaze pattern were also selected by identifying verbal and nonverbal cues that attracted participants' attention thus triggering the gaze shift.

Fig. 7. DIKABLIS eye tracking system has two separate cameras: the field camera looks to the front in order to capture the scene the participants are looking at, and an infrared camera captures a video of their left eyes.

TABLE II
ATTENTION OF PARTICIPANTS TOWARD PERSON A AND B WHILE SPEAKING, WHILE PERFORMING NONVERBAL CUES, AND THE AVERAGE FROM THE ENTIRE VIDEO (AVG. ATT. PART.)

Person	Speaking(%)	Non-verbal Cues(%)	Avg. Att. Part.(%)
Person A	41.8	20.3	54.4
Person B	32.5	23.1	43.6

The GCS parameters were extracted according to the target selection priorities of participants on the basis of the method described in Section V-F.

After the GCS parameters had been extracted through human gaze analysis and interpretation, the GCS-generated gazes were compared with the average gaze pattern of participants. The Kinect-acquired RGB-D data were used as input to the GCS module which generated a new video similar to the one obtained through the DIKABLIS eye tracker analysis software. A red circle identifying the FACE robot gaze point was streamed through YARP to the RC library. The GCS-generated video was annotated using ELAN with the same modalities used for the participant's video annotations. The error between the two gaze paths was calculated as an average of the absolute difference between the human gaze and GCS pattern functions.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

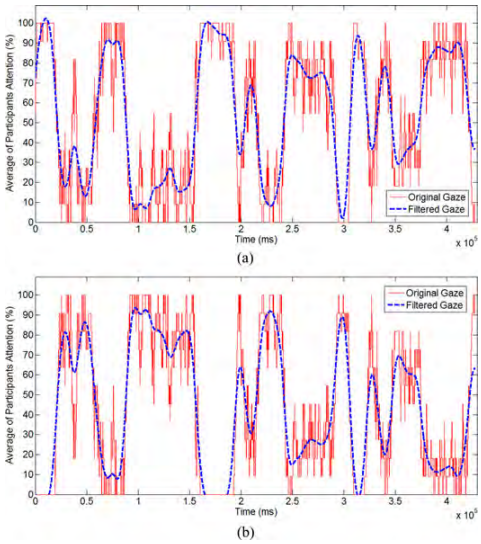


Fig. 8. Average attention on person A and person B in the recorded video. (a) Average attention of 11 participants on person A. (b) Average attention of 11 participants on person B.

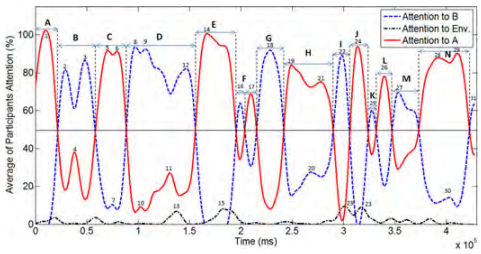


Fig. 9. Average participant attention on person A, person B, and the environment. The segments identify regions when the gaze is kept on a person (A or B). The peaks identify specific events that triggered the participant's attention.

E. Gaze Behavior Results

Fig. 8(a) shows the attention on person A, and Fig. 8(b) shows the attention on person B, respectively. In order to obtain the overall gaze behavior, saccadic gaze shifts were filtered through a second-order low-pass filter, which is shown as a dotted line in the figures.

As shown in Figs. 9 and 10, the average human gaze behavior can be divided into saccadic (high-frequency) and nonsaccadic (low-frequency) movements. For example, peak 1 of segment A shows that 100% of the participants looked at person A. Peak 1 corresponds to the instant when person A entered the room and initiated the conversation with the observer. Similarly at Peak 2 of segment B, person B entered the room while person A was still there. At this point, 82% of participants looked at person B and the rest of the participants kept their focus on person A. With the same methodology, we analyzed the entire average gaze pattern (see Fig. 9) by identifying various social verbal and nonverbal cues that attracted participants' attention. Social cues that were identified in the videos and associated peak numbers are reported in Table III.

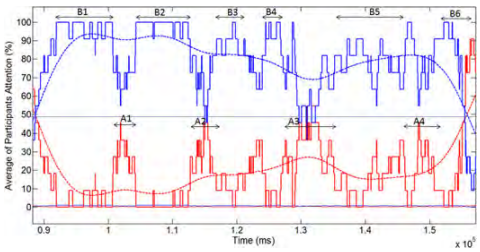


Fig. 10. Participants gaze shift between person A (in Section A1–A4) and person B (in Section B1–B6) in the video. Analyzing the corresponding videos demonstrates that peak points are associated with the verbal/nonverbal cues that person A and person B performed.

TABLE III
SOCIAL CUES IDENTIFIED IN THE AVERAGE GAZE PATTERN AND THEIR ASSOCIATED PEAK NUMBERS

Social Cue	Peak Number
Entering	1,2,14,16,17,25
Speaking	3,8,9,14,16,18,19,22,24,26,27,28,29
Leaving	12,15,21,22,29,31

TABLE IV
ANALYSIS OF SACCADIC GAZE BEHAVIOR

Social Cue	Section
Facial Expression	A,B,E,F,K
Body Gesture	C,D,G,H,M,N

To identify social cues that cause a saccadic gaze shift, the nonfiltered data were analyzed. Fig. 10 details the saccadic gaze shift during section D in Fig. 9. While almost all participants were attracted by person B (Sections B1– B6), some of them had several quick and short gaze shifts to person A (A1–A4). In addition to the entering, speaking, and leaving of social cues, analyzing the nonfiltered average gaze behavior of other cues which trigger saccadic gaze shift was conducted. Facial expressions, body gestures, and hand motions were selected as saccadic triggering cues. Saccadic triggering cues are reported together with the associated log segment in Table IV.

F. Gaze-Control System Parameter Estimation and Priorities Features

Gaze data analysis results showed that there is no generic gaze behavior that can be used to implement a unique model as a standard for *humanlike* gaze patterns. Especially in cases where speaking and hand or body motions occurred at the same time, participants demonstrated different gaze behaviors. However, through the analysis the maximum peaks in saccadic and

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE V
VERBAL AND NONVERBAL CUES IDENTIFIED AS ATTENTION TRIGGERS AND THEIR ASSOCIATED GAZE-CONTROL SYSTEM WEIGHT CALCULATED ON THE BASIS OF HUMAN-OBSERVED PRIORITIES

Priority	Social Cue	Assigned GCS Weight
1 (highest priority)	Entering	100
2	Speaking	100
3	Hand motion/body gesture	65
4	Leaving	55
5 (lowest priority)	Facial expressions	45

nonsaccadic gaze shifts of participants, the priorities of verbal and nonverbal cues in attracting human attention were estimated. The strongest cue attracting the attention of all participants was the new entry (person A/person B) who joined the interaction. Even when person A was speaking with a participant and at the same time person B arrived, all participants were attracted by the new entry (person B). Thus, the highest priority must be given to the new entry. The second priority is given to the auditory signals (speaker). Once person A/B started speaking, all participants were attracted by the speaker. It should be noted that if one person showed body gesture while another one was speaking, most participants were distracted by the body gesture for a very short time and gazed back to the speaker quickly, which shows the higher importance of auditory signals. In addition, a few participants were attracted by the speaker all the time and ignored the body gesture/hand motion of the other person. Therefore, the third priority goes to body gesture/hand motion. The last two priorities are given to the person leaving and facial expressions, respectively, which attracted less attention compared with other cues.

