Eleventh International Workshop
Modelling and Reasoning in Context

Jörg Cassens, Rebekah Wegener, Anders Kofod-Petersen

Digital ECAI 2020, Santiago de Compostela, Galicia, Spain

Context has been and remains a central topic in Artificial Intelligence in general. In terms of recent concerns within AI, context is crucial for understanding causation, for personalization and ethical AI and for the development of contextual AI. As well as these broader concerns, research on context is vital for developments within specific areas of AI:

Machine Learning & Knowledge Representation: In most cases, context can not be modelled a-priori but contextual information has to be inferred from data. In addition, contextual features might change over time, necessitating machine learning approaches for dynamic adaptation of context models and methods for reasoning with uncertainty.

Human-Centred AI: In Human-Computer Interaction, context is crucial for human-centred approaches to systems development. Because of the interdisciplinary and transdisciplinary nature of MRC, the workshop series is ideally suited to build bridges between these two closely related sub-fields of computer science and to facilitate the exchange of knowledge and methods for human-centred approaches.

Ethical & Responsible AI: Context is core to ethical and responsible approaches to AI, as reasoning about contextual parameters is inherent in human interpretation of ethical questions. Furthermore, explicit models of context can help mitigate the effects of algorithmic and data bias.

Responsible Personalization: Context is central to enabling a more collaborative partnership between humans and machines. But personalization brings risks to privacy and current methods embed and hide algorithmic bias and data biases. Making context explicit helps mitigating those effects.

Explainable AI: With a renewed interest in explainable systems, context is also increasingly important to identify user needs and system capabilities in providing explanations of system behaviour at runtime.

Autonomous Agents & Robotics: The concept of context is itself contextual and always pertains to the acting agent. Additionally, context is an important issue in autonomous systems, in particular if they are to be integrated in socio-technical environments with human actors.

Context is inherently an interdisciplinary topic that, besides AI and HCI, has clear relations to linguistics and semiotics, cognitive science and psychology, mathematics and philosophy as well as other areas such as sociology and anthropology. Given the recent interest in AI beyond the field, MRC acts as a bridge between these different communities and serves as a means for integrating models and findings from these different areas.

The Modelling and Reasoning in Context workshop series aims to bring together researchers & practitioners from different communities, both industry & academia, to study, understand, and explore issues surrounding context and to share problems, techniques and solutions across a broad range of areas. The workshop covers different understandings of what context is, a variety of approaches to learn about context from data, and different approaches to modelling. MRC is concerned with varying mechanisms and techniques for reasoning with context, storage of contextual information, and methods for enabling integration of context and application knowledge.

The organizers would like to thank all the authors for submitting their papers and the members of the program committee as well as the additional reviewers for their valuable review contribution.

Workshop website
mrc.kriwi.de

Hildesheim, August 2020
Jörg Cassens, Rebekah Wegener, Anders Kofod-Petersen

Workshop Chairs

- Jörg Cassens, IMAI, University of Hildesheim, Germany
- Rebekah Wegener, RWTH Aachen University, Germany *and* Audaxi, Sydney, Australia
- Anders Kofod-Petersen, Alexandra Institute, Copenhagen, Denmark *and* NTNU, Trondheim, Norway

Program Committee

- Juan Carlos Augusto – Middlesex University, UK
- Tobias Baur – Augsburg University, Germany
- Tarek Richard Besold – Alpha Health AI Lab, Catalonia, Spain
- Ioannis Chatzigiannakis – Sapienza University of Rome, Italy
- Henning Christiansen – Roskilde University, Denmark
- Sten Hanke – FH Joanneum, Graz, Austria
- Martin Christof Kindsmüller– Brandenburg University of Applied Sciences, Germany
- Christian Kohlschein – Germany
- Olya Kudina – TU Delft, The Netherlands
- David Leake – Indiana University Bloomington, USA
- Amy Loutfi – Örebro University, Sweden
- Ana Gabriela Maguitman – Universidad Nacional del Sur, Argentina
- Tobias Meisen – Bergische University Wuppertal, Germany
- Grzegorz J. Nalepa – AGH University, Kraków, Poland
- Stella Neumann – RWTH Aachen University, Germany
- Jeannie Marie Paterson – The University of Melbourne, Australia
- Michaela Reisinger – Austrian Institute of Technology, Vienna, Austria
- Myrthe L. Tielman – TU Delft, The Netherlands
- Harko Verhagen – Stockholm University, Sweden
- M. Birna van Riemsdijk – University of Twente, The Netherlands

