# The eyes have it

Mattheij, Ruud

*Publication date:*
2016

*Document Version*
Publisher's PDF, also known as Version of record

# The eyes have it

% Ruud Mattheij

SIKS Dissertation Series No. 2016-30. The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

TiCC Ph.D. Series No. 47.


Final Version as of September 6, 2016.

The cover of the Thesis is designed by the crafty and creative hands of Hans Westerbeek.

ISBN/EAN: 978 94 629 5485 4
Print: Uitgeverij BOXPress

# THE EYES HAVE IT

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan Tilburg University
op gezag van de rector magnificus,
prof. dr. E. H. L. Aarts,
in het openbaar te verdedigen ten overstaan van een
door het college voor promoties aangewezen commissie
in de aula van de Universiteit
op woensdag 5 oktober 2016 om 14.00 uur

door

RUDOLPHUS JOHANNES HUBERTUS MATTHEIJ,

geboren op 20 oktober 1987 te Venlo

Promotores:

Prof. dr. E. O. Postma

Prof. dr. H. J. van den Herik

Copromotor:

Dr. ir. P. H. M. Spronck

Overige leden van de promotiecommissie:

Prof. dr. A. P. J. van den Bosch

Prof. dr. V. Evers

Dr. J. R. C. Ham

Dr. C. L. Lisetti

Prof. dr. A. Plaat

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF DEFINITIONS

# 1 HOME IS THE PLACE TO GO

*"Home is a name, a word, it is a strong one; stronger than magician ever spoke, or spirit ever answered to, in the strongest conjuration."*
– Charles Dickens, *Martin Chuzzlewit*

Whether a single man takes one small step or mankind makes a giant leap, all endeavours require energy in some form. In the end, the very energy enabling these undertakings is often extracted from the energy resources produced by our planet. Since mankind's dependency on fossil fuels (such as gas, petroleum, and coal) increased over the centuries, the natural reserves are expected to be depleted in a future not too far away. Next to moving towards renewable energy sources, reducing our energy consumption and improving our methods to conserve energy are two important factors for our transition towards a sustainable society.

As reducing energy consumption may start at the household (see, for example, the work by Romero-Rodríguez, Zamudio Rodriguez, Flores, Sotelo-Figueroa, & Alcaraz, 2011), effective approaches towards energy conservation call for an intelligent environment that persuades its residents to change their energy consumption behaviour. To change the behaviour of its residents in the long term, the intelligent environment should provide its inhabitants with personalised feedback regarding their behaviour. Providing personalised feedback in a subtle and nonintrusive way can be achieved by employing a virtual person; a so-called "embodied agent". Employing a human-like appearance allows the intelligent environment to establish a social bond with a person. Establishing a social bond between the actuators of an intelligent environment (e.g., by means of a humanlike agent) and a person is a prerequisite for effective persuasion (Bailenson & Yee, 2005). A requirement for the establishment of the social bond between the person and the embodied agent, is the latter's ability to respond appropriately to a person's social signals (see, e.g., Vinciarelli et al., 2012; Breazeal & Scassellati, 2002).

This Thesis investigates novel algorithms that enable agents to perceive a person's non-verbal cues and gestures as accurately as possible. It allows

the agents to respond appropriately to a person's behaviour. The studies addressed in the Thesis are part of the *Persuasive Agents* research project (see Section 1.2), which explores the use of socially-aware virtual agents to persuade people to change their energy-consumption behaviour by providing them with subtle personalised feedback. Inspired by the magical paintings that litter the walls of the castle of Hogwarts, our ultimate goal is to develop smart, persuasive, and socially aware embodied agents that are able to engage in natural interactions with humans.

The remainder of this Chapter is as follows. Section 1.1 provides a general background of intelligent environments that can be used to influence the behaviour of their inhabitants. Subsequently, Section 1.2 presents the *Persuasive Agents* project. Section 1.3 then discusses the use of embodied agents to influence a person's behaviour. Next, Section 1.4 presents an interaction model describing the establishment of the social connection between humans and embodied agents. Section 1.5 describes the relevance of in-depth information when aiming to implement the interaction model in a household scene. Section 1.6 formulates the problem statement, including the resultant research questions and the corresponding research methodology used to answer them. Finally, Section 1.7 provides the structure of the Thesis.

## 1.1 INTELLIGENT ENVIRONMENTS

When Harry Potter walked through the dark hallways of Hogwarts, he was unaware that his school with its numerous magical paintings (cf. Rowling, 1997) bore remarkable similarities to modern visions of socially-aware virtual agents and intelligent environments. The technology enabling these environments (see, e.g., Vinciarelli, Pantic, & Bourlard, 2009), viz. "invisibly enhancing the world that already exists" (Weiser, 1997), offers numerous opportunities for new types and forms of human-computer interactions (see, e.g., Schmidt, Pfleging, Alt, Sahami, & Fitzpatrick, 2012; Sebe, 2009), such as computer systems that aim to influence a person's behaviour.

How should these intelligent environments be designed and deployed in a manner that facilitates more than it hinders their residents? The answer to this question should be guided by an emphasis on the social interaction between the residents and the intelligent environment. Recent progress (see, e.g., Murray-Smith, 2014; Pantic & Vinciarelli, 2014; Vinciarelli et al., 2012) in the automatic processing of affective and social signals enables intelligent environments and devices (1) to sense social cues, such as emotional facial expressions (e.g., distress, surprise) and emotional vocal expressions (tone of voice), and

**Figure 1.1:** An example of the deployment of an embodied agent in a domestic setting. The agent aims to persuade the household member to reduce his water consumption by providing him with personalised feedback about his energy consumption behaviour. The personalised feedback is presented using subtle facial expressions, e.g., by looking sad or angry when too much water is consumed.

(2) to respond appropriately to them. The envisioned intelligent environment consists of social signal sensors (cameras, microphones, and 3D scanners) and social actuators in the form of embodied virtual agents[1] or robots. The actuators emit social signals by means of virtually generated facial, vocal, and gestural expressions. The ultimate goal is to develop socially-aware virtual agents that are able to persuade people to reduce their energy consumption. Figure 1.1 shows an example of a human-like, virtual agent that aims to persuade a person to reduce his[2] water consumption by providing him with personalised feedback about his very behaviour, e.g., by looking sad or angry when too much water is consumed.

1 It is noted that embodied agents and virtual agents are, technically speaking, different concepts. An agent system is an abstract system that is able to make decisions based on empirical input; the correct designation of the agents described in this Thesis is *virtual embodied agents*. However, for the sake of readability, we will designate them as 'virtual agents', 'smart agents', 'embodied agents', or similar descriptions.

2 For brevity, 'he' and 'his' are used whenever 'he or she' and 'his or her' are meant.

## 1.2 THE PERSUASIVE AGENTS PROJECT

As reducing energy consumption may start at the household (see, e.g., the work by Romero-Rodríguez et al., 2011), early studies presented household members with general information about their energy consumption. The implicit assumption was that this would result in a voluntary change in the household members' energy-consumption behaviour. However, the results of more recent studies indicate this assumption to be false: providing individuals with general information regarding their energy consumption does lead to an increased awareness of the scarcity of resources, but it does not lead to actual changes in behaviour (see Abrahamse, Steg, Vlek, & Rothengatter, 2005). Based on earlier findings that indicated that personal feedback is more effective than general feedback (see, e.g., Midden, Meter, Weenig, & Zieverink, 1983), recent studies adopt a more promising approach to change a person's energy-consumption behaviour (see, e.g., the work by Ham, Midden, & Beute, 2009; Roubroeks, Midden, & Ham, 2009). These studies aim to change a person's short-term behaviour by providing him with automatically generated, personalised feedback regarding his behaviour. The belief that a person can be persuaded to adapt his behaviour in the long term by using intelligent environments, led to the launch of the *Persuasive Agents* project.

Since its establishment in 2007, the aim of the project is to develop novel techniques and autonomous systems that (1) persuade household members to reduce their energy consumption, and (2) support the conservation of the energy they have as much as possible. The systems collect information on consumption patterns through, e.g., power-consumption meters, and use that information to generate accurate feedback and suggestions. The main challenge of the research is to combine psychological and technological knowledge so as to identify and exploit successful human-embodied agent interactions. At the core of this project lies the belief that intelligent systems should stimulate people to adopt energy saving behaviour by means of persuasion, rather than by taking over control. The ultimate goal of the project is to develop embodied agents, often in virtual form on computer displays, but sometimes also as robotic interfaces, that are not annoying or obtrusive. The agents should be able to provide personalised and socially acceptable feedback with regard to saving energy to the inhabitants of intelligent environments. The implicit assumption of these studies is that the resulting reduction in energy consumption outweighs the actual costs of having and using such intelligent environments.

The Persuasive Agents project consists of a multidisciplinary group of researchers and practitioners from various fields and backgrounds: computer

scientists from Tilburg University[3], psychologists from Eindhoven University of Technology[4], and practitioners in smart home environments from the Smart Homes Foundation[5]. The research program is carried out under the stimulating leadership of Cees Midden and funded by Agentschap.nl under the EOS program for Long Term research.

## 1.3 PERSUASIVE EMBODIED AGENTS

Within the field of artificial intelligence, *agents* are autonomous parts of computer systems that possess some form of artificial intelligence (see, for example, Wooldridge, 2001; Neumann, 1958). It enables them to make autonomous decisions based on empirical input or past experiences. The designation *embodied* agent refers to agents with a recognisable form, e.g., in the form of a physically existing robot or in a mere virtual existence as a computer game character. An embodied form (e.g., a robot or game character) allows the agent to interact with human users in a natural way. The ability to interact in a natural way is a prerequisite when attempting to establish a social connection with a person. An example of a natural interaction in the context of the current project, is an agent that aims to persuade a person to change his energy consumption behaviour by providing him with personalised feedback about his very behaviour (see Figure 1.1; see, e.g., Vinciarelli et al., 2012; Bailenson & Yee, 2005; Breazeal & Scassellati, 2002).

When provided by an embodied agent, the effectiveness of personalised feedback in a persuasive context is enhanced when (1) participants perceive the feedback as *non-obtrusive*, and (2) the feedback is communicated in a *human-like* way. Below, these requirements are described in more detail.

First, personal feedback that is experienced as obtrusive may be regarded as a violation of the individual's autonomy (Brehm, 1989). In case of personalised feedback on energy consumption, this may give rise to an increase in energy consumption, rather than a decrease, an effect known as *psychological reactance* (Brehm, 1989). Providing individual feedback in a more subtle manner (see, e.g., Ham et al., 2009; Roubroeks et al., 2009), for example in the form of a smile or a nod, may therefore increase its effectiveness. Furthermore, employing human-like interfaces, such as "eyes in the wall" (Bateson, Nettle, & Roberts, 2006) or a talking head (see Figure 1.1), increases cooperative behaviour and leads to effective persuasion (André et al., 2011).

3 https://www.tilburguniversity.edu/research/institutes-and-research-groups/ticc
4 http://www.tue.nl/universiteit/faculteiten/industrial-engineering-innovation-sciences/onderzoek/onderzoeksgroepen/human-technology-interaction
5 http://www.smart-homes.nl

Second, persuasive agents should appear human-like, i.e., possess a certain degree of personality (Davies & Callaghan, 2012), and come across as credible, confident and non-threatening towards the users and their privacy (Tentori, Favela, & Rodriguez, 2006) in order to establish and maintain persuasive interaction.

To meet the two requirements, it is necessary that the agents are enriched with basic non-verbal characteristics, such as affective facial expressions and vocal prosody (see, e.g., Van den Broek, 2011; Esposito, 2009), which serve as carriers of social signals, such as attitudes, stands, and emotions. Moreover, non-verbal characteristics seem to play a crucial role in persuasive communication (e.g., Hogg & Reid, 2006; Hiltz, Johnson, & Turoff, 1986). Thus, their use is particularly relevant in the context of persuasive technology. Supplying agents with non-verbal cues makes them more appropriate for virtual reality applications and smart environments (Vinciarelli et al., 2012).

## 1.4 ESTABLISHING THE SOCIAL CONNECTION

A requirement to establish and maintain persuasive interaction, is the presence of a social connection between an embodied agent and its human counterpart (see, e.g., Dragone, Duffy, & O'Hare, 2005). The connection, which is highly similar to the social bond established in human-human interactions (see, e.g., Hari & Kujala, 2009; Miller, Downs, & Prentice, 1998), allows an agent to provide its human counterpart with personalised feedback regarding his behaviour. Figure 1.2 shows a model[6] of the envisioned social interactions between an embodied agent (left) and a person (right). The interactions result in the establishment of a social connection between the embodied agent and the person. In the model, the social bond between the agent and the person is established and maintained in three recursive stages. In the Figure, these stages are labelled A to C. They are represented as blue transparent ovals. In what follows, the individual stages are discussed in more detail.

STAGE A: ANALYSING BEHAVIOUR In stage A, the behaviour of the person is detected by analysing sensory data from an array of sensors, e.g., cameras and microphones. By utilising advanced artificial intelligence, including dedicated machine learning and data mining techniques, the agent is able to detect the person's mood, behaviour and responses.

6 Please note that the model itself is not validated in this Thesis. It merely serves as a guideline for the reader to illustrate the envisioned social interactions between humans and embodied agents.

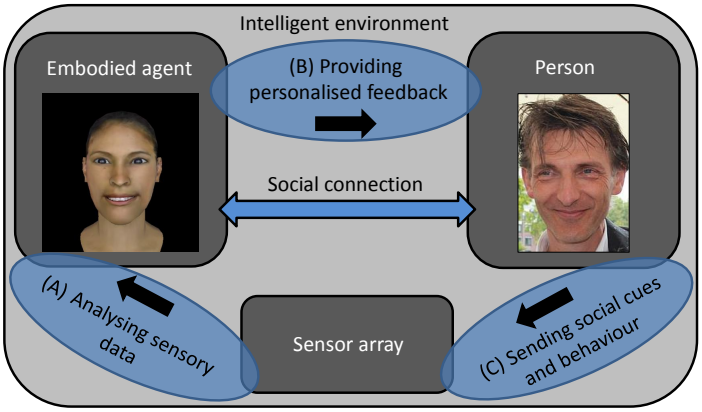



**Figure 1.2:** The model of the social interactions between humans and embodied agents. An agent detects the behaviour of the person in the intelligent environment (stage A), after which he provides the person with personalised feedback in the form of subtle social signals (stage B). Perceiving the feedback sent out by the agent, the person may adapt his behaviour accordingly (stage C), which can again be detected by the sensor array (stage A) and used for the next cycle of human-embodied agent interactions. As a result, a social bond between the person and the embodied agent is established.

STAGE B: PROVIDING PERSONALISED FEEDBACK Based on the analysis of the sensory data, the agent may decide to provide the person with personalised feedback. The feedback is presented as social cues, e.g., subtle changes in facial expressions or tone of voice, which are directed at the person.

STAGE C: SENDING SOCIAL CUES AND BEHAVIOUR Given the subtle nature of the feedback, the person perceives the feedback of the agent subconsciously - and therefore as nonintrusive. Perceiving the feedback sent out by the agent, the person may adapt his behaviour accordingly, or respond to it by using (1) verbal cues (e.g., voice), or (2) non-verbal cues, such as facial expressions, body pose, and gestures.

The adapted behaviour and responses of the person can then be detected by the sensor array (stage A), which completes a cycle of the recursive interactions. The social cues and behaviour of the person can be used as input for the next cycle. As a result, a social bond between the person and the embodied agent is established.

## 1.5 THE RELEVANCE OF DEPTH DATA

In the domain of human-agent interactions, it can be expected that enabling an agent to perceive a person's social (i.e., verbal and non-verbal) cues will allow the agent to respond more appropriately to the person's behaviour. Whereas *verbal* cues in general are difficult to detect and analyse due to their sensitivity to background noise, *non-verbal* cues (such as facial expressions and gestures) provide a rich and nowadays accessible source of information about a person's emotions, intentions, and actions.

To enable an agent to sense the non-verbal behaviour of its human communication partner, the agent requires sensors to perceive the world around the human. The agent sets the corresponding object detection algorithms to work to analyse and understand the person's behaviour. As embodied agents are likely to be deployed in noisy environments (i.e., environments with a large variety of objects, changing illumination conditions, and moving people, such as a household), the agents require state-of-the-art computer vision algorithms that are able to deal with the noisy nature of the environment.

Within the various fields of artificial intelligence, most object detection approaches (see, e.g., Khaligh-Razavi, 2014; Andreopoulos & Tsotsos, 2013) rely on visual features to segregate objects from their backgrounds (see, for example, De Croon, Postma, & Van den Herik, 2011; Bergboer, 2007; Lee & Nevatia, 2007). Visual features are extracted from visual data[7], e.g., RGB (Red Green Blue) images. While rich in detail, the main disadvantage of visual data is that it is sensitive to the illumination conditions (see, e.g., Rautaray & Agrawal, 2015; C. Zhang & Zhang, 2010; Zhao, Chellappa, Phillips, & Rosenfeld, 2003). Shadows, for example, may obscure objects from sight, making them difficult to detect.

While it is possible to reduce the sensitivity of visual features to illumination conditions (see, e.g., Qu, Tian, Han, & Tang, 2015; Huorong Ren, Yu, & Zhang, 2015; Shah & Kaushik, 2015; Son, Yoo, Kim, & Sohn, 2015), such improvements tend to result in an increase in computational complexity, and are therefore not ideal for agent systems that aim to operate in real-time. Thus, given the sensitive nature of visual data (and thereby the visual features extracted from it), using visual data as the main information source for automatic detection tasks is unpractical in noisy environments such as household scenes.

A requirement for effective object detection approaches in noisy environments is that they are insensitive to background noise. Thus, object segregation may be facilitated by using depth data rather than visual data. Exploiting

7 It is noted that there is a clear hierarchical difference between *(raw) data* and *information*. In this Thesis, we consider input data (e.g., an image) as *raw data*. Extracting features from input data results in cleaned raw data, i.e., *data*; predicting the corresponding class labels results in *information*, i.e., data with a meaning.

depth data allows for the extraction of depth features, which can be used as an alternative to the widely-used visual features. As depth features provide direct access to the third dimension, this enables object-background segregation even under noisy conditions (see, e.g., Brandão, Fernandes, & Clua, 2014; Tang, Sun, & Tan, 2014; Chan, Koh, & Lee, 2013). As such, using depth data as an additional - or even as the main - data source is highly relevant when aiming to achieve robust object detection. The use of depth data became feasible with the introduction of affordable depth sensors, such as the MICROSOFT KINECT device[8] (see, e.g., Dal Mutto, Zanuttigh, & Cortelazzo, 2012).

Although depth data is insensitive to illumination conditions, the depth images generated by the Kinect device still suffer from low image quality and resolution. This results in high levels of background noise in the depth data (see, e.g., Smisek, Jancosek, & Pajdla, 2013; Khoshelham & Elberink, 2012; Spinello & Arras, 2011). Object detection approaches that aim to incorporate depth data should therefore be able to deal with the background noise.

## 1.6 PROBLEM STATEMENT

When given a meaning, depth data is a robust and valuable source of information about a person's non-verbal cues. We call depth data with a meaning: *in-depth information*. Enhancing an agent's cognitive abilities by incorporating in-depth information is likely to increase the agent's ability to perceive human behaviour. In this Thesis, we will explore the possibilities to deploy in-depth information to detect the non-verbal cues of people. For this purpose, the problem statement of the Thesis is formulated as follows.

> **Problem statement:** *To what extent is it possible to detect human body parts and behaviour when using in-depth information?*

The problem statement is the point of departure for five separate research questions, which are presented in Subsection 1.6.1 on the next page. Answering the research questions to a sufficient degree may result in several contributions, which are envisaged in Subsection 1.6.2. Subsequently, the methodology employed to answer the research questions is described in Subsection 1.6.3.

8 For the sake of readability, we henceforth refer to the MICROSOFT KINECT device as "Kinect" or "Kinect device".

### 1.6.1 Research Questions

To answer the problem statement as described above, five research questions are formulated. Below, these research questions are listed and individually motivated.

Within the field of depth-based object detection, a well-known example of effective body part detection is proposed by Shotton and his collaborators (see Shotton, Girshick, et al., 2013; Shotton, Fitzgibbon, et al., 2013; Shotton et al., 2011). Hereafter, we will indicate these references for brevity as Shotton et al. (2013a,b; 2011). The teams guided by Shotton developed a state-of-the-art body part detector that is able to classify individual pixel locations as belonging to faces, body joints, and body parts. Their approach uses depth images that are generated by a Kinect device. Though able to achieve high detection speeds, their approach suffers from the low quality of the depth images. Deploying body part detection algorithms that are fast and insensitive to background noise (as discussed in Section 1.5), is highly relevant in the context of the current project. The first research question (RQ 1) therefore reads as follows.

> **Research question 1:** *How can we improve Shotton et al.'s body part detector in such a way that it enables fast and effective body part detection in noisy depth data?*

The answer to this research question is guided by the need for robust depth comparison features that enable effective object-background separation. The features should (1) enable a detector to deal efficiently with background noise, and (2) enable a high detection accuracy. With the help of the findings of RQ 1 we aim to develop the notion of Region Comparison features by which we are able to succeed with effective body part and gesture detection in noisy depth data. The Region Comparison features will guide our research. To evaluate the effectiveness of Region Comparison features for body part detection tasks, we perform a comparative evaluation of the RC features on several challenging object detection tasks. In the evaluation, the performance of the RC features is compared with the performance of the original approach as used by Shotton et al. (2013a,b; 2011). The second research question (RQ 2) thus reads as follows.

> **Research question 2:** *To what extent do Region Comparison features enable fast and accurate face and person detection in noisy depth images?*

Facilitating natural interactions between humans and embodied agents asks for advanced algorithms that are able of recognise a person's gestural cues. Developing and training gesture recognition algorithms require high quality corpora that contain annotated, visual and depth data recordings of people performing natural communicative gestures. However, to the best of our

knowledge there are no databases available that (1) contain visual and depth data recordings of natural gestures, and (2) are available for academic purposes. This leads us to formulate the third research question (RQ 3):

> **Research question** 3**:** *How do we develop an annotated database that incorporates visual and depth data recordings of natural human gestures?*

Enabling agents to perceive a person's social cues is a first step towards natural human-embodied agent interactions. Investigating the effectiveness of the Region Comparison features for accurate gesture recognition is thus highly relevant for the development of embodied agents that aim to engage in natural interactions with people. However, facilitating the actual interactions requires agents that are capable of perceiving a person's (natural) gestural cues. Hence, we will evaluate the performance of the Region Comparison features for effective gesture recognition. Our fourth research question (RQ 4) thus reads as follows.

> **Research question** 4**:** *To what extent do Region Comparison features enable accurate recognition of static gestures when using in-depth information?*

To establish the envisioned human-embodied agent interactions, we assume that it is possible to create a strong, social connection between humans and embodied agents, i.e., that humans are able to perceive an embodied agent as a communication partner. It is, however, unclear to what extent it is actually possible to create such social bonds between people and their virtual counterparts. Investigating to what extent such social bonds can be established may be guided by the work by Chartrand & Van Baaren (2009), who found that the process of imitation is an important social cue in human-human interactions. As such, examining the effect of virtual agents on the imitative behaviour of humans is highly relevant in this context. Given that mimicry is a form of imitation that is mostly unconscious and unintentional (Chartrand & Lakin, 2013), it is particularly interesting to investigate to what extent humans exhibit behavioural mimicry in the form of copying facial expressions and vocal characteristics when interacting with virtual agents. If humans, in fact, unknowingly imitate different non-verbal cues of the agent, it can be interpreted as an indicator of real social engagement. The fifth and last research question (RQ 5) therefore reads as follows.

> **Research question** 5**:** *To what extent do people mimic verbal and non-verbal cues sent out by an embodied agent?*

**Table 1.1:** Overview of the research approaches employed to investigate the individual research questions (RQs)

| RQ | Computational research | Behavioural research |
|---|---|---|
| **1** | ✓ | |
| **2** | ✓ | |
| **3** | ✓ | ✓ |
| **4** | ✓ | |
| **5** | ✓ | ✓ |

### 1.6.2 Research Objectives

Assuming that we are able to answer the research questions to a sufficient degree, we then arrive at the six research objectives of the Thesis. They are defined as follows.

1. The proposition of a set of effective depth comparison features.
2. The development of a state-of-the-art object detection algorithm that allows for fast and accurate body part detection in noisy depth images.
3. The development of an algorithm that recognises static fingerspelling signs using depth data.
4. Gaining advanced insights into the extent to which people are able to perceive a virtual person as a true communication partner.
5. The development of a challenging and publicly available database with annotated depth images of human body parts.
6. The development of an open source annotation tool for depth images.

In total, our research may result in a new set of features, two new algorithms, a new corpus, advanced insights, and a newly developed open source tool.

### 1.6.3 Research Methodology

Given that the investigation of the problem statement required a multidisciplinary approach that combined both behavioural research and computational science, the research methodology deployed in this Thesis is tuned to serve multidisciplinary research. It should be noted that the overlapping research

area of the behavioural and computational approach is flexible when answering the research questions. Combining the knowledge from both disciplines allows for the investigation of human behaviour, while it also enables fast and efficient processing and analysis of the experimental results. Table 1.1 provides an overview of the main research approaches that were employed to answer each individual research question. In general, the methodology employed to answer the problem statement and the research questions of the Thesis (as formulated in Subsection 1.6.1) consists of six separate stages.

1. Reviewing relevant scientific literature.
2. Designing and performing comparative experiments.
3. Analysing the results.
4. Formulating the resultant conclusions and discussing their implications.
5. Answering the research questions in detail.
6. Answering the problem statement.

## 1.7 STRUCTURE OF THE THESIS

The problem statement of the Thesis is investigated and discussed over the course of the next Chapters. Table 1.2, as shown on the next page, provides an overview of the problem statement and consecutive research questions, and the Chapters in which they are addressed. Below, the structure of the Chapters is presented in more detail.

***Chapter 1: There is no Place Like Home***

The Chapter proposes an interaction model that describes the process of establishing and maintaining social connections between humans and socially aware agent systems. It formulates the problem statement (PS) and five research questions: RQs 1, 2, 3, 4, and 5. Subsequently, the Chapter presents the six stage research methodology that is used to answer the research questions. Answering the research questions may lead to six individual research objectives.

***Chapter 2: In Depth Lies Truth***

A requirement for effective computer vision algorithms is that they are insensitive to variations in illumination conditions. In-depth information, which is extracted from depth data, is insensitive to changes in illumination conditions,

**Table 1.2:** Overview of the problem statement (PS) and the subsequent research questions (RQs), and the Chapters in which they are addressed.

| Chapter | PS | RQ 1 | RQ 2 | RQ 3 | RQ 4 | RQ 5 |
|---|---|---|---|---|---|---|
| **1** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **2** | | ✓ | | | | |
| **3** | | | ✓ | | | |
| **4** | | | | ✓ | | |
| **5** | | | | | ✓ | |
| **6** | | | | | | ✓ |
| **7** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **8** | | ✓ | ✓ | ✓ | ✓ | ✓ |

and may thus allow for robust object detection. Noisy depth measurements, however, may result in high levels of background noise in the depth data. This Chapter addresses RQ 1 by presenting the novel Region Comparison (RC) features. The features are likely to deal effectively with noisy depth data.

***Chapter 3: Through the Looking Glass***

As it is unclear to what extent the RC features actually enable fast and effective object detection in noisy depth data, this Chapter addresses RQ 2 by performing a comparative evaluation of the RC features on several challenging object detection tasks. In the evaluation, the performance of the RC features is compared with the performance of the state-of-the-art depth comparison features that are proposed by Shotton et al. (2013a,b; 2011).

***Chapter 4: Raising a Tiger***

This Chapter addresses RQ 3 by investigating to what extent it is possible to develop a corpus that contains annotated, visual and depth data recordings of people performing natural communicative gestures. To answer the research question, we present the Tilburg Gesture Research (TiGeR) Cub, a multimodal corpus that consists of dyadically interacting interlocutors. The interactions are recorded as visual data, depth data, and audio data. As such, the TiGeR Cub allows for detailed studies into the synthesis and automatic classification of human gesture.

***Chapter 5: Automatic Gesture Recognition From A to Y***

Having proven their worth for effective body part detection tasks, deploying RC features is highly relevant for embodied agents that aim to establish natural interactions with people. To enable natural interactions, it is imperative that agents are enriched with the ability to perceive gestural cues. Hence, this Chapter answers RQ 4 by investigating to what extent RC features are suitable for automatic approaches towards gesture recognition.

***Chapter 6: Mirror, Mirror on the Wall***

So far, our studies focused on increasing an agent's ability to perceive social cues and human behaviour, as this may allow agents to respond more appropriately to people. It is, however, unclear to what extent it is actually possible to establish a social connection between a person and an embodied agent, i.e., to what extent humans are able to perceive an embodied agent as an actual communication partner. As such, this Chapter addresses RQ 5 by investigating to what extent humans show mimicking behaviour when interacting with an emotionally expressive embodied agent.

***Chapter 7: Conclusions***

This Chapter combines the answers to the research questions into several conclusions. Based on the findings and conclusions, an answer to the problem statement is formulated.

***Chapter 8: General Discussion***

This Chapter discusses the findings that are presented in the Thesis, and their implications for the development of smart embodied agents. Subsequently, the Chapter discusses points of improvement. The Chapter concludes by formulating four recommendations for future research.

# 2 IN DEPTH LIES TRUTH

*"'Tis of great use to the Sailor to know the length of his Line, though he cannot with it fathom all the depths of the Ocean."*
– John Locke, *An Essay Concerning Humane Understanding*

In the domain of human-embodied agent interactions, increasing an agent's ability to perceive a person's non-verbal cues will allow the agent to respond appropriately to a person's behaviour. To perceive these social cues accurately, the agent needs a combination of sensors and machine-learning algorithms that extract meaningful information about the person's behaviour. Dedicated computer vision algorithms are at the core of the agent's ability to 'see' the person's gestures and facial expressions by detecting objects, such as the person's body parts and joints. A well-known example of an effective body part detection approach is proposed by Shotton et al. (2013a,b; 2011). They developed a state-of-the-art body part detector that classifies individual pixel locations as belonging to faces, body joints and body parts. Their approach uses depth images that are generated by a MICROSOFT KINECT device (see, e.g., Smisek et al., 2013). Though able to achieve high detection speeds, their approach suffers from the low quality of the depth images. Thus, a requirement for effective object detection algorithms (as discussed in Section 1.5) is that they are insensitive to background noise. This Chapter[9] outlines the need for robust depth comparison features that are (1) insensitive to background noise, and (2) able to maintain a high classification performance and detection speed. The Chapter then proposes a novel idea, viz. the Region Comparison (RC) features, which enable fast and robust human body part detection in noisy depth images.

The structure of the Chapter is in accordance with the description above. Section 2.1 presents depth data as a robust alternative to visual data. It also discusses the first principles and limitations of Shotton et al.'s state-of-the-art

9 This Chapter is based on work by R. J. H. Mattheij, K. Groeneveld, E. O. Postma, and H. Jaap van den Herik (2016); Depth-Based Detection with Region Comparison Features. Published in the *Journal of Visual Communication and Image Representation (JVCI)*.

body part detection algorithm. Subsequently, Section 2.2 presents and motivates the research question addressed in the Chapter. Section 2.3 reveals our contribution towards fast and robust object detection. Section 2.4 presents the work related to our approach. Finally, Section 2.5 concludes upon our contribution and answers the first research question.

## 2.1 TOWARDS ROBUST BODY PART DETECTION

In the last few years, the automatic detection of objects from digital video and image sources has gained considerable attention within the field of image analysis and understanding (see, e.g., Nanni, Lumini, Dominio, & Zanuttigh, 2014; Andreopoulos & Tsotsos, 2013; Jiang, Fischer, Kemal, & Shi, 2013). Many approaches towards object detection focus on extracting two-dimensional visual features (e.g., De Croon et al., 2011; Bergboer, 2007; Lee & Nevatia, 2007) to help to segregate objects from their backgrounds. Well-known visual features for object detection are the Haar-like features (Lienhart & Maydt, 2002) proposed by Viola and Jones (Viola, Jones, & Snow, 2005; Viola & Jones, 2001).

Despite the widespread and successful use of two-dimensional (2D) visual features in visual detection tasks, they have an important limitation: they typically respond to local visual transitions without being sensitive to the larger spatial context (see, e.g., Carlevaris-Bianco & Eustice, 2014). As a consequence, they are sensitive to factors that may influence scene properties locally, such as illumination conditions (see, e.g., C. Zhang & Zhang, 2010; Zhao et al., 2003). Bright lights, for example, may cause shadows (i.e., non-object contours) in the image. Local 2D visual features will respond to the contours of the shadows in the same way as to the contours of other, real objects. Typical situations in which 2D visual features fail are those where variations in the third dimension (depth) lead to shape deformations. In general, the failures are caused by object pose variations (e.g., Andreopoulos & Tsotsos, 2013; Liao, Jain, & Li, 2012).

A wide variety of methods attempts to overcome these sensitivities. The most frequently applied methods focus on extracting context-sensitive features (see, e.g., Bergboer, 2007). Although such approaches improve classification performance, they tend to be costly in terms of computational resources (J. Wu et al., 2013; Liao et al., 2012).

The remainder of this Section is as follows. Subsection 2.1.1 presents three-dimensional (3D) cues as a robust alternative to (2D) features. Subsequently, Subsection 2.1.2 describes the first principles of the KINECT device (a sensor array that can be used to capture depth data), while 2.1.3 deals with the state-

of-the-art body part detection algorithm by Shotton et al. (2013a,b) that is used to detect objects in 3D data.

### 2.1.1 From 2D Features to 3D Features

To overcome the limitations of 2D features, we add a third dimension by combining 2D spatial and 1D depth information into 3D features (see, e.g., Brandão et al., 2014; Tang et al., 2014; Baak, Müller, Bharaj, Seidel, & Theobalt, 2013; Chan et al., 2013; Riche, Mancas, Gosselin, & Dutoit, 2011). Depth cues then provide contextual information for a scene, which facilitates image segmentation (see, e.g., Jiang et al., 2013; Dal Mutto et al., 2012; Plagemann, Ganapathi, Koller, & Thrun, 2010; Hoiem, Efros, & Hebert, 2006). Visual objects, such as faces or persons, are actually much easier to distinguish in a 3D space than to recognise from a 2D image (e.g., Brunton, Salazar, Bolkart, & Wuhrer, 2014; Burgin, Pantofaru, & Smart, 2011). In recent years, the use of depth cues became feasible by the development of affordable depth sensors, such as KINECT device (see, e.g., Smisek et al., 2013). The depth cues captured by the depth sensors are represented as two 2D *depth images*, in which each pixel location describes the depth cue at that very specific location. As such, 2D depth images provide a 3D description of a scene.

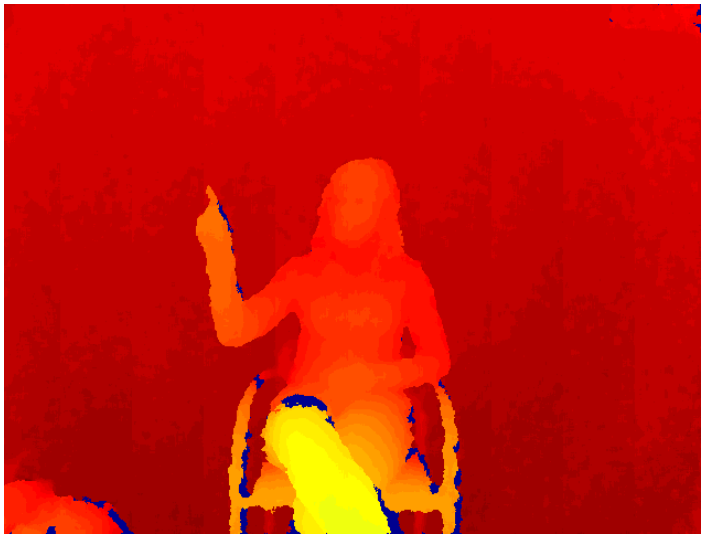
### 2.1.2 Capturing Depth with Microsoft Kinect

The MICROSOFT KINECT[10][11] (see, e.g., Smisek et al., 2013) device generates its depth images by (1) illuminating a spatial area with the Kinect's infrared laser, and (2) triangulating the corresponding depth with an infrared sensor (Z. Zhang, 2012). Using an infrared laser that passes through a diffraction grating, a grid of infrared dots is created. Given the known spatial distance between the Kinect's infrared laser and sensor, matching (A) the dots observed in an image with (B) the dots projected using the pattern from the diffraction grating, allows for effective depth triangulation. The resulting depth images have a resolution of $640 \times 480$ pixels. The pixel values of the depth images encode for the distance between an object and the Kinect device. A large depth value indicates a large distance between the object and the Kinect device, while a small depth value encodes for a small distance. On the next page, Figure 2.1 shows (in 2.1a) an example of a visual image that is captured with a Kinect device, and (in 2.1b) the corresponding depth image.

10 https://dev.windows.com/en-us/kinect

11 For the sake of readability, we henceforth refer to the MICROSOFT KINECT device as "Kinect" or "Kinect device".

a



b

**Figure 2.1:** An example of a visual image of a person (a), and the corresponding depth image (b). Note the (background) noise in the latter image, which is visualised as dark areas that can be seen at the edges of the objects in the depth image.

### 2.1.3 Shotton's Pixel Comparison Features

Using the Kinect device, Shotton et al. (2013a,b; 2011) proposed a depth-based body part detection algorithm that *selects* and classifies individual pixel locations in single depth images. Their method incorporates pixel-based depth comparison features. For the sake of readability, we refer to these features as the *Pixel Comparison* (PC) features. In what follows, we briefly discuss Shotton et al.'s (2013a,b) feature computation procedure.

Shotton et al. started their feature computation procedure by selecting a subset of random pixel locations from each individual depth image. For each pixel location P from this subset, the depth difference is computed by comparing the depth values at two randomly chosen offset locations Q and R. The offset locations are defined by the radius and angle with respect to point P. The radius is defined to be inversely proportional to the depth value at point P. A small depth value results in a larger radius for offset locations Q and R, and vice versa. This way, a scale-invariant measure of depth between two pixel locations is obtained. A single depth comparison between locations Q and R provides only a weak indication of the depth difference in a spatial area around point P. Repeating this measurement for other (randomly chosen) offset locations Q and R, however, provides a fair description of the depth difference in an area around the location of point P. Then, Shotton et al. classified the selected pixel locations in the subset as belonging to faces, body joints and body parts. Below, we discuss (1) the procedure of selecting and classification, and (2) the trade-off between speed and accuracy.

SELECTING AND CLASSIFYING There are two advantages of classifying individual pixel locations rather than image regions (e.g., by means of a sliding window): (1) the selection process allows for the detection of partially occluded objects, and (2) the classification process reduces the time required to process an entire depth image. Using pixel-based depth-comparison features makes their detector computationally efficient. In addition to these qualities, the detector works directly on the raw input depth data, i.e., without an image pre-processing stage to reduce noise in the data (cf. Förstner, 2000). Combining (1) efficient depth-comparison features and (2) the raw input depth image is relevant for fast and effective object detection, as it allows for a high detection speed. This enables a real-time operation.

SPEED VERSUS ACCURACY The detection speed, however, comes at the cost of accuracy. The classification accuracy is hampered by two limitations (see, e.g., Smisek et al., 2013; Khoshelham & Elberink, 2012; Spinello & Arras, 2011): (1) the limited quality of the depth images generated by the Kinect device, and (2) the limited resolution of the depth images.

The first limitation arises from the triangulation sensor that is incorporated in the Kinect device. Depending on the image geometry, parts of a scene may not be illuminated by the sensor's laser, i.e., the grid of infrared dots. These parts are therefore not captured by the infrared sensor, which results in empty regions in the depth image (cf. Khoshelham & Elberink, 2012). Figure 2.1b shows an example of a depth image that is captured with the Kinect device. Special attention should go to the (background) noise in the image, which is visualised as dark areas that can be seen at the edges of the objects in the depth image.

The second limitation is due to the point density of the Kinect device's sensor. Using its laser and depth sensor, the Kinect device generates a point cloud of triangulated depth measurements. The dimensions of the spatial area that are covered by the point cloud increase quadratically with the distance from the Kinect device. Hence, the resolution of the depth images generated by the Kinect device decreases with the distance (Khoshelham & Elberink, 2012). These two limitations result in noisy depth measurements. It calls for feature computation methods that are able to deal efficiently with the noisy nature of depth images.

## 2.2 IMPROVING SHOTTON'S DETECTOR

Shotton et al. (2013a,b; 2011) suggested that a larger computational budget may allow for the design of "potentially more powerful features based on, for example, depth integrals over regions, curvature, or more complex local descriptors" (see Shotton et al., 2013a). Alternatively, studies seeking to improve object detection in depth images (see, e.g., Han, Shao, Xu, & Shotton, 2013) can opt to use a larger computational budget to refine the input depth data itself by, for example, including (depth) image filters or other refinement techniques (e.g., Fanello et al., 2014; Vijayanagar, Loghman, & Kim, 2014; Wang, An, Zuo, You, & Zhang, 2014; S. Liu, Wang, Wang, & Pan, 2013). While deploying additional computational power is likely to increase the detector's accuracy, it may come at the cost of detection speed. This necessitates the development of local descriptors that are both fast and accurate. Hence, the research question addressed in this Chapter (RQ 1) reads as follows.

> RQ 1: *How can we improve Shotton et al.'s body part detector in such a way that it enables fast and effective body part detection in noisy depth data?*

In this Chapter, I propose an improvement of Shotton et al.'s pixel-based depth comparison features by introducing specialised region-based descriptors that do **not** require an increased computational budget: the Region Com-

parison (RC) features. I am inspired by the work by Papageorgiou, Oren, & Poggio (1998), and Viola & Jones (2001). So, the RC features are based on the well-known Haar-like region features (see, e.g., Lienhart & Maydt, 2002; Viola et al., 2005; Viola & Jones, 2001; Papageorgiou et al., 1998) and combined with the integral image representation (Crow, 1984) of depth images. As such, the RC features are able to detect depth transitions in adjacent regions of depth images.

## 2.3 REGION COMPARISON FEATURES

Below, Region Comparison (RC) features are introduced as our improvement of Shotton et al. (2013a,b; 2011)'s method. Their introduction and implementation aim to answer RQ 1. The RC features (as defined in Definition 2.1) translate depth transitions (i.e., depth contours or edges) over regions in a depth image into a numerical value, i.e., the *RC feature value*. The feature value provides an indication of the magnitude of the depth transition. The RC features are based on the well-known Haar wavelets (Guf & Jiang, 1996). They provide an indication of the direction and magnitude of depth transitions in an area of a depth image by comparing the depth differences over regions, i.e., large groups of pixels, instead of pixel pairs (as seen in, for instance, Shotton et al., 2013b). On the one hand, varying the dimensions of the regions over which the RC features are computed, allows for the description of depth transitions, *smaller or larger*. On the other hand, varying the relative positions of the regions towards each other allows for the *computation of the direction* of the depth transition.

**Definition 2.1: RC features**

RC features are two-dimensional filters that translate depth transitions over regions in a depth image into a numerical RC feature value, which describes the magnitude of a depth transition in an area of a depth image.

The advantage of *comparing regions* rather than individual pixel values (as seen in Shotton et al., 2013a,b, 2011) is that it allows to average over larger areas. As a result, RC features are less prone to local pixel noise. Averaging over larger regions, however, results in a loss of spatial precision. By virtue of the Viola-Jones approach (Viola & Jones, 2001), which combines (1) Haar wavelets, and (2) integral images, the RC features combine the best of both worlds. There are two advantages. Advantage 1 is that the RC features include the averaging

(summing) over large regions, which makes the features insensitive to local pixel noise. Advantage 2 is that the features also take individual pixel pairs, i.e., small regions, into account.

To extract the RC features for a pixel location, the sums of the pixel values enclosed in the rectangular regions around that pixel location is computed, after which the sums are subtracted from each other. The computation procedure of the RC feature is explained in more detail in Subsection 2.3.1. The spatial orientation of the regions of the RC features are predefined as combinations of symmetrically located rectangular regions in the depth image, i.e., the so-called *feature types* (see Subsection 2.3.2). The additional computational cost required to calculate the surfaces of the regions, i.e., the sum of the pixel values, is negligible when integral images are employed (cf. Fanelli, Dantone, Gall, Fossati, & Van Gool, 2013; Fanelli, Weise, Gall, & Van Gool, 2011). Thus, the RC features are computed using the *integral* depth image rather than the depth image itself. Combining RC feature values results in the creation of a RC feature vector, which provides a mathematical description of the depth transitions in the area around the selected pixel location (see Subsection 2.3.3).

### 2.3.1 Formal Definition

In this Subsection, Definition 2.1 is transformed into a formal definition, i.e., a mathematical description of the feature value. An RC feature value for pixel location $P(x, y)$ in a depth image is computed by first calculating the sums of the pixels enclosed by two[12] rectangular regions, and then subtracting these sums from each other (cf. Viola & Jones, 2001). Subtracting the sums of the areas results in a single feature value that indicates the depth difference over a region. The features are calculated using predefined dimensions for the rectangular regions and their relative positions to each other (see Subsection 2.3.2). The feature type depends on three variables, viz. (1) the parameter $r$ defining the size of the individual regions, (2) the number of rectangular regions $d$, and (3) the spatial configuration $i$ defining the orientation of the constituent rectangular regions. The resulting feature values thus provide (1) an indication of the direction and (2) the magnitude of the depth transition over an area around point $P$. Formally, the RC feature value of type $i$ at location $P$ in depth image $I$, $f_i(P, I)$, is defined as follows:

$$f_i(P, I) = \sum_{n=1}^{d(i)} S(A_n(i), r) - \sum_{n=1}^{d(i)} S(B_n(i), r),$$

12 Later, we will broaden the computation procedure by allowing more than two enclosing rectangular regions.

where $A_n(i)$ and $B_n(i)$ represent rectangular regions of feature type $i$. In our formalisation, we calculate sum $S(X_n(i), r)$ of the pixels enclosed by rectangular region $X_n(i)$ of size $r$, where $X_n$ encodes for region $A_n$ or $B_n$. In this definition, parameter $n$ represents the index number of the rectangular region: $n = \{1, 2, ...d(i)\}$. The maximum number of rectangular regions $d(i)$ is predefined by feature type $i$. Iterating over all regions of $X_n(i)$, we calculate the total sum of summed regions $S(X_1(i), r)$ to $S(X_{d(i)}(i), r)$. The feature value $f_i$ is then computed by subtracting the sums for the regions $A$ and $B$.

The rectangle image regions define the regions over which the depth difference is calculated. The value of $r$ determines the spatial scale of analysis. For a small value of $r$, the associated feature encodes depth transitions at a small scale, while large values of $r$ allow the associated feature to encode for depth transitions at a large scale.

### 2.3.2 Feature Types

The number of rectangular regions and their relative spatial positions in relation to each other are predefined in terms of feature types $i$ (see Definition 2.2). The feature types are based on the well-known Haar-like features as proposed by Papageorgiou, Oren, & Poggio (1998), and used by Viola & Jones (2001), and Lienhart & Maydt (2002).

**Definition 2.2: Feature types**

Feature types are predefined combinations of symmetrically located rectangular regions in a depth image that are used to compute the direction of a depth transition in an area of a depth image.