The identification of a set of parameters enables the GCS to generate in the FACE robot a similar gaze to that observed in humans; both saccadic and nonsaccadic movements triggering cues were considered. The weight parameter W_i of the GCS introduced in (2) was calculated considering a maximum value of 100, on the basis of the identified empirical priority order extracted, by analyzing the maximum peaks for each cue (see Fig. 10) during the video. The priority order and the assigned GCS weight are reported in Table V. In addition to Table V, we set a distinction factor for those features that cause a saccadic gaze shift. This enables the GCS to have both saccadic and nonsaccadic gaze behavior.

G. Human and Gaze-Control System-Generated Gaze Comparison

Fig. 11 compares the average gaze of participants with the GCS-generated pattern. The upper image shows the attention on person A; the lower image shows the attention on person B. The graph shows that the system follows the human gaze behavior for the entire duration of the video. The mean error shows that the system is able to replicate the average human gaze behavior with a replication factor of 89.4% throughout the video. When considering saccadic eyes movements [red continuous line in Fig. 8(a) and (b)], the accuracy rate of the GCS decreases to 75.2%, which is likely due to limitations in sensor detection range and speed in comparison with the human eye.

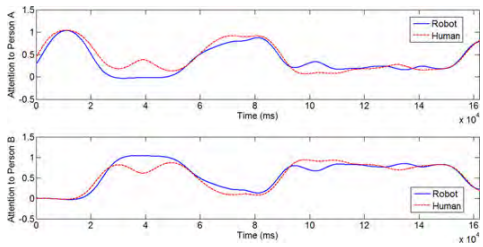


Fig. 11. Comparison of human and robot gaze behavior.

VI. DISCUSSION

In this paper, a context-dependent social GCS which enables the social humanoid robot FACE to display humanlike gaze behavior has been presented. The underlying attention mechanism of the implemented GCS used high-level social features, such as nonverbal and verbal cues, proxemics, an effective visual FOV, and the habituation effect, to determine where the robot should direct its attention.

Experimental results showed that the GCS is able to replicate average human behavior for both nonsaccadic (89.4% accuracy) and saccadic (75.2% accuracy) movements. The lower accuracy in the case of saccadic movement replication may be because of several points.

Diversity in human gaze behavior: Individual human gaze behavior is correlated especially in saccadic movements to factors such as personality, age, and gender [37]. Thus, gaze behavior is different from person to person. Our model only replicates the average gaze behavior of the participants in our experiment. These personal differences are not replicated in the GCS because of the extraction of average-based parameters. However, these differences are common in humans and consequently not perceived as being strange but more as personal and unique peculiarities.

Limitations of the input sensor used: Compared with the human eye, the Kinect sensor has a narrower FOV and a much lower resolution, which affects the attention computation of the GCS. The most influencing sensor limitation was probably the sensor range of the Kinect, which is between 800 and 4000 mm. Humans are able to see much further. Thus, the experiment participants were able to detect people who entered the room shown in the video when they were in their public space (see Fig. 2). The maximum sensor range of Kinect is similar to that of a human's social space.

Unmodeled human attention features: Although our human attention model already considers many features that guide human attention, there are still other unknown factors that we did not use in our model. For example, Fig. 9 shows the participants looking at the environment over time. In addition, there are further external features that guide the attention selection mechanism which we did not include in the current implementation. For example, taking into account the auditory information that comes from outside visual FOV and considering the intentions

of people during a social interaction could help the attention mechanism to generate a more natural humanlike social gaze behavior.

However, as shown in Fig. 11, the proposed implementation of the GCS is able to select the appropriate gaze target points at the right time, which is essential for the development of believable social robots.

ACKNOWLEDGMENT

The authors would like to thank A. Haslbeck and the staff of the Institute of Ergonomics at the Technical University of Munich, Germany, for providing the eye-tracking system, and M. B. Dehkordi for helping us in performing experiments.