Additional Reviewers

- David Aha – Naval Research Laboratory, USA
- Ilir Kola – TU Delft, The Netherlands
- Tim Miller – University of Melbourne, Australia

Contents

Affective Games Provide Controlable Context. Proposal of an Experimental Framework

Laura Żuchowska and **Krzysztof Kutt** and **Krzysztof Geleta** and **Szymon Bobek** and **Grzegorz J. Nalepa**[1]

Abstract. We propose an experimental framework for Affective Computing based of video games. We developed a set of specially designed mini-games, based of carefully selected game mechanics, to evoke emotions of participants of a larger experiment. We believe, that games provide a controllable yet overall ecological environment for studying emotions. We discuss how we used our mini-games as an important counterpart of classical visual and auditory stimuli. Furthermore, we present a software tool supporting the execution and evaluation of experiments of this kind.

1 INTRODUCTION

Emotions constitute an important context for interpretation of human behavior. Affective computing (AfC) is a field of study devoted to the computer-based analysis, modeling and synthesis of emotions [14]. In our work in this area, we focus on the use of wearable and mobile devices to support the acquisition and interpretation of bodily signals in order to the detect changes of affective states and possibly recognize the corresponding emotional states of subjects. We believe, that the context-aware systems paradigm considered in computer science, should take into the account the affective dimension [11]. Furthermore, the computer models should be personalized, i.e. take into the account individual differences of human behavior, as well as personality traits [8].

One of the principal challenges in the AfC experiments is the actual process to evoke individual emotions for the training and calibration of computer models. In the psychological literature, some of the typical experimental procedures assume the use of standardized visual and auditory stimuli that are supposed to evoke the specific emotions. From our perspective, such an approach is not sufficient as the experimental situation very often does not seem natural to the participant, furthermore it is not personalized. To tackle this challenge, in our work we employ computer games as the source of specific, rich, natural, yet controllable context to evoke emotions [12].

In this paper we present an experimental setup using affective games to evoke emotions of the participants. The principal contributions include: the design of original video games aimed at AfC experiments, a framework for configuration of experiments using such games, putting these two in the context of the BIRAFFE experiments we conducted.

The rest of the paper is organized as follows: In Section 2 we discuss the detailed motivation of our work. Then in Section 3 we describe an experiment in AfC we conducted to acquire data on the individual affective reactions. In this experiment we used a set of affective games we specifically developed for this task, as described in 4. Furthermore, we realized that in order to provide flexibility of such experiments, we should have a framework supporting the reconfiguration of such experiments for a range of game levels. We developed a prototype of such a framework, as described in Section 5. A short comparison with other solutions is provided in Section 6. In Section 7 we describe the evaluation of our work. We conclude the paper in Section 8.

2 MOTIVATION

Research on emotions requires, on the one hand, a controllable experimental environment to evoke and detect and emotions, on the other, natural conditions for experiments in order to minimize a possible discomfort for the participants. Video games seem to be a good trade-off between these two extreme requirements. Games allow to control the appearing stimuli and log everything that happened, especially the reaction times, Moreover, the environment is rich in stimuli and allows for user interaction with objects, including emotionally related interaction framed in the so-called Affective Loop [12].

"Regular" games, available on the market, do not meet the requirements of the experimental environment. First of all, they provide a (too) rich environment in which the player may do (too) many things. In such an environment, a very large sample size is needed to get the right statistical power to draw conclusions, which makes experiments difficult to conduct. Also, the use of machine learning methods will not be trivial, as there are many variables in such case, some of which will only be disruptive noise. Secondly, "regular" games do not allow for the evaluation of emotions too often. The player is constantly engaged in the game and interrupting it to complete the questionnaire will reduce the immersion of the game.

These issues have been observed in our previous experiments [11] including the BIRAFFE1 experiment [8]. To address them, a set of mini games, with restricted experimental conditions, was created. Each of them is built up on a very limited set of stimuli, with the aim of evoking a limited set of emotional reactions. The following sections describe an experiment called BIRAFFE2 (see Section 3) in which three such games were used (see Section 4).