Figure 2.2 (a-d) shows the basic feature types that are employed by the detector, and their associated number of constituent regions $d(i)$. The green rectangles represent the rectangular areas $A_n(i)$ and the blue rectangles represent the rectangular areas $B_n(i)$ as defined in eq. 2.3.1. Both are used for the computation of the RC features. In Figure 2.2, the red dot represents pixel location $P(x, y)$. The basic feature types enable the detector to calculate straightforward depth transitions in horizontal, vertical, diagonal and anti-diagonal orientations. Variations derived from the basic feature types result in specialised feature types, which are able to encode more complex *local* depth transitions (Figure 2.2, e - h), or *global* depth transitions (Figure 2.2, i - o).

We return to this topic in Chapters 3 and 5, but we already now provide the following forward pointers. Figure 2.2a shows an example of a basic feature

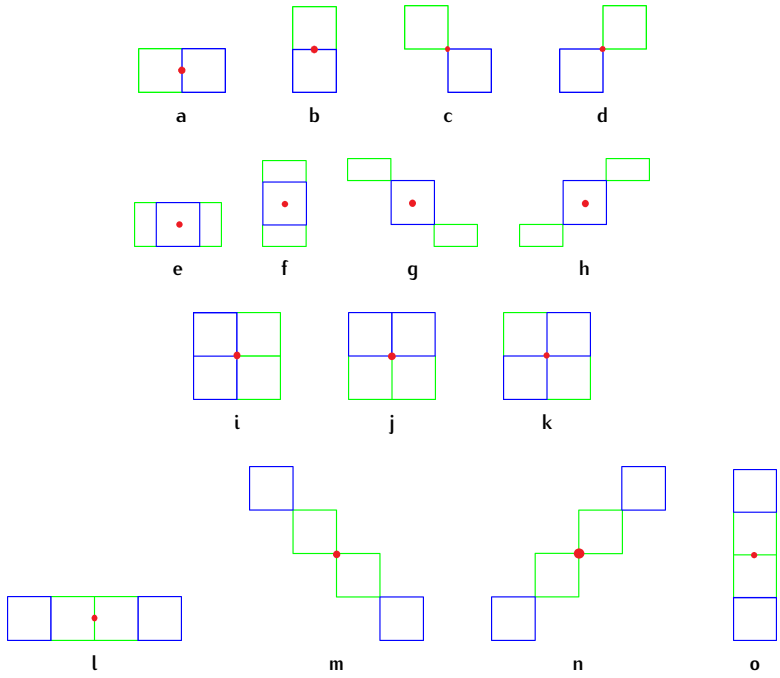



**Figure 2.2:** An enumeration of the Region Comparison (RC) feature types. The red dot indicates the pixel location in a depth image. The green and blue rectangles in each feature type represent the rectangular areas (regions) over which the RC features are computed. The basic feature types (a - d) allow for the computation of (a) horizontal, (b) vertical, (c) diagonal, and (d) anti-diagonal depth transitions. Combining several basic feature types results in specialised features types (e - o), which are able to encode more complex *local* depth transitions (e - h), or *global* depth transitions (i - o). The resulting feature values thus provide (1) an indication of the direction and (2) the magnitude of the depth transition over an area around the pixel location in a depth image.

type (represented as green rectangle and a blue rectangle; $d(i) = 1$), while Figure 2.2e shows an example of an specialised feature type (represented by two green rectangles $A$ and two blue rectangles $B$; $d(i) = 2$). The majority of the feature types (i.e., the ones shown in Figure 2.2, a - d, and i - o) consist of square rectangles of dimensions $r \times r$ (width $\times$ height), which results in $r^2$ pixel values per rectangle. We do note, however, that the derived variations (i.e., the ones shown in Figure 2.2, e - h) may include rectangles of alternative

width/height ratios. In those cases, the rectangles are created with dimensions $(0.5 \times r) \times r$ (Figure 2.2 e), or $r \times (0.5 \times r)$ (Figure 2.2, f - h).

Given feature type $i$, the spatial dimensions (see Definition 2.3) of the area over which the feature value is computed are defined by (1) the number of rectangular regions $d(i)$, and (2) the dimensions $r$ of the individual regions.

**Definition 2.3: Spatial dimensions**

The spatial dimensions of a feature type are defined as the dimensions of the two-dimensional area over which the depth transition is calculated. It is important to take into account that more than two rectangles can be used to enclose a region.

If a feature type consists of a number of small rectangles, it typically encodes for local depth transitions in a depth image. Similarly, feature types that are defined by means of large rectangles allow for the computation of depth features over larger areas, i.e., global depth transitions. Calculating local depth transitions is highly relevant for the detection and classification of small body parts (e.g., the individual fingers of a hand), while calculating global depth transitions is relevant for the recognition of larger body parts (e.g., a head, shoulder, or arms). Hence, feature types such as the ones shown in Figure 2.2 (e - h) are suitable to detect the local depth transitions that are associated with small objects, e.g., the fingers, while the feature types shown in Figure 2.2 (i - o) are suitable to detect global depth transitions, which are associated with larger objects, e.g., the head.

### 2.3.3 Feature Vector

Given a pixel location $P(x, y)$ in a depth image, the features for this point are calculated over the course of several iterations. In each iteration, the features are computed using the feature types as defined in Figure 2.2. The procedure results in a series of feature values - one feature value for each feature type employed in the iteration - which are concatenated in a feature vector (see Definition 2.4 on the next page). The feature types used to compute the features incorporate rectangular regions that enclose multiple pixels per rectangle. With each new iteration, the dimensions of the rectangles are increased: $r = \{1, 2, ..., r_{max}\}$. The feature vectors created after each iteration are then concatenated in the final RC feature vector. It provides an indication of (1) the orientation and (2) the extent of the depth differences in the near vicinity, as well as at a larger spatial distance around point $P$ (see Subsection 2.3.1). Cal-

culating the sum of the rectangular areas for all possible rectangle sizes up to $r_{max}$ can be done efficiently using the integral image representation.

**Definition 2.4: Feature vector**

A feature vector is defined as a collection of feature values. It provides a mathematical description of the direction and magnitude of the depth transitions in a region of a depth image.

## 2.4 RELATED WORK

The RC features deal effectively with background noise, without requiring additional computational power. They relate to several contributions in the fields of image refinement, computer vision and image understanding. In what follows, four related approaches are discussed. We characterise them briefly as methods that (1) actively counteract background noise in depth data, (2) extend the Viola-Jones detector, (3) propose generalisations of Shotton et al.'s method, and (4) incorporate the method proposed by Shotton et al. (2013a,b; 2011).

First, several approaches aiming to counteract background noise in depth data include advanced depth image filters or other refinement techniques (see, e.g., Vijayanagar et al., 2014; Wang et al., 2014; S. Liu et al., 2013). Although image refinement is likely to improve the quality of the input depth data, it comes at the cost of computational power. This may influence the prediction time negatively. An interesting approach was presented by Fanello et al. (2014) in the form of their 'filter forests'. Using location-dependent adaptive filters, their approach can be used to refine the quality of depth images. Such filters are computationally demanding and therefore not suitable for our goals. Inspired by their approach, our RC features incorporate a more straightforward - and computationally less demanding - way to filter noisy depth images.

Second, Nanni et al. (2014) aim to detect human faces by applying the well-known Viola-Jones detector (Viola & Jones, 2001) to visual (RGB - Red Green Blue) images. Aligned depth images are then used to validate the detection results. Although this approach does not deploy depth data as its main data source, using the Viola-Jones detector in this context provides an interesting element. Inspired by Nanni et al. and Viola and Jones, our approach incorporates Haar-like features (see, e.g., Lienhart & Maydt, 2002; Viola & Jones, 2001) to detect objects in depth images.

Third, the face-detection method proposed by Fanelli, Dantone, Gall, Fossati, & Van Gool (2013) operates on large, randomly selected patches in depth images (typically the size of a face), rather than on individual pixels (cf. Shotton et al, 2013b). Their method includes a decision forest for the automatic labelling of the patches. Using patches instead of individual pixels makes the method less prone to noise. Fanelli et al. suggest that using the integral image representation (Crow, 1984) of a depth image (rather than the depth image itself) may facilitate an efficient evaluation of the patches in the decision forest. Inspired by their suggestion, the RC features aim to describe individual pixel locations by computing depth comparison features over patches of various dimensions. The RC features can therefore be seen as a generalisation of the patch-based method by Fanelli et al. Contrary to the randomly selected patches proposed by Fanelli et al., the RC features provide an indication of the direction and the magnitude of depth transitions in a depth image. The RC features include small and large patches of depth images through a decomposition of the integral depth image. This ensures an efficient feature computation process, which may therefore result in short prediction times.

Fourth, Buys et al. (2014) incorporate the pixel-based depth comparison features that are proposed by Shotton et al. in their sophisticated method to detect human bodies and to estimate their pose in single depth images. They label pixels using a randomised decision forest classifier (Breiman, 2001). To deal with the noisy labels generated by their decision forest (which are partly due to the noisy nature of the individual pixels), Buys et al. perform a smoothing procedure on the pixel labels by means of a mode blur filter. In agreement with Buys et al., we acknowledge the importance of smoothing depth data to counteract the noise contained in depth images. In contrast to Buys et al.'s method for pixel comparison, the RC features do not require explicit smoothing. Instead, the RC features perform an implicit smoothing procedure by integrating over depth image regions of varying dimensions, rather than relying on individual pixels. Lacking the need for a post-hoc smoothing procedure is likely to contribute to the efficiency of our approach.

## 2.5 CHAPTER CONCLUSIONS

A requirement for effective object detection algorithms is that they are insensitive to background noise. Whereas visual data is sensitive to illumination conditions, depth data may provide a robust alternative. The main disadvantage of depth data, however, is the low quality and resolution of the depth images. Thus, so far depth images suffer from high levels of background noise. State-of-the-art approaches, such as the work by Shotton et al. (2013b), are sensitive to the background noise in depth data. This calls for improved feature com-

putation approaches that are able to deal efficiently with the noisy nature of depth images.

To answer RQ 1: *How can we improve Shotton et al.'s body part detector in such a way that it enables fast and effective body part detection in noisy depth data?*, I proposed a novel idea in this Chapter, viz. the Region Comparison (RC) features for robust object detection. The RC features provide an indication of (1) the direction and (2) the magnitude of depth transitions in an area of a depth image by comparing regions in a depth image rather than individual pixel values pairs (cf. Shotton et al., 2013a). Based on the theoretical description given in this Chapter, we may formulate the following Chapter conclusions.

- Conclusion 1: Comparing regions has a clear advantage over comparing individual pixel values in that comparing regions allows for averaging over larger areas.
- Conclusion 2: From Conclusion 1 we may conclude that our RC features are less prone to local pixel noise than the PC features.
- Conclusion 3: The RC features do not need an additional computational budget.

Whereas other attempts towards improved object detection required an increase in the computational budget available, the RC features aim to improve object detection without requiring additional computational budget. This can be achieved by calculating the RC features over the integral depth image, rather than over the depth image itself.

### Research Continuation

To investigate to what extent RC features contribute to fast and effective object detection in noisy depth images, the next Chapter presents a comparative evaluation of the RC features on three challenging object detection tasks. In the evaluation, the performance of the RC features is compared with the performance of the original approach used by Shotton et al. (2013a,b; 2011).

# 3 THROUGH THE LOOKING GLASS

*"Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!"*

– Lewis Carroll, *Alice Through the Looking Glass*

This Chapter[13] aims to present a comparative evaluation of the RC features on three challenging object detection tasks. To evaluate the results, the performance of the RC features is compared with the performance of the pixel-based depth comparison features that are proposed by Shotton et al. (2013a,b; 2011).

The course of this Chapter is as follows. First, Section 3.1 outlines the second research question and its evaluation procedure. Subsequently, Section 3.2 presents the *region comparison detector* which incorporates our RC features. Section 3.3 describes the procedure followed to evaluate the performance of the Region Comparison detector, after which Section 3.4 presents the results of our evaluation. The implications of the results are discussed in Section 3.5. Finally, Section 3.6 concludes upon our contribution and answers our second research question.

## 3.1 EVALUATING THE RC FEATURES

As mentioned in Conclusion 1 of Chapter 2, the Region Comparison features average over regions (i.e., large groups of pixels) in a depth image. However, this may counteract the image's background noise. Averaging over larger regions may result in a loss of spatial precision. Thus, RC features may be less sensitive to subtle depth differences. Yet, the RC features aim to prevent the loss of spatial precision by (1) averaging over large regions, which makes the

13 This Chapter is based on work by R. J. H. Mattheij, K. Groeneveld, E. O. Postma, and H. Jaap van den Herik (2016); Depth-Based Detection with Region Comparison Features. Published in the *Journal of Visual Communication and Image Representation (JVCI)*.

features insensitive to local pixel noise (see Conclusion 2 of Chapter 2), and (2) taking individual pixel pairs (i.e., small regions) into account, which allows the features to measure subtle local depth differences. Still, it is unclear to what extent RC features enable fast and accurate body part detection. To this end, the research question addressed in this Chapter (RQ 2) reads as follows.

> RQ 2: *To what extent do Region Comparison features enable fast and accurate face and person detection in noisy depth images?*

To answer this research question, we first define a body part detector that incorporates our RC features. Then, we compare its performance to a detector featuring Shotton et al.'s (2013a,b) Pixel Comparison (PC) features (see Subsection 2.1.3). In a comparative evaluation of the RC and PC features, both associated detectors are trained and evaluated on three challenging object detection experiments: two face detection tasks and a person detection task. There are two evaluation criteria. The first evaluation criterion is classification performance (see Definition 3.1). The second evaluation criterion is computational efficiency (as defined in Definition 3.2).

**Definition 3.1: Classification performance**

Classification performance is defined as the extent to which a detector is able to accurately detect objects in depth data.

**Definition 3.2: Computational efficiency**

Computational efficiency is defined in terms of the time required to process an entire depth image.

Both criteria are included in the definition of superiority of XX features over the YY features (see Definition 3.3), where XX and YY represent given feature sets (i.e., RC features or PC features). A higher classification accuracy corresponds to a higher classification performance, while a shorter processing time therefore corresponds to a higher computational efficiency. They are assessed to ensure that improvements in accuracy do not lead to insurmountable computational costs that prohibit real-time operation.

**Definition 3.3: Superior features**

XX features are defined to be superior to YY features when the detector incorporating the XX features outperforms the detector featuring the PC features on evaluation criterion 1, i.e., classification performance, and performs equally well or better on evaluation criterion 2, i.e., computational efficiency.

## 3.2 THE REGION COMPARISON DETECTOR

To investigate the effectiveness of the Region Comparison features, we define the *region comparison detector*, an object detector (see Definition 3.4) that incorporates our RC features for effective body part recognition tasks. It detects body parts in depth data by classifying individual pixel locations in a subset of random pixel locations, i.e., a point cloud (see Definition 3.5) as either belonging to an object (e.g., a face) or to the background.

**Definition 3.4: Object detector**

A detector is defined as a mechanism that detects objects by classifying parts of a depth image (e.g., individual pixel locations in a point cloud) as either belonging to an object (e.g., a face) or to the background.

**Definition 3.5: Point cloud**

A point cloud is defined as a subset of data points (i.e., a set of pixel locations) that is extracted from a depth image

The detector consists of two stages: (1) a *pre-processing* stage to compute our RC features for the individual pixel locations of the point cloud, and (2) a *classification* stage that uses a random decision forest classifier to predict the corresponding labels of the pixel locations. The labelled point cloud forms the final output of the classifier. Figure 3.1 shows a diagram of the region

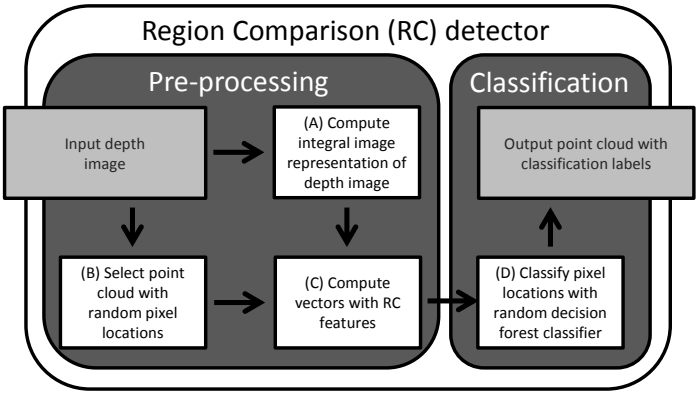



**Figure 3.1:** A diagram of the region comparison detector showing its pre-processing and classification stages (represented by grey rectangular areas), and the constituent sub-stages (represented as white boxes).

comparison detector. The pre-processing and classification stages are defined in Subsections 3.2.1 and Subsection 3.2.2, respectively. They are represented by dark grey, rectangular areas in Figure 3.1. Their constituent sub-stages (A to D) are represented by white boxes.

### 3.2.1 Pre-processing stage

The image pre-processing stage (see Definition 3.6) is the stage in which the input depth image is prepared for the classification process. In what follows, the sub-stages of the pre-processing stage are discussed in detail.

**Definition 3.6: Image pre-processing**

Image pre-processing is defined as the process that prepares an input image for the classification tasks, e.g., by extracting features from the input image.

First, the integral image representation of the input depth image is computed (sub-stage A, see Definition 3.7): the *integral* depth image. Then, a point cloud (see Definition 3.5) of pixel locations is selected at random from the input depth image (sub-stage B).

**Definition 3.7: Integral image representation**

The integral image representation is defined as an alternative image representation form in which an image is represented as the summed area table of the image. The integral image allows for an efficient computation of the sum of values in a rectangular subset of an image.

The advantage of selecting a subset of random pixel locations from the input image (as proposed by Shotton et al. (2011) is twofold: (1) it allows for the detection of partially occluded objects, and (2) it reduces the time required to process an entire depth image.

After selecting the pixel locations, the detector computes multiple RC features for each individual pixel location in the point cloud. The features are then combined into a single RC feature vector (see Definition 2.4), which provides a mathematical description for that particular pixel location (sub-stage C). The set of feature vectors (i.e., a single feature vector per pixel location in the point cloud) forms the input for the detector's classification stage. To extract the RC features for a pixel location, the sum of the pixel values enclosed in a region around that very pixel location is computed. This can be achieved efficiently by calculating the integral image representation of the depth image.

Figure 3.2 (a – d) shows an example of a visual image (Figure 3.2a) of a person, and the corresponding depth image (Figure 3.2b). Figures 3.2c and 3.2d show examples of RC feature types yielding a response, i.e., a depth transition over regions, for three randomly selected pixel locations (two in Figure 3.2c and one in Figure 3.2d). The red dot represents a pixel location, while the spatial positions of the green/blue rectangles (in these Figures also indicated by the capital letters A and B, respectively) represent the regions and direction over which a depth transition is measured; for two straightforward depth transitions (see Figure 3.2c), and for a more complex depth transition (see Figure 3.2d).

As stated in Subsection 2.3.2, the spatial dimensions of the area (over which the RC features are computed) are defined by (1) the number of rectangular regions d, and (2) the dimensions r of the individual regions. As such, feature types which consist of (a larger number of) large regions typically encode for more global depth transitions in a depth image, which are associated with larger body parts (e.g., a head, shoulder, or arms), or an entire person. Hence, the RC features for the region comparison detector are defined as a total of 11 different feature types, i.e., a combination of four basic feature types (see Figure 3.3, a – d) and seven specialised feature types (see Figure 3.3, e – k). In this Figure, the red dot indicates the pixel location in a depth image. The

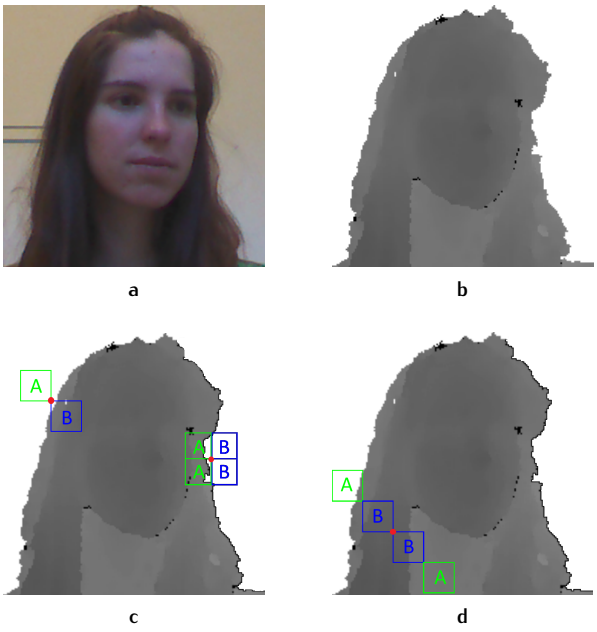



**Figure 3.2:** An example of a visual image (a) of a person, and the corresponding depth image (b). Two straightforward depth transitions are given in Figure 3.2c. A more complex depth transition is seen in Figure 3.2d. Additional information is given in the text.

green and blue rectangles in each feature type represent the rectangular areas (regions) over which the RC feature is computed. The basic feature types (a - d) allow for the computation of (a) horizontal, (b) vertical, (c) diagonal, and (d) anti-diagonal depth transitions. Combining several basic feature types results in specialised feature types (e - k), which are able to encode more complex (global) depth transitions.

### 3.2.2 Classification

In the classification stage, a random decision forest (RDF; cf. Breiman, 2001) is used to classify the RC feature vectors that are computed for the pixel locations in the point cloud (sub-stage D in Figure 3.1). After classifying a feature vector, the RDF maps a class label (OBJECT, NO OBJECT) onto the corresponding pixel location in the point cloud. The labelled point cloud forms the final output of the classifier. Given the output of the classifier, groups of pixel locations with similar labels provide an indication of the presence and location of a per-

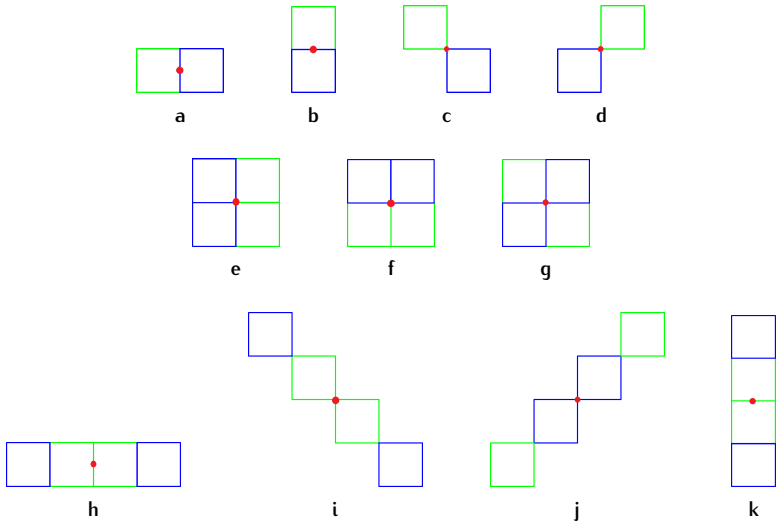



**Figure 3.3:** The 11 Region Comparison (RC) feature types that are deployed in our head and person detection tasks. The feature types are defined as a combination of four basic feature types (a – d), and seven specialised feature types that aim to describe global depth transitions (e – k). The explanation of the feature types is given in the text.

son's body parts in the image. In what follows, the classification algorithm is described briefly.

RDF classifiers are fast and effective multi-class classifiers that typically deploy an ensemble ("forest") of slightly different decision trees. They are suitable for various supervised machine-learning tasks, such as object classification tasks (see, e.g., Chang & Nam, 2013; Criminisi, Shotton, & Konukoglu, 2012). Each individual tree in an RDF classifier consists of multiple binary split nodes and leaf nodes. Individual split nodes compare single features from the feature vector with a threshold, branching left or right depending on the outcome of the comparison. The leaf nodes of the trees contain the prediction results. In a forest, the predictions of all constituent decision trees are averaged to obtain the final classification.

To grow the trees of a RDF classifier, each individual split node in a tree selects a random subset of features taken from the collection of candidate features from the training set. The number of features to select at random is (by default) the square root of the number of candidate features per pixel location. The best splitting candidate, i.e., the feature that best separates the subset of training examples, is selected as the split node's threshold. A tree

can be grown until each leaf node contains a limited number of observations, hence pruning the trees is not necessary.

Figures 3.4, 3.5, and 3.6 show a total of six examples from our test set, in which the RDF of the region comparison detector classified individual pixel positions as belonging to a head (see Figures 3.4 and 3.5) or person (see Figure 3.6). In these examples, a green dot represents a pixel location that is correctly classified as belonging to a head or a person. The examples show that groups of pixel locations with the same labels reveal the location of a person's head or body.

## 3.3 EVALUATION PROCEDURE

This Section describes the experiments performed to evaluate the performance of the RC features. In our evaluation procedure, we compare the performance of our Region Comparison (RC) features with a variant based on Shotton et al.'s (2013a,b) Pixel Comparison (PC) features. Thus, we use the same detector; one version uses the RC features, while the other version uses the PC features. The latter version is called the *pixel comparison detector*.[14] The aim of our experiments is to investigate to what extent our RC features enable fast and effective face and person detection in noisy depth images, as compared to the PC features. To perform our evaluation, the region comparison detector (see Section 3.2) and the pixel comparison detector are trained and evaluated on three quite different challenging datasets with noisy depth data (see below). Our ambition is to investigate (a) the difference between body part detection and person detection in depth images, and (b) the difference between object detection in smoothed and non-smoothed depth data. To satisfy our ambition we have compressed the number of experiments to three: two face detection tasks (smoothed and non-smoothed depth data) and one person detection task. As stated in Section 3.1, the evaluation investigates (1) the classification performance, and (2) the computational efficiency (i.e., the prediction speed) of the detectors.

The remainder of this Section is as follows. Subsection 3.3.1 describes the datasets that are used in the experiments, together with the criteria that we apply for the comparison of the experimental results concerning the PC and RC features. Then, we give the implementation details of both detectors in Subsection 3.3.2. Subsequently, we describe the experiments performed in

14 As there was no version of Shotton's algorithm available for academic purposes, we built an implementation of the body part detection algorithm as described in Shotton et al. (2013a,b; 2011).

Subsection 3.3.3 and the application of the criteria employed to evaluate their performances in Subsection 3.3.4.

### 3.3.1 Datasets and Criteria

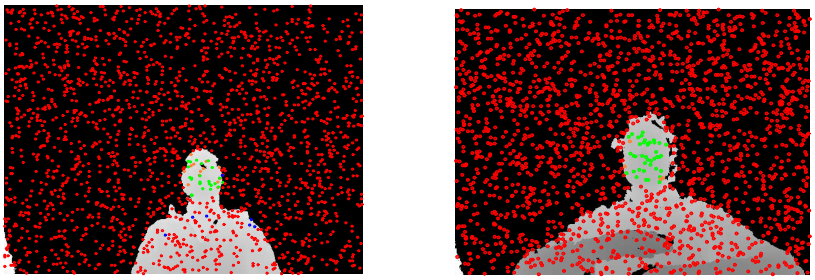
To assess to what extent the RC features are able to deal effectively with background noise in depth images, the region comparison detector and the pixel comparison detector are trained and evaluated on the following three publicly available databases with depth images.

1. *Biwi Kinect Head Pose Database* by Fanelli et al. (2013).
2. *RGB-D Face Database* by Høg et al. (2012).
3. *RGB-D People Dataset* by Spinello and Arras (2011).

The databases vary in (1) the amount of background noise (i.e., smoothed background and non-smoothed background), and (2) the objects captured in the depth data, i.e., human faces or entire humans. Figures 3.4, 3.5, and 3.6 show a total of six examples from the databases, in which the region comparison detector classified individual pixel positions as belonging to a head (see Figures 3.4 and 3.5) or person (see Figure 3.6). In these examples, a green dot represents a pixel location that is correctly classified as belonging to a head or a person, while a red dot represents a pixel location that is (correctly) dismissed by the detector. Orange dots indicate false negative predictions, while blue represents the false positive ones. Below, the datasets are reviewed briefly.

BIWI KINECT HEAD POSE DATABASE is developed by Fanelli et al. (2013). The dataset contains over $15,000$ visual (RGB) and depth (D) images of people with various head poses sitting in front of a Kinect device. It provides annotations in the form of masks that indicate the location of a person's face in a depth image. The masks use logical flags to indicate whether a pixel location belongs to a face or not. All depth images in this database have an image resolution of $640 \times 480$ pixels. The background of the depth data is removed using a threshold on the distance. The depth values are rescaled to an interval with values ranging from $0$ to $4,095$ (both inclusive). Removing the background is likely to result in a reduction of the amount of background noise in the depth images. On the next page, Figure 3.4 shows two examples from the *Biwi Kinect Head Pose Database*.

RGB-D FACE DATABASE is developed by Høg et al. (2012). The dataset contains $1,581$ visual (RGB) and depth (D) images of the heads and shoulders of human participants in various poses and with different facial expressions. As no annotations were provided for this database, each depth image in the subset



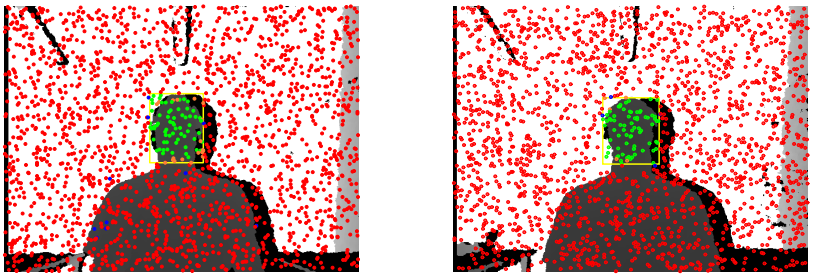
a b

**Figure 3.4:** Two examples of the classification results that are achieved by the region comparison detector on test images from the first head detection task, i.e., face detection in smoothed depth images.

was manually annotated[15] by selecting a rectangular area that encloses the person's face in the depth image. The boundaries of the annotation area were aligned with the left, top and right side of the face, and the lowest point of the person's lower jaw. Similar to the *Biwi Kinect Head Pose Database*, the image resolution of the depth images is $640 \times 480$ pixels. The depth values of the depth images range from 0 to $4,095$ (both inclusive). The background of the depth data is left intact. Figure 3.5 shows two examples of depth images from this database.

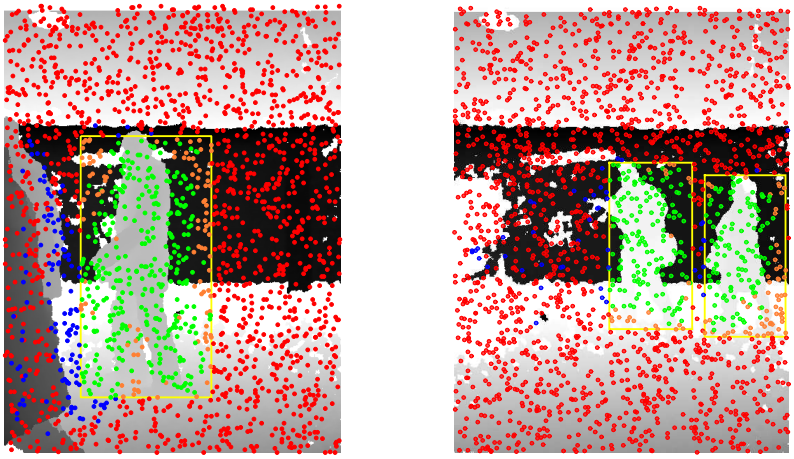
RGB-D PEOPLE DATASET is developed by Spinello and Arras (2011). The dataset contains over $3,000$ visual (RGB) and depth (D) images of mostly upright walking and standing people in a populated indoor environment, seen from different orientations and with different degrees of occlusions. This dataset is acquired in a university hall using three vertically mounted Kinect devices. In total, the dataset contains $1,133$ annotated depth images. As the Kinect devices used in this experiment were mounted vertically, the depth images in the dataset are rotated 90 angular degrees. Hence, the image resolution of the depth images is $480 \times 640$ pixels. The maximal distance between the Kinect device and the hand of the subject is 1.0 meter. The annotations consist of rectangular bounding boxes enclosing a person's body. Similar to the *RGB-D Face Database*, the background of the depth data is left intact. The depth values, however, are rescaled to an interval with values ranging from 0 to $4,095$ (both inclusive). Figure 3.6 shows two examples of depth images from this dataset.

15 The annotations for this database are available upon request from the author.



a b

**Figure 3.5:** Two examples of the classification results that are achieved by the region comparison detector on test images from the second head detection task, i.e., face detection in non-smoothed depth images.



a b

**Figure 3.6:** Two examples of the classification results that are achieved by the region comparison detector on test images from the person detection task, i.e., person detection in non-smoothed depth images.

For a proper evaluation of the performances of the PC features and the RC features, we use two evaluation criteria. They are identified on the next page.

CRITERIA The performance of the detectors will be quantified using two performance metrics: (1) a classification performance metric to report on the average per-class segmentation accuracy, and (2) a computational efficiency metric to measure the time required by a classifier to process an entire image. In our evaluation RC features are considered to outperform PC features when they achieve a higher average classification performance (evaluation criterion 1), without incurring an additional cost in terms of detection speed as compared to the PC features (evaluation criterion 2). Thus, we consider a feature set to be superior when the detector incorporating the RC features outperforms the detector featuring the PC features on evaluation criterion 1 and performs equally well or better on evaluation criterion 2. The performance metrics are defined in Subsection 3.3.4 together with their application on the results of the experiments.

### 3.3.2 Implementation Details

The experiments are described in Subsection 3.3.3. Unless specified otherwise, four types of parameters are used, viz. for (1) the selection of the random pixel locations and spatial search area, (2) the RC and PC features, (3) the RDF classifier, and (4) the implementation of the detectors.

THE SELECTION OF THE RANDOM PIXEL LOCATIONS AND SPATIAL SEARCH AREA For each depth image, a subset of $2,000$ random pixel locations is selected, for which the RC and PC features are computed. To ensure a fair comparison between both feature computation methods (i.e., RC vs. PC), both methods operate on exactly the same pixel locations.

The maximal dimensions of the spatial search area over which the PC detector computes its features are $150 \times 150$ pixels. The PC detector normalises the dimensions of the spatial search area based on the distance (depth value) at point $P(x, y)$. As a result, the search area is small for objects far from the Kinect device (high depth value at point $P(x, y)$) but large for objects close to the Kinect device (low depth value at point $P(x, y)$).

The spatial search area over which the RC features are computed is the same as the maximal (i.e, not normalised) search area used by the PC features. As such, the maximal dimensions of the rectangles incorporated by the RC features are $38 \times 38$ pixels. As the feature types with the largest spatial dimensions deploy 4 rectangles (positioned horizontally, vertically, or (anti)-diagonally next to each other), the resulting search area is $(4 \times 38) \times (4 \times 38) \approx 150 \times 150$ pixels. Contrary to the PC features, the search area used for the RC features is not normalised for the distance.

THE RC AND PC FEATURES The rectangle size parameter $r$ for the feature types that are used to compute the RC features, is defined as an integer value that increases with each iteration. In the first iteration, the value of $r$ is initiated at 1. After each iteration, the value of $r$ increases with step size 1, up to its maximum value of 38. Hence, the value of $r$ over the iterations is defined as: $r = \{1, 2, 3, ...38\}$. The resulting RC feature vectors may at most contain $11 \times 38 = 418$ unique elements for each pixel location. The parameters employed to compute the PC features are as specified in Shotton et al. (2013b). The resulting PC feature vectors contain $2,000$ unique elements for each pixel location.

THE RDF CLASSIFIER For the experiments, the MATLAB implementation of the random decision forest (the so-called "TreeBagger"[16]) is used. For the RC features, each split node of the forest selects a random subset of $\sqrt{418} \approx 20$ candidate features. For the PC features, each split node of the forest tests $\sqrt{2,000} \approx 44$ candidate features to find the best splitting threshold. Each tree of the random decision forest is trained until a minimum number of one observation per tree leaf is reached. The trees are not pruned.

THE IMPLEMENTATION OF THE DETECTORS Both detectors are implemented in MATLAB scripts. The implementations of the detectors are available upon request from the author. The entire training and evaluation procedure takes several days on a 50-core Linux calculation server.

### 3.3.3 Experiments

To evaluate the performance obtained by the RC features (as compared to the PC features), we perform three classification experiments. In the experiments, we train and evaluate the performance of both feature types (RC and PC) using publicly available databases. In the first experiment, the *Biwi Kinect Head Pose Database* (see Fanelli et al., 2013) is used to detect human faces in smoothed depth images. In the second experiment, the *RGB-D Face Database* (see Høg et al., 2012) is used to detect human faces in non-smoothed depth images. In the third experiment, the *RGB-D People Dataset* (see Spinello & Arras, 2011) is used to detect entire people in non-smoothed depth images. In what follows, the experimental setup and the individual experiments (experiment 1, experiment 2, experiment 3) are described briefly.

EXPERIMENTAL SETUP While the *Biwi Kinect Head Pose Database* and the *RGB-D Face Database* both contain depth images of annotated human faces, the depth images in the first database are likely to contain less background noise

16 http://nl.mathworks.com/help/stats/treebagger.html

than the depth images in the latter database. This is due to the removal of the background of the depth images in the *Biwi Kinect Head Pose Database*. Thus, training and evaluating the detectors on these databases in the first two experiments provides an indication of the extent to which the RC features are able to deal effectively with background noise. Compared to the face detection tasks, detecting an entire human is likely to be a more challenging task. To investigate whether our results also extend to more complex detection tasks, we therefore compare the performance of both detectors in the third experiment (the person detection task).

As the feature extraction procedure is computationally demanding (especially during the training procedure of the pixel comparison detector, due to the large feature vectors required for this detector), our experiments are performed using subsets of randomly selected depth images. The resulting subsets are considerably smaller than the original datasets. For all experiments, the generalisation performance is estimated using a 10-fold cross-validation procedure. The datasets used in our experiments are partitioned into separate training sets and test sets. In our experiments, the complexity of the RDF classifiers is not optimised prior to the experiment. Hence, we did not create a validation set. Moreover, the PC features are not optimised beyond the parameters provided by Shotton et al. (2013a,b; 2011). The RC features are optimised to match the parameters of the PC features, which ensures a fair comparison between both feature computation approaches.

EXPERIMENT 1 Experiment 1 deals with face detection in smoothed depth images. For the experiment, a random subset of 100 depth images is selected from the *Biwi Kinect Head Pose Database*. Using the 10-fold cross-validation procedure, individual folds are created that each consist of 90 training images and 10 test images. Inspired through the work by Shotton et al. (2011), a subset of 2,000 random pixel locations is selected from each depth image, and labelled in a binary fashion, i.e., FACE when a pixel is located within the region annotated as belonging to the face, or OTHER in all other cases. For each fold, the resulting dataset of experiment 1 consists of 180,000 training examples (pixels) and 20,000 test examples. For each individual training example, the RC and PC vectors are computed. Combined with the labels, the training examples are used to train RDF classifiers, with forests ranging from 1 tree up to 10 trees (both inclusive). The test examples and the corresponding labels are used to assess the generalisation performance of the detectors.

EXPERIMENT 2 Experiment 2 deals with face detection in non-smoothed depth images. The underlying idea of experiment 2 is as follows. We expect that RC features are suitable to deal with noise in depth data. Removing the background in depth images, however, would also reduce the amount of noise

in the depth data, which may influence the performance of the RC and PC features. Experiment 2 is performed on the *RGB-D Face Database*. Contrary to experiment 1, the current database contains depth images from which the background is not removed.

For experiment 2, a random subset of 93 depth images is selected. After performing 10-fold cross-validation, the resulting folds consist of 84 training images and 9 test images. Similar to experiment 1, a subset of 2,000 random pixel locations is selected from each depth image, and labelled accordingly. The resulting dataset contains 168,000 training examples and 18,000 test examples per fold. The training and evaluation procedure for the experiment is the same as in experiment 1.

EXPERIMENT 3 Experiment 3 deals with person detection in non-smoothed depth images. The underlying idea of experiment 3 is as follows. Whereas detecting human faces in depth images in general might be relatively easy due to the fairly consistent shape of the human face, detecting entire humans is likely to be a more challenging task.

For experiment 3, a subset of 100 random depth images is selected from the *RGB-D People Dataset*. The background of the depth images in this dataset is left intact. Similar to experiment 1 and 2, the resulting subset is divided into separate training and test sets using the 10-fold cross-validation procedure. We again select 2,000 random pixel locations from each depth image. Each individual fold therefore consists of 180,000 training examples and 20,000 test examples. Each example is labelled as either PERSON or OTHER. We deployed the same training and evaluation procedure of experiment 3 is the same as in experiment 1 and 2.

### 3.3.4 Performance Metrics

The introduction of the performance metrics took place at the end of Subsection 3.3.1. In what follows, we describe the details of the performance metrics that we use to quantify the classification performance and the computational efficiency of the feature sets. It can be seen as a follow-up on the implementation details (see Subsection 3.3.2).

CLASSIFICATION PERFORMANCE METRIC The detectors that are trained and evaluated in our experiments classify individual pixel locations as either belonging to an object (e.g., a face), or to the background. Given the binary nature of the experiments, we use (1) the *balanced accuracy* (as defined in Definition 3.8), (2) *precision* (see Definition 3.9), and (3) *recall* (see Definition 3.10) as our classification metrics to measure the classification performance of a detector.

As the class distributions of the datasets are highly skewed, i.e., a low percentage of FACE- or PERSON-samples versus a high percentage of OTHER-samples, the performance is likely to be biased towards the most frequent class in a dataset. To deal with the bias, Brodersen, O., Stephan, & Buhmann (2010), and Carrillo, Brodersen, & Castellanos (2014) proposed the use of the *balanced* accuracy as an alternative to the regular accuracy measurement. Thus, to handle the bias in the datasets we use the *balanced accuracy* as our first detection performance metric.

**Definition 3.8: Balanced accuracy**

The balanced accuracy is defined as the arithmetic mean of class-specific accuracies, which considers the recall of the positive and negative class.

**Definition 3.9: Precision**

Precision is defined as the percentage of instances recognized by the detector that are relevant.

**Definition 3.10: Recall**

Recall is defined as the percentage of relevant instances that are identified by the detector.

Moreover, we adopt two additional classification metrics to handle the bias: (4) the *F1-score* (see Definition 3.11, and, for example, the work by Powers (2011)), and (5) the *area under the receiver operating characteristic curve* (AUC) (see Definition 3.12, which is taken from, for example, Omary & Mtenzi (2010)).

**Definition 3.11: F1-score**

The F1-score is defined in terms of the weighted average of the precision and recall values of the positive class

**Definition 3.12: Area Under the Curve**

The Area Under the Curve (AUC) is defined as the relation between the true positive rate and the false positive rate of a classifier given various classification thresholds.

We provide an indication of the complexity of the classifier (see Definition 3.13) by measuring (6) the average number of levels per tree, and (7) the average number of leaf nodes per tree. The average number of levels per tree provides an indication of the (average) number of tests that are performed on each feature vector. Moreover, the combination of (a) the average number of levels, and (b) the average number of leaf nodes provides an indication of the efficiency of the tree. Efficient trees are typically characterised by a low number of levels, yet a relatively high number of leaf nodes.

**Definition 3.13: Complexity**

The complexity of the classifier is defined as the relation between the number of levels and the number of leaf nodes in a tree.

Here we note that in the presentation of the results (see, e.g., Tables 3.1 and 3.3, and Figure 3.7) we deviate from the enumeration of classification metrics provided above. For the Tables, either we use the order (1) balanced accuracy, (2) recall, (3) precision, and (4) F1-score, or we use the full Table for the AUC scores. For the Figures, we use the following order: (a) balanced accuracy, (b) precision and recall, (c) F1-scores, and (d) the AUC scores.

COMPUTATIONAL EFFICIENCY METRIC The computational efficiency metric measures the time required by a detector to process individual depth images (see Definition 3.14). A shorter classification time corresponds to a higher computational efficiency. Thus, the classification times provide an indication of the computational efficiency of a detector.

**Definition 3.14: Prediction time**

The classification time is defined as the time required to pre-process a single depth image, and classify the selected pixel locations.

## 3.4 EXPERIMENTAL RESULTS

In this Section, we describe the results of the experiments performed to evaluate the RC features and the PC features. Subsection 3.4.1 describes the results of experiment 1: the first face detection task, which uses a dataset with depth images from which the background is removed (smoothed) using a threshold. This experiment is considered as the benchmark experiment. Subsequently, Subsection 3.4.2 describes the results of the face detection task with non-smoothed depth images. Then, Subsection 3.4.3 describes the results of the person detection task with non-smoothed depth images.

### 3.4.1 Experiment 1: Face Detection, Smoothed background

The results of the face detection task with smoothed depth images are shown in Figures 3.7 and 3.8, and Tables 3.1, 3.2 and 3.3. In what follows, the results will be discussed in detail.

Figure 3.7 shows the performance for both feature computation methods (RC and PC) for ten sizes (i.e., number of trees per forest) of the RDF classifier for five metrics: (a) the balanced accuracy, (b) precision and recall, (c) F1-scores, and (d) the AUC scores. Please note that Figure 3.7b contains two metrics. Figure 3.8 shows (a) the average tree depth, and (b) the average number of leaf nodes per tree in the classifiers, which provides an indication of the complexity of the classifiers. Table 3.1 shows the minimum and maximum performances for the region comparison detector, while Table 3.2 shows this information for the pixel comparison detector. Table 3.3 shows the AUC values and the associated classification times (i.e., the computation times required to process an entire depth image), for ten sizes of the forest.

Figure 3.4 (see Subsection 3.3.1) shows two depth images from our test set, in which the region comparison detector recognized the location of a persons' head by classifying the pixel locations in the point cloud. In these examples, a green dot indicates a true positive prediction for a given pixel $\mathbf{p}$ $P(x, y)$, while a red dot indicates a true negative prediction. Orange represents false negative predictions, while blue represents the false positive ones. The black background of the images is the (visualized) result of the background removal (smoothing). Figure 3.13a (see Section 3.6) shows the AUC curve for the optimal detection parameters, i.e., a RDF classifier of 10 trees.