REFERENCES

[1] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robot. Auton. Syst.*, vol. 42, no. 3, pp. 143–166, 2003.
[2] D. Mazzei, N. Lazzeri, L. Billeci, R. Igliozzi, A. Mancini, A. Ahluwalia, F. Muratori, and D. De Rossi, "Development and evaluation of a social robot platform for therapy in autism," *Proc. IEEE Eng. Med. Biol. Soc.*, pp. 4515–4518, Aug. 2011.
[3] D. Mazzei, N. Lazzeri, D. Hanson, and D. De Rossi, "Hefes: An hybrid engine for facial expressions synthesis to control human-like androids and avatars," *IEEE RAS EMBS Int. Conf. Biomed. Robots Biomechatron.*, pp. 195–200, Jun. 2012.
[4] A. L. Rothenstein and J. K. Tsotsos, "Attention links sensing to recognition," *Image Vis. Comput.*, vol. 26, no. 1, pp. 114–126, Jan. 2008.
[5] L. Itti and C. Koch, "Computational modeling of visual attention," *Nat. Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.
[6] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annu. Rev. Neurosci.*, vol. 18, no. 1, pp. 193–222, Mar. 1995.
[7] K. Lamberts and R. Goldstone, *Handbook of Cognition*. Newbury Park, CA, USA: Sage, 2004.
[8] R. L. Canosa, "Real-world vision: Selective perception and task," *ACM Trans. Appl. Perception*, vol. 6, no. 2, p. 11, 2009.
[9] F. Baluch and L. Itti, "Mechanisms of top-down attention," *Trends Neurosci.*, vol. 34, no. 4, pp. 210–224, Apr. 2011.
[10] M. Carrasco, "Visual attention: The past 25 years," *Vis. Res.*, vol. 51, no. 13, pp. 1484–1525, Jul. 2011.
[11] D. Hanson, A. Olney, S. Prilliman, E. Mathews, M. Zielke, D. Hammons, R. Fernandez, and H. Stephanou, "Upending the uncanny valley," in *Proc. Nat. Conf. Artif. Intell.*, 2005, vol. 20, no. 4, pp. 1728–1729.
[12] D. Hanson, "Expanding the aesthetic possibilities for humanoid robots," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2005. Availble: http://androidscience.com/Publications/HansonUncannyIEEE_2005reduced.pdf
[13] D. F. Hanson, G. Pioggia, Y. Bar-Cohen, and D. De Rossi, "Androids: Application of EAP as artificial muscles to entertainment industry," in *Proc. SPIE 8th Annu. Int. Symp. Smart Struct. Mater.*, 2001, pp. 375–379.
[14] D. Mazzei, L. Billeci, A. Armato, N. Lazzeri, A. Cisternino, G. Pioggia, R. Igliozzi, F. Muratori, A. Ahluwalia, and D. De Rossi, "The face of autism," in *Proc. IEEE Int. Workshop Robots Human Interactive Commun.*, 2010, pp. 791–796.
[15] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proc. SPIE 48th Annu. Meet. Opt. Sci. Technol.*, 2004, pp. 64–78.
[16] T. J. Buschman and E. K. Miller, "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices," *Science*, vol. 315, no. 5820, pp. 1860–1862, 2007.
[17] S. E. Baron-Cohen, H. E. Tager-Flusberg, and D. J. Cohen, *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience*. London, U.K.: Oxford Univ. Press, 2000.
[18] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, pp. 97–136, 1980.
[19] J. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychonomic Bulletin Rev.*, vol. 1, no. 2, pp. 202–238, 1994.
[20] J. Wolfe and T. Horowitz, "What attributes guide the deployment of visual attention and how do they do it," *Nat. Rev. Neurosci.*, vol. 5, no. 6, pp. 495–501, Jun. 2004.
[21] J. M. Wolfe, *Guided Search 4.0.*. London, U.K.: Oxford Univ. Press, May 2007, pp. 99–119.
[22] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati, "Active vision for sociable robots," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 31, no. 5, pp. 443–453, Sep. 2001.
[23] G. Schillaci, S. Bodiroa, and V. V. Hafner, "Evaluating the effect of saliency detection and attention manipulation in human–robot interaction," *Int. J. Soc. Robot.*, vol. 5, no. 1, pp. 139–152, Jan. 2013.
[24] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," *Adv. Neural Inf. Process. Syst.*, vol. 20, pp. 241–248, 2008.
[25] A. Maki, P. Nordlund, and J.-O. Eklundh, "Attentional scene segmentation: Integrating depth and motion," *Comput. Vis. Image Understand.*, vol. 78, no. 3, pp. 351–373, Jun. 2000.
[26] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.
[27] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *J. Vis.*, vol. 11, no. 5, pp. 1–23, May 2011.
[28] B. Mutlu, J. Forlizzi, and J. Hodgins, "A storytelling robot: Modeling and evaluation of human-like gaze behavior," *Proc. IEEE Int. Conf. Humanoid Robots*, pp. 518–523, Dec. 2006.
[29] J. G. Trafton, M. D. Bugajska, B. R. Fransen, and R. M. Ratwani, "Integrating vision and audition within a cognitive architecture to track conversations," in *Proc. 3rd ACM/IEEE Int. Conf. Human Robot Interaction*, 2008, pp. 201–208.
[30] P. Holthaus, K. Pitsch, and S. Wachsmuth, "How can I help? Spatial attention strategies for a receptionist robot," *Int. J. Soc. Robot.*, vol. 3, no. 4, pp. 383–393, Nov. 2011.
[31] J. Goldring, M. Dorris, B. Corneil, P. Ballantyne, and D. Munoz, "Combined eye-head gaze shifts to visual and auditory targets in humans," *Exp. Brain Res.*, vol. 111, no. 1, pp. 68–78, Sep. 1996.
[32] S. Andrist, T. Pejsa, B. Mutlu, and M. Gleicher, "Designing effective gaze mechanisms for virtual agents," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 705–714.
[33] L. Itti, N. Dhavale, and F. Pighin, "Photorealistic attention-based gaze animation," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 521–524.
[34] J. Ruesch and W. Kees, *Nonverbal Communication: Notes on the Visual Perception of Human Relationships*. Berkeley, CA, USA: University of California Press, 1956.
[35] M. Argyle, *Bodily Communication*, 2nd ed. Evanston, IL, USA: Routledge, 1988.
[36] M. Argyle and J. Dean, "Eye-contact, distance and affiliation," *Sociometry*, vol. 28, no. 3, pp. 289–304, Sep. 1965.
[37] M. Argyle and M. Cook, "Gaze and mutual gaze." Cambridge, U.K. Cambridge Univ. Press, 1976.
[38] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt, *Eye Gaze Patterns in Conversations, There is More to Conversational Agents Than Meets the Eyes*. New York, NY, USA: ACM Press, 2001, pp. 301–308.
[39] E. T. Hall, *The Hidden Dimension*. New York, NY, USA: Anchor Books, 1969.
[40] B. W. Tatler, R. J. Baddeley, and B. T. Vincent, "The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task," *Vis. Res.*, vol. 46, no. 12, pp. 1857–1862, Jun. 2006.
[41] A. T. Bahill, D. Adler, and L. Stark, "Most naturally occurring human saccades have magnitudes of 15 degrees or less," *Investigative Ophthalmol. Vis. Sci.*, vol. 14, pp. 468–469, 1975.
[42] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, pp. 4–4, Nov. 2007.
[43] B. T. Vincent, R. Baddeley, A. Correani, T. Troscianko, and U. Leonards, "Do we look at lights? Using mixture modelling to distinguish between low- and high-level factors in natural image viewing," *Vis. Cogn.*, vol. 17, no. 6–7, pp. 856–879, Aug. 2009.
[44] R. F. Thompson and W. A. Spencer, "Habituation: A model phenomenon for the study of neuronal substrates of behavior," *Psychol. Rev.*, vol. 73, no. 1, pp. 16–43, 1966.
[45] G. Metta, P. Fitzpatrick, and L. Natale, "YARP: Yet another robot platform," *Int. J. Adv. Robot. Syst.*, vol. 3, no. 1, pp. 43–48, 2006.
[46] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," *Proc. 16th Int. Joint Conf. Artif. Intell.*, vol. 255, pp. 1146–1153, 1999.
[47] R. A. Hinde, *Non-verbal Communication*. Cambridge, U.K.: Cambridge Univ. Press, 1975.

[48] C. Küblbeck and A. Ernst, "Face detection and tracking in video sequences using the modifiedcensus transformation," *Image Vis. Comput.*, vol. 24, no. 6, pp. 564–572, Jun. 2006.
[49] F. IIS. (2009, Jun.). Cognitive systems. [Online]. Available: http://www.iis.fraunhofer.de/en/bf/bsy/fue/isyst.html/
[50] M. Turk and A. Pentland, *Face Recognition Using Eigenfaces.*. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1991, pp. 586–591.
[51] Kendon, *Gesture: Visible Action as Utterance*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[52] W. H. Zangemeister and L. Stark, "Gaze latency: Variable interactions of head and eye latency," *Exp. Neurol.*, vol. 75, no. 2, pp. 389–406, Feb. 1982.
[53] D. Hanson and Y. Bar-Cohen, New York, NY, USA: Springer, 2009
[54] D. F. Hanson and V. White, "Converging the capabilities of EAP artificial muscles and the requirements of bio-inspired robotics," *Proc. SPIE*, vol. 5385, pp. 29–40, Jul. 2004.
[55] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: A professional framework for multimodality research," in *Proc. Language Resources and Evaluation*, 2006, vol. 2006.