The BIRAFFE2 experiment has led to observation of further issues that need to be addressed when conducting game experiments. In particular, attention has been drawn to the fact that all mini games should generate event logs in a uniform format to avoid additional pre-processing steps when analysing the collected data. It is equally important to implement questionnaires directly in the games, at the end of each mini game. Filling out the questionnaires at the end of the gaming session makes the impressions fuzzy and the self-description may not be accurate enough.

Therefore, in parallel with the BIRAFFE experiments, a dedicated

[1] Jagiellonian University, Poland, email: krzysztof.kutt@uj.edu.pl, szymon.bobek@uj.edu.pl, grzegorz.j.nalepa@uj.edu.pl

framework was developed to automate the preparation of game-based affective experiments. It allows to generate an experiment template with questionnaires between different levels and provides a database-based logging interface. A detailed description of the framework can be found in Section 5.

3 THE BIRAFFE2 EXPERIMENT

The BIRAFFE2 study included 103 participants (33% female) between 18 and 26 (*M* = 21.63, *SD* = 1.32), recruited among students of the Artificial Intelligence Basics course at AGH University of Science and Technology, Kraków, Poland and their friends.

It is a revised version of a previous experiment called BIRAFFE1 (**Bi**o-**R**eactions **a**nd **F**aces **f**or **E**motion-based Personalization) described in [8].The aim of the study was to collect physiological data paired with behavioral data, which can then be used to develop models for prediction of emotions.

Behavioral data were twofold: from the part in which the subjects played three games (for details see Section 4) and from the classical experiment, in which sound and visual stimuli (from IADS [2] and IAPS [9] datasets respectively) were presented and then subjects were asked to assess what emotions they evoked. Specifically, the stimuli was presented for 6 seconds, what was followed by 6 seconds for affective rating with the use of custom widget with 2-dimensional space (valence and arousal). The whole behavioral data was collected as a set of logs in comma-separated (CSV) files.

Physiological signals, Electrocardiogram (ECG) and Electrodermal activity (EDA), were gathered using BITalino (r)evolution kit, as it is the most promising of cheap mobile hardware platforms (for comparison see [7]). Besides ECG and EDA, during the experiment also the following signals were collected: accelerometer and gyroscope from gamepad, facial images taken by webcam (every 250 milliseconds), screencast of the whole game session.

The whole protocol consisted of several phases:

1. NEO-FFI paper-and-pen questionnaire [15] for personality measurement (approx. 10 minutes),
2. Physiological devices setup (approx. 2 minutes),
3. Baseline signals recording (1 minute),
4. Instructions and training (approx. 5 minutes),
5. First part of stimuli presentation and rating (17.5 minutes),
6. Games session (up to 15 minutes in total),
7. Second part of stimuli presentation and rating (17.5 minutes),
8. Three paper-and-pen GEQ questionnaires [5] (one for each game) and gaming experience questionnaire (approx. 10 minutes).

The whole protocol lasts up to 75 minutes. Steps 3-7 were done on a PC. All of them were controlled by the Python 3.8 with the use of PsychoPy 3.2.4 library [13]. Participant interacted with the procedure only with a gamepad.

4 EVOKING EMOTIONS WITH AFFECTIVE GAMES

In order to support the game sessions of the experiment, three specific affective mini-games were created [16]. The aim for all the games was to create an immense amount of emotions in a short time. The main obstacle was the inability to create an intriguing story, therefore the whole section of narrative elements was discarded. The only way of building an affective project was to make different sets of games with a variety of mechanics and audiovisuals.

The simplest solution to create an emotional-changing environment was to revolve around the overall difficulty of games. While making the neutral, peaceful stage can relieve stress for the player, the loud and hard level can intensify the rage and increase heartbeat. Therefore, three genres have been selected: roguelike, platform, and maze. The first level is balanced to be an easy stage, supposed to develop energetic, happy emotions. On the contrary, the second level is extremely hard to beat, filled with traps, to give the sensation of unjust and fury. This juxtaposition is important to the study, given the sudden change. Last phase is neutral, without any emotion-boosting elements, it exists to check the player's decision-making, behavior and bodily changes due to previous irritation. Additionally, a proper collection of game patterns was implemented. For every stage, depending on what emotions it should boost, from that collection separate elements were chosen.