The results of the experiment show that the region comparison detector is able to achieve a significantly higher classification performance than the pixel comparison detector. The results also reveal that both detectors approach their optimal classification performance using a random decision forest of rather small dimensions, i.e., a forest consisting of only a limited number of trees (say, three to five). Training additional trees does not affect the balanced accuracy

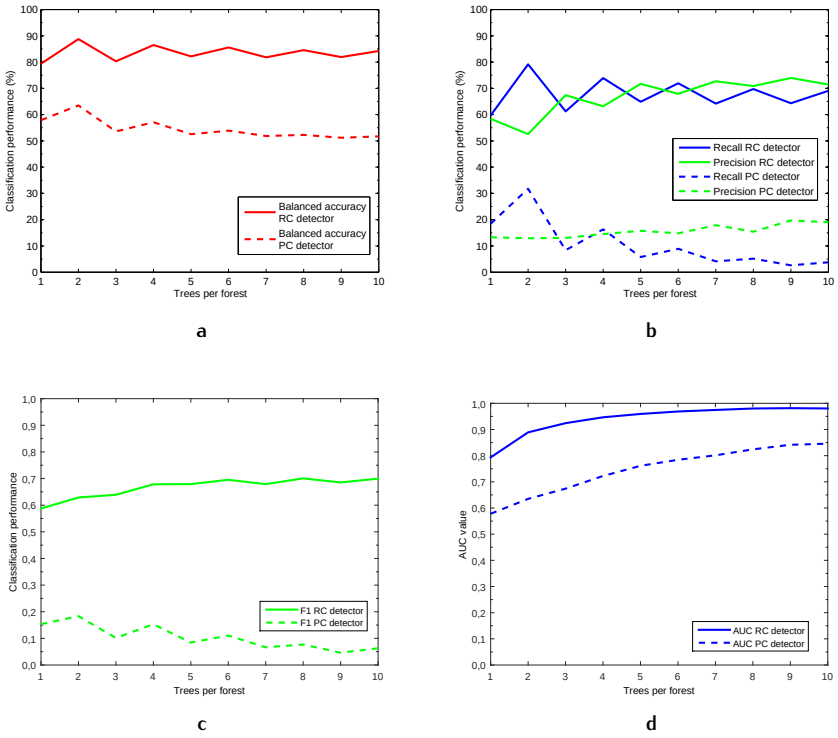



**Figure 3.7:** [Experiment 1: face detection] The classification performance for the first face detection task for ten sizes of the random decision forests: (a) balanced accuracy, (b) precision and recall, (c) F1-scores, and (d) the AUC scores (higher is better). The continuous line represents the performance obtained by the RC features, while the dotted line represents the performance obtained by the PC features. The x-axes of the graphs represent the number of trees in the RDF. The y-axes represent the classification performance. More details and interpretations are provided in the text.

(see Figure 3.7a) of the region comparison detector significantly, although it decreases slightly for the pixel comparison detector. When increasing the number of trees in the forest, recall and precision (Figure 3.7b), and the F1-score (Figure 3.7c) increase slightly for the region comparison detector, while the precision of the pixel comparison shows a slight increase, and even a decrease in recall and F1-score. For both detectors, the Area Under the Curve (AUC) score increases with the size of the forest, approaching its optimal score

**Table 3.1:** [Experiment 1 - RC features] The minimum and maximum scores for the balanced accuracy, recall, precision, and F1-scores obtained by the RC features in experiment 1.

| Performance metric | Min. score (SD) | Max. score (SD) |
|---|---|---|
| Balanced accuracy (%) | 79.4 (1.7) | 88.8 (1.5) |
| Recall (%) | 59.7 (3.5) | 79.1 (3.1) |
| Precision (%) | 52.6 (7.3) | 74.0 (7.4) |
| F1-score | 0.59 (0.04) | 0.70 (0.03) |

**Table 3.2:** [Experiment 1 - PC features] The minimum and maximum scores for the balanced accuracy, recall, precision, and F1-scores obtained by the PC features in experiment 1.

| Performance metric | Min. score (SD) | Max. score (SD) |
|---|---|---|
| Balanced accuracy (%) | 51.2 (0.4) | 63.5 (1.1) |
| Recall (%) | 2.63 (0.9) | 31.8 (2.1) |
| Precision (%) | 12.9 (1.8) | 19.7 (7.3) |
| F1-score | 0.05 (0.01) | 0.18 (0.02) |

using a forest of three trees and five trees for the region and pixel comparison detector, respectively (Figure 3.7d).

The results also show that the trees that are grown for the region comparison detector are significantly smaller than the ones that are grown for the pixel comparison detector. Figure 3.8 shows (in Figure 3.8a) the average tree depth, and (in Figure 3.8b) the average number of leaf nodes per tree for both detectors in experiment 1.

Averaged over all folds and dimensions in the experiment, the trees of the region comparison detector are 13.0 levels deep (SD = 0.21). The trees that are grown for the pixel comparison detector, however, reach an average depth of 17.6 levels (SD = 0.41). Moreover, our results show that the trees of the region comparison detector contain an average of 1,001 leaf nodes (SD = 23), while the trees of the pixel comparison detector contain an average of 2,015 leaf nodes (SD = 23).

These results imply that the random forests that are trained by means of the RC feature vectors, require, on average, a lower number of tests to perform the classification procedure than their pixel comparing counterparts. Our results thus indicate that, compared to the RDF classifiers that are trained by incorpo-

**Table 3.3:** [Experiment 1: face detection] The AUC scores and classification times per image for the both detectors while using RDF classifiers of ten sizes in experiment 1.

| Forest | RC features | | PC features | |
|---|---|---|---|---|
| size | AUC | Pred. time (s) (SD) | AUC | Pred. time (s) (SD) |
| 1 | 0.794 | 1.08 (0.16) | 0.578 | 0.01 (0.00) |
| 2 | 0.889 | 1.38 (0.01) | 0.635 | 0.01 (0.00) |
| 3 | 0.924 | 1.74 (0.01) | 0.674 | 0.01 (0.00) |
| 4 | 0.947 | 2.09 (0.02) | 0.722 | 0.01 (0.00) |
| 5 | 0.959 | 2.45 (0.01) | 0.761 | 0.02 (0.00) |
| 6 | 0.969 | 2.83 (0.02) | 0.784 | 0.02 (0.00) |
| 7 | 0.975 | 3.17 (0.03) | 0.801 | 0.02 (0.00) |
| 8 | 0.980 | 3.52 (0.02) | 0.824 | 0.02 (0.00) |
| 9 | 0.982 | 3.86 (0.02) | 0.842 | 0.02 (0.00) |
| 10 | 0.980 | 4.24 (0.01) | 0.846 | 0.02 (0.00) |

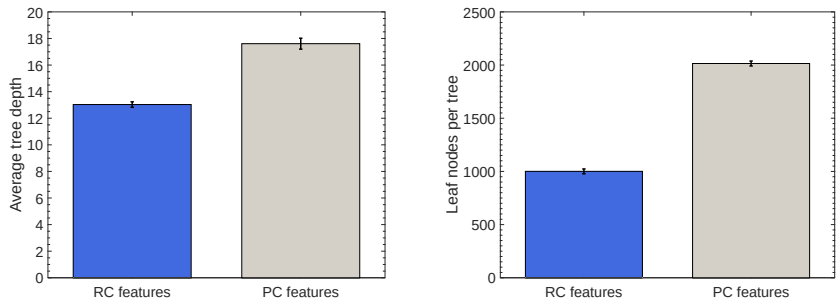



**Figure 3.8:** [Experiment 1: face detection] The bar plot of (a) the average tree depth, and (b) the average number of leaf nodes per tree for the region comparison detector (represented in blue) and the pixel comparison detector (represented in grey), and the corresponding error bars. The results are averaged over all trees and all folds.

rating the PC features, incorporating RC features leads to a lower complexity in the RDF classifiers.

The results of experiment 1 suggest that the RC features enable a significantly higher classification performance than the PC features. However, the average classification times (as shown in Table 3.3) indicate that the region comparison detector requires more time to process a depth image than the pixel comparison detector. This may (partially) be due to the time required to (1) compute the integral image representation, and (2) extract the RC features. While the results of experiment 1 show that RC features enable a higher classification performance in smoothed depth data, their classification performance comes at the cost of computational efficiency.

### 3.4.2 Experiment 2: Face Detection, non-smoothed background

We now investigate to what extent the pattern of results of experiment 1 are related to non-smoothed depth data. As noisy depth data is likely to result in erroneous depth measurements in the feature vectors, accurately separating the feature vectors may become a challenge. Classifiers that aim to separate feature vectors with erroneous feature values require additional tests to achieve an optimal separation of the data. This may lead to an increase in the number of split nodes and leaf nodes, i.e., an increase in the complexity of the classifiers. Increasing the number of tests that are performed on the input feature vectors may influence a detector's classification time negatively. This is assessed in our second face detection experiment, in which both detectors are trained and evaluated on a dataset with non-smoothed depth images.

For the second face detection task, the average classification performances are shown in Figures 3.9 and 3.10, and Tables 3.4, 3.5 and 3.6. The dataset used in the experiment contains depth images from which the background is left intact. In what follows, the results will be discussed in detail.

Figure 3.9 shows (a) the balanced accuracy, (b) precision and recall, (c) F1-scores, and (d) the AUC scores for both feature computation methods. Figure 3.10 provides an indication of the complexity of the classifiers by showing the average tree depth and the average number of leaf nodes per tree in the classifiers. Tables 3.4 and 3.5 show the minimum and maximum performances for the RC features and PC feature, respectively. Table 3.6 shows the AUC values and the associated classification times for ten sizes of the RDF classifier. Figure 3.5 (in Subsection 3.3.1) shows two examples of depth images from our test, in which the region comparison detector (featuring a RDF classifier of 10 trees) classified individual pixel locations. We remark the presence of the "depth shadow", i.e., empty parts in the depth image, on the right side of the person. It is a direct result of a part of a scene that is not illuminated by the laser of the Kinect device, and therefore not captured by its infrared sensor. Consequently, it results in the undefined (empty) regions that are described

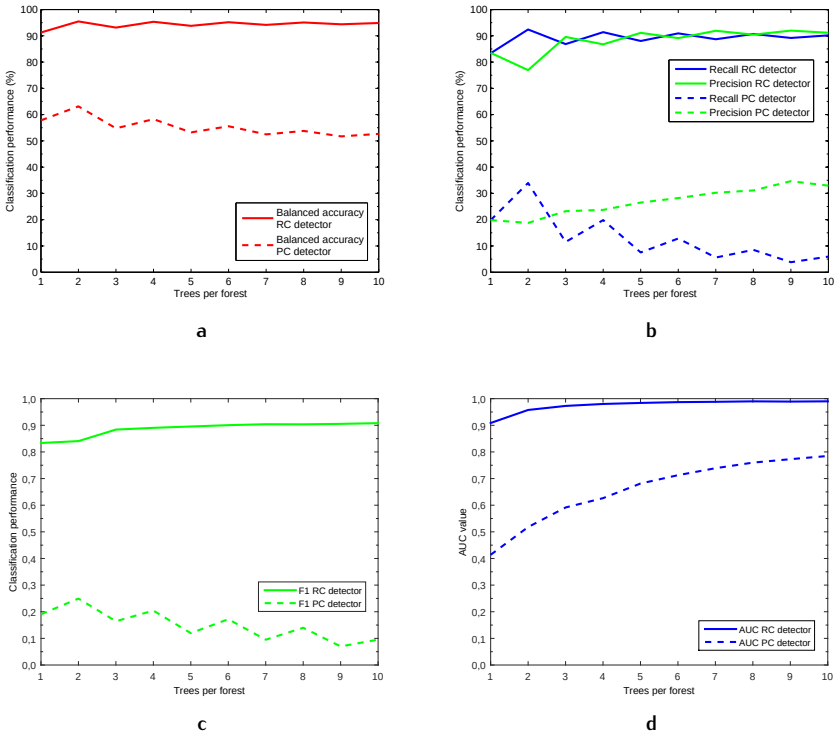



**Figure 3.9:** [Experiment 2: face detection] The classification performance for the second face detection task for ten sizes of the random decision forests: (a) balanced accuracy, (b) precision and recall, (c) F1-scores, and (d) the AUC scores (higher is better). The continuous line represents the performance obtained by the RC features, while the dotted line represents the performance obtained by the PC features. The x-axes of the graphs represent the number of trees in the RDF classifier. The y-axes represent the classification performance. More details and interpretations are provided in the text.

by, for example, (Khoshelham & Elberink, 2012). Figure 3.13b shows the AUC curve for the optimal detection parameters, i.e., a RDF classifier of 10 trees. The results of the experiment (Figure 3.9 and Table 3.4) show that the RC features again achieve a significantly higher classification performance than the PC features (Table 3.5), even though the number of training samples is slightly smaller than in experiment 1. Similar to the results of experiment 1, Figure 3.9 shows that both types of features approach their optimal classification perfor-

**Table 3.4:** [Experiment 2 - RC features] The minimum and maximum scores for the balanced accuracy, recall, precision, and F1-scores obtained by the RC features in experiment 2.

| Performance metric | Min. score (SD) | Max. score (SD) |
|---|---|---|
| Balanced accuracy (%) | 91.3 (2.0) | 95.5 (1.8) |
| Recall (%) | 83.4 (3.9) | 92.4 (3.5) |
| Precision (%) | 76.9 (4.7) | 92.0 (3.9) |
| F1-score | 0.83 (0.03) | 0.91 (0.03) |

**Table 3.5:** [Experiment 2 - PC features] The minimum and maximum scores for the balanced accuracy, recall, precision, and F1-scores obtained by the PC features in experiment 2.

| Performance metric | Min. score (SD) | Max. score (SD) |
|---|---|---|
| Balanced accuracy (%) | 51.7 (0.4) | 63.1 (1.3) |
| Recall (%) | 3.81 (0.9) | 33.9 (2.4) |
| Precision (%) | 18.8 (1.6) | 34.6 (6.0) |
| F1-score | 0.07 (0.02) | 0.24 (0.01) |

mance using a random decision forest of rather small dimensions (again, say three to five). Training additional trees does not affect the accuracy (see Figure 3.9a) of the RC features significantly, although it decreases slightly for the PC features. When increasing the size of the forest, recall, precision (Figure 3.9b), and the F1-score (Figure 3.9c) increase slightly for the RC features. The difference in performances of both feature types is also reflected in their AUC scores (Figure 3.9d).

Compared to experiment 1, the depth data used in experiment 2 contains a much higher amount of background noise. Following the increase in the amount of background noise, our results (see Figures 3.7d and 3.9d) show that the classification performance of the PC features decreases slightly. The performance of the RC features, however, shows a minor increase.

The average classification times (as shown in Table 3.6) indicate that, for experiment 2, the region comparison detector requires less time to process an entire depth image than the pixel comparison detector. In contrast to the results of experiment 1, the region comparison detector achieves a detection speed that is up to 2.5 times faster than the detection speed of the pixel comparison detector.

**Table 3.6:** [Experiment 2: face detection] The AUC scores and classification times per image for the both detectors while using RDF classifiers of ten sizes in experiment 2.

| Forest | RC features | | PC features | |
|---|---|---|---|---|
| size | AUC | Pred. time (s) (SD) | AUC | Pred. time (s) (SD) |
| 1 | 0.912 | 1.36 (0.23) | 0.415 | 7.55 (0.22) |
| 2 | 0.958 | 1.72 (0.02) | 0.507 | 8.17 (0.03) |
| 3 | 0.973 | 2.18 (0.03) | 0.583 | 8.82 (0.07) |
| 4 | 0.979 | 2.63 (0.02) | 0.641 | 9.46 (0.06) |
| 5 | 0.984 | 3.09 (0.02) | 0.680 | 10.1 (0.08) |
| 6 | 0.987 | 3.60 (0.05) | 0.719 | 10.7 (0.07) |
| 7 | 0.987 | 3.98 (0.04) | 0.745 | 11.4 (0.08) |
| 8 | 0.989 | 4.51 (0.04) | 0.761 | 12.1 (0.15) |
| 9 | 0.991 | 4.89 (0.09) | 0.782 | 12.7 (0.12) |
| 10 | 0.991 | 5.44 (0.08) | 0.784 | 13.3 (0.20) |

The difference in detection speed may partially be due to an increase in the complexity of the RDF classifier of the pixel comparison detector. Figure 3.10 shows (in Figure 3.10a on the next page) the average tree depth, and (in Figure 3.10b) the average number of leaf nodes per tree for both detectors in experiment 2.

An analysis of the complexity of the pixel comparison detector reveals that, while the classifier of the pixel comparison detector reaches an average depth of 17.6 levels (SD = 0.41) in experiment 1, the average tree depth of the classifier quadruples to 68.1 levels (SD = 2.36) in experiment 2. This increase is also reflected in the number of leaf nodes of the classifier, as they double from 2,015 (SD = 23) in experiment 1 to an average of 3880 leaf nodes (SD = 58) in experiment 2. The trees grown for the region comparison detector, however, now reach an average depth of 13.1 levels (SD = 0.22) and 997 leaf nodes (SD = 31), which translates to a minimal increase in average tree depth, or even a small decrease in the average number of leaf nodes.

The results imply that increased levels of background noise in depth data result in a significant increase in the number of tests that are performed by the RDF classifier of the pixel comparison detector. However, the results also imply that the RDF of the region comparison detector does not require additional tests to perform its classification task.

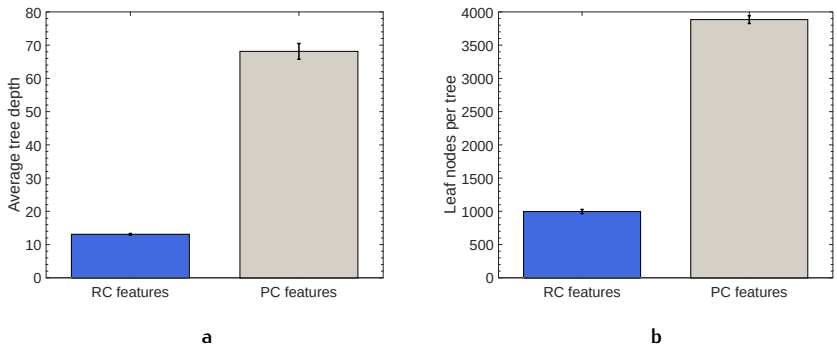



**Figure 3.10:** [Experiment 2: face detection] The bar plot of (a) the average tree depth, and (b) the average number of leaf nodes per tree for the region comparison detector (represented in blue) and the pixel comparison detector (represented in grey), and the corresponding error bars. The results are averaged over all trees and all folds.

The results of experiment 2 show a significant increase in the classification time of the pixel comparison detector, especially when compared to a relatively small increase in the classification times of the region comparison detector. The results show that the complexity of the RDF classifier employed by the pixel comparison detector increases significantly with the level of background noise in the depth data. It indicates that the pixel comparison detector is more sensitive to background noise than the region comparison detector. Thus, the results suggest that the RC features are better suited to handle background noise in depth images.

### 3.4.3 Experiment 3: Person Detection

We now turn to experiment 3 to assess the more complex task of person detection in depth images. While detecting human faces in depth images might be relatively easy, detecting entire humans is likely to be a more challenging task. The experiment explores whether and if so, to what extent, RC features outperform PC features in more complex detection tasks. The average classification performances obtained by RC features and PC features on the person detection task are shown in Figure 3.11 and 3.12, and Tables 3.7, 3.8 and 3.9. The dataset used in the experiment contains depth images with high levels of background noise. In what follows, the results will be discussed in detail.

Figure 3.11 shows the classification performance of both types of features for different sizes of the random decision forest: (a) the balanced accuracy, (b)

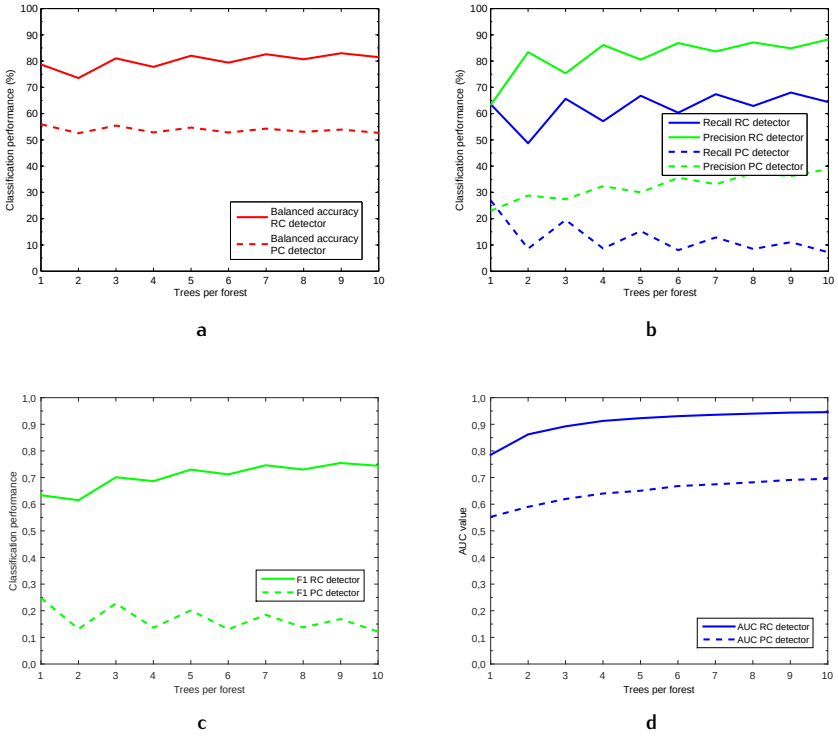



**Figure 3.11:** [Experiment 3: person detection] The classification performance for the person detection task for various sizes of the random decision forests: (a) balanced accuracy, (b) precision and recall, (c) F1-scores, and (d) the AUC scores (higher is better). The continuous line represents the performance obtained by the RC features, while the dotted line represents the performance obtained by the PC features. The x-axes of the graphs represent the number of trees in the RDF classifier. The y-axes represent the classification performance.

precision and recall, (c) F1-scores, and (d) the AUC scores. Figure 3.12 shows (a) the average tree depth, and (b) the average number of leaf nodes per tree in the classifiers, which provides an indication of the complexity of the classifiers. Tables 3.7 and 3.8 show the minimum and maximum performances of the aforementioned performance metrics for the RC features and PC features, respectively. Table 3.9 shows the performances expressed as the AUC of the detectors, versus the time required to process an entire depth image, i.e., the classification time.

**Table 3.7:** [Experiment 3 - RC features] The minimum and maximum scores for the balanced accuracy, recall, precision, and F1-scores obtained by the RC features in experiment 3.

| Performance metric | Min. score (SD) | Max. score (SD) |
|---|---|---|
| Balanced accuracy (%) | 73.6 (1.1) | 83.0 (1.0) |
| Recall (%) | 48.7 (2.1) | 68.0 (2.0) |
| Precision (%) | 63.3 (2.9) | 88.3 (2.2) |
| F1-score | 0.62 (0.02) | 0.76 (0.02) |

**Table 3.8:** [Experiment 3 - PC features] The minimum and maximum scores for the balanced accuracy, recall, precision, and F1-scores obtained by the PC features in experiment 3.

| Performance metric | Min. score (SD) | Max. score (SD) |
|---|---|---|
| Balanced accuracy (%) | 52.5 (0.6) | 56.0 (0.6) |
| Recall (%) | 7.21 (1.1) | 27.0 (1.3) |
| Precision (%) | 23.0 (1.8) | 38.9 (2.7) |
| F1-score | 0.12 (0.02) | 0.248 (0.01) |

Figure 3.6 (see Subsection 3.3.1) shows two examples of depth images from our test set, and the corresponding prediction results of the region comparison detector (using a RDF classifier of 10 trees). We remark that the region comparison detector is capable of detecting people at close range and at larger distances from the Kinect device. Figure 3.13c shows the AUC curve for the optimal detection parameters, i.e., a RDF classifier of 10 trees.

The results (see Figure 3.11) of the experiment show that the RC features again achieve a significantly higher classification performance than the PC features (see Tables 3.7 and 3.8). Both (balanced) accuracy and recall obtained by the RC features are largely independent of the number of trees in the RDF classifiers. The precision of the detector, however, increases significantly with the dimensions of the forest. The region comparison detector achieves its optimal AUC using a forest of 4 trees.

The classification times for the person detection task are listed in Table 3.9. The results are two-fold. On the one hand, the results show that the RC features allow for a considerably shorter classification time per image (which therefore results in a higher prediction speed) than the PC features. For instance, our results show that the detection speed obtained by the region com-

**Table 3.9:** [Experiment 3: person detection] The AUC scores and classification times per image for the both detectors while using RDF classifiers of ten sizes in experiment 3.

| Forest | region comparison detector | | pixel comparison detector | |
|---|---|---|---|---|
| size | AUC | Pred. time (s) (SD) | AUC | Pred. time (s) (SD) |
| 1 | 0.785 | 1.70 (0.22) | 0.552 | 8.07 (0.18) |
| 2 | 0.862 | 2.54 (0.03) | 0.590 | 9.21 (0.10) |
| 3 | 0.892 | 3.34 (0.04) | 0.620 | 10.3 (0.05) |
| 4 | 0.913 | 4.19 (0.05) | 0.640 | 11.5 (0.09) |
| 5 | 0.923 | 5.00 (0.09) | 0.651 | 12.6 (0.14) |
| 6 | 0.931 | 5.92 (0.06) | 0.668 | 13.7 (0.05) |
| 7 | 0.936 | 6.75 (0.17) | 0.675 | 15.0 (0.10) |
| 8 | 0.940 | 7.63 (0.07) | 0.682 | 16.1 (0.13) |
| 9 | 0.944 | 8.48 (0.12) | 0.691 | 17.3 (0.09) |
| 10 | 0.946 | 9.30 (0.17) | 0.695 | 18.3 (0.17) |

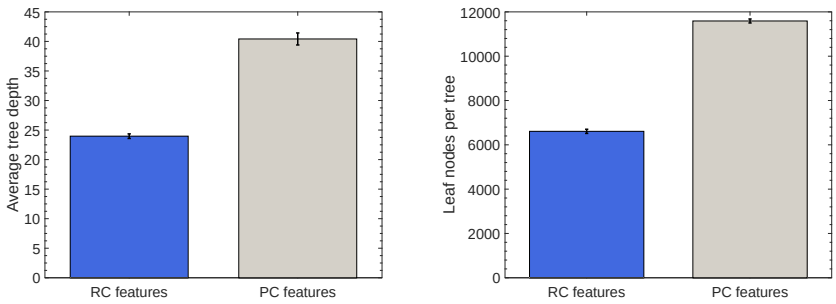



**Figure 3.12:** [Experiment 3: person detection] The bar plot of (a) the average tree depth, and (b) the average number of leaf nodes per tree for the region comparison detector (represented in blue) and the pixel comparison detector (represented in grey), and the corresponding error bars. The results are averaged over all trees and all folds.

parison detector is about two to four times higher than the detection speed obtained by its pixel comparing opponent. On the other hand, the results show

that both detectors require more time to classify the pixel locations than in experiment 1 and 2. Figure 3.12 shows that this is also reflected in the increased complexity of both detectors.

An analysis of the complexity of the detectors that are trained for experiment 3 (see Figure 3.12) reveals that the RDF classifier of the region comparison detector reaches an average depth of 24, 0 levels (SD = 0.39), versus an average depth of 40.4 levels (SD = 1.02) for the pixel comparison detector. The trees of the region comparison detector consist of, on average, 6, 610 leaf nodes (SD = 94), and 11, 600 leaf nodes (SD = 90) for the pixel comparison detector.

Compared to the results of experiment 2 (see Figure 3.10), the results show an increase in the number of levels of the RDF of the region comparison detector, but a decrease in the number of levels of its pixel comparing opponent. Yet, the results also indicate that the average number of leaf nodes increases significantly for either detector. This suggests an increase in the average number of tests per tree of either detector, which translates to an increase in complexity of both detectors. Although the complexity increases for both detectors, the RDF of the region comparison detector again achieves a lower complexity than the pixel comparison detector.

The results of experiment 3 indicate that the person detection task is, indeed, a harder task than the face detection tasks (experiment 1 and 2). For more complex classification tasks, the RC features seem to benefit from an increase in the number of trees. When detecting an entire person in non-smoothed depth images, RC features outperform PC features at a two to four-fold increase in processing speed. The results of experiment 3 thus show that the superiority of RC features also holds for the task of person detection.

## 3.5 DISCUSSION

In this Chapter, we evaluated the Region Comparison (RC) features that were proposed in Chapter 2. In our evaluation, we compared the classification performance and detection speed of a detector that incorporated our RC features with the performance achieved by a detector that incorporated Shotton et al.'s (2013a,b; 2011) Pixel Comparison (PC) features. The results of our evaluation of experiments with a non-smoothed noisy background reveal that our approach achieves a high detection accuracy without requiring an additional computational budget. The results of the empirical evaluation show that RC features do indeed outperform PC features in (1) classification performance, and (2) computational efficiency. This holds for face detection as well as for person detection.

This Section discusses the implications of the results in more detail. Subsection 3.5.1 discusses the relative superiority of the RC features over the PC features, while Subsection 3.5.2 addresses the number of samples that are used in our experiments. Subsequently, Subsection 3.5.3 discusses our future work and the steps to be taken before the region comparison detector can actually be employed for object detection tasks.

### 3.5.1 RC Features Combine the Best of Both Worlds

Given the superior results achieved by the RC features in our evaluation, the prevailing question still is: what precisely does explain the superiority of the RC features over PC features? As indicated in Section 2.1, handling the noise in depth images is an important challenge for object detection methods. Individual depth pixels may have incorrect values due to limited sensor resolution or false reflections. Comparing individual, incorrect pixel values may therefore lead to measurement errors. To counter the far-reaching effects of incorrect pixel values requires averaging over larger regions in the depth image. That is where RC features come in. However, averaging over pixel values results in a loss of spatial precision. Analogously to the Viola-Jones approach, which combines integral images and Haar wavelets, the RC features combine the best of both worlds. On the one hand, RC features include the averaging (summing) over large regions, which makes the features insensitive to local pixel noise, while on the other hand the RC features take individual pixel pairs (such as the PC features) into account. We believe that the balanced combination of global averaging and local precision explains the relative superiority of the RC features over the PC features. Moreover, the computational efficiency is a direct result of our use of the highly efficient integral image representation.

To arrive at this point, we incorporated the combination of RC features and the integral image in our region comparison detector. The detector computes feature vectors with RC features for each randomly selected pixel location in a depth image. The resulting RC feature vectors contain 418 elements. In contrast, the PC feature vectors (that are created using the PC features of Shotton et al. (2013a,b; 2011)) contain 2,000 elements per pixel location. It implies that calculating features over the same spatial area in a depth image results in RC feature vectors that are approximately 80% smaller than the feature vectors created for the PC features. As a result, the number of calculations required to create the individual feature vectors is likely to be in favour of the RC features. The additional computational cost required to compute the surface of the areas for the RC features is negligible when integral images are employed (cf. Fanelli et al., 2013, 2011). Moreover, shorter feature vectors contain fewer features that need to be tested by the random decision forest. This may there-

fore add to the computational efficiency of the RC features. One may argue that calculating the integral image representation of the depth image itself also requires computational power. However, the results of our evaluation reveal that processing an entire depth image using the RC features (by first calculating the integral image of a depth image and subsequently computing the RC feature vectors) takes less time than the time required to create and classify the PC feature vectors. This is partly due to the fact that the integral image representation is computed only once for each depth image. The time required to calculate the integral image is therefore likely to be compensated by the efficient feature computation process. We therefore argue that computing the integral image and the RC feature vectors can be achieved more efficiently than the PC feature vectors, which works in favour of the RC feature vectors.

As stated above, our results show that the RC features achieve performances superior to PC features. Of course, the next question is: how do RC features compare to other state-of-the-art methods? A direct comparison is difficult, because the RC and PC detectors assign labels to individual pixels, rather than to entire objects. Still, some indication may be given by relating our detection results to those obtained by Buys et al. (2014). As discussed in Section 2.4 (Related Work), Buys et al. (2014) developed a sophisticated method for human body pose detection by building on the PC features. Their person-classification performances range from 80 to 90%. Given our findings, we may expect that the detection accuracy of Buys et al.'s (2014) method would improve beyond 90% when the PC features would be replaced by RC features.

### 3.5.2 The Number of Samples Required

The PC features of Shotton et al. (Shotton, Girshick, et al., 2013) rely on the comparison of depth pixel pairs. They require many examples to encode objects uniquely against the background. As a case in point, in their experiments, Shotton et al. (2013a,b; 2011) use datasets of a size that are 150 to 9,000 times as large as the 100-image subsets that are used in our experiments. As each of our experiments took several days to complete on powerful 50-core calculation servers, the decision to use subsets of the original databases was motivated by computational considerations. The experiments reported in (Shotton et al., 2011), for example, relied on 1000-core servers, which are not available to us.

Using subsets of the databases, however, may give rise to two challenges. On the one hand, it may be the case that the performance obtained by PC features benefits from an increase in the number of training examples. On the other hand, using a small subset may result in overfitting. To investigate these challenges, we performed an additional exploratory experiment using a larger subset of depth images of the *Biwi Kinect Head Pose Database*. In the experiment, we trained and evaluated both detectors on a subset of 1,250 depth

images, using the same conditions as described in Section 3.3. The results of this experiment showed similar results as reported in Subsection 3.4.1. While increasing the size of the datasets (and thereby the number of training examples) may increase the classification performance of PC features, we expect that this will increase the classification performance of the RC features as well. Due to the results of the up-scaled experiments (in combination with the established cross-validation procedure), we feel confident that our results provide reliable estimates of the classification performance on large-scale datasets.

### 3.5.3 Future Work

Below, we recommend three instances of future work that may further improve the accuracy of the RC features.

First, we emphasise that the detection tasks performed in our evaluation procedure were limited to the automatic labelling of individual pixel locations as belonging to either an object (face/body) or to the background. Calculating features from pixel locations and classifying them accordingly are two important steps towards actual object detection. However, actual object detection requires an additional processing step which integrates the individual pixel labels into a higher-level detection of the object, i.e., the labelling of a larger region encompassing the face or body. We refrained from developing such a higher-level detection stage, because the focus of this study was on the evaluation of the RC features. Future work may therefore extend the region comparison detector with this stage. It is to be expected that the superiority of RC features (as compared to PC features) will be reflected in any higher-level detection method that takes the labels generated by the region comparison detector as input. Deploying the RC features is highly relevant for the development of embodied agents that aim to engage in natural interactions with the inhabitants of intelligent environments.

Second, we stress that the detectors that are used in the evaluation procedure were implemented as MATLAB scripts. While MATLAB allows for rapid prototype development, it is not optimised for speed. Implementing the RC features in a dedicated programming language such as C++ or Python may speed up their processing time. We expect that porting the RC features to a C++ or Python implementation is feasible and that the RC features are therefore likely to run on reasonable hardware. In this respect, future work is mainly a challenge from an engineering perspective. However, I am sure that in this phase, new ideas will arise that will make the RC features even faster.

Third, in this Chapter, we evaluated both feature computation methods on three publicly available databases with depth images: (1) the *Biwi Kinect Head Pose Database* by Fanelli et al. (2013), (2) the *RGB-D Face Database* by Høg et al. (2012), and (3) the *RGB-D People Dataset* by Spinello and Arras (2011). The

first two databases were created in controlled experiments, while the latter one was created by recording people walking in a semi-unrestricted space. Although capturing data under controlled conditions is likely to result in high quality recordings of a scene, it may also result in less natural behaviour by the participants in the experiment, such as consciously (and therefore clearly) performed gestural cues. As the goal of this Thesis is to facilitate *natural* interactions between humans and embodied agents, we argue that future work should include a training and evaluation procedure which evaluates the performance of the RC features on databases that contain depth data recordings of spontaneous human behaviour and unconsciously performed gestures.

## 3.6 CHAPTER CONCLUSIONS

The research question of this Chapter is RQ 2. It reads: *To what extent do Region Comparison features enable fast and accurate face and person detection in noisy depth images?* To answer the research question, this Chapter evaluates the RC features that were proposed in Chapter 2. In three different object detection tasks, a comparative evaluation investigated to what extent RC features contribute to fast and effective object detection in noisy depth images.

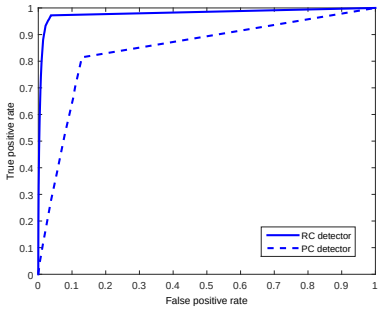
The results of the evaluation show that the RC features outperform the state-of-the-art PC features in classification performance and do so with the same (or even better) prediction speed, especially in noisy depth images. The RC features deal effectively with the background noise in depth images. They maintain precision in the depth images by sampling depth transitions on scales varying from small to large image regions. The RC features are able to provide an accurate indication of the direction and magnitude of the depth transitions in a depth image. Thus, they are able to perform fast and effective body part detection tasks in noisy depth data. Based on our results, we may provisionally conclude the following.

- Conclusion 1: The Region Comparison features contribute significantly to fast and effective face and person detection in noisy depth images.
- Conclusion 2: The RC features yield an improvement over PC features.
- Conclusion 3: The RC features are able to operate adequately with the same computational budget.

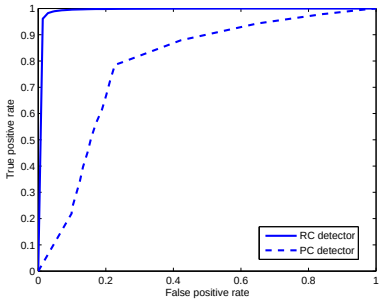
Employing RC features might be able to increase the classification performance of detectors based on the work by Shotton et al. (2013a,b; 2011), such as the body part detector by Buys et al. (2014).

### Research Continuation

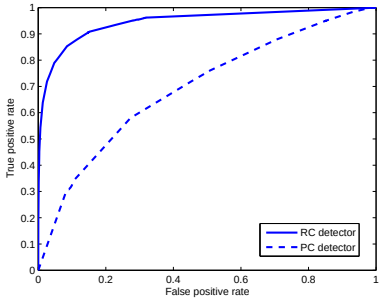
In the Chapter, our RC features have proven their value for face detection and person detection in noisy scenes. As such, deploying RC features is highly relevant for the development of embodied agents that aim to establish and maintain natural interactions with the inhabitants of intelligent environments. To enable such interactions, it is imperative that agents are enriched with the ability to perceive natural gestural cues. This requires a procedure which will train detectors (such as the region comparison detector) on depth data that contains spontaneous human behaviour and unconsciously performed gestures. However, to the best of our knowledge, there are no databases available in the public domain that (A) contain thoroughly annotated depth recordings of (B) people performing spontaneous and natural gestures. To meet these requirements, the next Chapter proposes, designs, and develops the *TiGeR Cub*, a new database with annotated depth images of people performing natural gestures.




a




b




c

**Figure 3.13:** The AUC (Area Under the Curve) graphs of the detectors when using their optimal detector parameters (i.e., the parameters resulting in the highest prediction performance; a forest of 10 trees). Figure 3.13a shows an example of the AUC from experiment 1 (face detection in smoothed depth images), while Figure 3.13b shows this for experiment 2 (face detection in non-smoothed depth images). Subsequently, Figure 3.13c shows an example of the AUC from experiment 3 (person detection in non-smoothed depth image).

# 4 RAISING A TIGER

*"Fie, fie upon her!*
*There's language in her eye, her cheek, her lip,*
*Nay, her foot speaks; her wanton spirits look out*
*At every joint and motive of her body."*
– William Shakespeare, *Troilus and Cressida*

Facilitating natural interactions between humans and embodied agents requires advanced algorithms that are able to recognise a person's gestural cues. Developing and training gesture recognition algorithms require high quality corpora that contain annotated, visual and depth data recordings of people performing natural communicative gestures. Ideally, such databases are available for the public domain. However, to the best of our knowledge, there are no databases available that meet these criteria. This Chapter designs and develops the Tilburg Gesture Research Cup database, or TiGeR Cub for short. The database contains annotated recordings of naturally interacting interlocutors. The interactions are recorded as a combination of visual data, depth data, and audio data. The TiGeR Cub therefore allows for detailed studies into automatic gesture recognition and human gesture synthesis.

The structure of the Chapter is as follows. First, Section 4.1 outlines the dire need to develop a new corpus for robust and accurate gesture recognition tasks. Subsequently, Section 4.2 provides an overview of related corpora in the field of natural human-human interactions. Then, Section 4.3 describes the experiment that has been performed to create our corpus, and the annotation procedure of the resulting corpus. Section 4.4 discusses the resulting TiGeR Cub corpus. Finally, Section 4.5 concludes on the creation of the TiGeR Cub corpus and answers our third research question.

## 4.1 TOWARDS A DATABASE WITH NATURAL GESTURES

In social interactions between humans, the process of communication is established by combining and aligning a person's social (i.e., verbal and non-verbal) expressions (see, e.g., Pickering & Garrod, 2004; McNeill, 1992). The synthesis and classification of the non-verbal cues that are employed in these interactions are studied extensively in various disciplines, such as psychology (see, e.g., Hinde, 1972), cognitive science (see, e.g., Manusov & Patterson, 2006), and artificial intelligence (see, e.g., Osawa & Imai, 2013).

The aim of this Thesis is to develop smart and socially aware embodied agents. To facilitate natural interactions between humans and embodied agents, the agents need accurate computer vision algorithms that are able to recognise a person's non-verbal cues. Thus, supporting the agents to recognise natural human gestures is highly relevant in the context of the current project.

As embodied agents are likely to be deployed in noisy environments (i.e., environments with a large variety of objects, changing illumination conditions, and moving people) the computer gesture recognition algorithms of the agents should be able to deal with the environment's background noise. Many approaches towards automatic gesture comprehension and recognition, however, incorporate visual data to perform their classification procedure. While rich in detail, the disadvantage of visual data is that it is sensitive to illumination conditions (see Section 1.5; see, e.g., C. Zhang & Zhang, 2010; Zhao et al., 2003). This may negatively influence the quality of the data.

Alternative data sources such as depth data (which is discussed extensively in Chapter 2 of this Thesis) may provide robust cues for accurate gesture recognition approaches, especially when combined with visual information (see, e.g., Jiang et al., 2013; Dal Mutto et al., 2012). As depth data is not bound to a light source, it is less sensitive to changes in illumination conditions. In the domain of automatic gesture recognition, the availability of low-cost depth sensors such as the Microsoft Kinect device (see, e.g., Smisek et al., 2013) enables recordings of human interactions in the form of multiple data streams, i.e., both depth and visual data (see, e.g., Dal Mutto et al., 2012). Incorporating alternative data sources such as depth data may thus increase the detection performance of gesture recognition algorithms.

Natural or - more specific - *dynamic* gestures can occur in any order, dimension, or shape. Moreover, the gestures that are performed by individuals can vary significantly per person (see McNeill, 1992). Thus, enabling agents to recognise natural gestures calls for the use of multimodal corpora that incorporate recordings of natural gestures. Well-known examples of such of corpora are (1) the SaGA corpus by Lücking, Bergmann, Hahn, Kopp, & Rieser (n.d.), which incorporates visual data recordings of dialogs between interlocutors that are engaged in a spatial communication task, and (2) the VACE cor-

pus by Chen et al. (2006), which contains multimodal cues from interlocutors gathered in a series of meetings. To combine the best of both worlds (i.e., data that is rich in detail, yet robust against changes in illumination conditions), such corpora should include both visual (RGB - Red Green Blue) and depth (D) data. Despite the clear need for detailed yet robust recordings of natural human gestures, there are, to the best of our knowledge, no databases available that:

1. contain recordings of people performing spontaneous behaviour and natural communicative gestural cues,
2. provide the recordings as RGB-D data,
3. provide clear annotations of (a part of) the data, and
4. are available for scientific purposes.

It calls for the development of a new database that meets the requirements as defined above. Thus, the research question addressed in this Chapter (RQ 3) reads as follows.

> RQ 3: *How do we develop an annotated database that incorporates visual and depth data recordings of natural human gestures?*

To answer the research question, the Chapter presents a multimodal corpus of social interactions between interlocutors. As evidence shows that interlocutors gesture more when talking about spatial topics (cf. Alibali, 2005), we performed an experiment in which participants fulfil two spatial event description tasks. A part of the recordings of the experimental tasks were annotated. This led to the creation of the Tilburg Gesture Research Cub corpus, or *TiGeR Cub* for short. The corpus contains 32 recordings of 16 dyadic interactions between interlocutors. Each recording contains movie sequences (audio included) of approximately 15 minutes and, on average, 27,000 frames with depth data per participant. Thus, the data is recorded as a combination of depth + visual + audio data. At present, we provide accurate annotations for individual body parts (e.g., the head, shoulders, and arms) for a subset of the depth data. The TiGeR Cub corpus is available for academic purposes upon request from the authors.

## 4.2 RELATED WORK

Over the last two decades, several multimodal databases have been proposed to study gesture recognition and human gesture synthesis (see, e.g., Guyon,

Athitsos, Jangyodsuk, & Escalante, 2014; Caridakis et al., 2013; L. Liu & Shao, 2013; Fothergill, Mentis, Kohli, & Nowozin, 2012; Lücking et al., n.d.). These corpora can roughly be divided into two categories: (1) corpora that focus on visual (RGB) recordings (see, e.g., Caridakis et al., 2013; Lücking et al., n.d.; L. Chen et al., 2006), and (2) databases that incorporate a combination of visual and depth (RGB-D) data (see, e.g., Guyon et al., 2014; L. Liu & Shao, 2013; Swift et al., 2012). While the TiGeR Cub database mainly relates to the corpora in the latter category, the design of the TiGeR Cub is inspired by several contributions in the field of human gesture study databases. In what follows, Subsection 4.2.1 briefly describes the related corpora. Subsequently, Subsection 4.2.2 discusses how these corpora inspired us to develop the TiGeR Cub.

### 4.2.1 Related Approaches

Below, we briefly describe four related corpora in the field of natural communicative gestures. The dialogs of the interlocutors are typically captured in controlled experiments. The corpora can be characterised as databases that contain (1) multimodal recordings of emotional gestures in multi-camera setups, (2) communication gestures captured in spatial event description tasks, (3) annotated RGB-D recordings of individuals performing sets of gestures, and (4) synchronised recordings of hand gesture sequences.

First, Caridakis et al. (2013) developed a corpus to study the use of emotional gestures in combination with other modalities, such as facial expressions and speech. Thus, they proposed a multimodal dataset that focusses on capturing hand gesture expressivity, but also includes the subjects' speech and facial expressions. The dataset contains visual data recordings of 51 subjects from 3 different countries. The hand gestures are captured by recording the subjects' bare hands, the movements of a Nintendo Wii remote controls, and a data glove. In their experiments, Caridakis et al. (2013) recorded the subjects in a dual camera setup. The first camera captures the subject's complete body, while the second camera captures a close-up of the subject's shoulder and head. The data is recorded as 25 frames per second video sequences. The resolution of the video material is $720 \times 576$ pixels.

Second, Lücking, Bergmann, Hahn, Kopp, & Rieser (n.d.) developed the thoroughly-annotated Bielefeld Speech and Gesture Alignment (SaGA) corpus. The database contains visual recordings of 25 dialogs between interlocutors who engage in a spatial communication task. In the experimental setup, the cameras capture multiple views: the router, the follower, and the entire scene. The data have been systematically annotated and evaluated in terms of interrater agreement. As a result, the annotations provide an indication of the order, dimension, and shape of the gestures.

Third, the ChaLearn Gesture Dataset developed by Guyon, Athitsos, Jangyodsuk, & Escalante (2014) contains over 54,000 hand and arm gestures from 20 subjects. The data was recorded as RGB-D data using a single Kinect device, with an image resolution of $240 \times 320$ pixels. In the majority of the recordings, the Kinect device focusses on the torsos of the subjects. The data comes with man-made annotations, i.e., temporal segmentation into individual gestures, alignment of RGB and depth images, and body part location. The corpus contains recordings of individuals performing sets of gestures, such as body language gestures, gesticulations performed to accompany speech, and signals (e.g., diving signals). The gestures are mostly performed by the arms and hands of the individuals.