Abolfazl Zaraki received the B.S. degree in electronics and electrical engineering from the University of Dezful, Dezful, Iran, in 2006. He received the Master's degree in engineering in mechatronics and automatic control from the University Technology of Malaysia, Skudai, Malaysia, in 2010. He has been working toward the Ph.D. degree in automatic robotics and bioengineering at the Research Center "E. Piaggio" University of Pisa, Pisa, Italy, since January 2011.

His current work focuses on developing a multimodal context-aware attention system for humanoid robots which regulates the dynamic attention of the robot, according to the salience of low-level visual features and high-level human features presented in the visual-auditory scene. His research interest includes the design and development of a system that controls the social behavior of a humanoid robot, in a dynamic human–robot interaction.

Daniele Mazzei received the Master's degree in biomedical engineering and the Ph.D. degree in automatic robotic and bioengineering from the University of Pisa, Pisa, Italy, in October 2006 and May 2010, respectively.

He is a Postdoctoral Researcher at the Research Center "E. Piaggio," University of Pisa. He is the Coordinator of the FACE Team at Research Center "E. Piaggio" which is developing a social robot cognitive systems for human–robot interaction studies. He is also developing a novel generation of social objects for the new paradigm of the Internet of Things. His current research interests include social robotic and human robot empathic interaction.

Manuel Giuliani received the Master's of Arts degree in computational linguistics from the Ludwig-Maximilian-University in Munich, Munich, Germany, in 2004, the Master's of Science and Ph.D. degrees in computer science from the Technical University Munich, Munich, in 2006 and 2011, respectively.

He worked on the European project JAST (Joint Action Science and Technology), the DFG-funded project AudiComm, which was part of the cluster of excellence "Cognition for Technical Systems," and is now part of the European project JAMES (Joint Action for Multimodal Embodied Social Systems). His research interests include social robotics, human–robot interaction, natural language processing, multimodal fusion, and robot architectures.

Danilo De Rossi received the graduation degree in chemical engineering from the University of Genoa, Genova, Italy, in 1976.

He has had teaching and research posts in Australia, Brazil, France, Japan, and USA. He joined the Faculty of Engineering in the University of Pisa, Pisa, Italy, in 1982. He is currently a Full Professor of Bioengineering. His scientific activities are related to the physics of organic and polymeric materials, and to the design of sensors and actuators for bioengineering and robotics. He is the author of more than 270 peer reviewed papers in international science journals and peer reviewed proceedings and is the coinventor of 14 patents, and the coauthor of eight books.

Appendix B

Appendix B - Publication: an Hybrid Engine for Facial Expressions Synthesis to control humanlike androids and avatars

HEFES: an Hybrid Engine for Facial Expressions Synthesis to control human-like androids and avatars

Daniele Mazzei, Nicole Lazzeri, David Hanson and Danilo De Rossi

***Abstract*— Nowadays advances in robotics and computer science have made possible the development of sociable and attractive robots. A challenging objective of the field of humanoid robotics is to make robots able to interact with people in a believable way. Recent studies have demonstrated that human-like robots with high similarity to human beings do not generate the sense of unease that is typically associated to human-like robots. For this reason designing of aesthetically appealing and socially attractive robots becomes necessary for realistic human-robot interactions.**

In this paper HEFES (Hybrid Engine for Facial Expressions Synthesis), an engine for generating and controlling facial expressions both on physical androids and 3D avatars is described. HEFES is part of a software library that controls a human robot called FACE (Facial Automaton for Conveying Emotions). HEFES was designed to allow users to create facial expressions without requiring artistic or animatronics skills and it is able to animate both FACE and its 3D replica.

The system was tested in human-robot interaction studies aimed to help children with autism to interpret their interlocutors' mood through facial expressions understanding.

I. INTRODUCTION

In the last years, more and more social robots have been developed due to rapid advances in hardware performance, computer graphics, robotics technology and Artificial Intelligence (AI).

There are various examples of social robots but it is possible to roughly classify them according to their aspect in two main categories: human-like and not human-like. Human-like social robots are usually associated to the pernicious myth that robots should not look or act like human beings in order to avoid the so-called 'Uncanny Valley' [1]. MacDorman and Ishiguro [2] explored observers' reactions to gradual morphing of robots and humans pictures and found a peak in judgments of the eeriness in the transition between robot and human-like robot pictures according to the Uncanny Valley hypothesis. Hanson [3] repeated this experiment morphing more attractive pictures and found that the peak of eeriness was much smoother, approaching to a flat line, in the transition between human-like robot and human beings pictures. This indicates that typical reactions due to the Uncanny Valley were present only in the transition between classic robots and cosmetically atypical human-like robots. Although more studies demonstrate the presence of the Uncanny Valley effect, it is possible to design and create human-like robots that are not uncanny using innovative technologies that integrate movies and cinema animation with make-up techniques [4].

The enhancement of the believability of human-like robots is not a pure aesthetic challenge. In order to create machines that look and act as humans, it is necessary to improve the robot's social and expressive capabilities in addition to the appearance. Therefore, facial expressiveness is one of the most important aspect to be analyzed in designing human-like robots since it is the major emotional communication channel used in interpersonal relationships together with facial and head micro movements [5].

Since the early 70's, facial synthesis and animation have raised a great interest among computer graphics researchers and numerous methods for modeling and animating human faces have been developed to reach more and more realistic results.

One of the first models for the synthesis of faces was developed by Parke [6], [7]. The Parke parametric model is based on two groups of parameters: conformation parameters which are related to the physical facial features, such as the shape of the mouth, nose, eyes, etc., and expression parameters which are related to facial actions such as wrinkling the forehand for anger or open the eyes wide for surprise.

Differently, physically-based models manipulate directly the geometry of the face to approximate real deformations caused by the muscles including skin layers and bones. Waters [8], using vectors and radial functions, developed a parameterized model based on facial muscles dynamic and skin elasticity.

Another approach used for creating facial expressions is based on interpolation methods. Interpolation-based engines use a mathematical function to specify smooth transitions between two or more basic facial positions in a defined time interval [9]. One, two or three-dimensional interpolations can be performed to create an optimized and realistic facial morphing. Although interpolations are fast methods, they are limited in the number of realistic facial configurations they can generate.