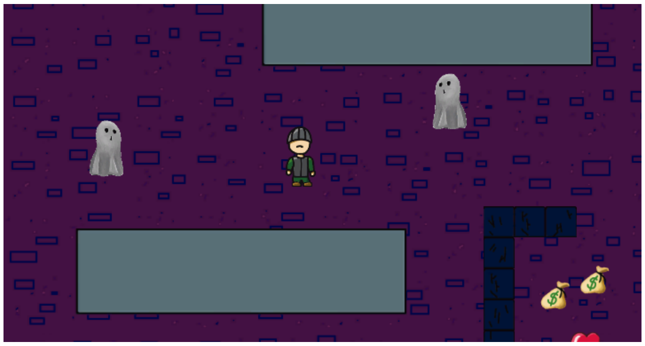


Figure 1. Stage 1: an example screen of the game

Stage one contains elements such as score tracking, weapons, enemies and looting. The finishing condition is elimination of all antagonists – no stress-inducting time limit was implemented. The difficulty in this level was balanced by setting the damage per second of the protagonist much higher than the one of the antagonist. While players can shoot up to 5 projectiles per second, enemies can shoot only one attack per second. Moreover, the speed of player's projectiles is 2.5 times higher for the default weapon. An additional blaster was placed on a map, giving the possibility for the user to eliminate the enemies even easier. Furthermore, in the case the subject is not used to playing games, health points can be increased by picking up heart-shaped objects. In order to unleash more fun and any form of achievement-getting sensation, a score tracker is incrementing when picking up money bags from the floor or from the killed enemy (the amount of bags dropped from antagonist is random).

In the platform game (stage 2) traps and time limit were implemented. Both of them are crucial in order to imply stress and rage. Until the end of the game, the player has to go through the whole level. However when the player dies, he respawns in the last checkpoint – a yellow flag with letter 'C'; when touched, a happy, although very distorted sound is played. There are two possible ways for the protagonist to lose: falling down off the stage, or stepping into a spike trap. Considering the fact that this level is supposed to be insanely hard to get through, two additional traps were implemented to basic blocks. The first type is an invisible block – before the protagonist collides with them, they are not to be seen in any way by the user. If the player dies after triggering the visibility, it is once again set to invisible. Similar mechanics is once again used for next type of traps – falling blocks. Once the collision with the user happens, blocks start to fall down.

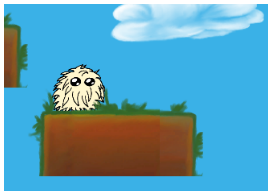


Figure 2. Stage 2: falling block before trigger

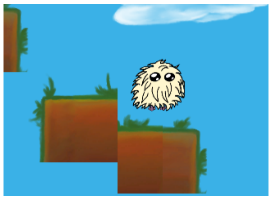


Figure 3. Stage 2: falling block after trigger

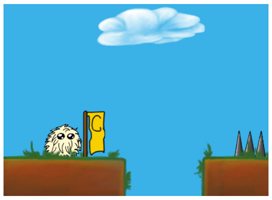


Figure 4. Stage 2: invisible block before trigger

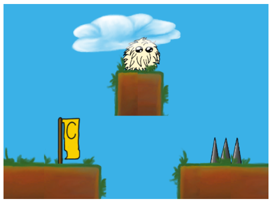


Figure 5. Stage 2: invisible block after trigger

For the last game in stage 3 memorizing the way through a maze is the only important part. No time nor score tracking is implemented. Visuals are very simple, no distracting elements were added. The choices made by the player are saved into logs, which will be discussed later on.

Size and shape of colliders were also adjusted to the game genre and difficulty intended. For the first scene, the collider for the protagonist is smaller than his real model. It removes the feeling of being hit before the projectile hits the player. On the contrary, in the second game colliders are too big. Player can get hit by a trap before he touches it with a model. This decision was made to enhance the irritation and the feeling of unjust. For the last level, colliders were adjusted to not hit the walls too often, so the movement will be pleasant and smooth. Another intentional difference in stage two from others is the protagonist's movement. It was designed similarly to the jumping mechanics, although it doesn't stop at a certain speed – the player's model is constantly given acceleration. This is a perfect example of poorly made mechanics, which are incredibly hard to control.

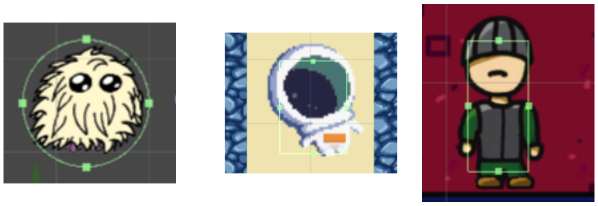


Figure 6. Stage 3: protagonist colliders for different stages.