Fourth, the SKIG (Sheffield KInect Gesture) dataset developed by L. Liu & Shao (2013) contains 1080 visual (RGB) and 1080 depth (D) hand gesture sequences that are collected from 6 subjects using a Kinect device. The dataset consists of 10 categories of descriptive hand gestures, such as circles, triangles, up-down, and right-left. The sequences are recorded under various illumination conditions and different backgrounds. Annotations are provided in the form of descriptions of the gesture sequences.

### 4.2.2 Inspiration for the TiGeR Cub

In what follows, we briefly discuss how the approaches described above inspired us to develop the TiGeR Cub.

First, the work by Caridakis et al. (2013) inspired us to develop a dual recording device setup to ensure that we capture high quality (i.e., high resolution) RGB-D data, we developed a static mount to which a Kinect device and a digital camcorder were connected. The recording devices focussed on the upper body of each participant, which allows for highly detailed recordings of a person's behaviour.

Second, inspired by Lücking, Bergmann, Hahn, Kopp, & Rieser (n.d.), we performed an experiment that incorporated spatial event description tasks. As evidence shows that interlocutors gesture more when talking about spatial topics (cf. Alibali, 2005), this allows us to capture natural communication gestures.

Third, inspired by the work by Guyon et al. (2014), the TiGeR Cub contains recordings of gestures that are performed by the arms and hands of the subjects. Moreover, we adopt their annotation procedure to annotate the locations of the individual body parts.

Fourth, L. Liu & Shao (2013) stress the importance to synchronise the visual and depth data recordings of the gestures. While the data in both the SKIG dataset by L. Liu & Shao (2013) and the ChaLearn Gesture Dataset by Guyon et al. (2014) are captured using a single Kinect device to record the

gestures, the experimental setup of the TiGeR Cub incorporates a combination of a Kinect device and a digital camcorder. Thus, inspired by the work by L. Liu & Shao (2013), we developed a mechanical solution on the mount that synchronises both recording devices.

## 4.3 EXPERIMENT

To create the TiGeR Cub, we performed an experiment that incorporated two event description tasks. The resulting recordings of these tasks are used to create the corpus. In what follows, Subsection 4.3.1 describes the physical setup that is used to create the corpus. Subsequently, Subsection 4.3.2 reviews the participants and the procedure employed to record the dialogues. Then, Subsection 4.3.3 presents the resulting database, while Subsection 4.3.4 discusses the annotations that are created for the corpus.

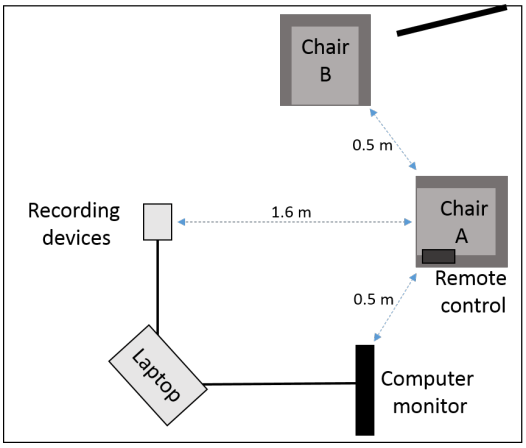
### 4.3.1 Experimental Setup

The TiGeR Cub contains visual (RGB) and depth (D) recordings of people engaging in natural communicative interactions, i.e., people describing a series of spatial events. The experimental setup is created in one of the offices at Tilburg University. To perform the experiment, we built an experimental setup that consists of the following.

- Two chairs in which the participants are seated.
- A computer monitor to show visual stimuli to the participants.
- A remote control to allow the participants to go to the next stimulus.
- A camera mount with a Kinect device and a digital camcorder.
- A laptop to store the recorded data.

Figure 4.1a shows an overview of the experimental setup. In this Figure, the first chair (chair A) is positioned with its back against the wall. A remote control is connected to the chair. This allows a seated participant to continue to the next stimulus in the experiment. Facing the chair, at a distance of 1.6 meters, a Sony HDR-XR550VE digital camcorder and a Microsoft Kinect device are mounted on a tripod mount at a height of 1.0 meters. At distances of 0.5 meters from the chair, a computer monitor and a second chair (chair B) are set up under 90 angular degrees on the left and right side, respectively.

The camcorder captures visual data at a resolution of $1,920 \times 1,080$ pixels (25 frames per second). The Kinect device captures depth data at a resolution

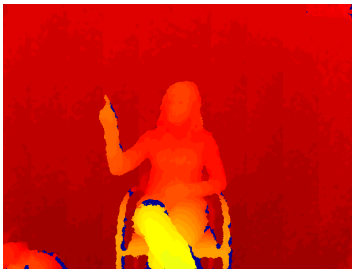



a

b

c

d



e

**Figure 4.1:** The experimental setup (a) and two photos (b & c) of the experiment that is performed to create the TiGeR Cub corpus. Figure 4.1d shows an example of the visual data that is captured in the experiment. Figure 4.1e shows the corresponding depth image. The participants shown in this example are actresses and are not included in the final corpus.

of $640 \times 480$ pixels (25 frames per second). The lenses of the camcorder and the Kinect device were mounted as close to each other as possible, to ensure that they recorded the same spatial scene. Figure 4.1 (b & c) shows two photos of the experiment; Figure 4.1d shows an example of the visual data that is captured in the experiment. Figure 4.1e shows the corresponding depth image. The participants shown in this example are actresses and are not included in the final corpus.

### 4.3.2 Methodology

To create the TiGeR Cub, we perform an experiment that consists of two event description tasks. In what follows, the event description tasks are discussed in more detail.

EVENT DESCRIPTION TASK 1 In the first event description task, the participants mimic the gestures that are shown in a series of video sequences. In this task, there is no verbal contact between the participants. For this task, we selected 21 hand signals from the set of internationally recognized scuba diving hand signals, as defined by the *CMAS Code of International Diving Signals*[17]. The scuba diving hand signals form a series of emblem gestures, i.e., specific gestures with a specific meaning, which are consciously used by the sender and consciously understood by the receiver (cf. McNeill, 1992). The collection of diving signals contains a series of emblems that are executed with only one hand, as well as emblems that are performed with both hands. Repeating the diving signals performed a function as an implicit warming-up for the second event description task in our experiment.

Our choice for diving hand signals is on the one hand motivated by the clearly defined beginning and end of the signal gestures. On the other hand, the gestural meaning of diving signals is (partially) derived from the gestural motion itself, instead of (for example) solely the pointing direction of the arms, hands, or fingers. We therefore argue that diving signals are a challenge for (computational) approaches towards gesture recognition, as such approaches are forced to take the gestural context into account.

EVENT DESCRIPTION TASK 2 In the second event description task, the participants describe the events occurring in a series of 5 video sequences of Tweety and Sylvester (cf. McNeill, 1992). The choice for the Tweety and Sylvester cartoons was motivated by the large number of spatial events occurring in the cartoons. As Alibali (2005) suggested that interlocutors gesture more when

17 CMAS stands for the *Confédération Mondiale des Activités Subaquatiques*. For a complete overview of the scuba diving hand signals, see http:www.cmas.orgdocument?sessionId=&fileId=2212&language=1

talking about spatial topics, we expect that using Tweety and Sylvester cartoons persuades participants to use spontaneous and natural spatial gestures to describe the aforementioned events.

Below, we describe the participants of the experiment in Subsection 4.3.2.1 and the procedure in 4.3.2.2.

#### 4.3.2.1 *Participants*

The participants are 32 first-year students from the Communication and Information Sciences curriculum at Tilburg University: 13 male participants, and 19 female participants. The students receive course credits for their participation in the experiment and were (afterwards) duly informed about the use of the recordings in the experiment. All participants are asked for their consent to share the recordings of their interactions in the experiment for academic purposes.

#### 4.3.2.2 *Procedure*

The procedure of the experiment consists of four successive stages: (1) a briefing stage, (2) the first event description task, (3) the second event description task, and (4) the debriefing stage of the experiment. The entire procedure, including instructions and debriefing, takes approximately one hour to complete. In what follows, the procedure and subsequent stages are described in more detail.

STAGE 1: BRIEFING Upon arrival in pairs of two, the experiment leader asks the participants to be seated. Subsequently, they are informed about the voluntary nature of the experiment. After the instructions, the experiment leader starts the recording devices and leaves the room.

STAGE 2: EVENT DESCRIPTION TASK I The video sequences contain a combination of 21 one-handed and two-handed diving signals. The diving signal are sequentially shown on the computer monitor. The individual diving signals are shown once to the participant. The computer monitor is only visible to the participant seated in chair A (henceforth referred to as participant A), but invisible to the participant seated in the second seat (henceforth referred to as participant B). Participant A mimics each gesture towards participant B. Participant B's aim is to guess the signal's presumed meaning. By guessing and writing down the meaning of the gesture, the continued mental focus of participant B's is ensured. The experiment leader is fetched upon completion of this stage. He then explains the second stage. After the instructions, he leaves the room.

STAGE 3: EVENT DESCRIPTION TASK II Following the completion of the first event description task, the second event description task is initiated. In this task, a series of 5 animated cartoon movie sequences of Sylvester and Tweety ("Canary Row") are shown to participant A. After each video sequence, participant A describes the events occurring in the sequence to participant B. The latter's aim is to answers a secret question, only known to this participant. This ensures the mental focus of participant B. Upon completion of this task, the experiment leader is fetched, after which both participants switch seats. Stages 2 and 3 are then repeated for participant B. To ensure the participant's genuine participation, we use a different secret question for the second task.

STAGE 4: DEBRIEFING Upon completion of stages 2 and 3 for both participants, the experiment leader debriefs the participants. Both participants are asked to sign the consent forms of the experiment. They then leave the room, after which the experiment is finished.

### 4.3.3 The TiGeR Cub

The resulting corpus contains 32 recordings of 16 dyadic interactions between participants. Each recording contains movie sequences (audio included) of $\approx$ 15 minutes and, on average, 27,000 frames with depth data per participant. For the corpus, we only used the recordings for which both participants gave their consent for the distribution of the recordings.

### 4.3.4 Annotations

The annotation procedure focussed on annotating the participants' body parts in the depth data generated in the second event description task. Annotating the extensive quantities of data generated in the experiment proved to be a challenge. This Section describes the procedure followed to annotate the data.

For the annotation procedure, the first 13 participants are selected from the original population of 32 participants. On average, the second event description tasks resulted in $\approx$ 13,000 depth images per participant, from which 50 depth frames are selected at random. The resulting subset contains (13 participants $\times$50 depth images per participant = ) 650 depth images, from which (if visible) the head, shoulders, upper arms, lower arms and hands are annotated. As a result, 12 body parts are annotated in each depth image. Inspired by the mask-based annotations that are used in the database by Fanelli et al. (2013), we use polygons to annotate the body parts in the depth images of the TiGeR Cub. The annotations are stored as the $(x, y)$-coordinates of the polygons in plain text files. To perform the actual annotations, we designed and built the

*AnnoTool*; an open source annotation tool for depth images. The AnnoTool is available for academic purposes upon request from the authors.

The annotation procedure is performed by a group of 52 first-year students from the Communication and Information Sciences curriculum at Tilburg University: 17 male students and 35 female students. The students who are participating as annotators are ignorant about the experiment and did not participate in the event description tasks. The students receive course credits for their participation as annotators. The 52 annotators are divided over 13 annotation teams of 4 annotators per team. The 13 subsets (with 50 depth images per subset) are divided over the annotation teams. Each annotation team is divided into two separate annotation groups. The first group performs the annotations of the head and both hands, while the second group annotates both upper and lower arms. The annotations are cross-validated within each group. Figures 4.2 and 4.3 on the next two pages show a total of four depth images from the TiGeR Cub and the corresponding annotations. In the Figures, the individual body parts are annotated as green polygons that follow the contours of the head, shoulders, upper and lower arms, and hands.

## 4.4 DISCUSSION

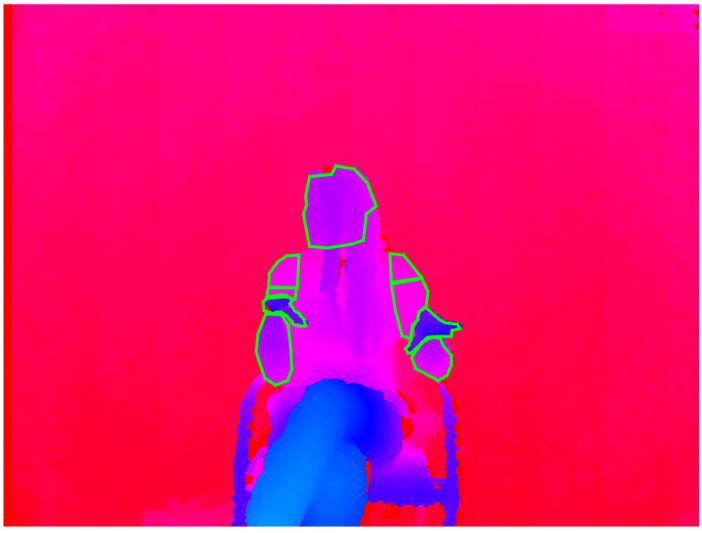
In this Chapter, we proposed the Tilburg Gesture Research Cup database, or TiGeR Cub for short. The corpus contains annotated RGB-D recordings of dialogues between interlocutors. To create the corpus, we performed an experiment in which the interlocutors were engaged in two spatial event description tasks. The TiGeR Cub allows for detailed studies into automatic gesture recognition and human gesture synthesis.

The remainder of this Section discusses the points of improvement of the corpus, as well as our future work. In what follows, Subsection 4.4.1 discusses the synchronisation of the RGB-D data. Then, Subsection 4.4.2 addresses the annotation procedure of the data. Finally, Subsection 4.4.3 discusses our future work on the TiGeR Cub.

### 4.4.1 Synchronising the RGB-D Data

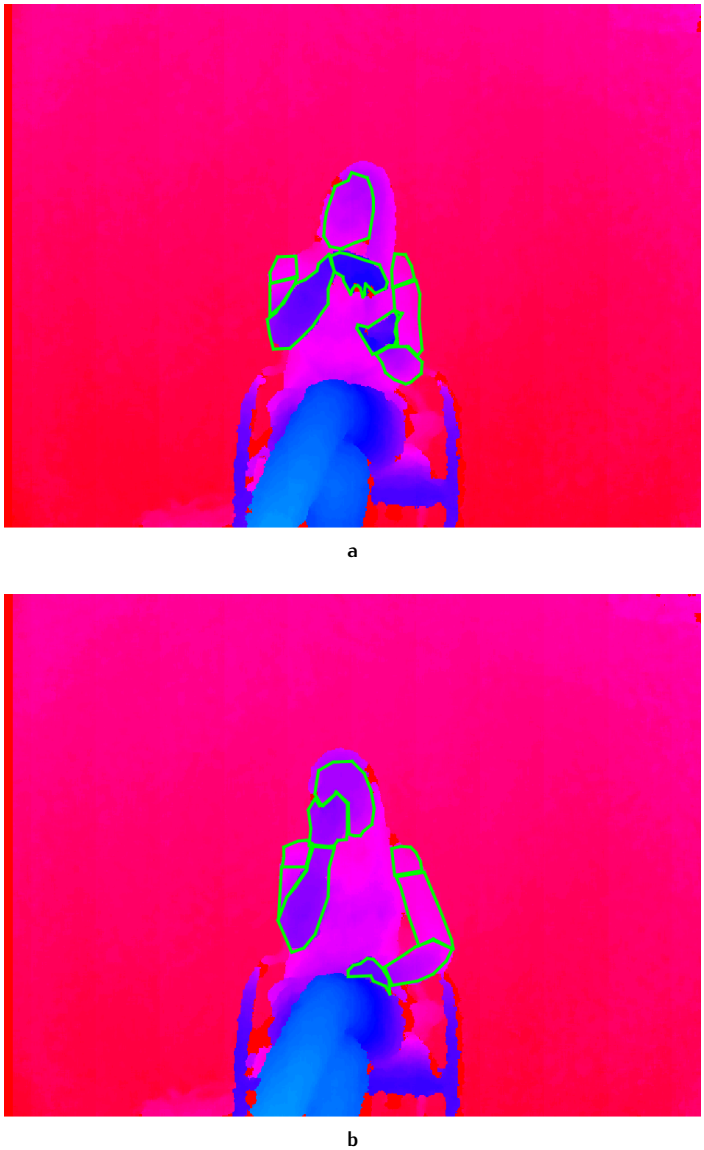
The aim of this Chapter is to develop a high quality corpus that contains visual and depth data recordings of people performing natural communicative gestures. High quality recordings are typically characterised by (1) a high image resolution, i.e., the detail an image holds, and (2) a high frame rate, i.e., the number of frames per second.

a



b

**Figure 4.2:** Two examples from the TiGeR Cub and (in green) their annotations. The body parts are annotated as polygons following the contours of the head, shoulders, upper and lower arms, and hands.



a

b

**Figure 4.3:** Two examples from the TiGeR Cub and (in green) their annotations. The body parts are annotated as polygons following the contours of the head, shoulders, upper and lower arms, and hands.

When attempting to capture high quality RGB-D recordings, approaches such as the work by Guyon et al. (2014) and L. Liu & Shao (2013) deploy Kinect devices to capture synchronous streams of visual (RGB) and depth (D) data. According to Khoshelham & Elberink (2012), the Kinect device captures visual data with a maximum image resolution of $1,280 \times 1,024$ pixels and depth data with a maximum image resolution of $640 \times 480$ pixels, at frame rates of up to 30 frames per second. However, due to limitations in the bandwidth of the hardware, the framerate tends to drop to a mere 15 frames per second when the maximum image resolution of both cameras is utilised. Drops in the frame rate result in less fluent recordings of, for example, a person's gestures.

To overcome this problem, we decided to build a combination of a Kinect device and a digital camcorder to capture the RGB-D data: the camcorder captures visual data with an image resolution of $1,920 \times 1,080$ pixels at 25 frames per second, while the Kinect device captures depth data with an image resolution of $640 \times 480$ pixels at 25 frames per second. On the one hand, this solution allows us to capture data with the highest image resolution available, which enables us to capture small objects (e.g., a person's fingers). On the other hand, using two recording devices allows us to maintain a high frame rate for both data streams. This enables us to capture fluent recordings of, for example, a person's gestural cues.

To synchronise both recording devices, we developed a synchronisation mechanism that activates both devices at the same time. However, empirical evidence suggests that our synchronisation mechanism functions with a small time delay. Thus, the visual and depth data streams are not entirely synchronised. While inconvenient, we argue that this will not influence the quality of the recordings severely, given that it can be corrected by removing a sequence of frames at the beginning of the visual data.

### 4.4.2 Annotating the Data

The aim of the Thesis is to facilitate natural interactions between humans and embodied agents. Thus, developing accurate computer vision algorithms that are able to recognise human gestures is highly relevant in the context of the current project (see Section 4.1). The training procedure of the algorithms requires high quality corpora that contain annotated recordings of people performing natural communicative gestures. As stated in Subsection 4.3.4, annotating the huge quantities of data gathered for the corpus proved to be a challenge. For our annotation procedure, we selected a subset of depth images from the second event description task. Inspired by the work by Guyon et al. (2014), we performed an annotation procedure in which we annotated individual body parts, e.g., the head, shoulders, and arms. As a result, the TiGeR Cub contains a total of 650 annotated depth images. In its current form,

the annotations can be used to train and evaluate modern day approaches towards, for example, automatic body part detection in depth data.

As the TiGeR Cub provides high quality recordings of people engaging in natural dialogues, the TiGeR Cub contains a rich set of examples of natural gestures and communicative interactions. Thus, the TiGeR Cub has the potential to provide a solid base to train and evaluate gesture recognition algorithms. This requires an extension of our current set of annotations, i.e., annotations that describe the order, dimension, and shape of the participants' (natural) gestures. Thus, inspired by the work by Lücking et al. (n.d.), we argue that the annotations of the TiGeR Cub should be extended with (1) annotations of the participants' body parts, and (2) annotations that describe the direction and magnitude of the participants' gestures.

### 4.4.3 Future Work

In the previous Chapter, we evaluated the performance of the region comparison detector on three challenging object detection tasks. In our experiments, we trained the region comparison detector to label individual pixel locations in depth images in a binary fashion, e.g, either as FACE when a pixel location belongs to a person's face, or OTHER in all other cases. As enabling the detector to recognise multiple body parts is highly relevant in the context of accurate gesture recognition, we argue that the region comparison detector should be extended with the ability to label pixel locations as belonging to, for example, the head, shoulders, upper and lower arms, hands, or background. Restricted by time constraints, we refrained from extending the detector with this ability. Thus, future work should focus on training and evaluating the performance of the region comparison detector on the body parts that are annotated in the TiGeR Cub. The results of the evaluation may provide a baseline performance for state-of-the-art approaches towards body part detection.

## 4.5 CHAPTER CONCLUSIONS

The research question of this Chapter is RQ 3. It reads: *How can we develop an annotated database that incorporates visual and depth data recordings of natural human gestures?* To answer the research question, this Chapter shows how the Tilburg Gesture Research (TiGeR) Cub has been developed. It is a multimodal corpus that consists of dyadically interacting interlocutors. The TiGeR Cub contains annotated visual + depth + audio recordings of the interactions. It is available for academic purposes. The answer to the research question resides in the experimental setup as given in Subsection 4.3.1 and in the methodology followed in the experiments (see Subsection 4.3.2). Both, setup and methodol-

ogy led to the development of an annotated database that incorporates visual and depth data recordings of natural human gestures. Of course, annotating the huge quantities of data remains a challenge. In its current form, the TiGeR Cub corpus provides a solid base to study the synthesis and classification of natural human gestures. Future work should focus on extending the corpus with additional annotations of the gestures.

### Research Continuation

When aiming to facilitate natural interactions between humans and embodied agents, the latter should be enriched with the ability to perceive the social cues of their human communication partners. Thus, developing approaches for effective gesture recognition is a first step towards natural human-embodied agent interactions. The next Chapter investigates to what extent RC features (as proposed in Chapter 2) are suitable for accurate (static) gesture recognition.

# 5 AUTOMATIC SIGN LANGUAGE RECOGNITION FROM A TO Y

*"Sign is a live, contemporaneous, visual-gestural language and consists of hand shapes, hand positioning, facial expressions, and body movements. Simply put, it is for me the most beautiful, immediate, and expressive of languages, because it incorporates the entire human body."*

– Myron Uhlberg, *Hands of My Father*

In the context of natural interactions between humans and embodied agents, the Region Comparison (RC - see Chapter 2) features have proven their value for fast and accurate body part detection tasks. This raises the question to what extent RC features are suitable to recognise (static) human gestures. To investigate their effectiveness, we propose and evaluate a novel detector that incorporates the RC features for effective static gesture recognition. The detector is trained and evaluated on a challenging dataset with fingerspelling signs of the American Sign Language (ASL).

The course of this Chapter is as follows. First, Section 5.1 further outlines the research question addressed in this Chapter. Then, Section 5.2 discusses the challenges that are to be faced when aiming to recognise static hand gestures in the American Sign Language. Section 5.3 presents the work related to our approach. Subsequently, Section 5.4 discusses the STAGE detector. Section 5.5 describes the evaluation procedure, in which the classification performance is assessed. Subsequently, Section 5.6 presents the results of our evaluation. The implications of the results are discussed in Section 5.7. Finally, Section 5.8 concludes upon our contribution and answers the fourth research question (RQ 4).

## 5.1 TOWARDS AUTOMATIC GESTURE RECOGNITION

Embodied agents aiming to engage in natural interactions with humans, require the ability to perceive and recognise a person's communicative gestural

cues. According to, e.g., Dixit & Agrawal (2015), gestures exist in two distinctive forms: (1) *dynamic* gestures, which involve direction and speed of motion to convey their meaning, and (2) *static* gestures, which involve arm and hand postures to represent a specific meaning. As constructing dynamic gestures requires motion in some form, perceiving those gestures requires computer vision techniques that are able to detect, track, and recognise the motion of the gestures (cf. Rautaray & Agrawal, 2015). Static gestures, however, derive their meaning from the shape of the hand. Approaches towards static gesture recognition may therefore focus on the recognition of the hand shape in single images, rather than recognising the hand motion in a sequence of images.

The studies described in this Thesis are part of a research project that aims to develop emotionally responsive agents. Providing embodied agents with the ability to perceive static gestures is a first step towards the automatic recognition of human gestures in general, and thus highly relevant in the context of the current project. In the previous Chapter, the combination of (1) the RC features, and (2) data depth has proven its value for robust body part detection tasks. This raises the question to what extent the combination of RC features and in-depth information is suitable for the recognition of static gestures. To this end, the research question addressed in the Chapter (RQ 4) reads as follows.

> RQ 4: *To what extent do Region Comparison features enable accurate recognition of static gestures when using in-depth information?*

To answer the research question, we perform a comparative evaluation to assess the effectiveness of the RC features in a gesture recognition task. We introduce an extension of the region comparison detector proposed in Chapter 3. Moreover, we claim that the extended detector, henceforth referred to as the *Static Gestures detector*, or STAGE detector for short, incorporates the RC features for effective gesture recognition in depth images. To investigate to what extent this claim holds, we perform a comparative evaluation of the STAGE detector on a dataset with static fingerspelling signs of the American Sign Language (ASL). The performance of the detector is compared with state-of-the-art approaches towards the automatic recognition of static gestures, i.e., automatic sign language recognition. For our experiments, we use the classification performance of the detector as its evaluation criterion. The classification performance is defined as the extent to which the STAGE detector is able to recognise the individual signs accurately. A higher detection accuracy corresponds to a higher classification performance. We establish straightforward outperforming the classification performance of the state-of-the-art as our decision criterion. We consider the STAGE detector (and therefore the RC features) to be superior to the state-of-the-art approaches when the STAGE detector outperforms its opponents in classification performance. Given that the

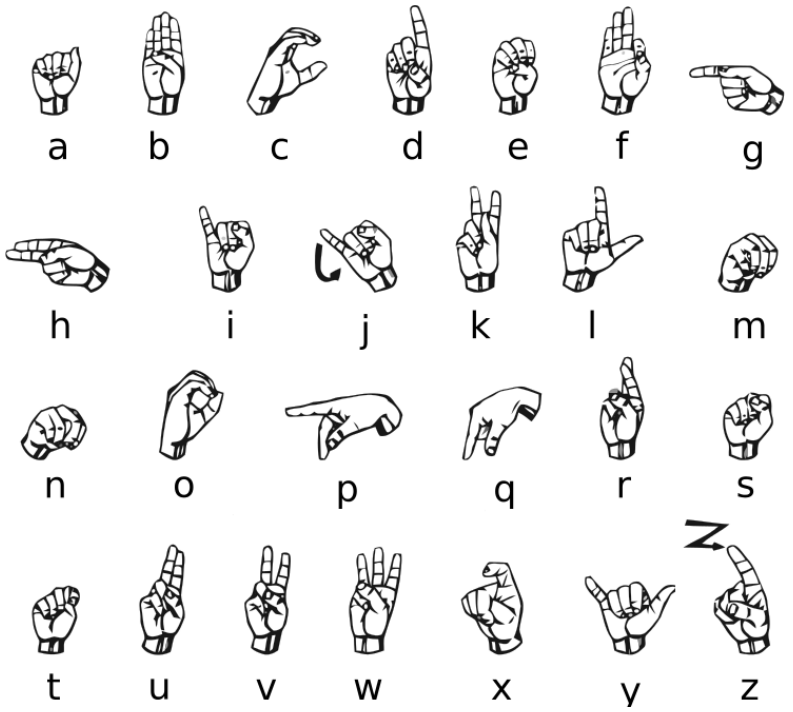



**Figure 5.1:** The fingerspelling signs of the ASL alphabet. We remark that all signs are constructed as static hand gestures, except from the signs for "J" and "Z", which involve hand motion for their gestural meaning. This overview of the American Sign Language alphabet is part of the image found at http://www.wpclipart.com/sign_language/American_Sign_La nguage_chart.png

RC features have proven their value for effective body part detection tasks, we expect that the STAGE detector is able to outperform the state-of-the-art approaches in classification performance.

## 5.2 THE AMERICAN SIGN LANGUAGE

The American Sign Language (or ASL - see, e.g., Battison & Baird, 1978) consists of a limited set of specific hand shapes that represent (1) the letters of the alphabet, and (2) several numerical values. The majority of the signs for the alphabet (i.e., 24 out of 26 signs) are constructed using static hand poses, while two signs require some form of motion. Figure 5.1 shows an overview

of the signs used to construct the ASL alphabet. We remark that all signs are constructed as static hand gestures, except from the signs for "J" and "Z", which involve hand motion for their gestural meaning. As the majority of the signs consists of static gestures, this allows us to focus on the recognition of the hand shape, rather than on the motion of the hand. Recognising the individual signs, however, remains a challenge. Pugeault & Bowden (2011) and Kuznetsova, Leal-Taixé, & Rosenhahn (2013) identified three main reasons why automatic sign language recognition is difficult:

1. the visual similarity between the signs;
2. the *inter-subject* variability in the production of the signs;
3. the *intra-subject* variability in the production of the signs.

To fully understand the challenging nature of automatic sign language recognition, we will discuss these difficulties in more detail.

First, several signs bear a strong visual resemblance (i.e., visual similarity; see Definition 5.1) with each other. For example, the signs for the letters A, S, and T are based on small variations of a closed fist; the meaning of the signs is derived from the position of the thumb. Similar signs increase the risk of confusion.

**Definition 5.1: Visual similarity**

Visual similarity is defined as the extent to which entities (e.g., a gesture) bear a visual resemblence with each other.

Second, variations in the ability to sign in a clear way lead to a high variability in the appearances of the signs, i.e., the inter-subject variability (see Definition 5.2). As a result, the construction of the signs is not consistent within a group of people.

**Definition 5.2: Inter-subject variability**

Inter-subject variability is defined as variations in the construction of a given entity (e.g., a gesture) between subjects.

Third, variations in hand pose (or, say, camera position) result in variations in the appearance of the individual signs. As a result, the construction of the

signs by individual people is not consistent over time, i.e., the intra-subject variability (see Definition 5.3).

**Definition 5.3: Intra-subject variability**

Inter-subject variability is defined as variations in the construction of a given entity (e.g., a gesture) within a single subject.

Figure 5.2 illustrates the visual resemblance and variability in the construction of the ASL signs. In this Figure, the columns represent the signs for the letters A (left column), S (middle column), and T (right column). The rows represent individual people constructing the signs for these letters. By comparing the different rows, it is evident that the visual resemblance between the signs leads to ambiguity between the signs. The way in which each person constructs the signs varies highly between people. This leads to a high inter-subject variability.

## 5.3 RELATED WORK

Recognizing human gestures (such as sign language) in visual data became an active field of research over the last twenty years (see, for example, Rautaray & Agrawal, 2015; Mitra & Acharya, 2007; Aggarwal & Cai, 1999; Y. Wu & Huang, 1999; Pavlovic, Sharma, & Huang, 1997). As such, modern day approaches towards gesture recognition can be divided into two computer vision-based categories (see L. Chen, Wang, Deng, & Ji, 2013): (1) approaches that primarily use visual data (see, e.g., Ghosh & Ari, 2015; Li, Yu, Wu, Su, & Ji, 2015), and (2) approaches that primarily rely on depth data for their recognition tasks (see, e.g., Brandão et al., 2014; Tang et al., 2014). While present-day approaches towards gesture recognition in visual data (the first category) achieve a near-perfect classification performance on the available datasets (see, e.g., the work by Li et al., 2015), they are sensitive to naturally occurring factors that may influence the quality of visual data. Variations in illumination conditions and skin tone, for example, may influence the detector's ability to separate the hand from its background negatively (cf. Rautaray & Agrawal, 2015). Advances in sensing technologies, however, allow for rapid improvements in the robustness and quality of gesture recognition solutions (see Kuznetsova et al., 2013) by, for example, using depth data for gesture detection tasks (the second category; see, e.g., Brandão et al., 2014; Tang et al., 2014).

When incorporating depth data, accurate gesture recognition approaches require feature extraction methods that are able to encode the local depth

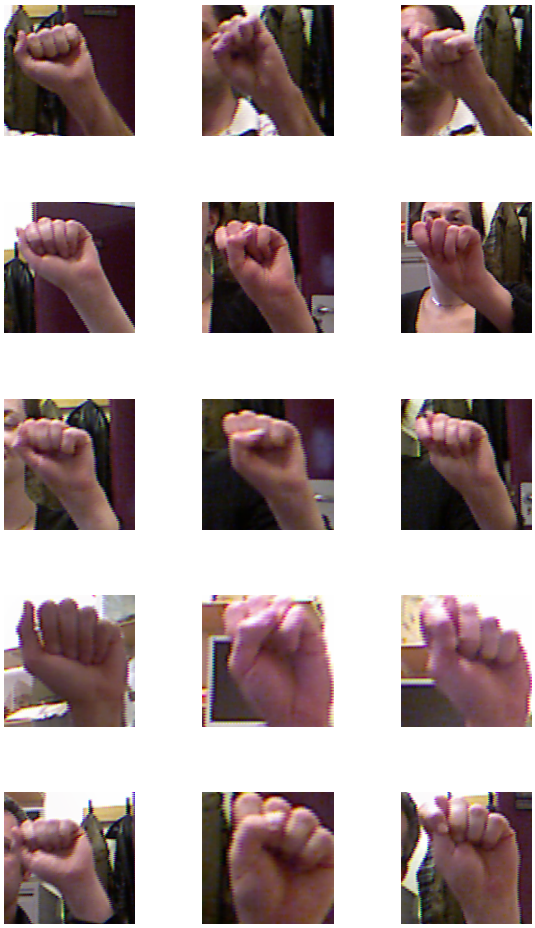


**Figure 5.2:** Illustration of the visual resemblance and variability in the ASL signs. In this Figure, the columns represent the signs for the letters A (left column), S (middle column), and T (right column). The rows represent five people producing the signs for these letters.

transitions which are typically associated with a person's hands. This requirement inspired us to develop an approach towards static gesture recognition that incorporates our RC features. Our approach relates to several contributions in the field of hand gesture recognition with depth cameras. Thus, it falls under the category of depth-based approaches towards gesture recognition.

In what follows, Subsection 5.3.1 describes four related approaches towards static gesture recognition. Subsequently, Subsection 5.3.2 discusses how these approaches inspired us to develop our approach.

### 5.3.1 Related Approaches

Below, we briefly describe four related approaches towards static gesture recognition. We characterise them as methods that (1) recognise parts of a hand by classifying individual pixel locations in a depth image, (2) use Gabor filters to compute local hand patterns, (3) separate the hand from its background using a threshold, and (4) use grid-like structures to compute distinctive descriptors for the input images.

First, Keskin, Kıraç, Kara, & Akarun (2013) propose a generalisation of the well-known body part detector that was proposed by Shotton et al. (2011). Keskin et al. aim to classify the individual hand skeleton parameters. Their approach aims to estimate the locations of the joints in the hand by performing per-pixel classifications. A random decision forest (see, e.g., Breiman, 2001) assigns the pixels to the individual parts of the hand, thus providing an estimation of the hand shapes and gestures.

Second, Pugeault & Bowden (2011) propose the use of multiscale filter banks with Gabor filters (see, e.g., Jain & Farrokhnia, 1990) to recognise static gestures in their ASL dataset. Gabor filters can be used to capture transitions in depth images, i.e., depth transitions. As such, Gabor filters can be used to describe local hand patterns in depth data. A random decision forest is then employed to perform the final classification. The disadvantage of using Gabor filters is that they require that the input images are rescaled to predefined dimensions.

Third, Pedersoli, Benini, Adami, & Leonardi (2014) aim to recognise static hand poses in their hand pose and gesture recognition framework. Their approach first segments the hand from the background of the input depth image by using (1) a mean shift segmentation, and (2) a hand palm detection procedure. Similar to Pugeault & Bowden (2011), Gabor filters are then used to extract features from the input depth data. A Support Vector Machine (SVM) classifier performs the final classification.

Fourth, Kapuscinski, Oszust, Wysocki, & Warchol, 2015) (2015) propose to perform gesture recognition on a point cloud of depth measurements. Their approach divides the point cloud into a grid-like structure of 3-D cells, for which unique descriptors are calculated. According to Kapuscinski et al., this allows for a more distinctive description of the depth image. The hand postures are classified using a nearest neighbour classifier.

### 5.3.2 Inspiration for Static Gesture Recognition

In what follows, we briefly discuss how the approaches described above inspired us to develop our approach towards static gesture recognition.

First, inspired by Keskin et al.'s (2013) approach, we adopt the idea to extract depth comparison features from individual pixel locations in a depth image. While Keskin et al. aim to recognise different parts of a hand by classifying individual pixel locations, our approach extracts depth descriptors from individual pixel locations to provide a description of the entire hand shape.

Second, following the idea by Pugeault & Bowden (2011), we use a multiscale filter bank to extract depth comparison features from the depth data. This allows us to encode depth transitions of various dimensions, i.e., local and global depth transitions.

Third, we adopt Pedersoli et al.'s (2014) idea to separate the hand from the background. This allows us to focus on the classification of the hand shape, which may allow for more accurate hand pose estimations.

Fourth, we adopt Kapuscinski et al.'s (2015) idea to divide the input scene into a grid-like structure and compute local descriptors for each element in the grid. The advantage of this approach is that we can compute a global descriptor by incorporating local information from several sampling locations.

## 5.4 THE STAGE DETECTOR

To investigate the effectiveness of the RC features for static gesture recognition (i.e., to what extent the RC features are able to recognise the fingerspelling signs of the American Sign Language), we propose a detector that incorporates our RC features for accurate static gesture recognition: the *static gestures detector*, or STAGE detector for short. The detector is an extension of the region comparison detector proposed in Chapter 3.

The STAGE detector recognises static gestures by estimating a person's hand pose in a depth image. To classify the hand shape, the detector first separates the hand from its background. This allows the detector to focus on the unique properties of the hand shape itself, i.e., the local depth differences that are typically associated with the hand pose. Then, the detector selects a subset of pixel locations in a grid-like structure from the depth image. For each pixel location in the subset, the detector computes multiple RC features. This provides a mathematical description of the entire hand, which is used for the final classification.

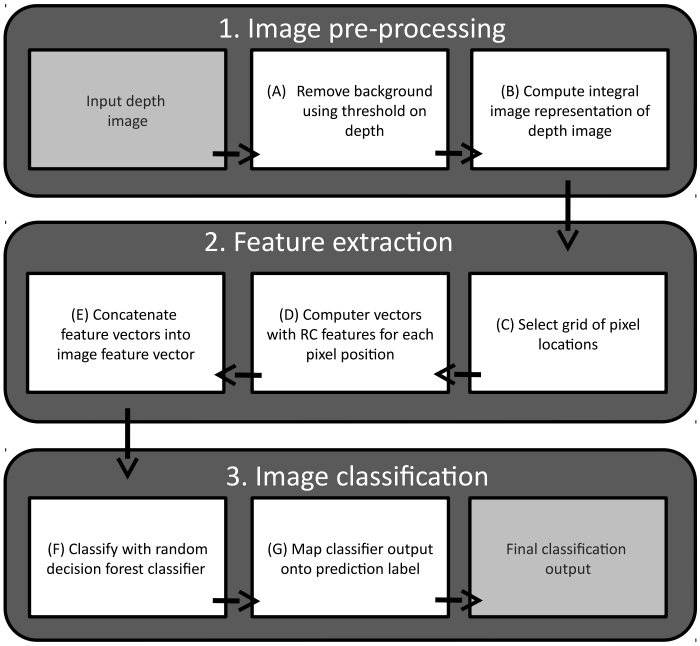



**Figure 5.3:** A diagram of the static gestures (STAGE) detector, showing its (1) image pre-processing stage, (2) feature extraction stage, and (3) classification stage, which are represented by grey rectangular areas, and the constituent sub-stages, which are represented as white boxes.

As such, the STAGE detector consists of three consecutive stages.

1. an *image pre-processing* stage to prepare the input image for the feature extraction process;
2. a *feature extraction* stage to extract and compute the RC features for the hand shape in the depth image;
3. a *classification* stage that incorporates a random decision forest classifier to classify the ensemble of RC features.

Figure 5.3 shows a diagram of the STAGE detector. In the Figure, the image pre-processing stage, feature extraction stage, and classification stage are represented by dark grey, rectangular areas. Their constituent sub-stages (A to G) are represented by white boxes. The input (step 1 in the first stage) and output (step 3 of the third stage) are represented as light grey boxes. All in all, the last step of the third stage forms the final classification.

The remainder of this Section presents the three individual stages. Subsection 5.4.1 discusses the image pre-processing stage of the detector. Subsequently, Subsection 5.4.2 discusses the feature extraction stage. Finally, Subsection 5.4.3 describes the classification stage.

### 5.4.1 Image Pre-processing

In the image pre-processing stage, the hand is separated from its background. Inspired by Pedersoli et al. (2014), we perform a separation procedure using a fixed threshold on the depth values (sub-stage A). An analysis of the dataset that is used in our experiments revealed that the hands of all subjects are present within a one meter range from the Microsoft Kinect device. Thus, we separate the hands from the background by disregarding all depth values beyond $1,000$ millimetres. The limitation of this solution is that it limits the maximal distance at which the hand gestures can be recognised. However, the low image resolution of the depth images makes it difficult to detect and segment small objects, such as a human hand (cf. Ren, Yuan, & Zhang, 2011). This is particularly the case when the distance between the Kinect device and the object increases (Khoshelham & Elberink, 2012). Thus, to achieve a high classification performance, we deem it necessary to use a fixed threshold on the depth values. After segmenting the background, the integral image representation of the depth image is computed (sub-stage B). Computing the integral image representation of the depth image ensures that the RC features can be computed efficiently. Figure 5.4 shows an example of a raw input depth image, before (Figure 5.4a), and after removing (Figure 5.4b) the background of the depth image.

### 5.4.2 Feature Extraction

After segmenting the hand from its background in the depth image, a subset of pixel locations is selected using a grid in the depth image (sub-stage C). For each point $P_n$ in the subset (where $n$ refers to the index of point $P$ in the subset, i.e., $n = \{1, 2, 3, ...n_{max}\}$), the feature computation procedure computes multiple RC features (sub-stage D) using distinctive feature types. Similar to the region comparison detector, the STAGE detector computes its features by calculating the sums of the pixel values enclosed in two or more rectangular regions around point $P$, and subtracting the sums from each other. The resulting features are combined into an RC feature vector that provides a (local) description of the depth differences in the area around point $P$. The resulting feature vectors for points $P_1$ to $P_{n_{max}}$ are then concatenated into a single feature vector that provides a description of the entire depth image (sub-stage E). Figure 5.4c shows an example of a depth image and the grid

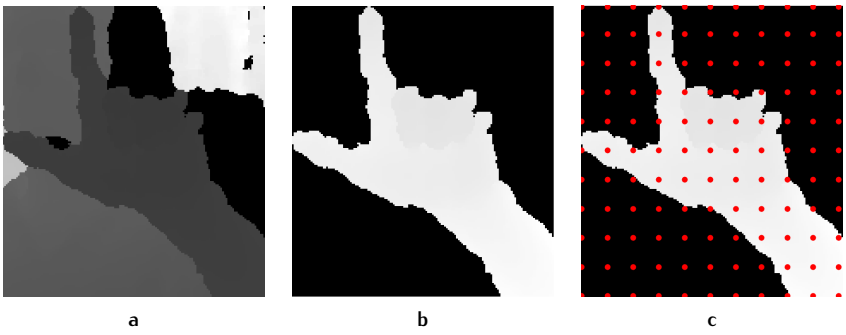



**Figure 5.4:** An example of a depth image that is to be classified by the STAGE detector. Figure 5.4a show the raw input depth image, while Figure 5.4b shows the same depth image after the removal of the background. Subsequently, Figure 5.4c shows the grid of pixels locations (displayed as red dots) for which the RC features are computed.

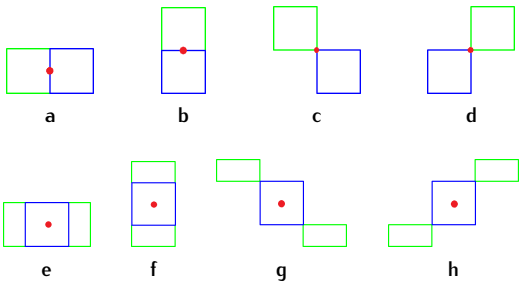



**Figure 5.5:** The 8 Region Comparison (RC) feature types that are used in the STAGE detector. The feature types are defined as a combination of four basic feature types (see Figure 5.5, a – d), and four specialised feature types that aim to describe very local depth transitions (Figure 5.5, e – h). The explanation of the feature types is given in the text.

that is used to select the subset of pixel locations. The red dots represent the pixel locations of points $P_1$ to $P_{n_{max}}$ in the grid, for which the RC feature vectors are computed.

As stated in Subsection 2.3.2, the RC feature types consist of a limited number of small rectangles that typically encode for local depth transitions in a depth image. Calculating features to describe local depth transitions is highly relevant for the detection and classification of small body parts, e.g., the individual fingers of a hand. To this end, the STAGE detector incorporates a total

of 8 different feature types, i.e., a combination of basic feature types (Figure 5.5, a - d) and four specialised feature types (Figure 5.5, e - h). In this Figure, the red dot represents point P in a grid of the depth image. The green and blue rectangles in each feature type represent the rectangular areas (regions) over which the RC feature (i.e., the depth differences) is computed. The basic feature types (a - d) allow for the computation of (a) horizontal, (b) vertical, (c) diagonal, and (d) anti-diagonal depth transitions. Variations derived from the basic feature types result in specialised feature types (e - k), which are able to encode more complex (local) depth transitions.

### 5.4.3 Image Classification

In the classification stage, a random decision forest (RDF) classifier is used to classify the depth images. To perform the actual classification, the STAGE detector trains and uses the standard random decision forest classifier (cf. Breiman, 2001) to classify the image feature vector for each depth image (sub-stage F). The prediction results of the RDF classifier are mapped onto a class label (sub-stage G), which forms the final classification output of the STAGE detector. Subsection 3.2.2 provides a more in-depth explanation of the RDF classifier.

## 5.5 EVALUATION PROCEDURE

This Section describes the experiments performed to evaluate the performance of the STAGE detector. The aim of the experiments is to investigate to what extent the STAGE detector (and therefore our RC features) enable accurate recognition of static gestures. To perform the evaluation, the detector is trained and evaluated on a challenging dataset with fingerspelling signs of the American Sign Language. Our evaluation focusses on the classification performance of the detector.

In what follows, Subsection 5.5.1 describes the dataset that is used in the experiments. Then, Subsection 5.5.2 discusses the performance metrics and criterion used to evaluate the performance of the detector. Subsequently, we give the implementation details of the detector in Subsection 5.5.3. Finally, the details of the experiments are discussed in Subsection 5.5.4.

### 5.5.1 Dataset

To assess to what extent RC features enable accurate sign language recognition, the detector is trained and evaluated on the publicly available *ASL*

*Fingerspelling Dataset*, which is developed by Pugeault & Bowden (2011). The dataset contains visual images (i.e., RGB - Red Green Blue) and depth data of 5 subjects constructing 24 static signs from the American Sign Language (ASL) alphabet. For our experiments, we only use the depth images in the dataset. Figure 5.1 shows an overview of the fingerspelling signs of the American Sign Language. We remark that Pugeault and Bowden excluded the signs for "J" and "Z" from the dataset, given that they involve hand motion for their gestural meaning.