All geometrically-based methods described above can generate difficulties in achieving realistic facial animations since they require artistic skills. On the other hand, animation skills are required only for creating a set of basic facial configurations since an interpolation space can be use to generate a wide set of new facial configurations starting from the basic ones.

In this work a facial animation engine called HEFES was implemented as fusion of a muscle-based facial animator and an intuitive interpolation system. The facial animation

Daniele Mazzei, Nicole Lazzeri and Danilo De Rossi are with Interdepartmental Research Center 'E. Piaggio', Faculty of Engineering - University of Pisa, Via Diotisalvi 2, 56126 Pisa, Italy. (mazzei@di.unipi.it)

David Hanson is with Hanson Robotics, Plano Tx, USA. (david@hansonrobotics.com)

system is based on the Facial Action Coding System (FACS) in order to make it compatible with both physical robots and 3D avatars and usable in different facial animation scenarios. The FACS is the most popular standard for describing facial behaviors in terms of muscular movements. The FACS is based on a detailed study of the facial muscles carried out by Ekman and Friesen in 1976 [10] and is aimed at classifying the facial muscular activity according to Action Units (AUs). AUs are defined as visually discernible component of facial movements which are generated through one or more underlying muscles. AUs can be used to describe all the possible movements that a human face can express. Therefore an expression is a combination of several AUs, each of them with their own intensity measured in 5 discrete levels (A:Trace, B:Slight, C:Marked pronounced, D:Severe, E:Extreme maximum).

II. Materials and Methods

A. FACE

FACE is a robotic face used as emotions conveying system (Fig. 1). The artificial skull is covered by a porous elastomer material called Frubber™ that requires less force to be stretched by servo motors than other solid materials [11]. FACE has 32 servo motors actuated degrees of freedom which are mapped on the major facial muscles to allow FACE to simulate facial expressions.

Fig. 1. FACE and the motor actuation system

FACE servo motors are positioned following the AUs disposition according to the FACS (Fig. 2) and its facial expressions consist of a combination of many AUs positions. Thanks to the fast response of the servo motors and the mechanical properties of the skin, FACE can generate realistic human expressions involving people in social interactions.

B. SYSTEM ARCHITECTURE

HEFES is a subsystem of the FACE control library deputed to the synthesis and animation of facial expressions and includes a set of tools for controlling FACE and its 3D avatar. HEFES includes four modules: synthesis, morphing, animation and display. The synthesis module is designed to allow

Fig. 2. Mapping between servo motors positions and Action Units of FACS

users to manually create basic facial expressions that are normalized and converted according to the FACS standard. The morphing module takes the normalized FACS-based expressions as input and generates an emotional interpolation space where expressions can be selected. The animation module merges concurrent requests from various control subsystems and creates a unique motion request resolving possible conflicts. Finally, the display module receives the facial motion request and converts it in movements according to the selected output display.

1) The synthesis module allows users to generate new facial expressions through the control of the selected emotional display, i.e. FACE robot or 3D avatar. Both modules provide a graphical user interface (GUI) with as many slider controls as the number of servo motors (FACE robot) or anchor points (3D avatar) which are present in the corresponding emotional display.

In the Robot editor, each slider defines a normalized range between 0 and 1 for moving the corresponding servo motor which is associated to an AU of the FACS. Using the Robot editor, the six basic expressions, i.e. happiness, sadness, anger, surprise, fear and disgust, defined as 'universally accepted' by Paul Ekman [12], [13], were manually created. According to the "Circumplex Model of Affect" theory [14], [15], each generated expression was saved as an XML file including the set of the AUs values, the expression name and the corresponding coordinates in terms of Pleasure and arousal. In the Circumplex Model of Affect expressions are associated with Pleasure that indicates the pleasant/unpleasant feelings and with Arousal which is related to a physiological activation.

The 3D virtual editor is a similar tool used to deform a facial mesh. The 3D editor is based on a user interface on which a set of slider controls is used to actuate various facial muscles. Expressions are directly rendered on the 3D avatar display and saved as XML files as in the Robot Editor.

2) The morphing module generates, on the base of the Posner's theory, an emotional interpolation space, called Emotional Cartesian Space (ECS) [16]. In the ECS the x coordinate represents the valence and the y coordinate represents the arousal. Each expression e(v, a) is consequently

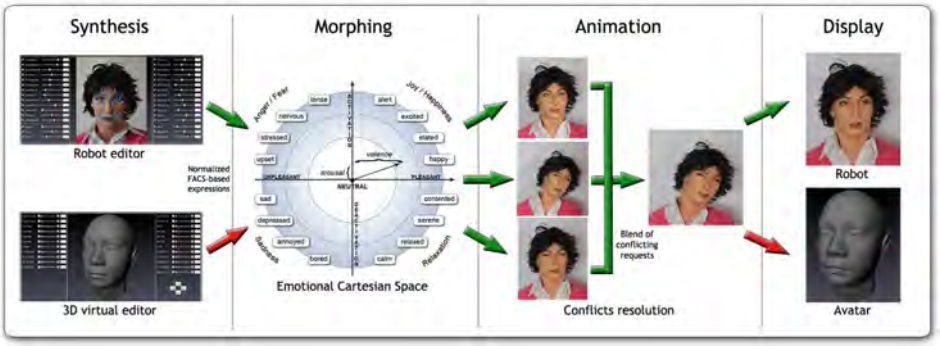


Fig. 3. The architecture of the facial animation system based on four main modules: synthesis, morphing, animation and display.

associated with a point in the valence-arousal plane where the neutral expression e(0, 0) is placed in the origin (Fig. 3, Morphing module). The morphing module takes the set of basic expressions as input and generates the ECS applying a shape-preserving piecewise cubic interpolation algorithm implemented in Matlab™. The output of the algorithm is a three-dimensional matrix composed of 32 planes corresponding to the 32 AUs. As shown in Fig. 4, each plane represents the space of the possible positions of a single AU where each point is identified by two the coordinates, valence and arousal. The coordinates of each plane range between -1 and 1 with a step of 0.1 therefore the generated ECS produces 21x21 new normalized FACS-based expressions that can be performed by the robot or the 3D avatar independently. Since the ECS is not a static space, each new expression manually created through the synthesis module can be used to refine the ECS including it in the set of expressions used by the interpolation algorithm. The possibility of updating the ECS with additional expressions allows users to continuously adjust the ECS covering zones in which the interpolation algorithm could require a more detailed description of the AUs (II-B.1).

3) The animation module is designed to combine and merge multiple requests coming from various modules which can run in parallel in the robot/avatar control library. The facial behavior of the robot or avatar is inherently concurrent since parallel requests could interest the same AU generating conflicts. Therefore the animation module is responsible for mixing movements, such as eye blinking or head turning, with requests of expressions. For example, eye blinking conflicts with the expression of amazement since normally amazed people react opening the eyes wide.