To boost the affective part of gameplay, sounds provided by *NIMH Center for the Study of Emotion and Attention* [2] were added to the every stage. They were proven to change the state of user's emotion by their degree of affectiveness. This level was separated into two values – intensity of the feeling (arousal) and pleasantness of a sound. Depending on these two values, proper sounds were chosen and included in the games.

Sound	Pleasure	Arousal
Puppy	2.88	4.91
Bees	2.16	7.03
Vomit	2.08	6.59
Babies cry	2.04	6.87
Baby cry	2.75	6.51
Scream	2.05	8.16
Child abuse	1.57	7.27
Applause	7.32	5.55
Rollercoaster	6.94	7.54
Colonial music	6.53	5.84
Bugle	6.32	6.35
Rock n roll	7.90	6.85
Funk music	6.94	5.87

Table 1. Affective sounds used in study

Furthermore, music themes and in-game sounds were recorded. The design was created with a view to expected emotions. First game's theme consists of electronic/rock music, sounds of picking items are clear, echo has been added to each sound. To keep the second level unbalanced and irritating, time signature for background music was disturbed – the last eighth note was erased. This gives an unsettling feeling, like someone has been playing off tempo. Additionally, each time the player dies increases the pitch and distortion effects for the background theme. Protagonist has a high-pitched voice, which gets more infuriating with every death. The sound of winning (which is hard to achieve, given the difficulty of the game) has a very disappointing and unsatisfying tone. Last level has a pleasant theme, edited to sound like old arcade, 8-bit music.

In order to get as much as possible from single gameplay about the state of players' emotions, additionally to their bodily functions, a proper context-gathering mechanism is required. It is implemented as a set of different event logs that are saved for each stage. Some information is constantly saved, no matter the level – the data about current player's position, ID and timestamp of an affective sound played in the background. For stage one, events such as killing an enemy, death, the amount of all objects picked up, current state of health and points are saved with the proper timestamp. Additionally, the amount of projectiles shot and their accuracy is recorded – this gives more insight on the aggressiveness and gaming experience. There are no enemies and pickable items in the second stage, therefore the distortion rate of music, number of deaths and the data about traps triggered is saved for every iteration. In the last stage, the amount of dead ends encountered and the data about going off the correct path is being saved.

All of the games were developed using the Unity Engine. It is a powerful environment, with tons of possibilities. One of many features used are previously mentioned colliders. The engine contains a variety of collider shapes, components and traits. For instance, Box Colliders were used not only as physical objects, but also as triggers in rooms for the first stage, logging in the third level etc. Animations in all games were handled through the Animator Controller feature. Another remarkable example for the possible power of Unity Engine is Camera - just a simple change in view can drastically change the

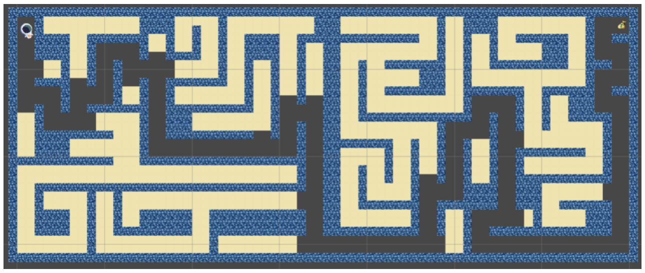


Figure 7. Illustration on how the data about wrong path is saved

perception of the player. While first-person cameras can increase the immersion with the protagonist, third-person cameras can give an insight on what's going on around the character, escalating the feeling of stress for the subject. Moreover, sounds and music have separate components - Audio Listener and Audio Source. Both of them have a simple mixer, included inside the engine. Those components create impressive opportunities for manipulation of emotion. The simplest example used in the game is the distortion attribute, for each death in the second stage.