The ASL dataset contains a total of 65,894 depth images, which are divided over the 5 subjects. As each subject constructs 24 fingerspelling signs, the dataset thus contains approximately 550 depth images per sign per subject. The dimensions of the depth images are approximately $100 \times 100$ pixels. The depth values of the depth images range from 0 to 4,095 (both inclusive). The labels of the fingerspelling signs are provided as plain text. Figure 5.4a shows an example of a depth image from the dataset.

### 5.5.2 Performance Metrics and Criterion

Below, the performance of the STAGE detector is quantified using two performance metrics: (1) a classification performance metric of the STAGE detector, and (2) the classification time metric. Moreover, we define the decision criterion that is applied for the comparison of the experimental results concerning the STAGE detector.

CLASSIFICATION PERFORMANCE METRICS The classification performance of the detector is quantified using the per-class detection accuracy of the detector. For a given size of the RDF classifier, the detection accuracy is calculated by averaging the detection accuracy over each fold. Within each fold, the detection accuracy is defined as the average accuracy achieved by the detector over all 24 fingerspelling signs.

CLASSIFICATION TIME METRIC The classification time metric measures the time required by a detector to identify the hand pose in a single depth image. A shorter classification time corresponds to a higher classification speed.

CRITERION Our decision criterion is established as straightforwardly outperforming the classification performance of the state-of-the-art. Thus, we consider the STAGE detector to be superior to the state-of-the-art approaches when the detector outperforms its opponents in classification performance. In Section 5.6 we will examine to what extent the impression holds. In what follows, the classification metrics and the classification time metrics are discussed.

### 5.5.3 Implementation Details

In the experiments, four types of parameters are used, viz. for (1) selecting the pixel locations and spatial search area, (2) the RC feature parameters, (3) the RDF classifier, and (4) the implementation of the detector. They are briefly discussed below.

SELECTING THE PIXEL LOCATIONS AND SPATIAL SEARCH AREA For each depth image, a subset of 121 pixel locations is selected in a $11 \times 11$ grid structure. The horizontal and vertical distance between the pixel locations is approximately 10 pixels. For each pixel position of the subset, a feature vector with RC features is computed. The maximal dimensions of the rectangles incorporated by the RC features are $24 \times 24$ pixels. The spatial search area over which the RC features are computed, is defined by the maximal dimensions of the rectangles incorporated by the RC features. As the feature types with the largest spatial dimensions (i.e., the ones shown in Figures 5.5g and 5.5h) incorporate 3 horizontally positioned rectangles and $(0.5 + 1 + 0.5 =)2$ vertically positioned rectangles, the maximal spatial search area of the RC features is $(3 \times 24) \times (2 \times 24) = 72 \times 48$ pixels.

RC FEATURE PARAMETERS The rectangle size parameter $r$ for the feature types that are used to compute the RC features, is defined as an integer value that increases over the course of 12 iterations. In the first iteration, the value of $r$ is initiated at 2. After each iteration, the value of $r$ increases with step size 2, up to its maximum value of 24. Hence, over the course of the iterations, the value of $r$ is defined as: $r = \{2, 4, 6, ...24\}$. The resulting RC feature vectors contain $(8$ feature types $\times 12$ iterations $=)96$ elements for each of the 121 pixel locations. After concatenating the feature vectors of the individual pixel locations, the resulting feature vector for the entire depth image contains $(121 \times 96 =)11,616$ elements.

RDF CLASSIFIER For our experiment, we use the MATLAB implementation of the random decision forest; the so-called TREEBAGGER[18]. For the RC features, each split node of the forest selects a random subset of $\sqrt{11,616} \approx 108$ candidate features. In the training procedure, the feature that best separates the observations is selected as the split node's test. Each tree of the random decision forest is trained until a minimum number of one observation per tree leaf is reached. The trees are not pruned.

IMPLEMENTATION OF THE DETECTOR The STAGE detector is implemented in MATLAB scripts. The implementations of the detector are publicly available

18 http://nl.mathworks.com/help/stats/treebagger.html

upon request. The entire training and evaluation procedure takes several days on a 50-core Linux calculation server.

### 5.5.4 Experimental Design

To evaluate the performance obtained by the ALS detector, we perform a classification experiment in which we evaluate the performance of the STAGE detector on the *ASL Fingerspelling Dataset*. This provides an indication of the extent to which the RC features are suitable for accurate static gestures recognition. In what follows, the experimental setup and the experiment are discussed.

EXPERIMENTAL SETUP In our experiment, we train and evaluate the STAGE detector on depth images from the *ASL Fingerspelling Dataset*; see Subsection 5.5.1. The dataset consists of five subjects constructing 24 fingerspelling signs each. To estimate the detector's generalisation performance, the data (i.e, the depth images) is separated into individual training sets and individual test sets.

To create the training and test sets, several approaches employ a validation procedure in which half of the data of each subject is selected at random and designated to the training set; the other half of the data is used as the test set; the so-called "50*v*s50" validation procedure (see, e.g., Kapuscinski et al., 2015; Li et al., 2015; Pugeault & Bowden, 2011). However, due to the way in which the data is divided into a training and test set, the 50*v*s50 validation method comes at the risk of classifying "seen" data samples. In casu, this means that the test data may be highly similar to the training data, given that it originates from the same subject in the dataset. While the classification performances achieved with the 50*v*s50 validation method reportedly surpass the performances achieved with cross-validation (see, e.g., Kapuscinski et al., 2015), it may result in a poor generalisation performance on depth images of unseen subjects. We therefore emphasise the requirement that the training and test data should be clearly partitioned using a suitable cross-validation procedure. To meet this requirement, we perform 5-fold cross-validation to create five folds with separate training and test sets. In each fold, the training set consists of 4 subjects from the dataset, while the test set consists of 1 subject.

In the experiment, the RC features are optimised for the small dimensions of the depth images by limiting the maximum size of the rectangles. As the complexity of the RDF classifier is not optimised prior to the experiment, we did not create a validation set. In the experiment, we measure (1) the classification performance of the STAGE detector, and (2) the time required to process an entire depth image. The highest average classification performance is used as the benchmark in the comparative evaluation.

EXPERIMENT Our evaluation procedure consists of (1) a *training* stage, and (2) an *evaluation* stage. In what follows, both stages are discussed in more detail.

In the training stage, the folds are used to train the STAGE detector using random decision forest (RDF) classifiers. On average, each fold consists of 52,715 training images and 13,179 test images. The images are labelled using the annotations that are provided by Pugeault and Bowden (2011). The training procedure is performed in several iterations. In each iteration, RDF classifiers are trained for the individual folds. The dimensions of the RDF classifiers are increased from 10 trees up to 100 trees (both inclusive), in steps of 10 trees. Additionally, we train an RDF classifier of 1,000 trees. Due to the computational power required to train the RDF classifiers, we refrained from training classifiers beyond 1,000 trees.

In the evaluation stage, we first assess the average classification performance of the STAGE detector over all folds. The test examples and the corresponding labels are used to evaluate the performance of the detector for different sizes of the RDF classifier. Repeating this procedure and averaging the classification performance over the individual folds result in an estimation of the classification performance on unseen data. Additionally, we investigate the variability within the dataset, and the extent to which the STAGE detector is able to distinguish the individual signs.

The highest average classification performance of the STAGE detector is compared with the performances of the state-of-the-art approaches that (1) include the *ASL Fingerspelling Dataset*, and (2) perform the 5-fold cross-validation procedure (and therefore meet the requirement regarding the partitioning of the dataset, as emphasised in the previous paragraph) to create their training and test sets. More specifically, we compare the classification performance of the STAGE detector with the highest performances as achieved by four competing approaches, i.e., the work by (1) Pugeault & Bowden (2011), (2) Kuznetsova et al. (2013), (3) Pedersoli et al. (2014), and (4) Kapuscinski et al. (2015). Similar to our approach, all aforementioned approaches experimentally optimised the parameters of their feature extraction methods prior to the experiment. The complexity of the classifiers was not optimised.

## 5.6 EXPERIMENTAL RESULTS

In this Section, we describe the experimental results of our evaluation. The aim of our experiment is to assess the classification performance of the STAGE detector. In what follows, Subsection 5.6.1 investigates the average classification performance of the STAGE detector over all folds, for several sizes of the RDF classifier. Then, Subsection 5.6.2 assesses the variability in the dataset.

Finally, Subsection 5.6.3 investigates to what extent the STAGE detector is able to distinguish the individual signs from each other.

### 5.6.1 Evaluating the classification performance

To investigate to what extent the dimensions of the forest influence the classification performance, we assess (1) the detector's classification performance (i.e., detection accuracy), and (2) its classification time using different sizes of the random forest. The results of our evaluation are shown in Figure 5.6.

Figure 5.6a shows the average detection accuracy over all folds for ten sizes of the RDF classifier. The results indicate that the detector approaches its optimal accuracy using a forest of rather small dimensions (say, 30 trees). Using a forest of 30 trees, the detector achieves a detection accuracy of 72% (SD = 16), up to 75% (SD = 16) for a forest of 100 trees. Increasing the size of the forest to 1,000 trees, results in an average detection accuracy of 77% (SD = 16). Due to the computational power required, we refrained from training forests that are larger than 1,000 trees. The results, however, weakly suggest that increasing the size of the random forest may result in a further small increase in classification performance.

Additionally, Figure 5.6b shows the classification times of the detector for ten sizes of the RDF classifier. Using a forest of 30 trees, the STAGE detector requires 0.8 seconds (SD = 0.1) to process an entire image, up to 17.6 seconds (SD = 1.2) for a forest of 1,000 trees. For the sake of readability, we did not include the latter in the Figure. The results suggest a rather linear relation between the size of the RDF classifier and the time required to process an entire depth image. In summary, the results of our experiment suggest that increasing the size of the forest leads to an increase in classification performance. However, the increase in detection performance comes at the cost of detection speed.

To complete this Subsection, we may conclude that the results of our experiment indicate that the STAGE detector achieves its highest average detection accuracy when using an RDF classifier of 1,000 trees. The results of our evaluation suggest that the STAGE detector outperforms the performance of the state-of-the-art approaches, i.e., the work by Pugeault & Bowden (2011), Kuznetsova, Leal-Taixé, & Rosenhahn (2013), Pedersoli et al. (2014), and Kapuscinski et al. (2015). Figure 5.7 shows the performance of the STAGE detector (using an RDF of 1,000 trees) in comparison with its competing approaches.

### 5.6.2 Assessing the Variability in the Dataset

To assess the variability in the dataset, we investigate the variations in detection accuracy scores between the individual folds. The *inter*-subject variability

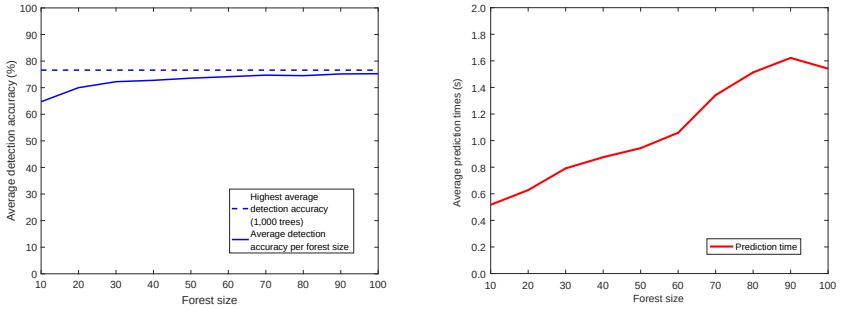



**Figure 5.6:** The average detection accuracy (a) and average classification times (b) of the STAGE detector over all folds, for ten sizes of the random forest (from 10 to 100). The first Figure shows that the detector approaches its highest classification performance of 77% (see dotted line, it is achieved by a forest of 1,000 trees; not visible in the Figure) using random forests of rather limited dimensions. The second Figure shows a linear relation between the size of the forest and the average classification time.

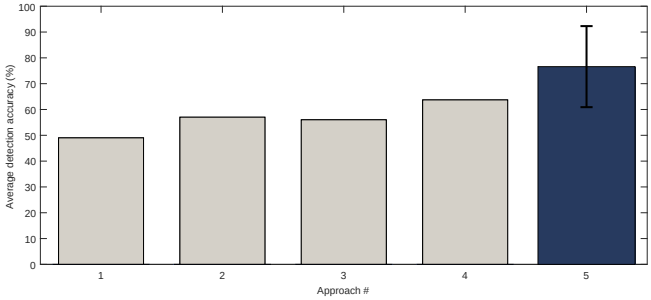



**Figure 5.7:** The detection accuracy achieved by (1) Pugeault & Bowden (2011), (2) Kuznetsova, Leal-Taixé, & Rosenhahn (2013), (3) Pedersoli et al. (2014), (4) Kapuscinski et al. (2015), and (5) the STAGE detector (including standard deviation). The STAGE detector uses a random decision forest of 1,000 trees. The results indicate that the STAGE detector achieves a higher detection accuracy.

(see Definition 5.2) describes the variations in the construction of the signs between different people, while the *intra*-subject variability (see Definition 5.3) describes the variations in hand pose or, say, camera position, which may result in variations in the appearance of the individual signs. To this end, we

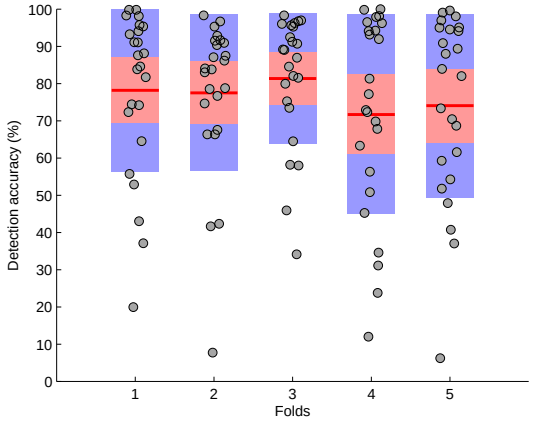



**Figure 5.8:** The per-fold detection accuracy for the gestures of the ASL dataset, using an RDF classifier of 1,000 trees. In this Figure, the grey dots represent the detection accuracy of the detector with respect to the individual gestures in each of the 5 folds. The horizontal lines (in this Figure shown in red) represent the mean per-fold detection accuracy of the detector.

investigate the performance of the STAGE detector yielding an RDF classifier of 1,000 trees. Figure 5.8[19] shows the per-fold detection accuracy of the detector. In this Figure, each fold represents a single subject in the dataset. In this Figure, the grey dots represent the detection accuracy of the detector with respect to the individual gestures in each fold. The red lines represent the mean per-fold detection accuracy, along with the 95% confidence interval (represented by purple bars), and 1 standard deviation (in this Figure represented by pink bars). The results of our evaluation suggest that there are no significant differences in construction between the individual subjects, i.e., a low level of variability in the dataset.

### 5.6.3 Detecting Individual Signs

To assess to what extent the STAGE detector is able to recognise the individual ASL signs, we investigate the classification performance of the STAGE detector with respect to the individual signs of the dataset. For our investigation, we again use an RDF classifier that consists of 1,000 trees. The results are shown in Figure 5.9 and Table 5.1.

19 Figures 5.8 and 5.9 are created using the MATLAB toolbox provided by R. Campbell at http://www.mathworks.com/matlabcentral/fileexchange/26508-notboxplot-alternative-to-box-plots

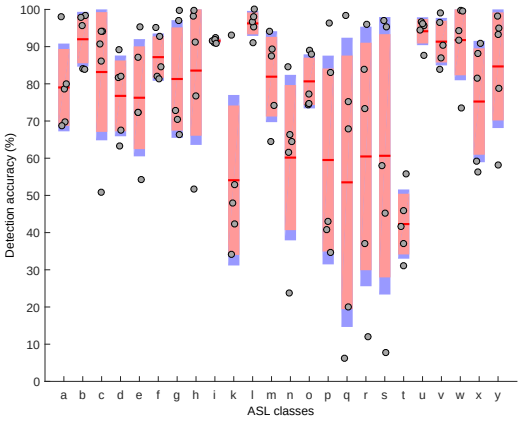



**Figure 5.9:** The per-class detection accuracy for the gestures of the ASL dataset, using an RDF classifier of 1,000 trees. The detector achieves a high detection accuracy for the majority of the individual signs, but a low accuracy for others signs.

Figure 5.9 shows the per-sign classification performance for the individual folds. In the Figure, the grey dots represent the detection accuracy of the detector for the individual folds. The red lines represent the mean per-sign detection accuracy, along with the 95% confidence interval (represented as purple bars), and 1 standard deviation (represented as pink bars). The results indicate that the STAGE detector achieves a high mean detection accuracy for the majority of the individual signs (e.g., the signs for the letters B, I, L, U, V, and W), i.e., well over 90%. The mean detection accuracy for several other signs (e.g., the signs for the letters K, N, and P to T), however, is considerably lower, i.e., far below the average of 77% accuracy.

Table 5.1 shows the distribution of the average detection scores over all folds. In the table, the rows represent the actual class labels (e.g., the letters A, B, ...) of the fingerspelling signs in the experiment. The columns represent the predicted class labels of the signs. The average detection accuracy (i.e., the percentage of correctly classified signs, averaged over all folds) is represented as rounded percentages. In the table, green represents a detection accuracy of at least 76%, while red represents a detector accuracy below 76%; higher is better. Orange indicates the most relevant misclassifications for the individual signs, i.e., the cases in which the detector wrongfully classifies 10% or more of the signs to this class label; lower is better. For example: in 15% of the cases, the detector predicts that a sign represents the letter N, while it actually represents the letter M. These results confirm that the detector achieves a high

**Table 5.1:** The distribution of the average detection scores over all folds. The rows represent the actual class labels of the fingerspelling signs, while the columns represent the predicted class labels. The mean detection scores are represented as rounded percentages. In the table, green represents a high detection accuracy, while red represents a low detector accuracy. Orange indicates detection errors above 10%.

| | A | B | C | D | E | F | G | H | I | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 79 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 6 | 0 | 0 | 2 | 0 | 3 | 7 | 0 | 0 | 0 | 0 | 0 |
| **B** | 1 | 92 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **C** | 0 | 0 | 84 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **D** | 0 | 0 | 0 | 77 | 1 | 1 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 1 | 0 | 0 | 7 | 0 | 0 | 2 | 0 | 0 | 2 | 1 |
| **E** | 0 | 0 | 2 | 0 | 76 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 5 | 1 | 0 | 0 | 8 | 4 | 0 | 0 | 0 | 1 | 0 |
| **F** | 0 | 1 | 0 | 0 | 0 | 87 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 1 | 4 | 0 | 0 |
| **G** | 0 | 0 | 0 | 0 | 0 | 0 | 81 | 6 | 0 | 3 | 0 | 0 | 0 | 0 | 1 | 3 | 0 | 2 | 2 | 0 | 0 | 0 | 1 | 0 |
| **H** | 1 | 0 | 0 | 0 | 0 | 0 | 10 | 84 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **I** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 92 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 2 |
| **K** | 0 | 0 | 0 | 6 | 0 | 0 | 7 | 0 | 1 | 53 | 2 | 0 | 0 | 0 | 1 | 0 | 7 | 0 | 0 | 3 | 14 | 2 | 3 | 0 |
| **L** | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 96 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **M** | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 82 | 7 | 0 | 0 | 0 | 0 | 5 | 2 | 0 | 0 | 0 | 0 | 0 |
| **N** | 3 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 15 | 61 | 0 | 0 | 0 | 0 | 1 | 15 | 0 | 0 | 0 | 3 | 0 |
| **O** | 0 | 0 | 1 | 0 | 7 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 81 | 3 | 2 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| **P** | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 4 | 1 | 1 | 0 | 0 | 3 | 4 | 60 | 13 | 0 | 0 | 2 | 0 | 0 | 4 | 2 | 0 |
| **Q** | 2 | 0 | 0 | 0 | 1 | 1 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 20 | 54 | 0 | 1 | 0 | 0 | 1 | 4 | 1 | 0 |
| **R** | 0 | 0 | 0 | 6 | 1 | 2 | 0 | 0 | 0 | 6 | 1 | 0 | 0 | 0 | 0 | 0 | 57 | 0 | 0 | 26 | 0 | 0 | 0 | 0 |
| **S** | 3 | 0 | 1 | 0 | 5 | 0 | 1 | 0 | 0 | 0 | 0 | 17 | 1 | 3 | 1 | 0 | 0 | 56 | 13 | 0 | 0 | 0 | 0 | 0 |
| **T** | 8 | 0 | 0 | 0 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 13 | 21 | 1 | 2 | 0 | 0 | 7 | 42 | 0 | 0 | 0 | 0 | 0 |
| **U** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 94 | 0 | 0 | 0 | 0 |
| **V** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 91 | 1 | 0 | 0 |
| **W** | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 3 | 90 | 0 | 0 |
| **X** | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 5 | 1 | 1 | 0 | 1 | 0 | 1 | 3 | 0 | 0 | 76 | 0 |
| **Y** | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 4 | 0 | 2 | 0 | 0 | 0 | 5 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 85 |

mean detection accuracy for the majority of the individual signs. The results, however, also indicate that the detector achieves relatively high error rates on a limited number of fingerspelling signs. The detector tends to confuse, for example, (1) the signs for the letters P and Q with each other, as well as (2) the signs for the letters M, N, S, and T.

In summary: there are no significant variations in detection accuracy *between* the individual folds (see Subsection 5.6.2), although the results do indicate that there are large differences in the detection accuracy achieved *within* the individual folds, i.e., between the individual signs. The results suggest that there are considerable differences in the way the individual subjects construct some of the individual signs.

## 5.7 DISCUSSION

In this Chapter, we investigated to what extent the RC features enable accurate gesture recognition in depth data. To investigate their effectiveness, we propose and evaluate the STAGE detector, which is an extension of the region comparison detector proposed in Chapter 3. The performance of the detector is evaluated on a dataset with depth images of the American Sign Language (ASL). The results of our evaluation reveal that the STAGE detector achieves a higher detection accuracy than the state-of-the-art approaches.

The remainder of this Section discusses the implications of the results in more detail. In what follows, Subsection 5.7.1 addresses the use of RC features for accurate gesture recognition. Subsequently, Subsection 5.7.2 discusses the differences in detection accuracy for the individual signs of the ASL alphabet. Finally, Subsection 5.7.3 discusses our future work, and the steps to be taken before the STAGE detector can actually be deployed to recognise human gestures.

### 5.7.1 Enabling Accurate Gesture Recognition

The results of our evaluation show that the STAGE detector outperforms its competing approaches in classification performance. This raises the question why the deployment of RC features enables the superior detector scores that are achieved by the STAGE detector. As stated in Subsection 5.2, recognising the individual signs remains a challenge due to three difficulties: (1) the visual similarity between the signs, (2) the *inter-subject* variability in the production of the signs, and (3) the *intra-subject* variability in the production of the signs. Approaches aiming to recognise the individual signs should thus be able to (1) take the global hand shape into account, and (2) detect subtle local differences between the individual signs. The STAGE detector computes its RC features using a grid-like structure of pixel locations, which enables the detector to calculate depth descriptors for the entire image. As the RC features are calculated over regions of various dimensions, this allows for the computation of *global* as well as *local* depth transitions. On the one hand, the RC features average over large regions in a depth image, which allows the detector to estimate the global hand shape. On the other hand, the features take local depth differences into account, which allows for the detector of subtle depth differences between the individual signs. We believe that the combination of global and local depth sampling explains the high performance of the RC features.

### 5.7.2 Recognising the Individual Signs

The results of our experiment indicate that the STAGE detector achieves a high mean detection accuracy for the majority of the individual signs. However, the results also show that the detector achieve a significantly lower classification performance for a limited number of fingerspelling signs. As a case in point, the detector tends to confuse the signs for the letters M, N, S, and T, which leads to lower detection scores for those signs. We argue that this is caused by the high degree of similarity between the signs. The signs for the letters M, N, S, and T, for example, are all based on a raised fist (see Figure 5.1). The position of the thumb constructs the meaning of the individual signs. This results in a high degree of visual similarity (and thus in very subtle differences) between the individual signs. Given the low quality of the depth data (see Subsection 2.1.3), we argue that increasing the quality of the input depth data may result in a higher classification performance of the STAGE detector.

### 5.7.3 Future Work

As two signs in the ASL alphabet involve motion for their gestural meaning, the evaluation procedure of the STAGE detector was limited to the recognition of the 24 static gestures from the ASL alphabet. As stated in Section 5.1, gestures exist in two distinctive forms: (1) *static* gestures and (2) *dynamic* gestures. The ability to recognise static gestures in the ASL alphabet is a first step towards accurate gesture recognition in general. As a next step, it is imperative that the STAGE detector is extended with the ability to (1) detect, (2) track, and (3) recognise dynamic gestures. We refrained from developing such extensions for the detector, because the focus of the experiment was on the evaluation of the RC features, rather than the detector's ability to detect and track gestures. Future work may therefore improve the detector with the aforementioned additions. It is to be expected that the performance of the RC features will be reflected in any detection approach that incorporates the features for hand pose recognition.

For our evaluation procedure, the STAGE detector was implemented as a combination of several MATLAB scripts. As stated in Subsection 3.5.3, the MATLAB environment is not optimised for speed. Implementing (parts of) the STAGE detector in a dedicated programming language (e.g., C++, Python, or equivalent) may speed up the processing time of the detector. Our evaluation results show that the detector is able to achieve a near-optimal classification performance, while requiring less than a second to process an entire depth images. We expect that developing a C++ or Python implementation of the

detector allows it to run in real time, i.e, to be able to process several frames per second on reasonable hardware.

This Chapter focusses on the recognition of static gestures in depth images. Future work may extend the STAGE detector to include visual data as well, similar to approaches such as the work by Li et al. (2015). As stated in Section 1.5, visual data is rich in detail, yet sensitive to external factors such as the illumination conditions. The advantage of using visual data (in addition to depth data) is that it may provide additional contextual information regarding small, individual parts of a hand, e.g., the fingers. Combining and aligning (1) visual features and (2) depth features may thus increase the general performance of the STAGE detector, provided that it this does not result in a significant increase in the computational budget required.

## 5.8 CHAPTER CONCLUSIONS

The research question of this Chapter is RQ 4: *To what extent do Region Comparison features enable accurate recognition of static gestures when using in-depth information?* To answer the research question, the Chapter evaluates the effectiveness of the RC features for accurate gesture recognition. To investigate the effectiveness of the RC features, we propose and evaluate the STAGE detector, which is an extension of the region comparison detector proposed in Chapter 3. It incorporates the RC features for accurate sign language recognition. The detection performance is evaluated on a dataset with depth images of the American Sign Language. The results of our evaluation reveal that the STAGE detector outperforms the state-of-the-art approaches in classification performance. The RC features enable for the computation of global depth transitions as well as local depth transitions. As the RC features are calculated over regions of various dimensions, this allows for the computation of global as well as local depth transitions. They allow the STAGE detector to estimate both the global hand shape as well as subtle depth differences between the individual signs. Based on our results, we may provisionally conclude the following.

- Conclusion 1: Due to a high degree of visual similarity between static gestures, identifying the individual signs proves to be a challenge.
- Conclusion 2: The RC features are able to distinguish subtle differences in depth data.
- Conclusion 3: The RC features contribute to accurate static gestures recognition in depth images.
- Conclusion 4: The RC features outperform the state-of-the-art in the field of static gesture recognition.

Future work may combine the RC features with visual features for an increase in classification performance.

### Research Continuation

In this Chapter, the RC features have shown their value for accurate gesture recognition, which is an important step towards embodied agents and intelligent environments that are able to perceive human behaviour and engage in natural interactions. As a next step in the establishment of natural interactions, it is imperative that smart agents are able to perceive on which objects (or topics) their human communication partners focus their attention. Hence, the next Chapter investigates to what extent RC features are suitable for accurate head pose estimation.

# 6 MIRROR, MIRROR ON THE WALL

*"The human is indissolubly linked with imitation; a human being only becomes human at all by imitating other human beings."*

– Theodor Adorno

The techniques investigated so far allow embodied agents to perceive and understand (some) social cues. The ability to understand these cues is a breakthrough step towards the establishment of a social connection between a person and an agent, which is a requirement for effective persuasion. It is, however, unclear to what extent it is actually possible to establish a social connection between humans and embodied agents, i.e., to what extent humans are able to perceive embodied agents as communication partners. As mimicking behaviour is widely considered to be a sign of a social connection between people, this Chapter[20] investigates the effect of agents on the mimicking behaviour of humans. This topic is highly relevant in the context of persuasive technology. Investigating a person's mimicking behaviour therefore provides an indication of the extent to which it is possible to establish a social bond between a human and a virtual agent. An experiment is conducted in which participants interact verbally with a virtual embodied agent. During the interaction, both the vocal pitch and the affective facial expressions of the agent are locally manipulated and the consecutive vocal and facial expressions of the participants registered. Computational analyses of the recorded expressions reveal vocal and facial mimicry as a sign of unconscious affect recognition and social connection.

The course of the Chapter is as follows. First, Section 6.1 outlines the relevance of people that unconscious imitate the cues sent out by embodied agents. Subsequently, the Section presents the details of RQ 5. Next, Section 6.2 describes the background of the experimental paradigm and the methodology used for the experiment. The results of the experiment are described in Section 6.3,

20 This Chapter is based on work by R. J. H. Mattheij, M. Postma-Nilsenová, and E. O. Postma (2015); Mirror, Mirror in the Wall: Is there mimicry in you all? Published in the *Journal of Ambient Intelligence and Smart Environments (JAISE)*.

while their implications for the development of embodied agents and intelligent environments in general are discussed in Section 6.4. We conclude upon our findings in Section 6.5.

## 6.1 SOCIAL SIGNALS AND EMBODIED AGENTS

Recent progress in the automatic processing of affective and social signals (see, e.g., Murray-Smith, 2014; Pantic & Vinciarelli, 2014; Vinciarelli et al., 2012) enables the employment of smart devices in intelligent environments. Capable of sensing social cues such as (1) emotional facial expressions (e.g., distress, surprise), and (2) emotional vocal expressions (tone of voice), these devices fulfil a role as the eyes and ears of the envisioned intelligent environment. Employing such smart devices enables the intelligent environment to perceive and understand the behaviour of people who are present, and, eventually, allows the environment to respond to their social cues in a contextually appropriate manner, i.e., responding in a natural way. Aiming to respond in a natural way, the intelligent environment may rely on actuators, i.e., virtual agents that emit social signals by means of virtually generated facial, vocal, and gestural expressions. The ultimate goal of these virtual agents is to provide personalised and socially acceptable feedback to humans, and to engage in natural interactions with them, without being experienced as annoying or obtrusive.

The course of the Section is as follows. Subsection 6.1.1 outlines the role of behavioural imitation in human-human interactions. Subsequently, Subsection 6.1.2 discusses the importance of imitation for embodied agents that aim to influence human behaviour. Finally, Subsection 6.1.3 addresses the research question of this Chapter, and the focus of the experiment performed to answer it.

### 6.1.1 Behavioural Imitation

A prerequisite for successful interactions between virtual agents and humans is twofold: (1) agents should respond appropriately to social signals, and (2) agents should evoke social signals. One of the most basic components of human-human interactions with respect to non-verbal behaviour is its automatic imitation (see, e.g., Louwerse, Dale, Bard, & Jeuniaux, 2012; Breazeal & Scassellati, 2002; Chartrand & Bargh, 1999).

Mitchell (1987) defines the concept of *imitation* as follows:

- something **C** (the copy) is produced by an organism and/or machine, where
- **C** is similar to something else **M** (the model)
- registration (or perception) of **M** is necessary for the production of **C**, and
- **C** is designed to be similar to **M**.

Apart from its crucial role in, for example, learning (see Dautenhahn, Nehaniv, & Alissandrakis, 2003), behavioural imitation and convergence (i.e., an increased similarity) creates (1) feelings of rapport, (2) empathy, and (3) social bonding (cf. Chartrand & van Baaren, 2009), which have a positive effect on social approval. Human behavioural studies show that speakers tend to mimic (1) verbal communication, e.g., the pitch in a person's voice (see, e.g., Looze, Oertel, Rauzy, & Campbell, 2011), and (2) non-verbal communication, such as facial expressions (see, e.g., Fischer, Becker, & Veenstra, 2012; Niedenthal, Brauer, Halberstadt, & Innes-Ker, 2001) and gestures (e.g., Holler & Wilkin, 2011).

### 6.1.2 Imitating Humans

With respect to artificial entities such as robots and embodied agents, three lines of experimental evidence suggest that positive effects can be achieved if the artificial entities engage in imitating humans.

First, Bailenson & Yee (2005) found that embodied agents gained social influence over participants when they mimicked the participants' head movements. Even though the participants did not explicitly notice the mimicry, the mimicking agents were more persuasive and received more positive trait ratings than their non-mimicking counterparts.

Second, Bevacqua, Hyniewska, & Pelachaud (2010) conducted an experiment in which an embodied agent in the role of the listener employed both unique smiles and mimicked smiles during an interaction with a human participant. The results show that participants smiled longer and more often when the embodied agent performed some smiling behaviour. Moreover, in both smiling conditions the agent was rated more positively than in the condition in which it never smiled. Hence, the results indicated that the frequency and duration of a participant's smile was influenced by the smiling behaviour of the embodied agent and that an embodied agent's behaviour can therefore positively influence a participant.

Third, Meltzoff, Brooks, Shon, & Rao, (2010) used behavioural imitation by robots as a way to bias children to treat the robots as psychological agents who can "perceive". In an experiment with 18-month-old infants, they showed that (1) humanoid robots which were observed to imitate the actions performed by the experimenter, and (2) the actions of the robots which were mimicked, were more likely to elicit gaze following. The outcome of the experiment suggests that (a) engaging in imitative behaviour is an important part of social interaction from an early age, and (b) that it can have an impact on humans' attention shifting.

### 6.1.3 Mimicking Embodied Agents

The imitation of a human sender's social signals leads to the agent being perceived as a socially appealing partner (see, e.g., Castellano, Mancini, Peters, & McOwan, 2012; Michalowski, Simmons, & Kozima, 2009; Bailenson & Yee, 2005). It is, however, unclear how fast, and in what manner, imitation takes place when the sender of the social cues is a human-like embodied agent. Given that imitation is recognized as an important social cue and as a factor affecting preferences and behaviours in commercial settings (see, e.g., Chartrand & Lakin, 2013; Stel, Mastop, & Strick, 2011; Tanner, Ferraro, Chartrand, Bettman, & Van Baren, 2008), examining the effect of agents on the imitative behaviour of humans is highly relevant in the context of persuasive technology. In particular, it is interesting to investigate to what extent humans exhibit behavioural mimicry in the form of copying facial expressions and vocal characteristics, which is a form of imitation that is mostly unconscious and unintentional (e.g., Chartrand & Lakin, 2013; Bell, Gustafson, & M., 2003). If humans, in fact, unknowingly imitate different non-verbal cues of the agent, their behaviour can be interpreted as an indicator of real social engagement. The research question of this Chapter therefore reads as follows.

> RQ 5: *To what extent do people mimic verbal and non-verbal cues sent out by an embodied agent?*

The experimental study described in this Chapter investigates to what extent humans display mimicking behaviour towards an emotionally responsive embodied agent, in particular its facial expressions and vocal characteristics. An earlier study examined global mimicking behaviour averaged over multiple interactions (Mattheij & Nilsenová, M. and Postma, E. O., 2013). The current study is directed at mimicry (see Definition 6.1) in individual interactions of male and female humans with a (female) embodied agent.

**Definition 6.1: Mimicry**

Mimicry is defined as a significant correlation between (1) the facial expressions of the participants and the embodied agent, or (2) the vocal pitch of the participants and the embodied agent, when the direction of the correlation points towards the embodied agent.

In the experiment, the participants are exposed to an embodied virtual agent that generates both verbal and non-verbal social signals. The experiment meets the conditions of non-obtrusiveness and human-like feedback (as described in Section 1.3) by deploying a realistically looking embodied agent. The agent meets the condition of non-obtrusiveness by having the agent communicate through both different facial expressions and subtle changes in the pitch of voice. It meets the condition of human-like feedback by using an embodied agent with a humanoid appearance, which incorporates emotional facial expressions that are modelled on the movement of human 'facial action units' (cf. Ekman & Friesen, 1978). To quantify the mimicking behaviour of the participants in our experiment, we consider significant correlations in (1) the facial expressions of the participants and the embodied agent, or in (2) the vocal pitch of the participants and the embodied agent as positive signs of mimicry, especially when the direction of the correlation points towards the embodied agent.

## 6.2 METHODOLOGY AND EXPERIMENT

This Section describes the experiment performed to measure the extent to which participants show mimicking behaviour when interacting with embodied agents. The aim of the experiment is to find evidence for mimicry in (1) emotional facial expressions, and (2) vocal pitch. The goal of the visual analysis was to find signs of mimicry in emotional facial expressions. The goal of the auditory analysis was to find a statistically significant change in the vocal pitch of participants in response to a low-pitched or high-pitched voice of the embodied agent.

In what follows, the methodology of the experiment is described. First, the background of the experiment is described in Subsection 6.2.1. Then, Subsection 6.2.2 presents the participant pool used for the experiment, as well as the design of the experiment. Subsequently, Subsection 6.2.3 describes the mate-

rial used. Finally, Subsection 6.2.4 describes the criteria and methods used to analyse the results (in the form of perceived data) from the experiment.

### 6.2.1 Background

The experiment consists of a verbal interaction between a participant and an embodied agent. The latter expresses multiple facial expressions and utilises different levels of vocal pitch. Facial expressions and vocal properties are the most important source of information about the speaker's emotions, mental states, and personality traits; the ability to process the information is at the core of *social intelligence* (cf. Vinciarelli et al., 2012). In particular, the human voice is a reliable indicator of social signals because the voice production mechanism directly reflects various physiological changes related to emotional responses (cf. Scherer, Johnstone, & Klasmeyer, 2003). Past research indicates that among the acoustic parameters that can be measured in the voice, the most robust effects can be found with respect to *pitch*, the perceptual correlate of fundamental frequency in the voice (cf. Juslin & Laukka, 2003). Given that strategic monitoring and regulation of facial expressions appears to be easier than conscious modification of affect-expressing vocal patterns, acoustic measurements possibly constitute a more reliable cue to social signals than visual data. As such, *visual* as well as *vocal* mimicry is examined in our experiment.

### 6.2.2 Participants and Design

In total, 73 participants (25 men and 48 women; mean age 20.1) were recruited from the Tilburg University student population. They received course credits for their participation. All participants were native speakers of Dutch. They participated in an interactive task that was presented as a word-association game with an embodied agent. The within-participant independent variables manipulated in the task are the agent's vocal pitch (High, Low) and seven facial expressions (Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise; (see Ekman & Friesen, 1978)). The dependent variables measured are the degree of participants' unconscious vocal and facial mimicry as indicators of social partnership.

### 6.2.3 Material

Commercially available software was used to create a realistic embodied agent with convincing facial expressions. In order to draw participants' attention to the facial expressions, the agent was designed as a female human head. The entire experiment consisted of (1) a training sequence of 3 trials, and (2) the ex-

perimental session (in a sequence of 28 trials). In each trial, the agent changes its neutral facial expression to one of the seven emotional expressions (duration: 1.0 second), while verbally producing a word. The word was produced with either a high or a low pitch of voice. Each trial lasted 2.5 seconds, including two 500 milliseconds transitional phases, i.e., a phase in which the agent changed its facial expression from neutral to one of the seven facial expressions, and vice versa. The task of the experimental participant was to react by the first verbal association that came to mind within a four-second time frame. Each combination of a facial expression and a pitch of voice was repeated twice ($7 \times 2$), which resulted in 28 trials per participant.

LEXICAL PROPERTIES The words that are used as primes in the experiment were selected from the list of the 100 most frequent Dutch nouns in the SoNaR corpus (cf. Oostdijk, Reynaert, Hoste, & Schuurman, 2013), which contains over 500 million contemporary Dutch words. To control for the duration of the stimuli, only disyllabic nouns were included in the list. The selected words were all semantically neutral, which excluded a possible effect of semantic and prosodic (in)congruence (e.g., a semantically positive word combined with a sad facial expression). Appendix A provides the full list of stimuli.

VOCAL CHARACTERISTICS The words were synthesized using the Text-Aloud text-to-speech (TTS) software which employs the L&H TTS3000[21] Dutch TTS engine. To fit the female agent used in the experiment, the TTS engine employed the L&H Karen voice package. The average pitch of the synthesized words was approximately 170 Hz which can be described as a low female voice. Using PRAAT (cf. Boersma & Weenink, 2012)[22], a randomly selected half of the stimulus material was raised in pitch with 40 Hz and resynthesized. Given that a normally hearing listener is able to distinguish speech sounds that differ in 5 Hz, the manipulation resulted in perceptually clearly distinguishable pitch variation.

FACIAL EXPRESSIONS The facial expressions in this experiment displayed the basic emotions identified by Ekman (cf. Ekman & Friesen, 1971), i.e., Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise, all presented four times in a random order. The expressions were created using *HapFACS* (Amini, Yasavur, & Lisetti, 2012), an API to generate dynamic 3D facial expressions based on the Facial Action Coding System (FACS) (Ekman & Friesen, 1978). HapFACS enables the creation of facial expressions by activating the relevant facial action units and the corresponding intensities.

21 http://www.ttsmaster.com/download
22 http://www.fon.hum.uva.nl/praat/

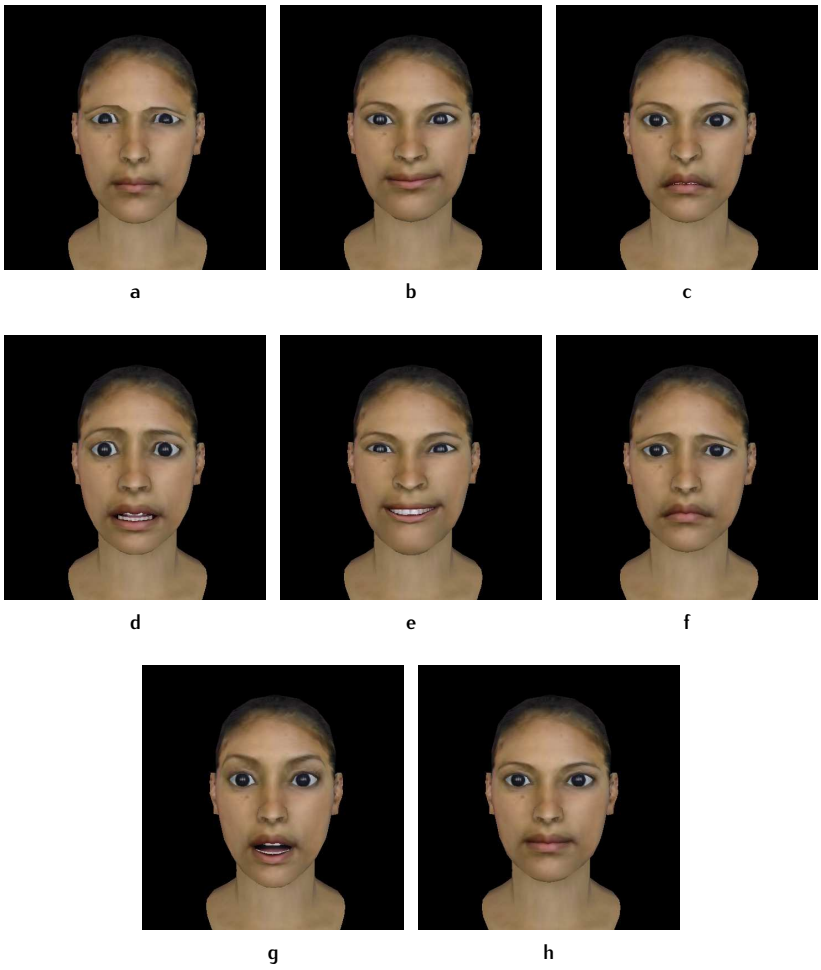



**Figure 6.1:** The eight facial expressions (seven emotional expressions and the neutral expression) employed by the embodied agent: (a) Anger, (b) Contempt, (c) Disgust, (d) Fear, (e) Happiness, (f) Sadness, (g) Surprise, and (h) Neutral.

Table 6.1 lists (1) the expressions, (2) the corresponding Action Units (AUs) and intensities (cf. Amini et al., 2012), and (3) the effect of activating the AU that were employed to create the facial expressions for the experiment. The individual AUs that are used to construct the facial expression are identified by unique coding numbers. Unless specified otherwise, all action units are activating bilaterally. Note that the intensities of the action units are expressed

**Table 6.1:** An overview of the combinations of action units and their intensities employed to create the facial expressions of the embodied agent. The intensities of the action units vary between A (lowest intensity) and E (highest intensity). Unless specified otherwise, all action units are activated bilaterally.

| Facial expression | Action Units | Effect |
|---|---|---|
| **Anger** | 4D | Brow lowerer |
| | 5E | Upper lid raiser |
| | 7C | Lid tightener |
| | 23C | Lip tightener |
| **Contempt** | 12B | Lip corner puller |
| | R14E | Dimpler right |
| **Disgust** | 9E | Nose wrinkler |
| | 15D | Lip corner depressor |
| | 16E | Lower lip depressor |
| **Fear** | 1D | Inner brow raiser |
| | 2D | Outer brow raiser |
| | 4D | Brow lowerer |
| | 5E | Upper lip raiser |
| | 20C | Lip stretcher |
| | 26E | Jaw drop |
| **Happiness** | 6C | Cheek raiser |
| | 12E | Lip corner puller |
| | 25C | Lip parts |
| **Sadness** | 1C | Inner brow raiser |
| | 4D | Brow lowerer |
| | 15D | Lip corner depressor |
| **Surprise** | 1C | Inner brow raiser |
| | 2D | Outer brow raiser |
| | 5B | Upper lid raiser |
| | 26E | Jaw drop |

as a degree between A (lowest intensity) and E (highest intensity). There is one exception: AU 14 is not activated bilaterally. In that case, the number is

preceded by a letter (R14) indicating the right (dimpler). The default intensity of an AU is 0 (zero), which corresponds to a deactivated Action Unit. Figure 6.1 displays the emotional expressions of the embodied virtual agent.

Using HapFACS, two instruction scripts were created for each of the seven basic emotions of the agent. Both scripts defined the changes of the intensities of the constituent AUs over time. For each of the agent's facial expressions, the first script specified the transition of the constituent AUs from a neutral expression to one of the seven emotional expressions, while the second set specified the transition in the reverse direction. The duration of both transitions were set to 500 milliseconds, which yields realistic facial-expression onsets and offsets.