The animation module receives as input a motion request, which is defined by a single AU or a combination of multiple AUs, with an associated priority. The animation engine is implemented as a Heap, a specialized tree-based data structure used to define a shared timer that is responsible for orchestrating the animation. The elements of the Heap,

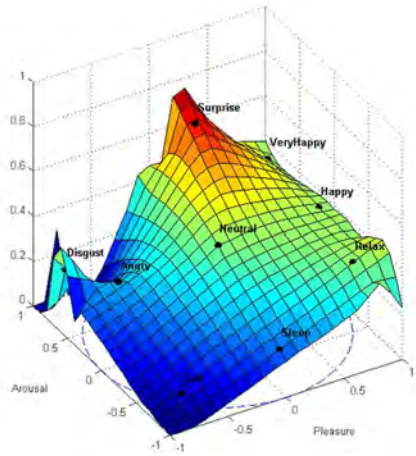


Fig. 4. The emotional Cartesian plane for the right eyebrow (motor #24 corresponding to the AU 1 in Fig. 2).

called Tasks, are ordered by their due time therefore the root of the Heap contains the first task to be executed. In the Heap there can be two types of tasks, Motion Task and Interpolator Task, that are handled in a different way. Both types of tasks are defined by the expiring time, the duration of the motion and the number of steps in which the task will be divided. A Motion Task also includes 32 AUs, each of them with their associated values and a priority. When a movement request is generated, a Motion Task is sent to the Animation Engine and inserted into the Heap which will be reordered according to the due time. The animation engine is always running to check whether some tasks into the Heap are expired. For each

expired task, the animation engine removes it from the Heap and executes it. If the task is a Motion Task, the animation engine calculates the amount of movement to be performed at the current step, stores the result in correspondence to the relative AU and reschedules the task into the Heap if the task is not completed. If the task is an Interpolation Task, the animation engine calculates the new animation state by blending all the steps, previously calculated, for each AUs according to their priority. At the end, the Interpolator Task is automatically rescheduled into the Heap with an expiring time of 40ms.

The output of the animation module is a motion task composed of 32 normalized AUs values that is sent to the emotional display module.

4) The display module represents the output of the system. We implemented two dedicated emotional displays: the FACE android and the 3D avatar. According to a calibration table, the FACE android display converts normalized AUs values into servo motor positions that are expressed as duty cycles in the range 500-2500. Each motor has a different range of movements due to its position inside the FACE. For this reason, the display module includes a control layer to avoid the exceeding the servo motor limits according to minimum and maximum values stored in the calibration tables.

The 3D avatar display is a facial animation system based on a physical model described in [17] that approximates the anatomy of the skin and the muscles. The model is based on a non-linear spring system which can simulate the dynamics of human face movements while the muscles are modeled as mesh of force deformed springs. Each skin point of the mesh is connected with its neighbors by non-linear springs. Human face includes a wide range of muscles types, e.g. rectangular, triangular, sheet, linear, sphincter. Since servo motors act as linear forces, the type of muscle satisfying this condition is the linear muscle that is specified by two points: the attachment point which is normally fixed and the insertion point which defines the area where the facial muscle performs its action. Facial muscle contractions pull the skin surface from the area of the muscle insertion point to the area of the muscle attachment point. When a facial muscle contracts, the facial skin points in the influence area of the muscle change their position according to the distance from the muscle attachment point and the elastic properties of the mass-spring system. Facial skin points not directly influenced by the muscle contraction are in a sort of unbalanced state that is stabilized through propagation of other unbalanced elastic forces.

The elastic model of the skin and the mathematical implementation of the muscles have been already developed while the manual mapping of the 3D mesh anchor points to AUs is still under development.

C. ANIMATION TOOL

Generally facial animation softwares are tools that require a certain level of knowledge in design, animation and anatomy. Often users only need to easily animate facial expressions without having these specific skills. Therefore the system was designed to be used both by experts in facial design and animation which can directly create or modify expressions and users that are interested in quickly designing HRI experimental protocols selecting a set of pre-configured expressions.

The ECS Timeline is a tool of the HEFES system that is intended to meet the needs of different users. The timeline is a Graphical User Interface (GUI) with two use modalities: ”Auto Mode” and ”Advanced Mode”. In Auto Mode, users can create sequences of expressions selecting the corresponding points in the ECS and dragging them into the timeline. Sequences can be saved, played and edited using the timeline control. When a sequence is reproduced, motion requests are sent to the animation module that resolves conflicts and forwards them to the robot or the avatar display. The ECS Timeline GUI includes a chart that visualizes the motors positions during an animation for a deeper understanding of the facial expression animation process (Fig. 5). In Advanced Mode, a sequence of expressions can be displayed as editable configurations of all AUs values in a multitrack graph where each AU is expressed as a motion track and can be manually edited. In the Advanced Mode is possible to use ECS expressions as starting point for creating more sophisticated animations in which single AUs can be adjusted in real-time.

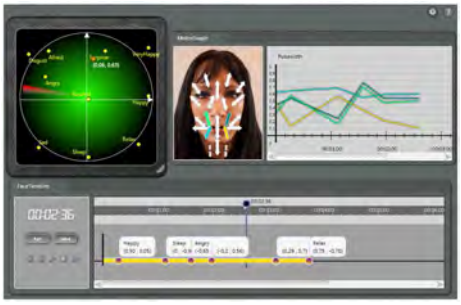


Fig. 5. The ECS Animation in the Auto Mode configuation.

III. RESULTS AND DISCUSSION

HEFES was used as emotions conveying system within the IDIA (Inquiry into Disruption of Intersubjective equipment in Autism spectrum disorders in childhood) project in collaboration with the IRCCS Stella Maris (Calambrone, Italy) [16], [18].

In particular, the ECS Animation tool was used by the psychologist in Auto Mode to easily design the therapeutic protocol creating facial animation paths without require FACE android direct motor configuration and calibration. The tool does not required skills in facial animation and human anatomy and allowed therapist to intuitively create therapeutic scenarios adding expressions to the timeline dragging them from the ECS. Moreover the Manual Mode

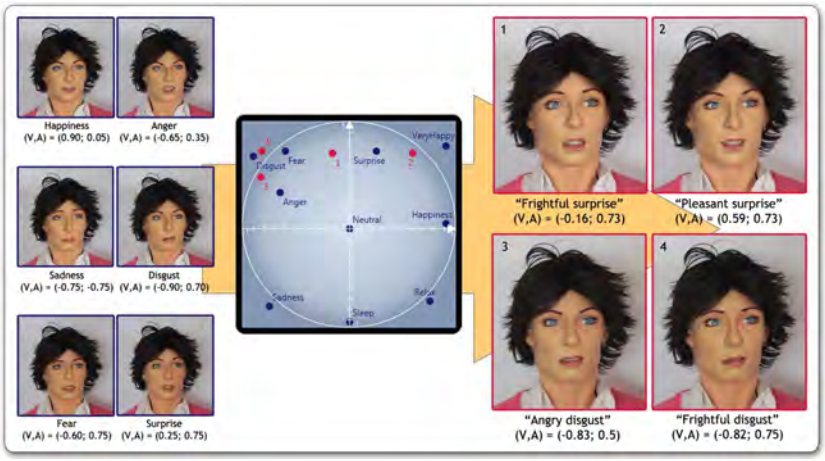


Fig. 6. The morphing module used for creating new 'mixed' expressions (right side) selecting (V,A) points (red dots) from the ECS. The module takes in input a set of basic expressions (left side) with their (V,A) values (blue dots).

configuration was used to create specific patterns of movements such as the turning of the head. Head movements was oriented to watch a little robot used by the therapist to test children's shared attention capabilities.