5 FRAMEWORK FOR GAME CONFIGURATION IN EXPERIMENTS

In order to get information from subjects about their feelings towards the games, the GEQ questionnaire was used [5]. This survey consists of three parts: The Core Questionnaire, The Social Presence Module and The Post-game Module. All of them contain important information about different sections of study. All of them involve questions about feelings, with a range of possible answers from 0 to 4. Zero means 'not at all', one means 'slightly', two is 'moderately', three is 'fairly' and the last, four is 'extremely'. First part of the survey has 33 questions about emotions and sensations felt during the game, for example: 'I was good at it' and 'I felt frustrated'. The Social Presence Module contains 17 questions, however it should only be taken when any form of social interactions were taken in game, whether it's another person or a simple non-playable character interaction. The last section involves 17 questions about the overall feeling of a subject after the game has been played, for instance: 'I felt satisfied' and 'I found it a waste of time'.

To make this questionnaire a part of study, and also to provide a unified context-logging mechanism, a software framework has been written. It is responsible for starting all mini-games and preparing the survey after each game. To install the plugin you need to copy .dll files and prefabs into Unity project. After restarting the editor, you should see “Feedback” menu in the menu bar and the configuration file in “/Resources” directory. In order to start using the plugin, you need to create an SQL database with tables for each survey form you want to include in your game. By default the plugin saves answers as integers, so each question should have a separate column of this type.

Everything is connected through a proper configuration file. It is required to set correct database provider and connection string. After that, the model classes can be generated by choosing “Generate model classes” button from “Feedback” menu. Pressing "Create survey form" button will open wizard that allows to choose different

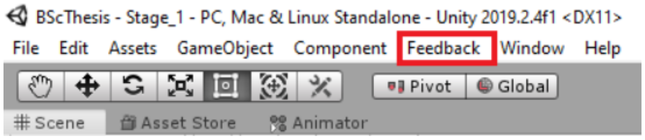


Figure 8. 'Feedback' option available in menu

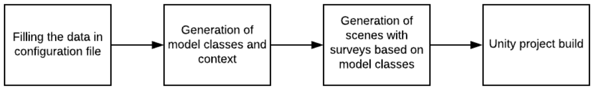
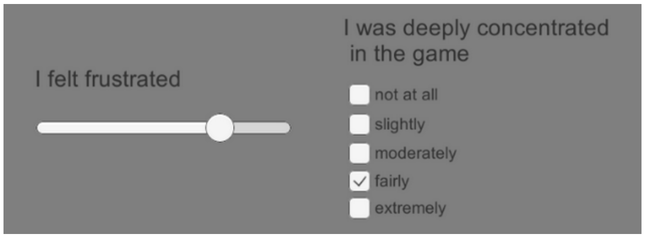

Filling the data in configuration file
Generation of model classes and context
Generation of scenes with surveys based on model classes
Unity project build

Figure 9. Plugin's use case diagram

types of questions (radio buttons, slider or dropdown). The plugin will create new scene with questions based on the table in database. After that, a request to manually add generated script-handling data persistence to an empty object in the scene will pop up. At the end, there is a possibility to change the text and position of questions in the scene. After building the project, a game is started and the survey pops up next. When all questions are answered, the framework sends all data to the local SQL database.



I felt frustrated

I was deeply concentrated in the game

□ not at all
□ slightly
□ moderately
☑ fairly
□ extremely

Figure 10. Survey example

This framework has a very high potential for further studies. Firstly, it gives an opportunity to create a multi-platform study. More computers would be available to use for a study. Furthermore, a mobile version could be implemented. This way, even more subjects would've taken part in a study. Another possible future usage is the adjustment of level difficulty for every game, dependent on answers in the survey. This would increase the affective part of study, as personal change in games would take place.

6 RELATED WORK

There are quite a few different frameworks for affective research. On the one hand, one can point out the tools used to build classical psychological experiments, like PsychoPy [13], OpenSeasame [10], or E-Prime[2] and on the other hand, the tools used for affective experiments with games, e.g. FILTWAM [1], iHEARu-PLAY [4], or emoCook [3].

The tools in the first group offer various widgets for collecting information from users, making it possible to transfer virtually any paper questionnaire to the electronic version. However, they do not allow one to control a stimulus-rich game environment. This problem has been addressed in the second group of tools, where affective interaction is carried out in games (e.g. educational game "emoCook"). Nevertheless, these solutions are prepared for specific applications and do not provide a general solution for affective experiments.

The framework described in this paper combines the advantages of these two groups. It both allows for the use of games as a research environment and is a general solution, allowing for the inclusion of any games (written in Unity) and any questionnaires (the application is not limited to the GEQ described in the article).