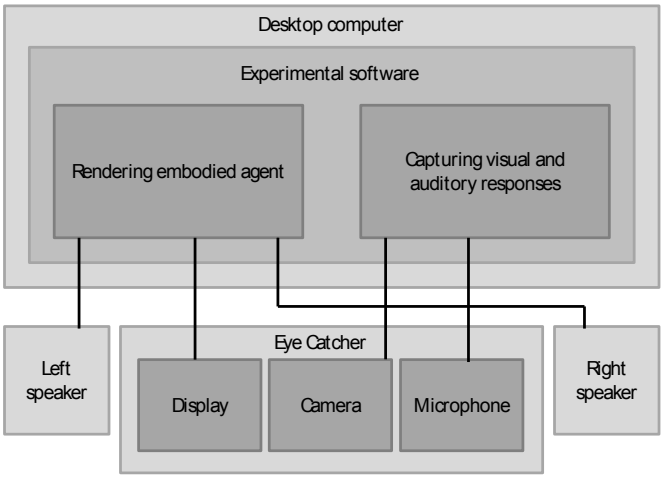
EXPERIMENTAL SETUP AND PROCEDURE The experimental setup consisted of a desktop computer[23] that ran the experimental software, two external audio speakers and an Eye Catcher device. The computer rendered the embodied agent[24] in full-screen mode against a black background. A computer mouse was connected to the computer. All other input devices were disabled. The Eye Catcher device[25] is commercially available teleconferencing equipment that was employed to display the embodied agent and capture the participants' visual and auditory responses using its internal camera and microphone. The resolution of the Eye Catcher is 800 $\times$ 600 pixels. The recordings were saved as a single AVI video file with a frame rate of 25 frames per second. The DV Video Encoder filter was employed to compress the video signal and the IMA ADPCM audio codec was used to compress the audio signal. The Eye Catcher and the speakers were positioned on a table in front of the participant at a distance of 40 cm and 30 cm, respectively; the computer was located on an adjacent table. Figure 6.2c shows a model participant watching the embodied agent.

Prior to each experimental session, the camera of the Eye Catcher was manually adjusted to capture a frontal view of the face of the participant. Participants were instructed to watch the screen and to respond to each word uttered by the agent with an arbitrarily associated word; the point of the interaction was to play an association game with the agent. Each session started by the presentation of a black screen with a red cross for a duration of 1 second. The location of the red cross corresponded to the location of the agent's eyes, drawing the participants' attention to the eye-region of the agent. At the end of the session, all participants filled out and signed an informed consent form, as well as a non-disclosure agreement to prevent them from informing other students of the content of the experiment.

23 The desktop computer employed in the experiment was a Dell Optiplex 740 with a 2.5 GHz dual-core processor and 4 GB RAM. All software ran on Windows XP.

24 http://www.haptek.com

25 http://www.qconferencing.eu/product/eye-catcher/




a

b

c

**Figure 6.2:** A schematic overview of the experimental setup (Figure 6.2a) and two pictures showing an impression of the experiment in which a participant engages in a conversation with an embodied agent (Figures 6.2b and 6.2c). The software developed for our experiment runs on a desktop computer. It renders the embodied agent, which is displayed on the Eye Catcher's screen. The agent's voice is produced through the external speakers. The participants' visual and vocal responses are captured using the Eye Catcher's built-in camera and microphone.

### 6.2.4 Criteria and Methods

This subsection first describes the criteria used to quantify the signs of mimicry. Then, it describes the methods for the visual and auditory analysis of the results.

CRITERIA FOR MIMICRY To quantify the mimicking behaviour of the participants in our experiment, we expect that the participants mimic the facial expressions or the pitch of voice of the embodied agents (see Definition 6.1). We therefore formulate two criteria: (1) a criterion to quantify facial mimicry, and (2) a criterion to quantify vocal mimicry

The visual analysis investigates the mean correlation between the facial expressions of the participant and the virtual agent. When a participant mimics the facial expressions of the agent, this should occur in a positive time lag starting after the initial facial expression. As such, our criterion to quantify facial mimicry (criterion 1) is defined as a positive correlation *after* the initialization of a facial expression by the agent. Correlations found before or somewhere after the agreed period are disregarded as facial mimicry.

The auditory analysis investigates the correlation between the mean pitch of the agent and the participants' responses. We expect that participants who mimic the vocal pitch of the embodied agent adapt their pitch to the pitch height of the agent. As such, we quantify the correlation between the pitch of the participants and the pitch of the agent as a significant shift in the participant's mean vocal pitch towards the pitch of the agent. We consider a significant shift in the participant's mean vocal pitch towards the pitch of the agent (criterion 2) a sign of vocal mimicry, while non-significant shifts in the mean vocal pitch are disregarded as signs of vocal mimicry.

VISUAL ANALYSIS The video sequences were analysed using the Computer Expression Recognition Toolbox (CERT) developed by Littlewort and colleagues (Littlewort et al., 2011). The video sequences of the embodied agent and of all participants were processed by CERT yielding time-series for the following seven emotional facial expressions: Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise. CERT computes scores on these expressions by estimating the presence of constituent FACS4.4 Action Units (AUs). On individual Action Units, CERT achieves an average estimation accuracy of almost 80% on a dataset of spontaneous facial expressions (Littlewort et al., 2011).

In the experiment, each participant performed 28 trials, in which 7 facial expressions were involved, i.e., 4 trials per facial expression per participant. Each of the 7 facial expressions obtained for each participant video, and the associated time-series obtained for the embodied agent, were submitted to time-dependent correlation analyses. The participant and embodied agent time-series were paired in the sense that they represented the same emotional expression. We employed Matlab's CROSSCORR function[26] to analyse mimicry in the individual human-agent interactions. The cross-correlation window size was 100 samples (about 4 seconds) and the lag ranged from $-100$ to $+100$ samples ($-4$ to $+4$ seconds). The emotions of participants mimicking those of the

26 http://nl.mathworks.com/help/econ/crosscorr.html

embodied agents should be reflected in a peak in the sample cross correlation in a positive time lag. Thus, a positive cross correlation in time indicates that an emotional expression of the agent is mimicked by the participant after some delay corresponding to the time lag. A time lag of 0 would indicate synchronous expressions of the agent and participant. Therefore, the time lag should be positive and within an appropriate range. Previous studies of facial mimicry report latencies ranging from 500 to 1500 milliseconds (Achaibou, Pourtois, Schwartz, & Vuilleumier, 2008; Sato & Yoshikawa, 2007; Dimberg & Thunberg, 1998), therefore we define peaks approximately within this range as probable signs of mimicry.

For each emotion and participant gender, the maximum cross correlation coefficient (peak), $CC_{max}$, and lag, $lag_{max}$, were computed from the average sample cross correlations. These were obtained by averaging over all participants with the same gender. It resulted in $7 \times 2$ (emotions $\times$ gender) pairs of $CC_max$ and $lag_{max}$ values as measures of visual mimicry.

AUDITORY ANALYSIS The audio recordings were analysed with the help of PRAAT 5.3.04 (Boersma & Weenink, 2012). The recordings were manually segmented. A visual and an auditory inspection were used to establish a speaker's pitch floor and ceiling in order to prevent pitch tracking errors due to octave jumps (cf. Boersma & Weenink, 2012). Generally, the range was set to 70 Hz - 250 Hz for male voices and 80 Hz - 400 Hz for female voices. Mean pitch of the voiced segments was estimated using the standard autocorrelation method (cf. Boersma, 1993). Only speech vocalisations were included in the acoustic measurements; non-speech sounds (e.g., filled pauses, laughter, and background noises) were filtered out. In total, pitch measurements were obtained for 1899 of the 2044 experimental trials (73 participants, 28 trials per participant); in 8(0.4%) of the trials, pitch was undefined, and in 137(6.7%) trials, the participants' output was missing or overlapped with the output of the embodied agent. The pitch measurements collected in the acoustic analysis were averaged per participant, participant gender, emotion (Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise) and pitch condition (High or Low Pitch, as used by the embodied agent in the prime preceding the participant's vocalisation).

## 6.3 EXPERIMENTAL RESULTS

The analysis of the video and audio recordings revealed evidence for mimicry. Below, the results for facial-expression mimicry are presented in Subsection 6.3.1. Then, Subsection 6.3.2 addresses the results for pitch mimicry.

### 6.3.1 Facial-Expression Mimicry

The main results for the analysis of facial expressions are listed in Tables 6.2 (female participants) and 6.3 (male participants). In both tables, the first column specifies the emotional expression under consideration, the second column displays the mean correlation at the location of the peak, mean $R_{max}$, and its standard deviation, and the third column displays the time lag $lag_{R_{max}}$ of the mean.

The clearest signs of mimicry are observed for the facial expressions of Disgust, Happiness, and Surprise (printed in boldface). For these emotions, the time lags of about $0.9 - 1.6$ seconds differ clearly from 0 with an average correlation of almost 0.5. There are no large differences in the intensity and delays of mimicry between females and males, suggesting that both respond similarly to the emotional expressions of the embodied conversational agent. Interestingly, for the facial expressions of the other emotions, the $R_{max}$ values are quite large. Given the short time lags ($< 500$ milliseconds), these facial responses are unlikely to be caused by mimicry and may reflect the actions of predictive mechanisms (cf. Kaufman & Johnston, 2014).

Figures 6.3 and 6.4 show a total of 14 graphs of the sample cross correlations for the seven emotions (rows) for female and male participants (left and right column, respectively). In each graph, the solid curve depicts the average correlation coefficient as a function of time lag. The shaded region represents the standard deviation from the mean. The dashed vertical line indicates the time lag at which the largest average correlation coefficient is obtained. Clearly, non-zero peaks are observed at positive time-lags. This indicates that the display of Disgust, Happiness or Surprise in the facial expression of the agent is likely to be followed by the display of the same emotional expression in the participant after about $0.9 - 1.6$ seconds in both females and males. Examination of the individual cross-correlation graphs for each participant revealed considerable individual differences in the degree of mimicry of facial expressions.

HUMANS MIMIC FACIAL EXPRESSIONS While we did not find clear signs of mimicry for all facial expressions, we did find signs of mimicry for the expressions of Disgust, Happiness and Surprise. The results indicate that participants unconsciously mimic several of the facial expressions expressed by the embodied agent.

### 6.3.2 Pitch Mimicry

The number of male and female participants ($> 20$ per between-subject cell, here: Gender) was judged to be sufficient to guarantee robustness to non-

**Table 6.2:** Main results of the visual analysis of facial-expression mimicry for female participants. The mean $R_{max}$ represents the mean correlation at the location of the peak and its standard deviation, while $lag_{R_{max}}$ represents the time lag of the mean. The clearest signs of facial expression mimicry are observed for the facial expressions of Disgust, Happiness and Surprise.

| Facial expression | mean $R_{max}$ (std) | $lag_{R_{max}}$ (secs) |
|---|---|---|
| Anger | 0.26 (0.18) | 0.28 |
| Contempt | 0.40 (0.19) | 0.08 |
| **Disgust** | 0.18 (0.22) | **0.88** |
| Fear | 0.30 (0.27) | 0.12 |
| **Happiness** | 0.24 (0.25) | **1.00** |
| Sadness | 0.29 (0.21) | 0.04 |
| **Surprise** | 0.34 (0.25) | **1.60** |

**Table 6.3:** Main results of the visual analysis of facial-expression mimicry for male participants. The mean $R_{max}$ represents the mean correlation at the location of the peak and its standard deviation, while $lag_{R_{max}}$ represents the time lag of the mean. The clearest signs of facial expression mimicry are observed for the facial expressions of Disgust, Happiness and Surprise.

| Facial expression | mean $R_{max}$ (std) | $lag_{R_{max}}$ (secs) |
|---|---|---|
| Anger | 0.26 (0.21) | 0.16 |
| Contempt | 0.43 (0.19) | 0.08 |
| **Disgust** | 0.23 (0.20) | **1.00** |
| Fear | 0.33 (0.27) | 0.12 |
| **Happiness** | 0.26 (0.24) | **1.20** |
| Sadness | 0.27 (0.24) | 0.08 |
| **Surprise** | 0.29 (0.24) | **1.40** |

normality (cf. Tabachnick & Fidell, 2007). A Box's M test (see Box, 1949) showed no significant difference between the Mean Pitch values in the responses of the male and female speakers to all stimuli, thus satisfying the assumption of homogeneity of the within-group covariance. Therefore, no adjustment of alpha levels due to unequal sample sizes of male and female speakers was necessary.

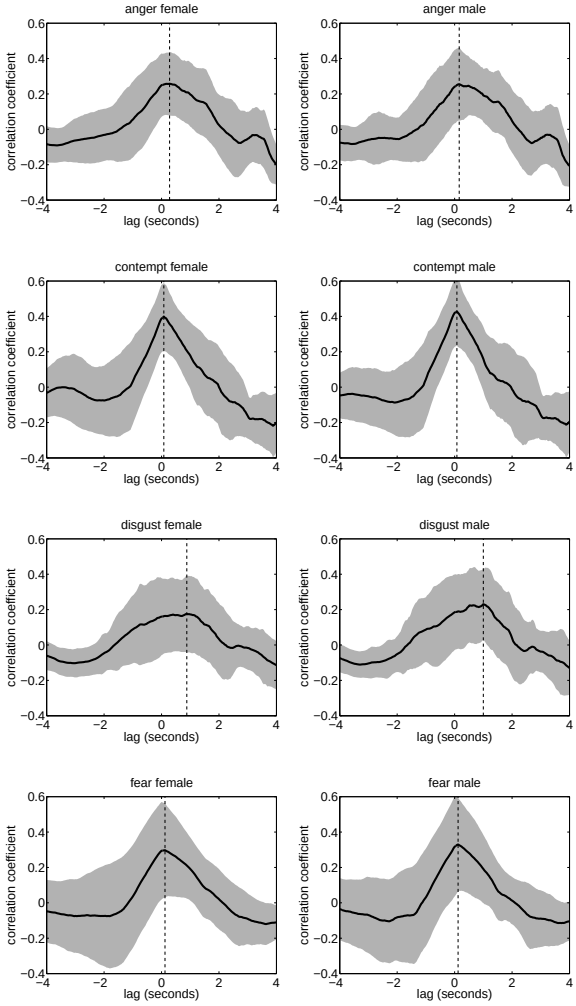



**Figure 6.3:** Overview of average correlation coefficients for the first four (Anger, Contempt, Disgust, and Fear) of the seven emotional expressions for female (left column) and male (right column) participants. In each plot, the solid curve depicts the average correlation coefficient as a function of time lag. The shaded region represents one standard deviation from the mean. The dashed vertical line indicates the time lag at which the largest average correlation coefficient is obtained.

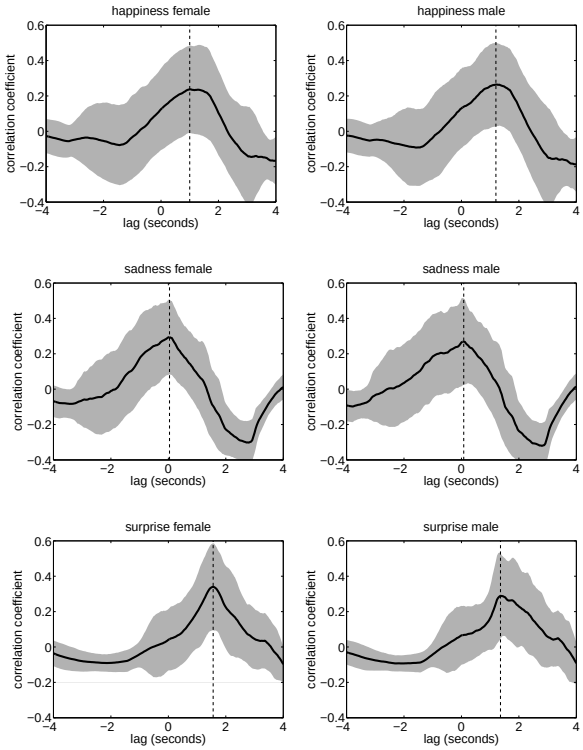



**Figure 6.4:** Overview of average correlation coefficients for the last three (Happiness, Sadness, and Surprise) of the seven emotional expressions for female (left column) and male (right column) participants. In each plot, the solid curve depicts the average correlation coefficient as a function of time lag. The shaded region represents one standard deviation from the mean. The dashed vertical line indicates the time lag at which the largest average correlation coefficient is obtained.

THE VARIABLE MEAN PITCH In both the within-participant conditions (High and Low Pitch), the variable Mean Pitch was not normally distributed (Shapiro-Wilk's (1965) test $< .0001$); therefore, a nonparametric test was used to compare the mean pitch values of the speaker's vocalisations in the two conditions. The Wilcoxon Signed Ranks test indicated that, overall, participants adapted their pitch to the pitch of the embodied agent (see Wilcoxon, 1945). Mean Pitch was lower in the vocalisations following a low pitch prime ($Mdn = 199$ Hz) compared to those uttered after a high pitch prime ($Mdn = 208$ Hz), $Z = -6.041$, $p < .0001$. The effect size measure suggested a large effect of

**Table 6.4:** Median values (in Hz) and main statistical results of the auditory analysis of pitch mimicry for the low and high pitch conditions. There are clear signs of pitch mimicry are observed for all facial expressions except Happiness.

| Facial expression | Low Pitch | High Pitch | Z | Significance |
|---|---|---|---|---|
| **Anger** | **201** | **206** | -2.837 | .005 |
| **Contempt** | **201** | **205** | -3.427 | .001 |
| **Disgust** | **194** | **206** | -5.132 | <.001 |
| **Fear** | **198** | **209** | -3.916 | <.001 |
| Happiness | 202 | 202 | -0.237 | .813 |
| **Sadness** | **201** | **211** | -3.916 | .003 |
| **Surprise** | **196** | **198** | -3.137 | .002 |

the experimental manipulation, $r = -.71$. Low-pitched agents induce lower pitched responses and high-pitched agents induce higher pitched responses.

POST HOC SPLIT-FILE ANALYSIS To compare the effect on male and female speakers, a post hoc split-file analysis was performed. The Wilcoxon Signed Ranks tests revealed a slightly higher effect for the male speakers ($Z = -3.969$, $p < .0001$, $r = -.79$) compared to the female speakers ($Z = -4.749$, $p < .0001$, $r = -.69$). In comparison to female participants, male participants showed a slightly higher tendency to shift their average pitch towards the pitch of the agent. In order to explore the effect of the agent's emotional expression on pitch mimicry, a series of nonparametric paired-samples tests was performed for Mean Pitch values collected in the two within-participant conditions (High and Low Pitch) for each of the seven emotions separately. Wilcoxon Signed Ranks tests revealed a significant difference for Anger, Contempt, Disgust, Fear, Sadness and Surprise, but not for Happiness. As shown in Table 6.4, for the six emotions where the difference was present (printed in bold), participants mimicked the pitch height of the agent by lowering their pitch in the Low-Pitch condition and raising it in the High-Pitch condition. Figure 6.5 displays the same results. For each emotion, the median pitch of the participants in response to the agent with a low-pitched voice (white bar) or high-pitched voice (grey bar) is shown. With the exception of Happiness, all emotions show a clear shift in the median vocal pitch of the participants towards the pitch of the agent.

HUMANS SHOW SIGNS OF VOCAL IMITATION The results of this experiment indicate that low-pitched agent vocalisation induced lower pitched responses

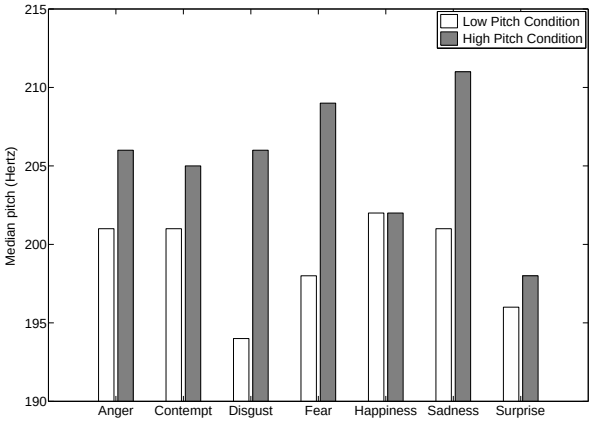



**Figure 6.5:** The main results of the auditory analysis at the level of emotional expressions. Seen from left to right, the bar plots show the results for the six basic emotions. For each emotion, the two bars show the median vocal pitch of the participant in the Low Pitch condition (white bar) and High Pitch condition (grey bar). For all emotions, except Happiness, participants adapt their vocal pitch to the pitch of the agent (please note the truncated range on the vertical axis).

by the participants, while a high-pitched vocalisation induced higher pitched responses. Our results therefore indicate that humans show signs of vocal mimicry when interacting with embodied agents. The implications of the results collected both in the visual and auditory domain are discussed in the next Section.

## 6.4 DISCUSSION

In this Chapter, we investigated to what extent people show mimicking behaviour when interacting with an emotionally expressive virtual agent. We quantify the mimicking behaviour of the participants as a significant correlation in (1) the facial expressions between the participants and the embodied agent, and (2) the vocal pitch between the participants and the embodied agent. Our results indicate that people unconsciously mimic both facial and vocal cues that are emitted by an embodied agent.

In what follows, the implications of the results are discussed in more details. Subsection 6.4.1 discusses the implications of human mimicking behaviour for the establishment of a social connection between humans and embodied agents. Subsequently, Subsection 6.4.2 discusses future extensions of our research.

### 6.4.1 Mimicry as a Sign of a Social Connection

The results of our experimental investigation support the notion of virtual embodied agents as social communication partners for humans in intelligent environments. In general, local manipulation of behavioural cues expressed in the visual and auditory domain led to changes in facial expressions and in vocal behaviour in the participants. Given that such changes are a sign of social connection in human-human interactions, our results suggest the establishment of a social connection between the participant and the agent.

With respect to facial expressions, the clear signs of facial-expression mimicry for the emotions of Happiness and Surprise point at a social connection. It is not clear why these two emotions gave rise to facial mimicry, whereas the others seem to fail to do so (with the exception of Disgust). Possibly, the details of the virtual embodied agent (appearance, level of realism) and task setting (not all participants seemed equally engaged in the task) determine the extent to which emotions evoke facial-expression mimicry. A second option that can be explored in future research is that positive emotions give rise to higher degrees of facial imitation, as suggested in Chartrand & Lakin (2013).

In the auditory domain, participants were quick to adapt the pitch of their voice to the perceptible change in the pitch of the agent. However, a more detailed analysis of pitch changes following different emotional facial expressions revealed no significant effect of the vocal manipulation after expressions of Happiness. This result is particularly striking in view of the pronounced imitation of the agent's visual expression in this emotion condition. A possible explanation is that the strong visual imitation of a happy expression actually resulted in higher pitch (viz. the median values reported in Table 6.4), given that high pitch is typically associated with positively valenced aroused states.

### 6.4.2 Future work

The results presented above may obscure the fact that there appear to be large individual differences in the way participants responded to the agent. For instance, female participants seem to be more responsive in terms of facial expressions than male participants, with the opposite effect found in the auditory domain. An important challenge for future study is the identification of

individual interaction styles and the automatic adaptation of agent behaviour to the preferred interaction style of the participant.

A recent point that needs to be addressed in the future is the cultural context of the interaction and the interpretation of different facial expressions by participants of different origins. Although many studies of emotional expressions assume a high amount of cultural homogeneity, there is growing evidence of cultural differences in non-verbal communication (see, e.g., Elfenbein, 2013; Jack, Caldara, & Schyns, 2012).

## 6.5 CHAPTER CONCLUSIONS

The research question of this Chapter is RQ 5. It reads: *To what extent do people mimic verbal and non-verbal cues sent out by an embodied agent?* To this end, an experiment was conducted to measure behavioural mimicry in individual interactions between humans and a human-like embodied agent. In the experiment, participants were exposed to an embodied agent that generated both verbal and non-verbal social signals. The embodied agent, equipped with a humanoid appearance, communicated through both different facial expressions and subtle changes in the pitch of voice.

The results of the experiment reveal that local manipulations of behavioural cues expressed in the visual and auditory domain led to (1) significant changes in several facial expressions, and (2) observable changes in the vocal behaviour of the participants. The changes in the participants' facial expressions and vocal behaviour are directed towards the embodied agent, i.e., mimicking behaviour. In human interactions, such changes are a sign of a social connection. The answer to the research question is therefore that humans do exhibit behavioural mimicry when interacting with an embodied agent, by matching both the facial expressions and the pitch of voice of the embodied agent. These results imply that humans are able to perceive virtual agents as potential communication partners. Based on our results, we may provisionally conclude that humans are able to establish a social connection with an human-like embodied agent.

# 7 CONCLUSIONS

*"I think and think for months and years. Ninety-nine times, the conclusion is false. The hundredth time I am right."*

– Albert Einstein

The work presented in this Thesis is part of the *Persuasive Agents* research project. As explained in Chapter 1, the project explores the use of socially aware virtual agents that persuade people to change their energy-consumption behaviour by providing them with subtle personalised feedback. Enhancing the ability of a virtual agent to perceive human non-verbal behaviour increases its ability to act as a persuasive agent. Thus, the Thesis investigates novel methods that enable agents to perceive a person's non-verbal cues and gestures as accurately as possible.

The structure of the Chapter is as follows. Section 7.1 answers the five research questions on the basis of the work in the Thesis. Subsequently, Section 7.2 formulates our conclusion to the problem statement.

## 7.1 ANSWERS TO THE RESEARCH QUESTIONS

In this Section, we provide the answers to the individual research questions.

> **Research question 1:** *How can we improve Shotton et al.'s body part detector in such a way that it enables fast and effective body part detection in noisy depth data?*

The first research question is investigated in Chapter 2. There, we proposed the use of region comparison (RC) features for fast and effective object detection in noisy depth data. The features provide a robust alternative to the pixel comparison (PC) features that were proposed by Shotton et al. (2013a,b; 2011). Based on the theoretical description given in the Chapter, we may conclude

that (1) comparing regions in a depth image has a clear advantage over comparing individual pixel values in that comparing regions allows for averaging over larger areas. We may further conclude that (2) the RC features are less prone to local pixel noise than the PC features, and (3) the RC features do not need an additional computational budget.

> **Research question** 2**:** *To what extent do Region Comparison features enable fast and accurate face and person detection in noisy depth images?*

The answer to the second research question is derived from Chapter 3, where we performed a comparative evaluation to investigate to what extent RC features contribute to fast and effective object detection in noisy depth images. From our empirical results we observe that the RC features outperform the state-of-the-art PC features in both classification performance and prediction speed. Our results reveal that the RC deal effectively with the background noise in depth images. The RC features provide an accurate indication of the direction and magnitude of the depth transitions in a depth image. We may therefore conclude that (1) RC features contribute significantly to fast and effective face and person detection in noisy depth images, (2) the RC features yield an improvement over PC features, and (3) the RC features are able to operate with the same computational budget.

> **Research question** 3**:** *How do we develop an annotated database that incorporates visual and depth data recordings of natural human gestures?*

To answer the third research question, Chapter 4 shows how the Tilburg Gesture Research (TiGeR) Cub has been developed. It is a multimodal corpus that consists of annotated, visual, depth, and audio recordings of dyadically interacting interlocutors. So, we may conclude that the answer to the research question resides in the experimental setup as given in Subsection 4.3.1 and in the methodology followed in the experiments (see Subsection 4.3.2). Both, setup and methodology led to the development of an annotated database that incorporates visual and depth data recordings of natural human gestures. Of course, annotating the huge quantities of data remains a challenge.

> **Research question** 4**:** *To what extent do Region Comparison features enable accurate recognition of static gestures when using in-depth information?*

The fourth research question is addressed in Chapter 5. In the Chapter, we evaluate the effectiveness of the RC features for accurate static gesture recognition. To perform the evaluation, we proposed a detector that incorporates the RC features for effective gesture recognition. The performance of the detector is evaluated on a dataset with depth images of static American Sign

Language (ASL) signs. Based on our results, we may conclude that (1) due to a high degree of visual similarity between static gestures, identifying the individual signs proves to be a challenge. We may further conclude that (2) the RC features are able to distinguish subtle differences in depth data, (3) the RC features contribute to accurate static gestures recognition in depth images, and (4) the RC features outperform the state-of-the-art in the field of static gesture recognition.

> **Research question** 5**:** *To what extent do people mimic verbal and non-verbal cues sent out by an embodied agent?*

Chapter 6 investigates the fifth research question. In the Chapter, we conducted an experiment to measure behavioural mimicry in individual interactions between humans and a human-like embodied agent. In the experiment, participants were exposed to an embodied agent that generated both verbal and non-verbal social signals. The embodied agent, equipped with a humanoid appearance, communicated through both different facial expressions and subtle changes in the pitch of voice. Our results revealed that local manipulation of behavioural cues expressed in the visual and auditory domain led to (1) significant changes in several facial expressions, and (2) observable changes in the vocal behaviour of the participants, i.e., mimicking behaviour. Our results therefore imply that humans are able to perceive virtual agents as potential communication partners. Based on our empirical results, we may conclude that humans are able to establish a social connection with an human-like embodied agent.

## 7.2 ANSWER TO THE PROBLEM STATEMENT

In this Section, we provide an answer to the problem statement. Our answer is based on the overall results reported in the thesis.

In Chapter 1 we outlined the importance of smart embodied agents that are able to establish and maintain a social connection between a person and the agent by using social signals such as affective facial expressions and vocal prosody. A requirement for the establishment of the social bond between the person and the embodied agent, is the latter's ability to respond appropriately to a person's social signals. In the Chapter, we identified depth data as a robust source of information. As depth data provides contextual information for a scene, it facilitates effective foreground-background segmentation. Therefore, depth data may be a robust alternative to the widely-used visual data. This calls for the development of novel computer vison algorithms that employ in-

depth information for the detection of human body parts and behaviour. As such, the problem statement of this Thesis was defined as follows.

> **Problem statement:** *To what extent is it possible to detect human body parts and behaviour when using in-depth information?*

In the Thesis, we found that the use of depth data for object detection purposes is hampered by two limitations: (1) the limited quality of the depth images, and (2) the limited resolution of the depth images, with result in noisy depth images. To deal with these challenges, we proposed the use of RC features. The RC features average over larger regions in a depth image, making them less prone to local pixel noise. By outperforming several state-of-the-art competing methods in a series of experiments, the features have proven their worth for fast and effective body part and gesture recognition in noisy depth data.

Seeing that the RC features are able to deal effectively with the background noise in the depth data without leading to insurmountable computational costs, we may conclude that is it possible to perform accurate human body parts and behaviour recognition by means of in-depth information that is encoded by RC features. We may further conclude that using the RC features for human body part and behaviour recognition tasks may enhance an agent's cognitive abilities. Our findings and their implications for the design of embodied agents are discussed in the next Chapter.

# 8 GENERAL DISCUSSION

*"The aim of argument, or of discussion, should not be victory, but progress."*
– Joseph Joubert

To establish the envisioned interactions between humans and embodied agents, we assume that it is possible to create a strong social connection between a person and an agent. To initiate the actual interactions, the agents use advanced artificial intelligence techniques to analyse a person's non-verbal behaviour. As the agents are likely to be deployed in noisy environments, i.e., environments with a large variety of objects, changing illumination conditions, and moving people, this necessitates the use of state-of-the-art computer vision algorithms that (1) allow for accurate behaviour recognition, and (2) are able to deal with the noisy nature of the environment. To meet the requirements, we proposed six objectives that enable the agents to perceive a person's non-verbal cues and gestures more accurately (see Subsection 1.6.2). Moreover, we investigated to what extent our assumption regarding the social bond between a human and a virtual agent holds. This Chapter discusses our findings, as well as their implications for the design of embodied agents.

The course of this Chapter is as follows. First, Section 8.1 reflects upon the implications of our findings for the design of smart agents and intelligent environments. Then, Section 8.2 discusses the points of improvement of our studies. Finally, Section 8.3 presents pointers to future work.

## 8.1 TOWARDS SOCIALLY AWARE EMBODIED AGENTS

As stated in Chapter 1, the aim of the Thesis is to facilitate natural interactions between humans and embodied agents. Enabling the agents to perceive a person's social cues is a first step towards natural human-embodied agent interactions.

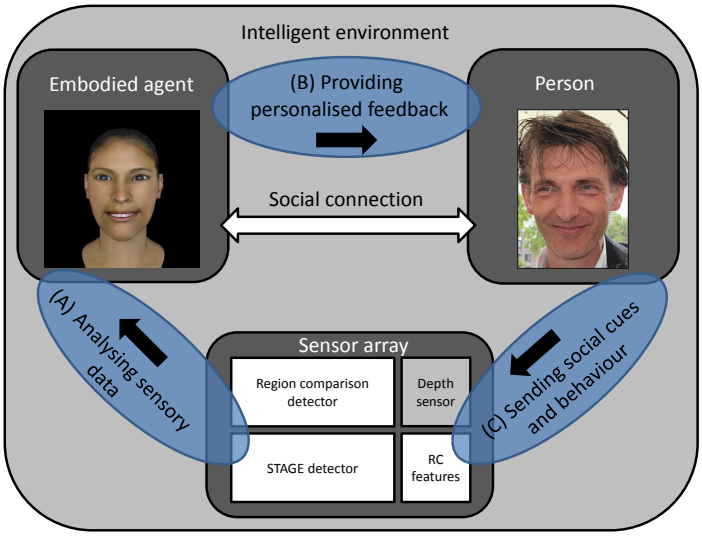



**Figure 8.1:** A model describing the envisioned interactions between humans and embodied agents, and the establishment of the corresponding social bond. In this Figure, our main contributions to the establishment of the interactions (i.e., the object and gesture recognition algorithms) are represented as light grey rectangles. This Figure is a modified reproduction the interaction model that is presented in Figure 1.2. We refer to Section 1.4 for a detailed description of the interaction model.

Based on the findings in the Thesis, our contributions are two-fold. On the one hand, we introduce a set of novel computer vision algorithms that allow the agents to perceive a person's non-verbal cues and gestures accurately. On the other hand, we investigate to what extent it is possible to establish a social bond between a human and a virtual agent. In this Section, we reflect upon our objectives, as well as their implications for the design of smart embodied agents and intelligent environments.

In Section 1.4, we presented an interaction model to describe the envisioned human-embodied agent interactions (see Figure 1.2). To provide the proper context for our research objectives (see Subsection 1.6.2), we extend the interaction model by including our main objectives. Figure 8.1 shows the extended interaction model. Please note that the model itself is not validated in this Thesis. It merely serves as a guideline for the reader to illustrate the envisioned social interactions between humans and embodied agents.

In the Figure, the first three objectives are represented as white rectangles.

1. The RC features, i.e., robust depth comparison features that are proposed in Chapter 2.
2. The region comparison detector that is proposed in Chapter 3, which allows for fast and effective body part and person detection in depth data.
3. The STAGE detector (as proposed in Chapter 5), which is able to recognise static gestures in depth images.

Our fourth research objective follows from our findings in Chapter 6. In that Chapter, we gained

4. advanced insights into the extent to which people are able to perceive a virtual person as a true communication partner.

The findings of Chapter 6 provide an indicater of the degree to which it is possible to establish a social bond between a person and an embodied agent. In the Figure, the social bond is represented by the white bi-directional arrow. Our last two research objectives (not shown in the Figure; see Chapter 4) concern the development of

5. the TiGeR Cub corpus, which contains annotated RGB-D recordings of naturally interacting interlocutors, and
6. the AnnoTool, which ensures detailed annotations of the data in the corpus.

The first three objectives allow for fast and effective body part and gesture recognition, which enable agents to perceive a person's behaviour and gestural cues more accurately. Our objectives enable agents to respond properly to a person's behaviour, e.g., by interpreting a person's gestures, or following him with its gaze. The fourth objective investigates to what extent humans are able to perceive embodied agents as communication partners. It provides an indication of the extent to which is possible to establish a social connection between humans and embodied agents. Thus, it provides insights into the extent to which social signals that are sent out by an embodied agents can influence a person's behaviour. We therefore state that objectives 1 to 4 are directly relevant for the development of socially adaptive agents (see, e.g., Ben Youssef et al., 2015; Van Welbergen, Ding, Sattler, Pelachaud, & Kopp, 2015). Thus, the objectives may ultimately increase the human-likeness of embodied agents (see, e.g., Gris, Rivera, & Novick, 2015; Wargnier et al., 2015).

The fifth and sixth objective are the TiGeR Cub corpus and our annotation tool. The corpus can, for example, be used to (1) train and evaluate machine

learning algorithms for gesture recognition tasks, and (2) to study the synthesis of gestures. The annotation tool can be used to create detailed annotations of objects and body parts in depth data. While they do not contribute directly to the development of socially aware agents, we argue that both objectives are relevant for the development of the next generation of gesture and behaviour recognition approaches. As the agents require accurate computer vision algorithms to detect a person's behaviour, our last two objectives support the development of accurate gesture recognition approaches, making them relevant for the development of socially aware embodied agents.

## 8.2 POINTS OF IMPROVEMENT

In this Section, we discuss three points of improvement and provide two alternative methods to further enhance the accuracy and robustness of the proposed approach. In what follows, Subsection 8.2.1 discusses the use of depth data for accurate behaviour recognition. Subsequently, Subsection 8.2.2 reflects upon the selection of the features types of the RC features. Then, Subsection 8.2.3 discusses the use of RC features for body part and gesture recognition detection tasks. Finally, Subsection 8.2.4 discusses two alternative methods.

### 8.2.1 The Use of Depth Data

The agents are likely to be deployed in noisy environments, i.e., environments with a large variety of objects, changing illumination conditions, and moving people. Thus, the agents require state-of-the-art computer vision algorithms that are able to handle the noisy nature of the environment. Many present-day approaches towards automatic object detection, however, rely on visual data as their main source of information (see, e.g., Q. Chen et al., 2015; Khalligh-Razavi, 2014; Andreopoulos & Tsotsos, 2013). While rich in detail, the disadvantage of visual data is that it is sensitive to the illumination conditions, such as shadows or bright lights (see, e.g., Qu et al., 2015; Rautaray & Agrawal, 2015; Shah & Kaushik, 2015). Shadows, for example, may obscure objects from sight, which may make them difficult to detect. Thus, the quality of information that is extracted from visual data suffers from the illumination conditions present. This makes the use of visual data unpractical in noisy scenes.

In Chapter 2, we introduced depth data as an alternative to visual data. As depth data combines spatial and depth cues (see, e.g., Brandão et al., 2014; Tang et al., 2014), depth data provides contextual information for a scene, which facilitates image segmentation (see, e.g., Brunton et al., 2014; Jiang et al., 2013). Moreover, depth cues are invariant to the illumination conditions in

a scene. Our motivation to use depth data as the main source of information for the computer vision algorithms of the agents is thus two-fold. On the one hand, depth data allows for effective foreground-background segmentation, which enables computer vision algorithms to separate a person's body parts accurately from their background. On the other hand, the quality of depth data is not influenced by the illumination conditions present, which makes depth data a suitable source of information in noisy environments.

The ability to perform accurate object detection in depth data is hampered by two limitations: (1) the limited quality of the depth images, and (2) the limited resolution of the depth images, with result in noisy depth images. Our RC features are able to deal with the noise effectively by averaging over larger regions in the images, which enable high detection performances. Using a combination of visual (RGB) and depth (D) data, however, may combine the best of both worlds: (1) the high level of detail from visual data, and (2) the ease with which objects can be segmented from their background in depth data. We therefore argue that the use of RGB-D data over the use of solely depth data may increase the detection performance of RC-based approaches even further. We do remark, however, that the process of combining and aligning both types of data may result in an increase in computational complexity, which may negatively influence the operating time of the detection algorithm. Thus, combining visual and depth data is only feasible when it does not result in a insurmountable increase in operating speed and time.

### 8.2.2 The Search for RC Features

In Chapter 2, we proposed our RC features for robust and effective object detection in depth images. The design of the RC features was inspired by the work by Lienhart & Maydt (2002), Viola & Jones (2001), and Papageorgiou et al. (1998). In a non-exhaustive search, we proposed a set of 15 predefined RC feature types, i.e., spatially oriented combinations of symmetrically located rectangular regions in a depth image. We first defined a set of 4 basic feature types, which encode straightforward depth transitions in horizontal, vertical, diagonal and anti-diagonal orientations. Based on the basic feature types, we defined 11 specialised feature types, which are able to encode more complex depth transitions. In general, we can state that a feature type that consist of small rectangles typically encodes for local the depth transitions. Similarly, feature types that consist of large rectangles typically encode for global depth transitions.

Our design of the feature types was limited to the design of two, three, and four-rectangle RC features. Moreover, we emphasise that we deployed optimised sets of RC features in our experimental tasks, i.e., RC features that are specialised in encoding depth transitions over larger regions for the body part

detection task (see Chapter 3), and RC features that typically encode subtle depth differences for the sign language recognition task (see Chapter 5). We assume that designing more complex feature types (e.g., by incorporating (1) rectangles or polygons instead of squares, or (2) complex spatial combinations of the regions) may provide more accurate encodings of complex depth transitions in a depth image. However, using an extended set of feature types may result in an overall increase in computational complexity, which in turn may result in a decrease of the algorithm's operating time. We therefore argue that an investigation into the development of additional feature types will lead to the development of better optimised RC feature types. However, we estimate it as unlikely that the development of additional feature types will cause a significant increase in overall detection performance.

### 8.2.3 From Body Part Detection to Gesture Recognition

Embodied agents that aim to engage in natural interactions with humans require the ability to detect a person's body parts and gestural cues. To meet this requirement, we proposed the region comparison detector (see Chapter 3) for accurate body part detection, and the STAGE detector (see Chapter 5) for effective gesture recognition.

As stated in Section 5.1, gestures exist in two distinctive forms: (1) *dynamic* gestures, which involve direction and speed of motion to convey their meaning, and (2) *static* gestures, which involve arm and hand postures to represent a specific meaning (see, e.g., Dixit & Agrawal, 2015). In its current form, the STAGE detector is able to recognise (static) gestures by analysing the hand shape. Recognising dynamic gestures requires computer vision techniques that are able to (1) detect, (2) track, and (3) recognise the motion of the gestures (cf. Rautaray & Agrawal, 2015). Thus, to enable the STAGE detector - and thereby embodied agents in general - to recognise dynamic gestures, the detector should be extended by incorporating these techniques. We believe that this can easily be achieved by combining (1) the region comparison, and (2) the STAGE detector into a detector that first identifies a person's body parts, and then classifies their shape to recognise the gestures. Combining information about the location and shape of body parts in a sequence of images will then enable the agents to recognise dynamic gestures. Moreover, such a combination will broaden the capabilities of the STAGE detector and let the detector recognise both static and dynamic gestures.

### 8.2.4 Alternative Methods

In what follows, we discuss two alternative methods that are related to the work presented in the Thesis.

First, we evaluated the effectiveness of the RC features in comparison with the state-of-the-art PC features (see Chapter 3) that were developed by Shotton et al. (2013a,b; 2011). Future studies, however, may compare the effectiveness of the RC features with other depth-based features. Promising competing approaches are the HOG features (see, e.g., Dalal & Triggs, 2005) and their derivatives, such as the HOD features that were proposed by Spinello & Arras (Spinello & Arras, 2011). Although, some of our informal comparative evaluations seem to suggest that the RC features outperform the HOD features, future work should be directed at a systematic comparative evaluation of RC and HOD features. Similarly, the performances obtained by RC features may be compared with recently developed schemes, such as the work by Su, Liu, Xu, Li, & Ji (2015) and C. Zhang & Tian (2015).

Second, nowadays convolutional neural networks represent the state-of-the-art in computer vision and machine learning approaches. In the Thesis, we did not examine deep learning methods in combination with depth images. The large body of work available on this domain (Eitel, Springenberg, Spinello, Riedmiller, & Burgard, 2015; Lenz, Lee, & Saxena, 2015; Oberweger, Wohlhart, & Lepetit, 2015; Schmidhuber, 2015; Wohlhart & Lepetit, 2015), however, suggests that a great gain in performance can be obtained by using deep learning, either on the raw depth data or on the Haar feature-encoded depth images. We expect that deep learning-based methods (i.e., approaches that learn the most suitable features, instead of using predefined features) may achieve superior classification scores. However, given that deep learning is computationally intensive, we also expect that our approach outperforms the majority of deep learning methods in detection speed. Determining the trade-off between accuracy and efficiency for deep learning and RC-feature based approaches is left to future study.

## 8.3 REALISING THE INTERACTION MODEL

In the Thesis we described the blueprints for crucial components of a model that addressed the interactions between humans and embodied agents. A full-fletched model, however, requires four main additions. In this Section, we elaborate on the additions.

First, our objectives provide embodied agents with the ability to detect a person's body parts and gestures (see stage A in Figure 8.1). As such, our objectives allow the agents to perceive a person's non-verbal cues. However, the agents described in the interaction model do not possess the ability to recognise a person's facial expressions. To extend the agents with this ability, we suggest that the agents are enriched with software packages that are able to recognise a person's facial expressions. Well-know examples of (commercially

available) facial expression recognition software are the *Computer Expression Recognitino Toolbox* (CERT; Littlewort et al., 2011) and *IntraFace* (see Chu, De la Torre, & Cohn, 2013).

Second, we acknowledge that the sensor array (see stage A) could be extended with sensors that are able to measure a person's behaviour beyond his regular social interactions. Thus, we suggest that the sensing capabilities of the agents are improved by incorporating sensors of the intelligent environment itself, e.g., water and energy usage sensors.

Third, in stage B of the interaction model (see Figure 8.1), the agent incorporates a series of simple affective facial expressions (e.g., fear and happiness; see Chapter 6) to respond to a person's behaviour. We expect that incorporating more complex facial expressions will increase the human-likeness of agents even further (see, e.g., Gris et al., 2015). Human-likeness is highly relevant for the development of socially aware embodied agents.

Fourth, based on the social behaviour of a person, stage B describes the creation of the agent's response. If you want to influence a person's behaviour subtly, you need to provide the person with personalised feedback regarding his behaviour. The feedback should encourage a person to show the desired behaviour. As such, stage B in the interaction model should be extended by (1) a clear goal, i.e., the desired behaviour of the person, and (2) the corresponding specifications of the expression-feedback. In the long term, this may influence a person to show the desired behaviour, which is one of the goals of our project.

# REFERENCES


Abrahamse, W., Steg, L., Vlek, C., & Rothengatter, T. (2005). A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*, *25*(3), 273 - 291. doi: 10.1016/j.jenvp.2005.08.002

Achaibou, A., Pourtois, G., Schwartz, S., & Vuilleumier, P. (2008). Simultaneous recording of {EEG} and facial muscle reactions during spontaneous emotional mimicry. *Neuropsychologia*, *46*(4), 1104 - 1113. doi: 10.1016/j.neuropsychologia.2007.10.019

Aggarwal, J. K., & Cai, Q. (1999). Human motion analysis: A review. *Computer Vision and Image Understanding*, *73*(3), 428 - 440. doi: 10.1006/cviu.1998.0744

Alibali, M. W. (2005). Gesture in Spatial Cognition: Expressing, Communicating, and Thinking About Spatial Information. *Spatial Cognition & Computation*, *5*(4), 37–41. doi: 10.1207/s15427633scc0504_2

Amini, R., Yasavur, U., & Lisetti, C. (2012). Hapfacs 1.0: Software/api for generating facs-based facial expressions. , 17–17. doi: 10.1145/2491599.2491616

André, E., Bevacqua, E., Heylen, D. K. J., Niewiadomski, R., Pelachaud, C., Peters, C., ... Rehm, M. (2011). Non-verbal persuasion and communication in an affective agent. In R. Cowie, C. Pelachaud, & P. Petta (Eds.), *Emotion oriented systems. the humaine handbook* (pp. 585–608). London: Springer Verlag. doi: 10.1007/978-3-642-15184-2_30

Andreopoulos, A., & Tsotsos, J. K. (2013). 50 years of object recognition: Directions forward. *Computer Vision and Image Understanding*, *117*(8), 827–891. doi: 10.1016/j.cviu.2013.04.005

Baak, A., Müller, M., Bharaj, G., Seidel, H.-P., & Theobalt, C. (2013). A data-driven approach for real-time full body pose reconstruction from a depth camera. In A. Fossati, J. Gall, H. Grabner, X. Ren, & K. Konolige (Eds.), *Consumer depth cameras for computer vision* (p. 71-98). Springer London. doi: 10.1007/978-1-4471-4640-7_5

Bailenson, J. N., & Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science*, *16*(10), 814-819. doi: 10.1111/j.1467-9280.2005.01619.x

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*(3), 412-414. doi: 10 .1098/rsbl.2006.0509

Battison, R., & Baird, E. (1978). *Lexical borrowing in american sign language*. Linstok Press.