Recent study demonstrated that people with Autism Spectrum Disorders (ASDs) do not perceive robots as machine but as "artificial partners" [19]. On the base of this theory the IDIA project aimed to the study of alternative ASD treatment protocol involving robots, avatars and other advanced technologies. One of the purposes of the protocol was to verify the capability of the FACE android to convey emotions to children with ASD. Figure 6 shows examples of expressions generated by the morphing module. It takes the six basic expressions as input (expressions on the left side of the figure corresponding to the blue dots in the ECS) and generates 'half-way' expressions (right side of the figure corresponding to the red dots in the ECS) by clicking on the ECS. All these generated expressions are identified by their corresponding pleasure and arousal coordinates.

FACE base protocol was tested on a panel of normally developing children and children with Autism Spectrum Disorders (ASDs) (aged 6-12 years).

The test was conducted on a panel of 5 children with ADSs and 15 normally developing interacting with the robot individually under therapist supervision. The protocol was divided in phases and one of these concerned evaluating the accuracy of emotional recognition and imitation skills. In this phase children were asked to recognize, label and then imitate a set of facial expressions performed by the robot and subsequently by the psychologist. The sequence of expressions included happiness, anger, sadness, disgust, fear and surprise. Moreover, the protocol included a phase called "free play" where the ECS tool was directly used by the psychologist to control the FACE android in real-time.

The subjects' answers in labeling an expression were scored as correct or wrong by a therapist and used for calculating the percentage of correct expressions recognition. As shown in Fig. 7 both children with ASDs and normally developing children were able to label Happiness, Anger and Sadness performed by FACE and by the psychologist without errors. Otherwise Fear, Disgust and Surprise performed by FACE and by the psychologist have not been labeled correctly, especially by subjects with ASDs. Fear, Disgust and Surprise are emotions which convey empathy not only through stereotypical facial expressions but also with body movements and vocalizations. The affective content of this emotions is consequently dramatically reduced if expressed only through facial expressions.

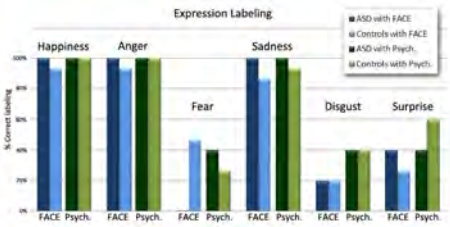


Fig. 7. Results of the labeling phase for ASD and control subjects observing FACE and psychologist expressions.

In conclusion HEFES allows operators and psychologists to easily model and generate expressions following the

current standards of facial animations. The morphing module provides a continuous emotional space where it is possible to select a wide range of expressions, most of them difficult to be manually generated. The possibility to continuously add new expressions to the ECS interpolator allows users to refine the expressions generation system for reaching a high expressiveness level without requiring animation or artistic skills.

Through HEFES is possible to control robot or avatar creating affective based human-robot interaction scenarios on which different emotions can be conveyed. Facial expressions performed by FACE and by the psychologist have been labeled by children with ASDs and normally developed children with the same score. This analysis demonstrates that the system is able to correctly generate human-like facial expressions.

IV. FUTURE WORKS

HEFES was designed to be used both with a physical robot and with a 3D avatar. The actual state of the 3D editor includes the algorithm to animate the facial mesh according to the model described in Sec. II and the definition of some anchor points. In future all the AUs will be mapped on the 3D avatar mesh for a complete control of the avatar. HEFES will be used to study how human beings perceive facial expressions and emotion expressed by a physical robot in comparison with its 3D avatar for understanding if the physical appearance has an emphatic component in conveying emotions.

Moreover the synthesis module will include the control of facial micro movements and head dynamics that are associated with human moods. For example, blinking frequency and head speed are considered to be indicators of discomfort. These micro movements will be designed and controlled using an approach similar to the one designed for facial expressions. A set of basic head and facial micro movements will be generated and associated with corresponding behaviors according to their pleasure and arousal coordinates. The set of basic behaviors will be used as input of the morphing module which will generate a Behavioral Cartesian Space (BCS). Future experiment on emotion labeling and recognition will be conducted including the facial micro movement generator and a face tracking algorithm in order to investigate the contribute of this affective related activities on emotions conveying FACE capabilities.

REFERENCES

[1] M. Mori, "Bukimi no tani (the uncanny valley)," *Energy*, 1970.

[2] K. F. MacDorman and H. Ishiguro, "The uncanny advantage of using androids in cognitive and social science research," *Interaction Studies*, vol. 7, no. 3, pp. 297–337, 2006.

[3] D. Hanson, "Exploring the aesthetic range for humanoid robots," in *Proceedings of the ICCS CogSci 2006 Symposium Toward Social Mechanisms of Android Science*. Citeseer, 2006, p. 1620.

[4] H. Ishiguro, "Android science - toward a new cross-interdisciplinary framework," *Development*, vol. 28, pp. 118–127, 2007.

[5] P. Ekman, "Facial expression and emotion," *American Psychologist*, pp. 384–392, 1993.

[6] F. I. Parke, "Computer generated animation of faces," in *ACM '72: Proceedings of the ACM annual conference*. New York, NY, USA: ACM, 1972, pp. 451–457.

[7] F. I. Parke, "A parametric model for human faces," Ph.D. dissertation, The University of Utah, 1974.

[8] K. Waters, "A muscle model for animation three-dimensional facial expression," *SIGGRAPH Computer Graphics*, vol. 21, pp. 17–24, August 1987.

[9] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, "Synthesizing realistic facial expressions from photographs," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '98. New York, NY, USA: ACM, 1998, pp. 75–84.

[10] P. Ekman and W. V. Friesen, "Measuring facial movement," *Journal of Nonverbal Behavior*, vol. 1, no. 1, pp. 56–75, Sep. 1976.

[11] D. Hanson, "Expanding the design domain of humanoid robots," in *Proceedings of ICCS CogSci Conference, special session on Android Science*, 2006.