7 EVALUATION

The motivation to introduce a few short mini-games was better control over the emotions evoked during the experiment. The assumption was that each game aim is to evoke specific emotions using a small number of stimuli. These assumptions were confirmed by the results of the GEQ questionnaire.

Revised list of GEQ factors [6][3] was used for analysis. A series of one-way ANOVAs was conducted to evaluate the differences between games. Post-hoc comparisons were done using the Tukey HSD test. Analysis was performed in Python with scipy[4] and statsmodels[5] libraries.

The strongest effects can be observed for the second level, which should give the sensation of unjust and fury. It was connected with significantly higher *Negativity* ($M = 2.85$), significantly lower *Positive Affect* ($M = 1.14$) and significantly lower *Competence*[6] ($M = 0.86$) than the two other stages (*Negativity*: $M = 1.11$ and $M = 0.70$, *Positive Affect*: $M = 2.53$ and $M = 2.49$, *Competence*: $M = 2.42$ and $M = 2.78$, for Stage 1 and Stage 3 respectively).

Stage 1, designed as an easy stage connected with positive emotions, and Stage 3, designed as emotionally neutral, were both evaluated as the ones with the higher *Positive Affect* (there were no significant differences between them). Neutrality of the third level is revealed with the lowest *Negativity* ($M = 0.70$; significantly lower than the first level, $M = 1.11$).

Flow, indicating whether or not players have lost control of their time in the game, was significantly lowest in the third level ($M = 1.35$) than the other two levels ($M = 1.99$ and $M = 2.02$), indicating that for this factor the most important is the fact that emotions are evoked, no matter whether they are positive or negative. Finally, *Immersion*, the subjective connection to the game, was low for all levels ($M = 1.75$, $M = 1.23$, $M = 1.66$ for levels 1-3 respectively), which is also consistent with the assumptions. The games were too short for the players to get fully involved.

We tested the framework on several platforms including Windows, Linux and mobile platforms running Android operating systems. It ran correctly on all of the platforms[7] proving its portability between most popular operating systems. We also tested it with different databases including remote MySQL databases and SQLite database for Android systems, where in both cases it worked correctly. While experiments presented in this paper did not use the framework, they will be used by us as a baseline for future evaluation of the framework.

8 FUTURE WORK AND SUMMARY

In the paper we presented our recent work conducted as a part of the BIRAFFE2 experiment in Affective Computing. As a novel part of the experiment we developed three specially designed mini-games, based of carefully selected game mechanics. We believe, that games provide a controllable yet overall ecological environment for studying emotions. We used these games as an important counterpart of classical visual and auditory stimuli during the experiment to evoke emotions of participants. Moreover, we presented a software tool, with a built-in context-logging mechanism, supporting the execution, automation and evaluation of experiments of this kind.

In the future, we would like to develop our work in several directions. First of all, based on the analysis of the results of the experiment, we will continue the development of new games with improved mechanics to fine tune the evocation of emotions. Ultimately, we expect games will help us in developing computer-based personalized models of emotions to be used in different applications. Furthermore, based on the future findings, we would like to study the aspects of emotional adaptation and personalization in games using the machine learning methods. Finally, our current setup is ready to be used not just in desktop games, but also on mobile devices. We will explore this direction, as mobile games not only constitute a very important market for games, but also offer new opportunities for interaction.

REFERENCES

[1] Kiavash Bahreini, Rob Nadolski, and Wim Westera, 'FILTWAM - A framework for online affective computing in serious games', in *Fourth International Conference on Games and Virtual Worlds for Serious Applications, VS-GAMES 2012, Genoa, Italy, October 29-31, 2012*, eds., Alessandro De Gloria and Sara de Freitas, volume 15 of *Procedia Computer Science*, pp. 45–52. Elsevier, (2012).

[2] Margaret M. Bradley and Peter J. Lang, 'The international affective digitized sounds (2nd edition; iads-2): Affective ratings of sounds and instruction manual. technical report B-3', Technical report, University of Florida, Gainsville, FL, (2007).

[3] Jose Maria Garcia-Garcia, Victor M. Ruiz Penichet, María Dolores Lozano, Juan Enrique Garrido, and Effie Lai-Chong Law, 'Multimodal affective computing to enhance the user experience of educational software applications', *Mobile Information Systems*, **2018**, 8751426:1–8751426:10, (2018).