Bell, L., Gustafson, J., & M., H. (2003). Prosodic adaptation in human-computer interaction. In *Proceedings of the international congress of phonetic sciences* (pp. 833–836). (Retrieved from: http://www.speech.kth.se/ctt/publications/papers03/icphs03_ 2453.pdf)

Ben Youssef, A., Chollet, M., Jones, H., Sabouret, N., Pelachaud, C., & Ochs, M. (2015). Towards a socially adaptive virtual agent. In W.-P. Brinkman, J. Broekens, & D. Heylen (Eds.), *Intelligent virtual agents* (Vol. 9238, p. 3-16). Springer International Publishing. doi: 10.1007/978-3-319-21996-7_1

Bergboer, N. H. (2007). *Context-based image analysis*. Maastricht, the Netherlands: Universiteit Maastricht. Ph.D thesis. (Retrieved from: http://arno.unimaas.nl/show.cgi?fid=9175)

Bevacqua, E., Hyniewska, S., & Pelachaud, C. (2010). Evaluation of a virtual listener's smiling behavior. In *Proceedings of the international conference on computer animation and social agents.* (Retrieved from: http://biblio.telecom-paristech.fr/cgi-bin/download.cgi?id=10738)

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences* (pp. 97–110). (Retrieved from: http://www.fon.hum.uva.nl/paul/papers/Proceedings_1993. pdf)

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer (version 5.3.04) [computer program]. (Retrieved from: http://www.praat.org)

Box, G. E. P. (1949). A general distribution theory for a class of likelyhood criteria. *Biometrika*, *36*(3-4), 317-346. doi: 10.1093/biomet/36.3-4.317

Brandão, A., Fernandes, L. A. F., & Clua, E. (2014). M5aie - a method for body part detection and tracking using rgb-d images. In *Proceedings of the international conference on computer vision theory and applications* (pp. 367–377). doi: 10.5220/0004738003670377

Breazeal, C., & Scassellati, B. (2002). Robots that imitate humans. *Trends in Cognitive Sciences*, *6*(11), 481 - 487. doi: 10.1016/S1364-6613(02)02016-8

Brehm, J. W. (1989). Psychological reactance: Theory and applications. *Advances in Consumer Research*, *16*, 72–75. (Retrieved from: http://www.acrwebsite.org/search/view-conference-proceedings.aspx?Id=6883)

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/a:1010933404324

Brodersen, K. H., O., C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *Proceedings of the international conference on pattern recognition* (pp. 3121–3124). doi: 10.1109/icpr.2010.764

Brunton, A., Salazar, A., Bolkart, T., & Wuhrer, S. (2014). Review of statistical shape spaces for 3d data with comparative analysis for human faces. *Computer Vision and Image Understanding*, *128*(0), 1–17. doi: 10.1016/j.cviu.2014.05.005

Burgin, W., Pantofaru, C., & Smart, W. D. (2011). Using depth information to improve face detection. In *Proceedings of the international conference on human-robot interaction* (pp. 119–120). New York, NY: ACM. doi: 10.1145/1957656.1957690

Buys, K., Cagniart, C., Baksheev, A., Laet, T. d., Schutter, J. d., & Pantofaru, C. (2014). An adaptable system for rgb-d based human body detection and pose estimation. *Journal of Visual Communication and Image Representation*, *25*(1), 39–52. doi: 10.1016/j.jvcir.2013.03.011

Caridakis, G., Wagner, J., Raouzaiou, A., Lingenfelser, F., Karpouzis, K., & Andre, E. (2013). A cross-cultural, multimodal, affective corpus for gesture expressivity analysis. *Journal on Multimodal User Interfaces*, *7*(1-2), 121-134. doi: 10.1007/s12193-012-0112-x

Carlevaris-Bianco, N., & Eustice, R. M. (2014). Learning visual feature descriptors for dynamic lighting conditions. In *Proceedings of the ieee/rsj international conference on intelligent robots and systems* (p. 2769-2776). doi: 10.1109/iros.2014.6942941

Carrillo, H., Brodersen, K. H., & Castellanos, J. A. (2014). Probabilistic performance evaluation for multiclass classification using the posterior balanced accuracy. In M. A. Armada, A. Sanfeliu, & M. Ferre (Eds.), *Robot2013: First iberian robotics conference* (Vol. 252, pp. 347–361). Springer International Pub-

lishing. doi: 10.1007/978-3-319-03413-3_25

Castellano, G., Mancini, M., Peters, C., & McOwan, P. (2012). Expressive copying behavior for social agents: A perceptual analysis. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *42*(3), 776-783. doi: 10.1109/TSMCA.2011.2172415

Chan, K.-C., Koh, C.-K., & Lee, C. S. G. (2013). A 3d-point-cloud feature for human-pose estimation. In *Proceedings of the ieee international conference on robotics and automation* (p. 1623-1628). doi: 10.1109/icra.2013.6630787

Chang, J. Y., & Nam, S. W. (2013). Fast random-forest-based human pose estimation using a multi-scale and cascade approach. *ETRI Journal*, *35*(6), 949–959. doi: 10.4218/etrij.13.2013.0063

Chartrand, T. L., & Bargh, A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893–910. doi: 10.1037/0022-3514.76.6.893

Chartrand, T. L., & Lakin, J. L. (2013). The antecedents and consequences of human behavioral mimicry. *Annual Review of Psychology*, *64*(1), 285-308. doi: 10.1146/annurev-psych-113011-143754

Chartrand, T. L., & van Baaren, R. (2009). Human mimicry. In M. P. Zanna (Ed.), (Vol. 41, p. 219 - 274). Academic Press. doi: 10.1016/S0065-2601(08)00405-x

Chen, L., Rose, R. T., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., ... Huang, T. (2006). Vace multimodal meeting corpus. In S. Renals & S. Bengio (Eds.), *Machine learning for multimodal interaction* (Vol. 3869, p. 40-51). Springer Berlin Heidelberg. doi: 10.1007/11677482_4

Chen, L., Wang, F., Deng, H., & Ji, K. (2013). A survey on hand gesture recognition. In *Proceedings of the international conference on computer sciences and applications* (p. 313-316). doi: 10.1109/CSA.2013.79

Chen, Q., Song, Z., Dong, J., Huang, Z., Hua, Y., & Yan, S. (2015). Contextualizing object detection and classification. , *37*(1), 13-27. doi: 10.1109/tpami.2014.2343217

Chu, W.-S., De la Torre, F., & Cohn, J. F. (2013). Selective transfer machine for personalized facial action unit detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (p. 3515-3522). doi: 10.1109/cvpr.2013.451

Criminisi, A., Shotton, J., & Konukoglu, E. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, *7*(2&3), 81–227. doi: 10.1561/0600000035

Crow, F. C. (1984). Summed-area tables for texture mapping. In *Proceedings of the annual conference on computer graphics and interactive techniques* (pp. 207–212). New York, NY: ACM. doi: 10.1145/800031.808600

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the ieee computer society conference on computer vision and pattern recognition conference on* (Vol. 1, p. 886-893). doi: 10.1109/cvpr.2005.177

Dal Mutto, C., Zanuttigh, P., & Cortelazzo, G. M. (2012). Scene segmentation and video matting assisted by depth data. In *Time-of-flight cameras and microsoft kinect* (pp. 93–105). Springer US. doi: 10.1007/978-1-4614-3807-6_6

Dautenhahn, K., Nehaniv, C. L., & Alissandrakis, A. (2003). Learning by experience from others — social learning and imitation in animals and robots. In R. Kühn, R. Menzel, W. Menzel, U. Ratsch, M. M. Richter, & I.-O. Stamatescu (Eds.), *Adaptivity and learning* (p. 217-241). Springer Berlin Heidelberg. doi: 10.1007/978-3-662-05594-6_17

Davies, M., & Callaghan, V. (2012). iworlds: Generating artificial control systems for simulated humans using virtual worlds and intelligent environment. *Journal of Ambient Intelligent and Smart Environments*, *4*(1), 5–27. doi: 10.3233/ais-2011-0129

De Croon, G. C. H. E., Postma, E. O., & Van den Herik, H. J. (2011). Adaptive gaze control for object detection. *Cognitive Computation*, *3*(1), 264–278. doi: 10.1007/s12559-010-9093-9

Dimberg, U., & Thunberg, M. (1998). Rapid facial reactions to emotional facial expressions. *Scandinavian Journal of Psychology*, *39*(1), 39–45. doi: 10.1111/1467-9450.00054

Dixit, V., & Agrawal, A. (2015). Real-time hand tracking for dynamic gesture recognition. In *Proceedings of fourth international conference on soft computing for problem solving* (Vol. 336, p. 153-164). doi: 10.1007/978-81-322-2220-0_12

Dragone, M., Duffy, B. R., & O'Hare, G. M. P. (2005, Aug). Social interaction between robots, avatars humans. In *Proceedings of the ieee international workshop on robot and human interactive communication* (p. 24-29). doi: 10.1109/ROMAN.2005.1513751

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M. A., & Burgard, W. (2015). Multimodal deep learning for robust RGB-D object recognition. *Computer Vision and Pattern Recognition*. (Retrieved from: http://arxiv.org/abs/1507.06821)

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124-129. doi: 10.1037/h0030377

Ekman, P., & Friesen, W. V. (1978). Facial action coding system: A technique for the measurement of facial movement.

Elfenbein, H. A. (2013). Nonverbal dialects and accents in facial expressions of emotion. *Emotion Review*, *5*(1), 90-96. doi: 10.1177/1754073912451332

Esposito, A. (2009). The perceptual and cognitive role of visual and auditory channels in conveying emotional information. *Cognitive Computation*, *1*(3), 268-278. doi: 10.1007/s12559-009-9017-8

Fanelli, G., Dantone, M., Gall, J., Fossati, A., & Van Gool, L. (2013). Random forests for real time 3d face analysis. *International Journal of Computer Vision*, *101*(3), 437–458. doi: 10.1007/s11263-012-0549-0

Fanelli, G., Weise, T., Gall, J., & Van Gool, L. (2011). Real time head pose estimation from consumer depth cameras. In R. Mester & M. Felsberg (Eds.), *Pattern recognition* (Vol. 6835, p. 101-110). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-23123-0_11

Fanello, S., Keskin, C., Kohli, P., Izadi, S., Shotton, J., Criminisi, A., ... Paek, T. (2014). Filter forests for learning data-dependent convolutional kernels. In *Proceedings of the conference on computer vision and pattern recognition* (p. 1709-1716). doi: 10.1109/CVPR.2014.221

Fischer, A., Becker, D., & Veenstra, L. (2012). Emotional mimicry in social context: The case of disgust and pride. *Frontiers in Psychology*, *3*(475). doi: 10.3389/fpsyg.2012.00475

Förstner, W. (2000). Image preprocessing for feature extraction in digital intensity, color and range images. In A. Dermanis, A. Grün, & F. Sansò (Eds.), *Geomatic method for the analysis of data in the earth sciences* (Vol. 95, p. 165-189). Springer Berlin Heidelberg. doi: 10.1007/3-540-45597-3_4

Fothergill, S., Mentis, H., Kohli, P., & Nowozin, S. (2012). Instructing people for training gestural interactive systems. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1737–1746). New York, NY: ACM.

doi: 10.1145/2207676.2208303

Ghosh, D. K., & Ari, S. (2015). Static hand gesture recognition using mixture of features and svm classifier.
(Retrieved from: http://dspace.nitrkl.ac.in/dspace/bitstream/2080/2289/1/3a.pdf)

Gris, I., Rivera, D. A., & Novick, D. (2015). Animation guidelines for believable embodied conversational agent gestures. In R. Shumaker & S. Lackey (Eds.), *Virtual, augmented and mixed reality* (Vol. 9179, p. 197-205). Springer International Publishing. doi: 10.1007/978-3-319-21067-4_21

Guf, J., & Jiang, W. (1996). The haar wavelets operational matrix of integration. *International Journal of Systems Science*, *27*(7), 623–628. doi: 10.1080/00207729608929258

Guyon, I., Athitsos, V., Jangyodsuk, P., & Escalante, H. J. (2014). The chalearn gesture dataset (cgd 2011). *Machine Vision Applications*, *25*(8), 1929–1951. doi: 10.1007/s00138-014-0596-3

Ham, J., Midden, C., & Beute, F. (2009). Can ambient persuasive technology persuade unconsciously? using subliminal feedback to influence energy consumption ratings of household appliances. In *Proceedings of the international conference on persuasive technology* (pp. 29:1–29:6). New York, NY: ACM. doi: 10.1145/1541948.1541988

Han, J., Shao, L., Xu, D., & Shotton, J. (2013). Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics*, *43*(5). doi: 10.1109/TCYB.2013.2265378

Hari, R., & Kujala, M. V. (2009). Brain basis of human social interaction: from concepts to brain imaging. *Physiological reviews*, *89*(2), 453–479. doi: 10.1152/physrev.00041.2007

Hiltz, S. R., Johnson, K., & Turoff, M. (1986). Experiments in group decision making communication process and outcome in face-to-face versus computerized conferences. *Human Communication Research*, *13*(2), 225–252. doi: 10.1111/j.1468-2958.1986.tb00104.x

Hinde, R. A. (1972). *Non-verbal communication*. Cambridge University Press.

Høg, R. I., Jasek, P., Rofidal, C., Nasrollahi, K., Moeslund, T. B., & Tranchet, G. (2012). An rgb-d database using microsoft's kinect for windows for face detection. In *Proceedings of the ieee international conference on signal image technology and internet based systems* (pp. 42–46). IEEE. doi: 10.1109/SITIS

.2012.17

Hogg, M. A., & Reid, S. A. (2006). Social identity, self-categorization, and the communication of group norms. *Communication Theory*, *16*(1), 7–30. doi: 10.1111/j.1468-2885.2006.00003.x

Hoiem, D., Efros, A. A., & Hebert, M. (2006). Putting objects in perspective. In *Proceedings of the ieee computer society conference on computer vision and pattern recognition - volume 2* (pp. 2137–2144). Washington, DC: IEEE Computer Society. doi: 10.1109/cvpr.2006.232

Holler, J., & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, *35*(2), 133–153. doi: 10.1007/s10919-011-0105-6

Huorong Ren, H., Yu, P., & Zhang, P. (2015). Illumination invariant feature extraction using relative gradient difference. *Journal for Light and Electron Optics*. doi: 10.1016/j.ijleo.2015.08.198

Jack, R. E., Caldara, R., & Schyns, P. G. (2012). Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *Journal of Experimental Psychology: General*, *141*(1), 19–25. doi: 10.1037/a0023463

Jain, A. K., & Farrokhnia, F. (1990). Unsupervised texture segmentation using gabor filters. In *Proceedings of the ieee international conference on systems, man and cybernetics* (p. 14-19). doi: 10.1109/ICSMC.1990.142050

Jiang, F., Fischer, M., Kemal, H. E., & Shi, B. E. (2013). Combining texture and stereo disparity cues for real-time face detection. *Signal Processing: Image Communication*, *28*(9), 1100 - 1113. doi: 10.1016/j.image.2013.07.006

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770–814. doi: 10.1037/0033-2909.129.5.770

Kapuscinski, T., Oszust, M., Wysocki, M., & Warchol, D. (2015). Recognition of hand gestures observed by depth cameras. *International Journal of Advanced Robotic Systems*, *12*. doi: 10.5772/60091

Kaufman, J., & Johnston, P. J. (2014). Facial motion engages predictive visual mechanisms. *PLoS ONE*, *9*(3). doi: 10.1371/journal.pone.0091038

Keskin, C., Kıraç, F., Kara, Y. E., & Akarun, L. (2013). Real time hand pose estimation using depth sensors. In (pp. 119–137). doi: 10.1007/978-1-4471-4640-7_7

Khaligh-Razavi, S.-M. (2014). What you need to know about the state-of-the-art computational models of object-vision: A tour through the models. *arXiv preprint*. (Retrieved from: http://arxiv.org/ftp/arxiv/papers/1407/1407.2776.pdf)

Khoshelham, K., & Elberink, S. O. (2012). Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, *12*(2), 1437–1454. doi: 10.3390/s120201437

Kuznetsova, A., Leal-Taixé, L., & Rosenhahn, B. (2013). Real-time sign language recognition using a consumer depth camera. *IEEE International Conference on Computer Vision Workshops*, 83-90. doi: 10.1109/ICCVW.2013.18

Lee, M. W., & Nevatia, R. (2007). Body part detection for human pose estimation and tracking. In *Proceedings of the ieee workshop on motion and video computing* (p. 23-23). doi: 10.1109/wmvc.2007.10

Lenz, I., Lee, H., & Saxena, A. (2015). Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, *34*(4-5), 705–724. doi: 10.1177/0278364914549607

Li, S.-Z., Yu, B., Wu, W., Su, S.-Z., & Ji, R.-R. (2015). Feature learning based on sae–pca network for human gesture recognition in rgbd images. *Neurocomputing*, *151, Part 2*(0), 565 - 573. doi: 10.1016/j.neucom.2014.06.086

Liao, S., Jain, A. K., & Li, S. Z. (2012). *Unconstrained face detection* (Tech. Rep. No. MSU-CSE-12-15). Department of Computer Science, Michigan State University. (Retrieved from: http://www.cse.msu.edu/biometrics/Publications/Face/Liao JainLi_UnconstrainedFaceDetection_TechReport.pdf)

Lienhart, R., & Maydt, J. (2002). An extended set of haar-like features for rapid object detection. In *Proceedings of the international conference on image processing* (Vol. 1, p. I-900-I-903 vol.1). doi: 10.1109/ICIP.2002.1038171

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). The computer expression recognition toolbox (cert). In *Proceedings of the ieee international conference on automatic face gesture recognition and workshops* (p. 298-305). doi: 10.1109/FG.2011.5771414

Liu, L., & Shao, L. (2013). Learning discriminative representations from rgb-d video data. In *Proceedings of the international joint conference on artificial intelligence* (pp. 1493–1500). AAAI Press.

Liu, S., Wang, Y., Wang, H., & Pan, C. (2013). Kinect depth inpainting via graph laplacian with tv21 regularization. In *Proceedings of the asian conference on pattern recognition* (p. 251-255). doi: 10.1109/ACPR.2013.35

Looze, C. d., Oertel, C., Rauzy, S., & Campbell, N. (2011). Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In *Proceedings of the international congress of phonetic sciences* (pp. 1294–1297). (Retrieved from: http://www.sscnet.ucla.edu/cbd/bios/royaumont.pdf)

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, *36*(8), 1404–1426. doi: 10.1111/j.1551-6709.2012.01269.x

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (n.d.). In M. Kipp, J.-P. Martin, P. Paggio, & D. Heylen (Eds.), *Proceedings of the international conference on language resources and evaluation.*

Manusov, V., & Patterson, M. L. (2006). *The sage handbook of nonverbal communication*. SAGE Publications.

Mattheij, R. J. H., Groeneveld, K., Postma, E. O., & Van den Herik, H. J. (2016). Depth-based detection with region comparison features. *Journal of Visual Communication and Image Representation*, *38*, 82–99. doi: 10.1016/j.jvcir.2016.02.008

Mattheij, R. J. H., & Nilsenová, M. and Postma, E. O. (2013). Vocal and facial imitation of humans interacting with virtual agents. In *Proceedings of the humaine association conference on affective computing and intelligent interaction* (p. 815-820). doi: 10.1109/ACII.2013.152

Mattheij, R. J. H., Postma-Nilsenová, M., & Postma, E. O. (2015). Mirror mirror on the wall: Is there mimicry in you all? *Journal of Ambient Intelligence and Smart Environments*, *7*(2), 121–132. doi: 10.3233/AIS-150311

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

Meltzoff, A. N., Brooks, R., Shon, A. P., & Rao, R. P. N. (2010). "social" robots are psychological agents for infants: A test of gaze following. *Neural Networks*, *23*(8–9), 966 - 972. doi: 10.1016/j.neunet.2010.09.005

Michalowski, M. P., Simmons, R., & Kozima, H. (2009). Rhythmic attention in child-robot dance play. In *Proceedings of the ieee international symposium on robot and human interactive communication* (p. 816-821). doi: 10.1109/ROMAN.2009.5326143

Midden, C. J. H., Meter, J. F., Weenig, M. H., & Zieverink, H. J. A. (1983). Using feedback, reinforcement and information to reduce energy consumption in households: A field-experiment. *Journal of Economic Psychology*, *3*(1), 65 - 86. doi: 10.1016/0167-4870(83)90058-2

Miller, D. T., Downs, J. S., & Prentice, D. A. (1998). Minimal conditions for the creation of a unit relationship: the social bond between birthdaymates. *European Journal of Social Psychology*, *28*(3), 475–481. doi: 10.1002/(SICI)1099-0992(199805/06)28:3<475::AID-EJSP881>3.0.CO;2-M

Mitchell, R. W. (1987). A comparative-developmental approach to understanding imitation. In P. P. G. Bateson & P. H. Klopfer (Eds.), *Perspectives in ethology* (p. 183-215). Springer US. doi: 10.1007/978-1-4613-1815-6_7

Mitra, S., & Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *37*(3), 311-324. doi: 10.1109/TSMCC.2007.893280

Murray-Smith, R. (2014). *Mobile social signal processing: First international workshop, mssp 2010, lisbon, portugal, september 7, 2010, invited papers*. Springer.

Nanni, L., Lumini, A., Dominio, F., & Zanuttigh, P. (2014). Effective and precise face detection based on color and depth data. *Applied Computing and Informatics*, *10*(1–2), 1 - 13. doi: 10.1016/j.aci.2014.04.001

Neumann, J. v. (1958). *The computer and the brain*. New Haven, CT: Yale University Press.

Niedenthal, P. M., Brauer, M., Halberstadt, J. B., & Innes-Ker, . H. (2001). When did her smile drop? facial mimicry and the influences of emotional state on the detection of change in emotional expression. *Cognition and Emotion*, *15*(6), 853-864. doi: 10.1080/02699930143000194

Oberweger, M., Wohlhart, P., & Lepetit, V. (2015). Hands deep in deep learning for hand pose estimation. *Computer Vision and Pattern Recognition*. (Retrieved from: http://arxiv.org/abs/1502.06807)

Omary, Z., & Mtenzi, F. (2010). Machine learning approach to identifying the dataset threshold for the performance estimators in supervised learning. *International Journal for Infonomics (IJI)*, *3*(3). (Retrieved from: https:www.zotero.orggroupspesquisatesteperformanceitemsite mKey5AMP5SXE)

Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written dutch. In

P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for dutch* (p. 219-247). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-30910-6_13

Osawa, H., & Imai, M. (2013). Researching nonverbal communication strategies in human-robot interaction. In J. Filipe & A. Fred (Eds.), *Agents and artificial intelligence* (Vol. 358, p. 417-432). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-36907-0_28

Pantic, M., & Vinciarelli, A. (2014). Social signal processing. In (p. 84). Oxford University Press.

Papageorgiou, C. P., Oren, M., & Poggio, T. (1998). A general framework for object detection. In *Proceedings of the international conference on computer vision* (pp. 555–562). doi: 10.1109/ICCV.1998.710772

Pavlovic, V. I., Sharma, R., & Huang, T. S. (1997). Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 677-695. doi: 10.1109/34.598226

Pedersoli, F., Benini, S., Adami, N., & Leonardi, R. (2014). Xkin: an open source framework for hand pose and gesture recognition using kinect. *The Visual Computer*, *30*(10), 1107-1122. doi: 10.1007/s00371-014-0921-x

Pickering, M. J., & Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190. (Retrieved from: http:journals.cambridge.orgdownload.php?file=%2FBBS%2FBBS27_02%2FS0140525X04000056a.pdf)

Plagemann, C., Ganapathi, V., Koller, D., & Thrun, S. (2010). Real-time identification and localization of body parts from depth images. In *Proceedings of the ieee international conference on robotics and automation* (pp. 3108–3113). doi: 10.1109/robot.2010.5509559

Powers, D. M. W. (2011). *Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation* (Tech. Rep. No. SIE-07-001). School of Informatics and Engineering, Flinders University. (Retrieved from: http:david.wardpowers.infoBMEvaluation_SIETR.pdf)

Pugeault, N., & Bowden, R. (2011). Spelling it out: Real-time asl fingerspelling recognition. In *Proceedings of the ieee international conference on computer vision workshops* (p. 1114-1119). doi: 10.1109/ICCVW.2011.6130290

Qu, L., Tian, J., Han, Z., & Tang, Y. (2015). Pixel-wise orthogonal decomposition for color illumination invariant and shadow-free image. *Optics Express*, *23*(3), 2220–2239. doi: 10.1364/oe.23.002220

Rautaray, S. S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, *43*(1), 1-54. doi: 10.1007/s10462-012-9356-9

Ren, Z., Yuan, J., & Zhang, Z. (2011). Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In *Proceedings of the acm international conference on multimedia* (pp. 1093–1096). ACM. doi: 10.1145/2072298.2071946

Riche, N., Mancas, M., Gosselin, B., & Dutoit, T. (2011). 3d saliency for abnormal motion selection: The role of the depth map. In J. Crowley, B. Draper, & M. Thonnat (Eds.), *Computer vision systems* (Vol. 6962, pp. 143–152). Springer Berlin / Heidelberg. doi: 10.1007/978-3-642-23968-7_15

Romero-Rodríguez, W. J. G., Zamudio Rodriguez, V. M., Flores, R. B., Sotelo-Figueroa, M. A., & Alcaraz, J. A. S. (2011). Comparative study of bso and ga for the optimizing energy in ambient intelligence. In I. Batyrshin & G. Sidorov (Eds.), *Advances in soft computing* (Vol. 7095, p. 177-188). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-25330-0_16

Roubroeks, M., Midden, C., & Ham, J. (2009). Does it make a difference who tells you what to do? exploring the effect of social agency on psychological reactance. In *Proceedings of the international conference on persuasive technology* (pp. 15:1–15:6). New York, NY: ACM. doi: 10.1145/1541948.1541970

Rowling, J. K. (1997). *Harry potter and the philosopher's stone*. eM Publications.

Sato, W., & Yoshikawa, S. (2007). Spontaneous facial mimicry in response to dynamic facial expressions. *Cognition*, *104*(1), 1 - 18. doi: 10.1016/j.cognition.2006.05.001

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. Davidson, K. Scherer, & H. Goldsmith (Eds.), *Handbook of the affective sciences* (pp. 433–456). New York and Oxford: Oxford University Press. (Retrieved from: http://www.affectivesciences.org/system/files/biblio/2003_Scherer_HdbAffSci_Vocal.pdf)

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85 - 117. doi: 10.1016/j.neunet.2014.09.003

Schmidt, A., Pfleging, B., Alt, F., Sahami, A., & Fitzpatrick, G. (2012). Interacting with 21st-century computers. *IEEE Pervasive Computing*, *11*(1), 22-31. doi: 10.1109/mprv.2011.81

Sebe, N. (2009). Multimodal interfaces: Challenges and perspectives. *Journal of Ambient Intelligence and Smart Environments*, *1*(1), 23–30. (Retrieved from: http://disi.unitn.it/~sebe/publications/aiso03.pdf)

Shah, Z. H., & Kaushik, V. (2015). Performance analysis of canny edge detection for illumination invariant facial expression recognition. In *Proceedings of the international conference on industrial instrumentation and control* (p. 584-589). doi: 10.1109/IIC.2015.7150809

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3), 591–611. (Retrieved from: http://links.jstor.org/sici?sici=0006-3444%28196512%2952%3A3%2F4%3C591%3AAAOVTF%3E2.0.CO%3B2-B)

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., ... Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1297–1304). Washington, DC: IEEE Computer Society. doi: 10.1109/cvpr.2011.5995316

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., ... Blake, A. (2013). Real-time human pose recognition in parts from single depth images. In R. Cipolla, S. Battiato, & G. M. Farinella (Eds.), *Machine learning for computer vision* (Vol. 411, pp. 119–135). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-28661-2_5

Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., ... Blake, A. (2013). Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2821–2840. doi: 10.1109/tpami.2012.241

Smisek, J., Jancosek, M., & Pajdla, T. (2013). 3d with kinect. In A. Fossati, J. Gall, H. Grabner, X. Ren, & K. Konolige (Eds.), *Consumer depth cameras for computer vision* (p. 3-25). Springer London. doi: 10.1007/978-1-4471-4640-7_1

Son, J., Yoo, H., Kim, S., & Sohn, K. (2015). Real-time illumination invariant lane detection for lane departure warning system. *Expert Systems with Applications*, *42*(4), 1816 - 1824. doi: 10.1016/j.eswa.2014.10.024

Spinello, L., & Arras, K. O. (2011). People detection in rgb-d data. In *Proceedings of the international conference on intelligent robots and systems* (pp. 3838–3843). doi: 10.1109/iros.2011.6095074

Stel, M., Mastop, J., & Strick, M. (2011). The impact of mimicking on attitudes toward products presented in tv commercials. *Social Influence*, *6*(3), 142-152. doi: 10.1080/15534510.2011.580978

Su, S.-Z., Liu, Z.-H., Xu, S.-P., Li, S.-Z., & Ji, R. (2015). Sparse auto-encoder based feature learning for human body detection in depth image. *Signal Processing*, *112*, 43 - 52. doi: 10.1016/j.sigpro.2014.11.003

Swift, M., Ferguson, G., Galescu, L., Chu, Y., Harman, C., Jung, H., ... Kautz, H. (2012). A multimodal corpus for integrated language and action. In *Proceedings of the international workshop on multimodal corpora for machine learning.* (Retrieved from: http:www.cs.rochester.edu˜kautzpaperslrec2012MMC.pdf)

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (6th edition).

Tang, Y., Sun, Z., & Tan, T. (2014). Slice representation of range data for head pose estimation. *Computer Vision and Image Understanding*, *128*(0), 18–35. doi: 10.1016/j.cviu.2014.05.008

Tanner, R. J., Ferraro, R., Chartrand, T. L., Bettman, J. R., & Van Baren, R. (2008). Of chameleons and consumption: The impact of mimicry on choice and preferences. *Journal of Consumer Research*, *47*, 754–766. (Retrieved from: https://faculty.fuqua.duke.edu/ jrb12/bio/Jim/mimic_final. pdf)

Tentori, M., Favela, J., & Rodriguez, M. D. (2006). Privacy-aware autonomous agents for pervasive healthcare. *IEEE Intelligent Systems*, *21*(6), 55-62. doi: 10.1109/mis.2006.118

Van den Broek, E. L. (2011). *Affective signal processing (asp): Unraveling the mystery of emotions.* Enschede, the Netherlands: University of Twente. Ph.D thesis. (SIKS Dissertation series no. 2011-30) doi: 10.3233/AIS-2011-0131

Van Welbergen, H., Ding, Y., Sattler, K., Pelachaud, C., & Kopp, S. (2015). Real-time visual prosody for interactive virtual agents. In W.-P. Brinkman, J. Broekens, & D. Heylen (Eds.), *Intelligent virtual agents* (Vol. 9238, p. 139-151). Springer International Publishing. doi: 10.1007/978-3-319-21996-7_16

Vijayanagar, K. R., Loghman, M., & Kim, J. (2014). Real-time refinement of kinect depth maps using multi-resolution anisotropic diffusion. *Mobile*

*Networks and Applications*, *19*(3), 414–425. doi: 10.1007/s11036-013-0458-7

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743–1759. doi: 10.1016/j.imavis.2008.11.007

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., & Schroeder, M. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, *3*(1), 69-87. doi: 10.1109/t-affc.2011.27

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the ieee computer society conference on computer vision and pattern recognition* (Vol. 1, pp. 511–518). Los Alamitos, CA: IEEE. doi: 10.1109/cvpr.2001.990517

Viola, P., Jones, M., & Snow, D. (2005). Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, *63*(2), 153-161. doi: 10.1007/s11263-005-6644-8

Wang, J., An, P., Zuo, Y., You, Z., & Zhang, Z. (2014). High accuracy hole filling for kinect depth maps. *Optoelectronic Imaging and Multimedia Technology III*, *9273*, 92732L-92732L-17. doi: 10.1117/12.2071437

Wargnier, P., Malaisé, A., Jacquemot, J., Benveniste, S., Jouvelot, P., Pino, M., & Rigaud, A.-S. (2015). Towards Attention Monitoring of Older Adults with Cognitive Impairment During Interaction with an Embodied Conversational Agent. In *Proceedings of the ieee workshop on virtual and augmented assistive technology* (p. 23 - 28). doi: 10.1109/vaat.2015.7155406

Weiser, M. (1997). The computer for the 21st century. *Scientific American*, *265*(3), 94–104. doi: 10.1145/329124.329126

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80-83. (Retrieved from: http://www.jstor.org/stable/3001968)

Wohlhart, P., & Lepetit, V. (2015). Learning descriptors for object recognition and 3d pose estimation. *Computer Vision and Pattern Recognition*. (Retrieved from: http://arxiv.org/abs/1502.05908)

Wooldridge, M. J. (2001). *Introduction to multiagent systems*. New York, NY: John Wiley & Sons, Inc.

Wu, J., Cui, Z., Sheng, V. S., Zhao, P., Su, D., & Gong, S. (2013). A comparative study of sift and its variants. *Measurement Science Review*, *13*(3), 122–131. doi: 10.2478/msr-2013-0021

Wu, Y., & Huang, T. S. (1999). Vision-based gesture recognition: A review. In A. Braffort, R. Gherbi, S. Gibet, D. Teil, & J. Richardson (Eds.), *Gesture-based communication in human-computer interaction* (Vol. 1739, p. 103-115). Springer Berlin Heidelberg. doi: 10.1007/3-540-46616-9_10

Zhang, C., & Tian, Y. (2015). Histogram of 3d facets: A depth descriptor for human action and hand gesture recognition. *Computer Vision and Image Understanding*, *139*, 29 - 39. doi: 10.1016/j.cviu.2015.05.010

Zhang, C., & Zhang, Z. (2010). *A survey of recent advances in face detection* (Tech. Rep. No. MSR-TR-2010-66). Microsoft Research. (Retrieved from: http://131.107.65.14/pubs/132077/facedetsurvey.pdf)

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE MultiMedia*, *19*(2), 4-10. doi: 10.1109/mmul.2012.24

Zhao, W., Chellappa, R., Phillips, P., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, *35*, 399–458.

# A OVERVIEW OF LEXICAL STIMULI

The Appendix provides an overview of the words that were used as primes in the experiment described in Chapter 6. The words were selected from the list of the 100 most frequent Dutch nouns in the SoNaR corpus (Oostdijk et al., 2013). In what follows, we present the list of stimuli. The order of the stimuli is the same as in the experiment.

1. Hoofdstuk
2. Koning
3. Muziek
4. Wedstrijd
5. Voertuig
6. Brandweer
7. Gebruik
8. Seizoen
9. Oorzaak
10. Prinses
11. Probleem
12. Gebaar
13. Huisdier
14. Miljoen
15. Regel
16. Procent
17. Miljard
18. Moment
19. Begin
20. Gedrag
21. Foto
22. Partij
23. Vader
24. Talent
25. Beker
26. Water
27. Auto
28. Moeder
29. Motor
30. Zuster

In this overview, the first three words were used for the training sequence; the remaining words were used in the actual experiment.

# B ACRONYMS AND ABBREVIATIONS

The Appendix provides a list of acronyms and abbreviations that are used in the Thesis.

**ASL** American Sign Language

**AUC** Area Under the Curve

**CMAS** Confédération Mondiale des Activités Subaquatiques

**D** Depth

**HOD** Histogram of Oriented Depths

**HOG** Histogram of Oriented Gradients

**PC** Pixel Comparison

**RC** Region Comparison

**RGB** Red Green Blue

**RGB-D** Red Green Blue + Depth

**STAGE** Static Gestures

**TiGeR** Tilburg Gesture Recognition

# SUMMARY

As reducing energy consumption may start at the household, effective approaches towards energy conservation call for the development of an intelligent environment that persuades its residents to change their energy consumption behaviour. To change their behaviour in the long term, the intelligent environment should provide its residents with personalised feedback regarding their behaviour. Providing personalised feedback in a subtle and nonintrusive way can be achieved by employing a persuasive virtual person; a so-called "embodied agent". Enhancing the ability of an agent to perceive human behaviour accurately increases its ability to establish a social bond with a person, which, in turn, allows the agent to act as a persuasive agent.

As the agents are likely to be deployed in noisy environments, they require state-of-the-art computer vision algorithms that are able to handle the noisy nature of the environment. Many present-day approaches towards automatic object detection, however, rely on visual data as their main source of information. While rich in detail, the disadvantage of visual data is that it is sensitive to the illumination conditions, such as shadows or bright lights. Depth data, however, is insensitive to the illumination conditions. Object segregation may therefore be facilitated by using depth data rather than visual data.

Thus, enhancing an agent's cognitive abilities by incorporating in-depth information is likely to increase its ability to perceive human behaviour. As such, the Thesis explored the possibilities to deploy in-depth information to detect the non-verbal cues of people. Moreover, the Thesis investigated to what extent it is possible to establish a social bond between a human and a virtual agent.

The problem statement of the Thesis reads as follows: *To what extent is it possible to detect human body parts and behaviour when using in-depth information?* The problem statement is the point of departure for five separate research questions: (RQ 1) *How can we improve Shotton et al.'s body part detector in such a way that it enables fast and effective body part detection in noisy depth data?*, (RQ 2) *To what extent do Region Comparison features enable fast and accurate face and person detection in noisy depth images?*, (RQ 3) *How do we develop an annotated database that incorporates visual and depth data recordings of natural human gestures?*, (RQ 4) *To what extent do Region Comparison features enable accurate recognition of static gestures when using in-depth information?*, and (RQ 5) *To what extent do people mimic verbal and non-verbal cues sent out by an embodied agent?* The answers to the research questions enable us to formulate our conclusion to the problem statement.

Chapter 1 first introduces the concepts of intelligent environments and embodied agents. Subsequently, the Chapter presents an interaction model that describes the establishment of a social connection between humans and embodied agents. As enhancing an agent's ability to perceive human behaviour increases its ability to establish a social bond with a person, the Chapter then introduces the use of depth data for effective behaviour recognition. Finally, the Chapter formulates the problem statement, including the resultant research questions and the corresponding research methodology used to answer them.

Chapter 2 answers RQ 1. The Chapter first elaborates on the use of depth data as a robust alternative to visual data. Then, the Chapter discusses the first principles and limitations of Shotton et al.'s state-of-the-art body part detection algorithm, which incorporates their Pixel Comparison (PC) features. Inspired by Shotton et al.'s detector, the Chapter presents our contribution to fast and robust object detection, i.e., the Region Comparison (RC) features. Finally, the Chapter presents and discusses the work related to our approach.

Chapter 3 answers RQ 2. First, the Chapter presents the *region comparison detector*, which incorporates our RC features for effective body part detection. In a comparative evaluation of the RC and PC features, both associated detectors are then trained and evaluated on three challenging object detection experiments: two face detection tasks and a person detection task. Finally, the Chapter presents the results of the evaluation, and discusses their implications. The results show that the RC features outperform the RC features in both detection performance and computational efficiency.

Chapter 4 answers RQ 3. Guided by the dire need for a new corpus with detailed recordings of natural human gestures, the Chapter discusses recent multimodal databases that have been proposed to study automatic gesture recognition. Subsequently, the Chapter describes the creation of the TiGeR Cub corpus, i.e., a novel database with depth recordings of two interacting interlocutors, and the procedure followed to annotate the data.

Chapter 5 answers RQ 4. The Chapter first discusses the challenges that are to be faced when aiming to recognise static hand gestures in the American Sign Language. After discussing four recent related approaches in the field of static gesture recognition, the Chapter presents the STAGE detector, which incorporates the RC features for automatic gesture recognition. In a comparative evaluation, the performance of the STAGE detector is assessed and compared to the performance of four state-of-the-art approaches towards static gesture recognition. The results of the evaluation show that the STAGE detector outperforms all competing approaches towards static sign language recognition in detection performance.

Chapter 6 answers RQ 5. First, the Chapter outlines the relevance of mimicking the cues sent out by embodied agents. Then, the Chapter describes the

methodology and results of an experiment in which we investigate whether, and if so to what extent, humans unconsciously imitate the cues sent out by embodied agents. The results show that humans unconsciously imitate both verbal and non-verbal cues sent out by embodied agents. Finally, the Chapter discusses the implications of the results for the development of embodied agents and intelligent environments in general.

Chapter 7 answers the five research questions on the basis of the work in the Thesis, which are then used to formulate our conclusions, and the answer to the problem statement. Based on the answers to our research questions, we may conclude that it is possible to perform accurate human body parts and behaviour recognition by means of in-depth information that is encoded by RC features. We may further conclude that using the RC features for human body part and behaviour recognition tasks may enhance an agent's cognitive abilities.

Chapter 8 completes the Thesis by discussing our findings and conclusions, as well as their implications for the design of embodied agents. The Chapter first reflects upon the implications of our findings for the design of smart embodied agents and intelligent environments. Subsequently, the Chapter formulates three points of improvement of our studies, and provides two alternative methods to further enhance the accuracy and robustness of the proposed approach. Finally, the Chapter presents four pointers to future work.

# CURRICULUM VITAE

Ruud Mattheij was born on October 20, 1987 in Venlo, the Netherlands. He completed his secondary education (VWO, Natuur en Gezondheid) at BC Schöndeln in Roermond. In 2009, Ruud obtained his Bachelor's degree in Computer Science with a specialisation in Software Engineering at the Fontys University of Applied Sciences in Eindhoven. Fascinated by the interactions between humans and computers, Ruud pursued further specialisation in the human aspects of information technology by studying Communication and Information Sciences (CIS) at Tilburg University. His focussed on courses in the domain of artificial intelligence, such as computer vision, data mining and machine learning. He obtained his Master's degree in June 2011.

Immediately thereafter, Ruud started as a Ph.D. researcher at the Tilburg Center for Cognition and Communication (TiCC) and as an academic teacher at the department of Communication and Information Sciences (DCI), both at Tilburg University. Supervised by prof. dr. Eric Postma, prof. dr. H. Jaap van den Herik, and dr. ir. Pieter Spronck, he investigated and developed novel techniques that enable computers to perceive a person's gestural cues. His research was part of the *Persuasive Agents* project. The ultimate goal of the project was to develop smart, persuasive, and socially aware embodied agents that are able to engage in natural interactions with humans. In his role as academic teacher, Ruud taught various data processing courses.

Currently, Ruud is working as a Data Scientist within the Oncology Solutions group at Philips Research. As a Data Scientist, he designs, builds, and tests image registration systems for the next generation of adaptive radiation therapy solutions. Additionally, he helps to understand and tap into the potential of big data analytics. As such, he contributes to the development of novel techniques that will help medical experts to find and fight diseases such as cancer.

# LIST OF PUBLICATIONS

Some ideas and figures in the Thesis previously appeared in the following publications.

## JOURNAL PUBLICATIONS

Mattheij, R. J. H., Groeneveld, K., Postma, E. O., & Van den Herik, H. J. (2016). Depth-based detection with region comparison features. *Journal of Visual Communication and Image Representation, 38*, 82-99.

Mattheij, R. J. H., Postma-Nilsenová, M., & Postma, E. O. (2015). Mirror, mirror in the wall: Is there mimicry in you all? *Journal of Ambient Intelligence and Smart Environments, 7*(2), 121-132.

## CONFERENCE PROCEEDINGS

Mattheij, R. J. H., Nilsenová, M., & Postma, E. O. (2013). Vocal and facial imitation of humans interacting with virtual agents. In *Proceedings of the 5th international conference of Affective Computing and Intelligent Interaction* (p. 815-820).

Mattheij, R. J. H., & Postma, E. O. (2013). Feature-based hand detection in visual images. In *Proceedings of TiGeR 2013: The combined meeting of the 10th international Gesture Workshop (GW) and the 3rd Gesture and Speech in Interaction (GESPIN) conference*.

Mattheij, R. J. H., Postma, E. O., Van den Hurk, Y., & Spronck, P. H. M. (2012). Depth-based detection using haar-like features. In *Proceedings of the 24th BENELUX Conference on Artificial Intelligence (BNAIC)*.

Mattheij, R. J. H., & and Postma, E. O. (2012). The eyes have it: Towards enhancing sustainability. In *Proceedings of the 6th international conference on persuasive technology (PERSUASIVE)*.

## BOOK CHAPTERS

Mattheij, R. J. H., Szilvasi, L., De Beer, L., Rakiman, K., & Shahid, S. (2011). GooGreen: Towards Increasing the Environmental Awareness of Households. In *Human-Computer Interaction, Part III, HCII* (p. 500-509).

## POSTER PRESENTATIONS

Mattheij, R. J. H., Groeneveld, K., Postma, E. O., & Van den Herik, H. J. (2015). Improved Body-Part Detection with Microsoft Kinect. Presented at *Symposium on Intelligent Machines*, Nijmegen, The Netherlands.

## POPULAR SCIENTIFIC

Artificial intelligence: Ruud Mattheij at TEDxTilburgUniversity (2014) - a TEDx talk about the use of artificial intelligence for the development of humanlike embodied agents, which are able to engage in natural interactions.

# ACKNOWLEDGEMENTS

Men zegt weleens, dat het schrijven van het dankwoord een van de grootste uitdagingen vormt bij het voltooien van een proefschrift. Persoonlijk vond ik deze uitdaging voornamelijk in het benutten van de beperkte ruimte die ik kon gebruiken om alle mensen te bedanken, die mij de afgelopen jaren geïnspireerd, gesteund en geholpen hebben. Indien je je naam niet terug vindt in dit dankwoord, lieve lezer, wanhoop dan niet - *you're in my heart, and in my mind*.

Bij dezen wil ik graag mijn promotoren bedanken voor hun begeleiding, de talloze brainstormsessies en vele fijne en leerzame gesprekken. Eric, jouw optimisme bleek van onschatbare waarde tijdens de fijne, maar soms ook lastige momenten in de afgelopen jaren. Sorry voor alle pranks en plagerijtjes, en natuurlijk dat ene incidentje met de rode kaarten op je deur. En met dat nieuwsbericht. Oh, en een paar andere geintjes, natuurlijk. Jaap, jouw begeleiding heeft het proefschrift echt naar een hoger niveau weten te tillen. Je weet feilloos tot de kern van de zaak door te dringen en de zwakke plekken in mijn onderzoek te benoemen. Je hebt me regelmatig geholpen om, zogezegd, mijn intuïtie te programmeren. Alhoewel ik in het begin even moest wennen aan je stijl, zou ik je, zonder een seconde te aarzelen, zo weer als promotor willen. Pieter, ik heb genoten van alle fijne brainstormsessies en begeleidingsmomenten die we de afgelopen jaren gehad hebben. Je hebt een prettig kritische blik, en hebt me regelmatig weten te inspireren om heel praktisch na te denken over de uitdagingen van de wetenschap.