[12] P. Ekman, "Are there basic emotions?" *Psychological Review*, vol. 99, no. 3, pp. 550–553, Jul 1992.

[13] P. Ekman, *Handbook of Cognition and Emotion: 3 Basic emotions*. New York: John Wiley & Sons Ltd, 1999, ch. 3, pp. 45–60.

[14] J. A. Russell, "The circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[15] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[16] D. Mazzei, L. Billeci, A. Armato, N. Lazzeri, A. Cisternino, G. Pioggia, R. Igliozzi, F. Muratori, A. Ahluwalia, and D. De Rossi, "The face of autism," in *RO-MAN 2009. The 18th IEEE International Symposium on Robot and Human Interactive Communication, 2009*, 2010, pp. 791–796.

[17] Y. Zhang, E. C. Prakash, and E. Sung, "Real-time physically-based facial expression animation using mass-spring system," in *Computer Graphics International 2001*, ser. CGI '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 347–350.

[18] D. Mazzei, N. Lazzeri, L. Billeci, R. Igliozzi, A. Mancini, A. Ahluwalia, F. Muratori, and D. De Rossi, "Development and evaluation of a social robot platform for therapy in autism," in *EMBC 2011. The 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 4515–4518.

[19] J. Scholtz, "Theory and evaluation of human robot interactions," in *Proc. 36th Annual Hawaii Int System Sciences Conf*, 2003.

Appendix C

Appendix C - Dikabilis Eye-tracking Device Data-sheet

Dikablis

World´s fastest and most reliable way of getting eye-tracking results

The only Eye-Tracking System worldwide that offers unlimited mobility via the so called „Inside-out Head-Position-Measurement“ and which combines these features with the automated Processing of the recorded Gaze-Data in any environment.

Ergoneers
Ergonomic Engineers

Technical data

Dikablis Cable

- cable length up to: 50m
- power supply: 12V/cigarette lighter; 230V or battery pack (battery life: 2h)

Dikablis Wireless

- transmission distance: 500m
- weight transmitter: 0,25kg
- weight receiver: 1,5kg
- power supply transmitter: 12V/cigarette lighter; 230V or battery pack (battery life: 2h)
- power supply receiver: 12V/cigarette lighter (adapter); 230V

Dikablis Wireless Plus

- tranmission distance: 5.000m
- weight transmitter: 2,5kg
- weight receiver: 5,5kg
- power supply transmitter: 12V/cigarette lighter; 230V or battery pack (battery life: 2h)
- power supply receiver: 12V/cigarette lighter; 230V

Recording Laptop

- operating system: Microsoft® Windows XP
- processor: Intel Core 2 Duo CPU 2,4 GHz
- working memory: 2 Gb RAM
- recording time: 65h
- weight: 2kg
- power supply: 12V/cigarette lighter (adapter), 230V or battery pack (battery life: 3h)

Head-Unit

- lightweight, comfortable, easy to adjust
- design: also suits people who wear glasses
- field Cam: color or black/white; resolution 380 TVL
- eye Cam: resolution: 380 TVL
- visual range: (can be adjusted by changing the objective):
  - horizontal: from 50° up to 115°
  - vertical: from 40° up to 90°
- gaze position accuracy: 0.5 degrees visual angle
- tracking resolution of pupil: 0,10 degree visual angle
- frequency: PAL (50Hz interlaced)
- head movement: unlimited
- weight: 69g
- power consumption: 320mA
- power supply:
  - 230V
  - 12V/cigarette lighter
  battery pack (battery life: 2h)
- mobility:
  - complete freedom of movement
  - wireless data transfer up to 5.000 meters

Ergoneers GmbH
Mozartstraße 8 ½, D-85077 Manching

Tel. 08459/331364 **Fax.** 08171/965306

www.ergoneers.com
Dikablis@ergoneers.com

Dikablis

World´s fastest and most reliable way of getting eye-tracking results

The only Eye-Tracking System worldwide that offers unlimited mobility via the so called „Inside-out Head-Position-Measurement“ and which combines these features with the automated Processing of the recorded Gaze-Data in any environment.

Ergoneers
Ergonomic Engineers

Data Output

Realtime Output

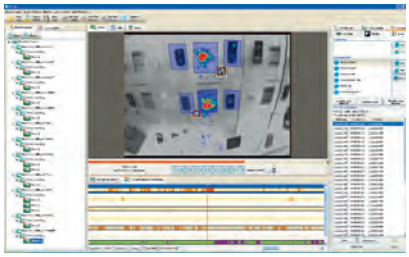
- Video of Scene Camera
- Video of Eye Camera
- Absolute and relative timeline
- Calibration settings
- x- and y- coordinates of the center of the pupil in relation to the zero point of the eye camera
- x- and y- coordinates of the center of the pupil in relation to the zero point of the scene camera (calculated with the calibration settings)
- height, width and size of the pupil
- fixation coordinates in a world coordinate system (when eye control module is purchased)
- start and end of task intervals and moments of events (when marked via triggers)

Data output after data analysis

- fixation coordinates in a world coordinate system
- Glance durations to all defined areas of interest (start time, duration, end time)
- Task interval durations
- Workload Glance Metrics:
  - Horizontal search activity
- Area of Interest based Glance metrics:
  - Total glance time to all defined areas of interest
  - Number of glances to all defined areas of interest
  - Mean glance duration to all defined areas of interest
  - Percentaged glance proportion to all defined areas of interest
  - Fixation frequency for all defined areas of interest
  - Maximum glance duration to all defined areas of interest
  - Minimum glance duration to all defined areas of interest
- Graphical data output:
  - Single HeatMaps
  - Multi HeatMaps
  - Gaze flow diagrams

D-Lab

Data Analysis



USPs:

- Live Eye-Tracking (with Dikablis Cable, Dikablis Wireless and Dikablis Wireless Plus)
  - Subject´s gaze behavior can be observed in absolute realtime
  - Relevant events can be marked directly online
  - Subject´s gaze behavior can be replayed immediately after the experiment and be included in retrospective think-aloud
- Fully Automated gaze data analysis
  - Because of inside-out Head-position measurement due to Marker detection
    - Works in any environment
    - Works for small objects like mobile phones up to big areas like supermarkets
  - Enables:
    - Autmated Area of Interest based analysis
    - 3D visualizations (e.g. Single HeatMap, Multi HeatMap)
- Eye controlled interaction with eye control module
  - Get world coordinates in realtime
  - Setup glance based interaction with any kind of device (e.g.: computer, displays, touchscreens, TV, and so on)
- Synchronous recording and analysis of up to 4 video streams and any kind of TCP/IP network data stream (with video & external data module)
- Quick set-up and calibration: the whole system can be set up and used in less than 10 minutes in every environment
- 100% data availability due to saving of all raw data. Enables: Re-Calibration and offline eye-detection improvement after the experiment