De ervaring leert, dat wetenschap het beste (lees: het aangenaamst) bedreven kan worden onder het genot van veel koffie. Heel. Veel. Koffie. De liefde voor het zwarte goud werd gelukkig volledig gedeeld door Bart en Nanne (twee van de gastronomen die TiCC rijk is) en, natuurlijk, door mijn kamergenoot Rick. Naast deze koffieliefhebbers zijn er nog vele andere collega's die het leven in het Dante gebouw bijzonder aangenaam maakten. Ik wil een aantal van hen graag benoemen. Bij dezen; Ruud K. (''Ruud No. 2''), aangezien ik altijd enorm heb genoten van je brede interesses, je plagerijen en scherpe humor. Anja, want je bent toch een beetje de onofficiële TiCC mama van iedereen. Martijn B., vanwege je gortdroge humor die de lunch nog een stukje vrolijker maakte. Rein, aangezien je altijd een grijns op m'n gezicht wist te toveren. Emmelyn, omdat je altijd voor iedereen klaar staat, met goed advies en fijne gesprekken. Jacqueline, ik word altijd zo enorm vrolijk van jouw ent-

housiasme voor alles dat met technologie te maken heeft. Lauraine, Eva, en Joke, omdat we allemaal weten dat jullie het kloppende hart binnen TiCC vormen. Martijn G., omdat je een ver-bovengemiddeld-leuk (en significant grappig) persoon bent. Yu, for being Prince Yu, and the semi-official TiCC ninja. Yueqiao, for your kindness and wise words. Alain, omdat wetenschap nou eenmaal aangenamer is met een Helmondse insteek. Maaike, vanwege je aanstekelijke enthousiasme. Naomi, omdat je mensen altijd een hart onder de riem weet te steken. Emiel, want je goede muzieksmaak blijft me voor altijd bij. Max, omdat je sarcastische gevoel voor humor zo enorm aanstekelijk werkt. Fons en Marc, omdat jullie Vlaamse accent en bijbehorende taalgebruik iedere vergadering tot een waar genot maakten. Ingrid, because you know how to make a person smile. Maria en Joost, omdat jullie passie voor onderwijs zo aanstekelijk werkt. Phoebe, for your kindness (and cookies)! Janneke, Anne-Marie, Marije en Monique, omdat ik zelden zulke gepassioneerde mensen gezien heb. Het is een waar genot geweest om met jullie samen te werken. Mariek, omdat je liefde voor theater ervoor zorgde, dat we een lekker bizarre scène konden spelen tijdens dat beruchte TiCC lipdub filmpje. Per, omdat je altijd in me geloofde. Suleman, because you inspired me to pursue a scientific career in the first place. Carel, vanwege je onnavolgbare gevoel voor humor en hilarische reisverslagen, die menig schrijfsessie onderbroken hebben. Kiek, omdat je me gewoon lekker voor de klas liet staan. Juliette, stiekem ben je helemaal niet zo streng - eigenlijk ben je gewoon ronduit cool. Hille, het blijft geweldig om als ware piraten een TEDx podium te beklimmen om even wat foto's te kunnen maken. Alex, Menno, Véronique, Paul, Sander, en alle andere collega's: omdat jullie enorm fijne mensen zijn!

Naast de onvoorwaardelijke steun die ik heb mogen ontvangen van alle collega's binnen TiCC, hebben ook veel mensen buiten de muren van de universiteit direct of indirect bijgedragen aan het voltooien van dit proefschrift. Ik wil hen bij dezen dan ook graag de revue laten passeren.

Allereerst wil ik mijn familie bedanken voor hun onvermoeibare steun en opbeurende woorden. Alhoewel het een beetje cliché klinkt - en ik nou niet echt het type ben voor emotionele uitspraken - ben ik jullie, papa en mama, mijn broertje Paul en mijn zusje Jeanne, bijzonder dankbaar voor alles dat juliie voor me gedaan hebben, en voor het feit dat jullie er altijd voor mij zijn. Als bedankje wil ik jullie best nog wel een paar keer redden tijdens het duiken.

Daarnaast wil ik graag mijn vriendin bedanken voor alle energie die ze me geeft. Lieve Ineke, mijn favoriete piraatje, je hebt in korte tijd mijn hart veroverd en mijn hoofd op hol gebracht. Ik dank je voor alle mooie momenten die we al samen gehad hebben, en die we nog gaan beleven. Op naar meer

springkussens, oude gevangenissen, nieuwe uitdagingen, en vooral heel erg veel *arrharrr*.

Bij de afronding van dit proefschrift werd ik bijgestaan door twee wetenschappelijke bodyguards, c.q., mijn paranimfen Hans en Alwin. Ik wil jullie hartelijke danken voor alle informele discussies, creatieve oplossingen en een gezonde dosis sarcastische humor, maar vooral: *thanks for watching my back*.

Ook noem ik hier graag Vera, Michelle, Maartje, Ronald, Daniël en Tom vanwege de vele keren dat we samen op het podium gestaan hebben, en de hechte vriendschap die daaruit ontstaan is. Vita, ik dank je voor je vriendschap, en natuurlijk je geweldige danslessen. Rina, dank je wel voor het regelmatig versturen van digitale eenhoorns en regenbogen; wetenschap wordt er een stuk kleurrijker door. Marjet, onze skeelertochten naar - en speciaalbiertjes in - het middeleeuwse plaatsje Dongen is inmiddels gelukkig een rijke traditie geworden. Dineke, de combinatie van je gevoel voor theater en je interesse in technologie leverde, gecombineerd met een snufje praktische rebelsheid, regelmatig geheel nieuwe inzichten op. Ik hoop nog vele malen de degens met je te kunnen kruisen, zowel in professionele zin, als op het podium.

Het schrijven van een proefschrift voelt soms als een wetenschappelijke achtbaanrit. Ik heb het voltooien van deze rit voor een belangrijk deel te danken aan het doorzettingsvermogen, dat ik geleerd heb tijdens de welhaast filosofische Krav Maga lessen van Lex, Daan, Jack en Wendy, waarvoor ik jullie zeer dankbaar ben. Daarnaast wil ik mijn naamgenoot en sparringspartner Ruud bedanken voor alle keren dat hij me hielp bij het toegepast filosoferen, en voor de keren dat ik menig filosofisch inzicht op hem af mocht reageren.

Zoals de oplettende lezer ongetwijfeld gemerkt heeft, heb ik naast wetenschap nog een passie: improvisatie-theater. Deze passie blijkt volledig te worden gedeeld door de fantastische RLG community. Last, but certainly not least, wil ik dan ook graag mijn dank uitspreken richting Rik, Jasper, Cees, Nikie, Rowan, Naduah, Buddy, Ramir, Gees, Esther, Ciska, Justine, Han, Saïd, Rogier, Sebas, Natalie, Brian, Robert, Felix, Bert, Doris, Joeri, Milly, Cor, Alwin, Nathanja, Bas, Okke, Bernd, Tessa, Danny, Cendy, Judit, Paula, Ilya, Shilton, Ayrton, Tony, Björn, Dennis, Nico - en natuurlijk iedereen die hier nog niet genoemd wordt. Het is een eer om schouder aan schouder met jullie te mogen staan, en samen met jullie door de verlaten gangen van de bekende, 140 jaar oude locatie te kunnen rennen. Het bouwen van de RLG belevingen behoort tot de mooiste ervaringen van de afgelopen jaren. Ik weet zeker dat de Keizer trots zou zijn. After all, "Niks is onmogelijk, er is slechts een gebrek aan capaciteit". *Sir, yes sir!*

# SIKS DISSERTATION SERIES

**2016-30** Ruud Mattheij (TiU), *The eyes have it.*
**2016-29** Nicolas Höning (TUD), *Peak reduction in decentralised electricity systems - Markets and prices for flexible planning.*
**2016-28** Mingxin Zhang (TUD), *Large-scale agent-based social simulation - A study on epidemic prediction and control.*
**2016-27** Wen Li (TUD), *Understanding geo-spatial information on social media.*
**2016-26** Dilhan Thilakarathne (VU), *In or out of control: Exploring computational models to study the role of human awareness and control in behavioural choices, with applications in aviation and energy management domains.*
**2016-25** Julia Kiseleva (TU/e), *Using contextual information to understand searching and browsing behavior.*
**2016-24** Brend Wanders (UT), *Repurposing and probabilistic integration of data: An iterative and data model independent approach.*
**2016-23** Fei Cai (UVA), *Query auto completion in information retrieval.*
**2016-22** Grace Lewis (VU), *Software architecture strategies for cyber-foraging systems.*
**2016-21** Alejandro Moreno Célleri (UT), *From traditional to interactive playspaces: Automatic analysis of player behavior in the interactive tag playground.*
**2016-20** Daan Odijk (UVA), *Context & semantics in news & web search.*
**2016-19** Julia Efremova (Tu/e), *Mining social structures from genealogical data.*
**2016-18** Albert Meroño Peñuela (VU), *Refining statistical data on the web.*
**2016-17** Berend Weel (VU), *Towards embodied evolution of robot organisms.*
**2016-16** Guangliang Li (UVA), *Socially intelligent autonomous agents that learn from human reward.*
**2016-15** Steffen Michels (RUN), *Hybrid probabilistic logics - Theoretical aspects, algorithms and experiments.*
**2016-14** Ravi Khadka (UU), *Revisiting legacy software system modernization.*
**2016-13** Nana Baah Gyan (VU), *The web, speech technologies and rural development in West Africa - An ICT4D approach.*
**2016-12** Max Knobbout (UU), *Logics for modelling and verifying normative multi-agent systems.*
**2016-11** Anne Schuth (UVA), *Search engines that learn from their users.*
**2016-10** George Karafotias (VU), *Parameter control for evolutionary algorithms.*
**2016-09** Archana Nottamkandath (VU), *Trusting crowdsourced information on cultural artefacts.*
**2016-08** Matje van de Camp (TiU), *A link to the past: Constructing historical social networks from unstructured data.*
**2016-07** Jeroen de Man (VU), *Measuring and modeling negative emotions for virtual training.*
**2016-06** Michel Wilson (TUD), *Robust scheduling in an uncertain environment.*
**2016-05** Evgeny Sherkhonov (UVA), *Expanded acyclic queries: Containment and an application in explaining missing answers.*
**2016-04** Laurens Rietveld (VU), *Publishing and consuming linked data.*
**2016-03** Maya Sappelli (RUN), *Knowledge work in context: User centered knowledge worker support.*
**2016-02** Michiel Christiaan Meulendijk (UU), *Optimizing medication reviews through decision support: Prescribing a better pill to swallow.*
**2016-01** Syed Saiden Abbas (RUN), *Recognition of shapes by humans and machines.*
**2015-35** Jungxao Xu (TUD), *Affective body language of humanoid robots: Perception and effects in human robot interaction.*
**2015-34** Victor de Graaf (UT), *Geo-social recommender systems.*
**2015-33** Frederik Schadd (TUD), *Ontology mapping with auxiliary resources.*
**2015-32** Jerome Gard (UL), *Corporate venture management in SMEs.*
**2015-31** Yakup Koç (TUD), *On the robustness of power grids.*
**2015-30** Kiavash Bahreini (OU), *Real-time multimodal emotion recognition in E-Learning.*
**2015-29** Hendrik Baier (UM), *Monte-Carlo tree search enhancements for one-player and two-player domains.*
**2015-28** Janet Bagorogoza(TiU), *Knowledge management and high performance: The Uganda financial institutions model for HPO.*
**2015-27** Sándor Héman (CWI), *Updating compressed column stores.*
**2015-26** Alexander Hogenboom (EUR), *Sentiment analysis of text guided by semantics and structure.*
**2015-25** Steven Woudenberg (UU), *Bayesian tools for early disease detection.*
**2015-24** Richard Berendsen (UVA), *Finding people, papers, and posts: Vertical search algorithms and evaluation.*
**2015-23** Luit Gazendam (VU), *Cataloguer support in cultural heritage.*
**2015-21** Sibren Fetter (OUN), *Using peer-support to expand and stabilize online learning.*
**2015-20** Loïs Vanhée (UU), *Using culture and values to support flexible coordination.*
**2015-19** Bernardo Tabuenca (OUN), *Ubiquitous technology for lifelong learners.*
**2015-18** Holger Pirk (CWI), *Waste not, want not! - Managing relational data in asymmetric memories.*
**2015-17** André van Cleeff (UT), *Physical and digital security mechanisms: Properties, combinations and trade-offs.*
**2015-16** Changyun Wei (UT), *Cognitive coordination for cooperative multi-robot teamwork.*

**2015-15** Klaas Andries de Graaf (VU), *Ontology-based software architecture documentation.*
**2015-14** Bart van Straalen (UT), *A cognitive approach to modeling bad news conversations.*
**2015-13** Giuseppe Procaccianti (VU), *Energy-efficient software.*
**2015-12** Julie M. Birkholz (VU), *Modi operandi of social network dynamics: The effect of context on scientific collaboration networks.*
**2015-11** Yongming Luo (TU/e), *Designing algorithms for big graph datasets: A study of computing bisimulation and joins.*
**2015-10** Henry Hermans (OUN), *OpenU: Design of an integrated system to support lifelong learning.*
**2015-09** Randy Klaassen (UT), *HCI perspectives on behavior change support systems.*
**2015-08** Jie Jiang (TUD), *Organizational compliance: An agent-based model for designing and evaluating organizational interactions.*
**2015-07** Maria-Hendrike Peetz (UvA), *Time-aware online reputation analysis.*
**2015-06** Farideh Heidari (TUD), *Business process quality computation - Computing non-functional requirements to improve business processes.*
**2015-05** Christoph Bösch (UT), *Cryptographically enforced search pattern hiding.*
**2015-04** Howard Spoelstra (OUN), *Collaborations in open learning environments.*
**2015-03** Twan van Laarhoven (RUN), *Machine learning for network data.*
**2015-02** Faiza Bukhsh (TiU), *Smart auditing: Innovative compliance checking in customs controls.*
**2015-01** Niels Netten (UvA), *Machine learning for relevance of information in crisis response.*
**2014-47** Shangsong Liang (UVA), *Fusion and diversification in information retrieval.*
**2014-46** Ke Tao (TUD), *Social web data analytics: Relevance, redundancy, diversity.*
**2014-45** Birgit Schmitz (OUN), *Mobile games for learning: A pattern-based approach.*
**2014-44** Paulien Meesters (TiU), *Intelligent blauw. Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.*
**2014-43** Kevin Vlaanderen (UU), *Supporting process improvement using method increments.*
**2014-42** Carsten Eijckhof (CWI/TUD), *Contextual multidimensional relevance models.*
**2014-41** Frederic Hogenboom (EUR), *Automated detection of financial events in news text.*
**2014-40** Walter Omona (RUN), *A framework for knowledge management using ICT in higher education.*
**2014-39** Jasmina Maric (TiU), *Web communities, immigration, and social capital.*
**2014-38** Danny Plass-Oude Bos (UT), *Making brain-computer interfaces better: Improving usability through post-processing.*
**2014-37** Maral Dadvar (UT), *Experts and machines united against cyberbullying.*
**2014-36** Joos Buijs (TU/e), *Flexible evolutionary algorithms for mining structured process models.*
**2014-35** Joost van Ooijen (UU), *Cognitive agents in virtual worlds: A middleware design approach.*
**2014-34** Christina Manteli (VU), *The effect of governance in global software development: Analyzing transactive memory systems.*
**2014-33** Tesfa Tegegne (RUN), *Service discovery in eHealth.*
**2014-32** Naser Ayat (UvA), *On entity resolution in probabilistic data.*
**2014-31** Leo van Moergestel (UU), *Agent technology in agile multiparallel manufacturing and product support.*
**2014-30** Peter de Cock (TiU), *Anticipating criminal behaviour.*
**2014-29** Jaap Kabbedijk (UU), *Variability in multi-tenant enterprise software.*
**2014-28** Anna Chmielowiec (VU), *Decentralized k-clique matching.*
**2014-27** Rui Jorge Almeida (EUR), *Conditional density models integrating fuzzy and probabilistic representations of uncertainty.*
**2014-26** Tim Baarslag (TUD), *What to bid and when to stop.*
**2014-25** Martijn Lappenschaar (RUN), *New network models for the analysis of disease interaction.*
**2014-24** Davide Ceolin (VU), *Trusting semi-structured web data.*
**2014-23** Eleftherios Sidirourgos (UvA/CWI), *Space efficient indexes for the big data era.*
**2014-22** Marieke Peeters (UU), *Personalized educational games - Developing agent-supported scenario-based training.*
**2014-21** Kassidy Clark (TUD), *Negotiation and monitoring in open environments.*
**2014-20** Mena Habib (UT), *Named entity extraction and disambiguation for informal text: The missing link.*
**2014-19** Vinicius Ramos (TU/e), *Adaptive hypermedia courses: Qualitative and quantitative evaluation and tool support.*
**2014-18** Mattijs Ghijsen (UVA), *Methods and models for the design and study of dynamic agent organizations.*
**2014-17** Kathrin Dentler (VU), *Computing healthcare quality indicators automatically: Secondary use of patient data and semantic interoperability.*
**2014-16** Krystyna Milian (VU), *Supporting trial recruitment and design by automatically interpreting eligibility criteria.*
**2014-15** Natalya Mogles (VU), *Agent-based analysis and support of human functioning in complex socio-technical systems: Applications in safety and healthcare.*
**2014-14** Yangyang Shi (TUD), *Language models with meta-information.*
**2014-13** Arlette van Wissen (VU), *Agent-based support for behavior change: Models and applications in health and safety Domains.*
**2014-12** Willem van Willigen (VU), *Look ma, no hands: Aspects of autonomous vehicle control.*
**2014-11** Janneke van der Zwaan (TUD), *An empathic virtual buddy for social support.*
**2014-10** Ivan Salvador Razo Zapata (VU), *Service value networks.*
**2014-09** Philip Jackson (TiU), *Toward human-level artificial intelligence: Representation and computation of meaning in natural language.*

**2014-08** Samur Araujo (TUD), *Data integration over distributed and heterogeneous data endpoints.*
**2014-07** Arya Adriansyah (TU/e), *Aligning observed and modeled behavior.*
**2014-06** Damian Tamburri (VU), *Supporting networked software development.*
**2014-05** Jurriaan van Reijsen (UU), *Knowledge perspectives on advancing dynamic capability.*
**2014-04** Hanna Jochmann-Mannak (UT), *Websites for children: Search strategies and interface design - Three studies on children's search performance and evaluation.*
**2014-03** Sergio Raul Duarte Torres (UT), *Information retrieval for children: Search behavior and solutions.*
**2014-02** Fiona Tuliyano (RUN), *Combining system dynamics with a domain modeling method.*
**2014-01** Nicola Barile (UU), *Studies in learning monotone models from data.*
**2013-43** Marc Bron (UVA), *Exploration and contextualization through interaction and concepts.*
**2013-42** Léon Planken (TUD), *Algorithms for simple temporal reasoning.*
**2013-41** Jochem Liem (UVA), *Supporting the conceptual modelling of dynamic systems: A knowledge engineering perspective on qualitative reasoning.*
**2013-40** Pim Nijssen (UM), *Monte-Carlo tree search for multi-player games.*
**2013-39** Joop de Jong (TUD), *A method for enterprise ontology based design of enterprise information systems.*
**2013-38** Eelco den Heijer (VU), *Autonomous evolutionary art.*
**2013-37** Dirk Börner (OUN), *Ambient learning displays.*
**2013-36** Than Lam Hoang (TU/e), *Pattern mining in data streams.*
**2013-35** Abdallah El Ali (UvA), *Minimal mobile human computer interaction.*
**2013-34** Kien Tjin-Kam-Jet (UT), *Distributed deep web search.*
**2013-33** Qi Gao (TUD), *User modeling and personalization in the microblogging sphere.*
**2013-32** Kamakshi Rajagopal (OUN), *Networking for learning: The role of networking in a lifelong learner's professional development.*
**2013-31** Dinh Khoa Nguyen (TiU), *Blueprint model and language for engineering cloud applications.*
**2013-30** Joyce Nakatumba (TU/e), *Resource-aware business process management: Analysis and support.*
**2013-29** Iwan de Kok (UT), *Listening heads.*
**2013-28** Frans van der Sluis (UT), *When complexity becomes interesting: An inquiry into the information eXperience.*
**2013-27** Mohammad Huq (UT), *Inference-based framework managing data provenance.*
**2013-26** Alireza Zarghami (UT), *Architectural support for dynamic homecare service provisioning.*
**2013-25** Agnieszka Anna Latoszek-Berendsen (UM), *Intention-based decision support: A new way of representing and implementing clinical guidelines in a decision support system.*
**2013-24** Haitham Bou Ammar (UM), *Automated transfer in reinforcement learning.*
**2013-23** Patricio de Alencar Silva (TiU), *Value activity monitoring.*
**2013-22** Tom Claassen (RUN), *Causal discovery and logic.*
**2013-21** Sander Wubben (TiU), *Text-to-text generation by monolingual machine translation.*
**2013-20** Katja Hofmann (UvA), *Fast and reliable online learning to rank for information retrieval.*
**2013-19** Renze Steenhuizen (TUD), *Coordinated multi-agent planning and scheduling.*
**2013-18** Jeroen Janssens (TiU), *Outlier selection and one-class classification.*
**2013-17** Koen Kok (VU), *The PowerMatcher: Smart coordination for the smart electricity grid.*
**2013-16** Eric Kok (UU), *Exploring the practical benefits of argumentation in multi-agent deliberation.*
**2013-15** Daniel Hennes (UM), *Multiagent learning - Dynamic games and applications.*
**2013-14** Jafar Tanha (UVA), *Ensemble approaches to semi-supervised learning.*
**2013-13** Mohammad Safiri (UT), *Service tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly.*
**2013-12** Marian Razavian (VU), *Knowledge-driven migration to services.*
**2013-11** Evangelos Pournaras (TUD), *Multi-level reconfigurable self-organization in overlay services.*
**2013-10** Jeewanie Jayasinghe Arachchige (TiU), *A unified modeling framework for service design.*
**2013-09** Fabio Gori (RUN), *Metagenomic data analysis: Computational methods and applications.*
**2013-08** Robbert-Jan Merk (VU), *Making enemies: Cognitive modeling for opponent agents in fighter pilot simulators.*
**2013-07** Giel van Lankveld (TiU), *Quantifying individual player differences.*
**2013-06** Romulo Goncalves (CWI), *The data cyclotron: Juggling data and queries for a data warehouse audience.*
**2013-05** Dulce Pumareja (UT), *Groupware requirements evolutions patterns.*
**2013-04** Chetan Yadati (TUD), *Coordinating autonomous planning and scheduling.*
**2013-03** Szymon Klarman (VU), *Reasoning with contexts in description logics.*
**2013-02** Erietta Liarou (CWI), *MonetDB/DataCell: Leveraging the column-store database technology for efficient and scalable stream processing.*
**2013-01** Viorel Milea (EUR), *News analytics for financial decision support.*
**2012-51** Jeroen de Jong (TUD), *Heuristics in dynamic sceduling: A practical framework with a case study in elevator dispatching.*
**2012-50** Steven van Kervel (TUD), *Ontologogy driven enterprise information systems engineering.*
**2012-49** Michael Kaisers (UM), *Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions.*
**2012-48** Jorn Bakker (TU/e), *Handling abrupt changes in evolving time-series data.*

**2012-47** Manos Tsagkias (UVA), *Mining social media: Tracking content and predicting behavior.*
**2012-46** Simon Carter (UVA), *Exploration and exploitation of multilingual data for statistical machine translation.*
**2012-45** Benedikt Kratz (TiU), *A model and language for business-aware transactions.*
**2012-44** Anna Tordai (VU), *On combining alignment techniques.*
**2012-42** Dominique Verpoorten (OU), *Reflection amplifiers in self-regulated Learning.*
**2012-41** Sebastian Kelle (OU), *Game design patterns for learning.*
**2012-40** Agus Gunawan (TiU), *Information access for SMEs in Indonesia.*
**2012-39** Hassan Fatemi (UT), *Risk-aware design of value and coordination networks.*
**2012-38** Selmar Smit (VU), *Parameter tuning and scientific testing in evolutionary algorithms.*
**2012-37** Agnes Nakakawa (RUN), *A collaboration process for enterprise architecture creation.*
**2012-36** Denis Ssebugwawo (RUN), *Analysis and evaluation of collaborative modeling processes.*
**2012-35** Evert Haasdijk (VU), *Never too old to learn - On-line evolution of controllers in swarm- and modular robotics.*
**2012-34** Pavol Jancura (RUN), *Evolutionary analysis in PPI networks and applications.*
**2012-33** Rory Sie (OUN), *Coalitions in Cooperation Networks (COCOON).*
**2012-32** Wietske Visser (TUD), *Qualitative multi-criteria preference representation and reasoning.*
**2012-31** Emily Bagarukayo (RUN), *A learning by construction approach for higher order cognitive skills improvement, building capacity and infrastructure.*
**2012-30** Alina Pommeranz (TUD), *Designing human-centered systems for reflective decision making.*
**2012-29** Almer Tigelaar (UT), *Peer-to-peer information retrieval.*
**2012-28** Nancy Pascall (TiU), *Engendering technology empowering women.*
**2012-27** Hayrettin Gurkok (UT), *Mind the sheep! User experience evaluation & brain-computer interface games.*
**2012-26** Emile de Maat (UVA), *Making sense of legal text.*
**2012-25** Silja Eckartz (UT), *Managing the business case development in inter-organizational IT projects: A methodology and its application.*
**2012-24** Laurens van der Werff (UT), *Evaluation of noisy transcripts for spoken document retrieval.*
**2012-23** Christian Muehl (UT), *Toward affective brain-computer interfaces: Exploring the neurophysiology of affect during human media interaction.*
**2012-22** Thijs Vis (TiU), *Intelligence, politie en veiligheidsdienst: verenigbare grootheden?.*
**2012-21** Roberto Cornacchia (TUD), *Querying sparse matrices for information retrieval.*
**2012-20** Ali Bahramisharif (RUN), *Covert visual spatial attention: A robust paradigm for brain-computer interfacing.*
**2012-19** Helen Schonenberg (TU/e), *What's next? Operational support for business process execution.*
**2012-18** Eltjo Poort (VU), *Improving solution architecting practices.*
**2012-17** Amal Elgammal (TiU), *Towards a comprehensive framework for business process compliance.*
**2012-16** Fiemke Both (VU), *Helping people by understanding them - Ambient agents supporting task execution and depression treatment.*
**2012-15** Natalie van der Wal (VU), *Social agents. Agent-based modelling of integrated internal and social dynamics of cognitive and affective processes.*
**2012-14** Evgeny Knutov (TU/e), *Generic adaptation framework for unifying adaptive web-based systems.*
**2012-13** Suleman Shahid (TiU), *Fun and face: Exploring non-verbal expressions of emotion during playful interactions.*
**2012-12** Kees van der Sluijs (TU/e), *Model driven design and data integration in semantic web information systems.*
**2012-11** J.C.B. Rantham Prabhakara (TU/e), *Process mining in the large: Preprocessing, discovery, and diagnostics.*
**2012-10** David Smits (TU/e), *Towards a generic distributed adaptive hypermedia environment.*
**2012-09** Ricardo Neisse (UT), *Trust and privacy management support for context-aware service platforms.*
**2012-08** Gerben de Vries (UVA), *Kernel methods for vessel trajectories.*
**2012-07** Rianne van Lambalgen (VU), *When the going gets tough: Exploring agent-based models of human performance under demanding conditions.*
**2012-06** Wolfgang Reinhardt (OU), *Awareness support for knowledge workers in research networks.*
**2012-05** Marijn Plomp (UU), *Maturing interorganisational information systems.*
**2012-04** Jurriaan Souer (UU), *Development of content management system-based web applications.*
**2012-03** Adam Vanya (VU), *Supporting architecture evolution by mining software repositories.*
**2012-02** Muhammad Umair (VU), *Adaptivity, emotion, and rationality in human and ambient agent models.*
**2012-01** Terry Kakeeto (TiU), *Relationship marketing for SMEs in Uganda.*
**2011-49** Andreea Niculescu (UT), *Conversational interfaces for task-oriented spoken dialogues: Design aspects influencing interaction quality.*
**2011-48** Mark Ter Maat (UT), *Response selection and turn-taking for a sensitive artificial listening agent.*
**2011-47** Azizi Bin Ab Aziz (VU), *Exploring computational models for intelligent support of persons with depression.*
**2011-46** Beibei Hu (TUD), *Towards contextualized information delivery: A rule-based architecture for the domain of mobile police work.*
**2011-45** Herman Stehouwer (TiU), *Statistical language models for alternative sequence selection.*
**2011-44** Boris Reuderink (UT), *Robust brain-computer interfaces.*
**2011-43** Henk van der Schuur (UU), *Process improvement through software operation knowledge.*
**2011-42** Michal Sindlar (UU), *Explaining behavior through mental state attribution.*
**2011-41** Luan Ibraimi (UT), *Cryptographically enforced distributed data access control.*

**2011-40** Viktor Clerc (VU), *Architectural knowledge management in global software development.*
**2011-39** Joost Westra (UU), *Organizing adaptation using agents in serious games.*
**2011-38** Nyree Lemmens (UM), *Bee-inspired distributed optimization.*
**2011-37** Adriana Burlutiu (RUN), *Machine learning for pairwise data: Applications for preference learning and supervised network inference.*
**2011-36** Erik van der Spek (UU), *Experiments in serious game design: A cognitive approach.*
**2011-35** Maaike Harbers (UU), *Explaining agent behavior in virtual training.*
**2011-34** Paolo Turrini (UU), *Strategic reasoning in interdependence: Logical and game-theoretical investigations.*
**2011-33** Tom van der Weide (UU), *Arguing to motivate decisions.*
**2011-32** Nees-Jan van Eck (EUR), *Methodological advances in bibliometric mapping of science.*
**2011-31** Ludo Waltman (EUR), *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality.*
**2011-30** Egon van den Broek (UT), *Affective Signal Processing (ASP): Unraveling the mystery of emotions.*
**2011-29** Faisal Kamiran (TU/e), *Discrimination-aware classification.*
**2011-28** Rianne Kaptein(UVA), *Effective focused retrieval by exploiting query context and document structure.*
**2011-27** Aniel Bhulai (VU), *Dynamic website optimization through autonomous management of design patterns.*
**2011-26** Matthijs Aart Pontier (VU), *Virtual agents for human communication - Emotion regulation and involvement: Distance trade-offs in embodied conversational agents and robots.*
**2011-25** Syed Waqar ul Qounain Jaffry (VU), *Analysis and validation of models for trust dynamics.*
**2011-24** Herwin van Welbergen (UT), *Behavior generation for interpersonal coordination with virtual humans on specifying, scheduling and realizing multimodal virtual human behavior.*
**2011-23** Wouter Weerkamp (UVA), *Finding people and their utterances in social media.*
**2011-22** Junte Zhang (UVA), *System evaluation of archival description and access.*
**2011-21** Linda Terlouw (TUD), *Modularization and specification of service-oriented systems.*
**2011-20** Qing Gu (VU), *Guiding service-oriented software engineering - A view-based approach.*
**2011-19** Ellen Rusman (OU), *The mind's eye on personal profiles.*
**2011-18** Mark Ponsen (UM), *Strategic Decision-Making in complex games.*
**2011-17** Jiyin He (UVA), *Exploring topic structure: Coherence, diversity and relatedness.*
**2011-16** Maarten Schadd (UM), *Selective search in games of different complexity.*
**2011-15** Marijn Koolen (UvA), *The meaning of structure: The value of link evidence for information retrieval.*
**2011-14** Milan Lovric (EUR), *Behavioral finance and agent-based artificial markets.*
**2011-13** Xiaoyu Mao (TiU), *Airport under control - Multiagent scheduling for airport ground handling.*
**2011-12** Carmen Bratosin (TU/e), *Grid architecture for distributed process mining.*
**2011-11** Dhaval Vyas (UT), *Designing for awareness: An experience-focused HCI perspective.*
**2011-10** Bart Bogaert (TiU), *Cloud content contention.*
**2011-09** Tim de Jong (OU), *Contextualised mobile media for learning.*
**2011-08** Nieske Vergunst (UU), *BDI-based generation of robust task-oriented dialogues.*
**2011-07** Yujia Cao (UT), *Multimodal information presentation for high load human computer interaction.*
**2011-06** Yiwen Wang (TU/e), *Semantically-enhanced recommendations in cultural heritage.*
**2011-05** Base van der Raadt (VU), *Enterprise architecture coming of age - Increasing the performance of an emerging discipline.*
**2011-04** Hado van Hasselt (UU), *Insights in reinforcement learning - Formal analysis and empirical evaluation of temporal-difference.*
**2011-03** Jan Martijn van der Werf (TU/e), *Compositional design and verification of domponent-based information systems.*
**2011-02** Nick Tinnemeier (UU), *Organizing agent organizations. Syntax and operational semantics of an organization-oriented programming language.*
**2011-01** Botond Cseke (RUN), *Variational algorithms for Bayesian inference in latent Gaussian models.*
**2010-53** Edgar Meij (UVA), *Combining concepts and language models for information access.*
**2010-52** Peter-Paul van Maanen (VU), *Adaptive support for human-computer teams: Exploring the use of cognitive models of trust and attention.*
**2010-51** Alia Khairia Amin (CWI), *Understanding and supporting information seeking tasks in multiple sources.*
**2010-50** Bouke Huurnink (UVA), *Search in audiovisual broadcast archives.*
**2010-49** Jahn-Takeshi Saito (UM), *Solving difficult game positions.*
**2010-47** Chen Li (UT), *Mining process model variants: Challenges, techniques, examples.*
**2010-46** Vincent Pijpers (VU), *e3alignment: Exploring inter-organizational business-ICT alignment.*
**2010-45** Vasilios Andrikopoulos (TiU), *A theory and model for the evolution of software services.*
**2010-44** Pieter Bellekens (TU/e), *An approach towards context-sensitive and user-adapted access to heterogeneous data sources, illustrated in the television domain.*
**2010-43** Peter van Kranenburg (UU), *A computational approach to content-based retrieval of folk song melodies.*
**2010-42** Sybren de Kinderen (VU), *Needs-driven service bundling in a multi-supplier setting - The computational e3-service approach.*
**2010-41** Guillaume Chaslot (UM), *Monte-Carlo tree search.*
**2010-40** Mark van Assem (VU), *Converting and integrating vocabularies for the semantic web.*
**2010-39** Ghazanfar Farooq Siddiqui (VU), *Integrative modeling of emotions in virtual agents.*

**2010-38** Dirk Fahland (TU/e), *From scenarios to components.*
**2010-37** Niels Lohmann (TU/e), *Correctness of services and their composition.*
**2010-36** Jose Janssen (OU), *Paving the way for lifelong learning: Facilitating competence development through a learning path specification.*
**2010-35** Dolf Trieschnigg (UT), *Proof of concept: Concept-based biomedical information retrieval.*
**2010-34** Teduh Dirgahayu (UT), *Interaction design in service compositions.*
**2010-33** Robin Aly (UT), *Modeling representation uncertainty in concept-based multimedia retrieval.*
**2010-32** Marcel Hiel (TiU), *An adaptive service oriented architecture: Automatically solving interoperability problems.*
**2010-31** Victor de Boer (UVA), *Ontology enrichment from heterogeneous sources on the web.*
**2010-30** Marieke van Erp (TiU), *Accessing natural history - Discoveries in data cleaning, structuring, and retrieval.*
**2010-29** Stratos Idreos(CWI), *Database cracking: Towards auto-tuning database kernels.*
**2010-28** Arne Koopman (UU), *Characteristic relational patterns.*
**2010-27** Marten Voulon (UL), *Automatisch contracteren.*
**2010-26** Ying Zhang (CWI), *XRPC: Efficient distributed query processing on heterogeneous XQuery engines.*
**2010-25** Zulfiqar Ali Memon (VU), *Modelling human-awareness for ambient agents: A human mindreading perspective.*
**2010-24** Dmytro Tykhonov, *Designing generic and efficient negotiation strategies.*
**2010-23** Bas Steunebrink (UU), *The logical structure of emotions.*
**2010-22** Michiel Hildebrand (CWI), *End-user support for access to heterogeneous linked data.*
**2010-21** Harold van Heerde (UT), *Privacy-aware data management by means of data degradation.*
**2010-20** Ivo Swartjes (UT), *Whose story is it anyway? How improv informs agency and authorship of emergent narrative.*
**2010-19** Henriette Cramer (UvA), *People's responses to autonomous and adaptive systems.*
**2010-18** Charlotte Gerritsen (VU), *Caught in the act: Investigating crime by agent-based simulation.*
**2010-17** Spyros Kotoulas (VU), *Scalable discovery of networked resources: Algorithms, infrastructure, applications.*
**2010-16** Sicco Verwer (TUD), *Efficient identification of timed automata, theory and practice.*
**2010-15** Lianne Bodenstaff (UT), *Managing dependency relations in inter-organizational models.*
**2010-14** Sander van Splunter (VU), *Automated web service reconfiguration.*
**2010-13** Gianluigi Folino (RUN), *High performance data mining using bio-inspired techniques.*
**2010-12** Susan van den Braak (UU), *Sensemaking software for crime analysis.*
**2010-11** Adriaan Ter Mors (TUD), *The world according to MARP: Multi-Agent Route Planning.*
**2010-10** Rebecca Ong (UL), *Mobile communication and protection of children.*
**2010-09** Hugo Kielman (UL), *Politiële gegevensverwerking en privacy - Naar een effectieve waarborging.*
**2010-08** Krzysztof Siewicz (UL), *Towards an improved regulatory framework of free software - Protecting user freedoms in a world of software communities and eGovernments.*
**2010-07** Wim Fikkert (UT), *Gesture interaction at a distance.*
**2010-06** Sander Bakkes (TiU), *Rapid adaptation of video game AI.*
**2010-05** Claudia Hauff (UT), *Predicting the effectiveness of queries and retrieval systems.*
**2010-04** Olga Kulyk (UT), *Do you know what I know? Situational awareness of co-located teams in multidisplay environments.*
**2010-03** Joost Geurts (CWI), *A document engineering model and processing framework for multimedia documents.*
**2010-02** Ingo Wassink (UT), *Work flows in life science.*
**2010-01** Matthijs van Leeuwen (UU), *Patterns that matter.*
**2009-46** Loredana Afanasiev (UvA), *Querying XML: Benchmarks and recursion.*
**2009-45** Jilles Vreeken (UU), *Making pattern mining useful.*
**2009-44** Roberto Santana Tapia (UT), *Assessing business-IT alignment in networked organizations.*
**2009-43** Virginia Nunes Leal Franqueira (UT), *Finding multi-step attacks in computer networks using heuristic search and mobile ambients.*
**2009-42** Toine Bogers (TiU), *Recommender systems for social bookmarking.*
**2009-41** Igor Berezhnyy (TiU), *Digital analysis of paintings.*
**2009-40** Stephan Raaijmakers (TiU), *Multinomial language learning: Investigations into the geometry of language.*
**2009-39** Christian Stahl (TU/e, Humboldt-Universitaet zu Berlin), *Service substitution – A behavioral approach based on Petri nets.*
**2009-38** Riina Vuorikari (OU), *Tags and self-organisation: A metadata ecology for learning resources in a multilingual context.*
**2009-37** Hendrik Drachsler (OUN), *Navigation support for learners in informal learning networks.*
**2009-36** Marco Kalz (OUN), *Placement support for learners in learning networks.*
**2009-35** Wouter Koelewijn (UL), *Privacy en politiegegevens - Over geautomatiseerde normatieve informatie-uitwisseling.*
**2009-34** Inge van de Weerd (UU), *Advancing in software product management: An incremental method engineering approach.*
**2009-33** Khiet Truong (UT), *How does real affect affect affect recognition in speech?.*
**2009-32** Rik Farenhorst (VU) and Remco de Boer (VU), *Architectural knowledge management: Supporting architects and auditors.*
**2009-31** Sofiya Katrenko (UVA), *A closer look at learning relations from text.*
**2009-30** Marcin Zukowski (CWI), *Balancing vectorized query execution with bandwidth-optimized storage.*

**2009-29** Stanislav Pokraev (UT), *Model-driven semantic integration of service-oriented applications.*
**2009-28** Sander Evers (UT), *Sensor data management with probabilistic models.*
**2009-27** Christian Glahn (OU), *Contextual support of social engagement and reflection on the web.*
**2009-26** Fernando Koch (UU), *An agent-based model for the development of intelligent mobile services.*
**2009-25** Alex van Ballegooij (CWI), *RAM: Array database management through relational mapping.*
**2009-24** Annerieke Heuvelink (VU), *Cognitive models for training simulations.*
**2009-23** Peter Hofgesang (VU), *Modelling web usage in a changing environment.*
**2009-22** Pavel Serdyukov (UT), *Search for expertise: Going beyond direct evidence.*
**2009-21** Stijn Vanderlooy (UM), *Ranking and reliable classification.*
**2009-20** Bob van der Vecht (UU), *Adjustable autonomy: Controling influences on decision making.*
**2009-19** Valentin Robu (CWI), *Modeling preferences, strategic reasoning and collaboration in agent-mediated electronic markets.*
**2009-18** Fabian Groffen (CWI), *Armada, An evolving database system.*
**2009-17** Laurens van der Maaten (TiU), *Feature extraction from visual data.*
**2009-16** Fritz Reul (TiU), *New architectures in computer chess.*
**2009-15** Rinke Hoekstra (UVA), *Ontology representation - Design patterns and ontologies that make sense.*
**2009-14** Maksym Korotkiy (VU), *From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA).*
**2009-13** Steven de Jong (UM), *Fairness in multi-agent systems.*
**2009-12** Peter Massuthe (TU/e, Humboldt-Universitaet zu Berlin), *Operating guidelines for services.*
**2009-11** Alexander Boer (UVA), *Legal theory, sources of law & the semantic web.*
**2009-10** Jan Wielemaker (UVA), *Logic programming for knowledge-intensive interactive applications.*
**2009-09** Benjamin Kanagwa (RUN), *Design, discovery and construction of service-oriented systems.*
**2009-08** Volker Nannen (VU), *Evolutionary agent-based policy analysis in dynamic environments.*
**2009-07** Ronald Poppe (UT), *Discriminative vision-based recovery and recognition of human motion.*
**2009-06** Muhammad Subianto (UU), *Understanding classification.*
**2009-05** Sietse Overbeek (RUN), *Bridging supply and demand for knowledge intensive iasks - Based on knowledge, cognition, and quality.*
**2009-04** Josephine Nabukenya (RUN), *Improving the quality of organisational policy making using collaboration engineering.*
**2009-03** Hans Stol (TiU), *A framework for evidence-based policy making using IT.*
**2009-02** Willem Robert van Hage (VU), *Evaluating ontology-alignment techniques.*
**2009-01** Rasa Jurgelenaite (RUN), *Symmetric causal independence models.*

# TICC PH.D. SERIES

**47** Ruud Mattheij, *The eyes have it (2016)*
**46** Rick Tillman, *Language matters: The influence of language and language use on cognition (2016)*
**45** Annemarie Quispel, *Data for all: How designers and laymen use and evaluate information visualizations (2016)*
**44** Matje van de Camp, *A link to the past: Constructing historical social networks from unstructured data (2016)*
**43** Hans Westerbeek, *Visual realism: Exploring effects on memory, language production, comprehension, and preference (2016)*
**42** Janet Bagorogoza, *Knowledge management and high performance: The Uganda financial institutions models for HPO (2015)*
**41** Elisabeth Lubinga, *Stop HIV/AIDS. Start talking? The effects of rhetorical figures in health messages on conversations among South African adolescents (2015)*
**40** Marieke Hoetjes, *Talking hands: Reference in speech, gesture and sign (2015)*
**39** Sterling Hutchinson, *How symbolic and embodied representations work in concert (2015)*
**38** Mandy Visser, *Better use your head: How people learn to signal emotions in social contexts (2015)*
**37** Pauline Meesters, *Intelligent Blauw (2014)*
**36** Jasmina Maric, *Web communities, immigration and social capital (2014)*
**35** Constantijn Kaland, *Prosodic marking of semantic contrasts: Do speakers adapt to addressees? (2014)*
**34** Peter de Kock, *Anticipating criminal behaviour (2014)*
**33** Jorrig Vogels, *Referential choices in language production: The role of accessibility (2014)*
**32** Philip Jackson, *Toward human-level artificial intelligence: Representation and computation of meaning in natural language (2014)*
**31** Douglas Mastin, *Exploring infant engagement, language socialization and vocabulary development: A study of rural and urban communities in Mozambique (2013)*
**30** Ruud Koolen, *Need I say more? On overspecification in definite reference (2013)*
**29** Lisanne van Weelden, *Metaphor in good shape (2013)*
**28** Martijn Balsters, *Expression and perception of emotions: The case of depression, sadness and fear (2013)*
**27** Jeroen Janssens, *Outlier selection and one-class classification (2013)*
**26** Sander Wubben, *Text-to-text generation using monolingual machine translation (2013)*
**25** Giel van Lankveld, *Quantifying individual player differences (2013)*
**24** Agus Gunawan, *Information access for SMEs in Indonesia (2012)*
**23** Nancy Pascall, *Engendering technology empowering women (2012)*
**22** Thijs Vis, *Intelligence, politie en veiligheidsdienst: Verenigbare grootheden? (2012)*
**21** Suleman Shahid, *Fun & face: Exploring non-verbal expressions of emotion during playful interactions (2012)*
**20** Terry Kakeeto-Aelen, *Relationship marketing for SMEs in Uganda (2012)*
**19** Herman Stehouwer, *Statistical language models for alternative sequence selection (2011)*
**18** Lisette Mol, *Language in the hands (2011)*
**17** Olga Petukhova, *Multidimensional dialogue modelling (2011)*
**16** Xiaoyu Mao, *Airport under control (2011)*
**15** Bart Bogaert, *Cloud content contention (2011)*
**14** Edwin Commandeur, *Implicit causality and implicit consequentiality in language comprehension (2010)*
**13** Marieke van Erp, *Accessing natural history: Discoveries in data cleaning, structuring, and retrieval (2010)*
**12** Maria Mos, *Complex lexical items (2010)*
**11** Sander Bakkes, *Rapid adaptation of video game AI (2010)*
**10** Toine Bogers, *Recommender systems for social bookmarking (2009)*
**9** Igor Berezhnoy, *Digital analysis of paintings (2009)*
**8** Stephan Raaijmakers, *Multinomial language learning (2009)*
**7** Laurens van der Maaten, *Feature extraction from visual data (2009)*
**6** Fritz Reul, *New architectures in computer chess (2009)*
**5** Sander Canisius, *Structured prediction for natural language processing (2009)*
**4** Jeroen Geertzen, *Dialogue act recognition and prediction (2009)*
**3** Hans Stol, *A framework for evidence-based policy making using IT (2009)*
**2** Ben Torben-Nielsen, *Dendritic morphology: Function shapes structure (2008)*
**1** Pashiera Barkhuysen, *Audiovisual prosody in interaction (2008)*