[4] Simone Hantke, Florian Eyben, Tobias Appel, and Björn W. Schuller, 'ihearu-play: Introducing a game for crowdsourced data collection for affective computing', in *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21-24, 2015*, pp. 891–897. IEEE Computer Society, (2015).

[5] Wijnand A. IJsselsteijn, Yvonne A. W. de Kort, and Karolien Poels, *The Game Experience Questionnaire*, Technische Universiteit Eindhoven, 2013.

[6] Daniel M. Johnson, M. John Gardner, and Ryan Perry, 'Validation of two game experience scales: The player experience of need satisfaction (PENS) and game experience questionnaire (GEQ)', *Int. J. Hum. Comput. Stud.*, **118**, 38–46, (2018).

[7] Krzysztof Kutt, Wojciech Binek, Piotr Misiak, Grzegorz J. Nalepa, and Szymon Bobek, 'Towards the development of sensor platform for processing physiological data from wearable sensors', in *Artificial Intelligence and Soft Computing - 17th International Conference, ICAISC*

[2] See: `https://pstnet.com/products/e-prime/`.

[3] The values are ranging from 0 (not at all) to 4 (extremely) for each factor.

[4] See: `https://www.scipy.org/`.

[5] See: `https://www.statsmodels.org/`.

[6] *Competence* reflects how well players judged their own performance against the game's goals.

[7] The only requirement is to use Unity build 2019.2.19f1

2018, Zakopane, Poland, June 3-7, 2018, Proceedings, Part II, pp. 168–178, (2018).

[8] Krzysztof Kutt, Dominika Drążyk, Paweł Jemioło, Szymon Bobek, Barbara Giżycka, Víctor Rodríguez Fernández, and Grzegorz J. Nalepa, 'BIRAFFE: Bio-reactions and faces for emotion-based personalization', in *AfCAI 2019: Workshop on Affective Computing and Context Awareness in Ambient Intelligence*, volume 2609 of *CEUR Workshop Proceedings*. CEUR-WS.org, (2020).

[9] Peter J. Lang, Margaret M. Bradley, and B. N. Cuthbert, 'International affective picture system (iaps): Affective ratings of pictures and instruction manual. technical report B-3', Technical report, The Center for Research in Psychophysiology, University of Florida, Gainsville, FL, (2008).

[10] Sebastiaan Mathôt, Daniel Schreij, and Jan Theeuwes, 'OpenSesame: An open-source, graphical experiment builder for the social sciences', *Behavior Research Methods*, **44**(2), 314–324, (2012).

[11] Grzegorz J. Nalepa, Krzysztof Kutt, and Szymon Bobek, 'Mobile platform for affective context-aware systems', *Future Generation Computer Systems*, **92**, 490–503, (mar 2019).

[12] Grzegorz J. Nalepa, Krzysztof Kutt, Barbara Giżycka, Paweł Jemioło, and Szymon Bobek, 'Analysis and use of the emotional context with wearable devices for games and intelligent assistants', *Sensors*, **19**(11), 2509, (2019).

[13] Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv, 'Psychopy2: Experiments in behavior made easy', *Behavior Research Methods*, **51**(1), 195–203, (2019).

[14] Rosalind W. Picard, *Affective Computing*, MIT Press, Cambridge, MA, 1997.

[15] Bogdan Zawadzki, Jan Strelau, Piotr Szczepaniak, and Magdalena Śliwińska, *Inwentarz osobowości NEO-FFI Costy i McCrae. Polska adaptacja*, Pracownia Testów Psychologicznych, Warszawa, 1998.

[16] Laura Żuchowska, *Game Design with Unity for Affective Games*, BSc thesis, AGH University of Science and Technology, 2020. Supervisor: G.J. Nalepa.

Grouping Situations Based on their Psychological Characteristics Gives Insight into Personal Values

Ilir Kola[1] and **Catholijn M. Jonker**[2] and **Myrthe L. Tielman**[3] and **M. Birna van Riemsdijk**[4]

Abstract. Support agents are investigated more and more as a way of assisting people in carrying out daily tasks. Support agents should be flexible in adapting their support to what their user needs. Research suggests that the situation someone is in affects their behaviour, however its effect has not been incorporated in the decision making of support agents. Modelling the characteristics of situations explicitly and studying their effect on internal perceptions of the user, such as their personal values, would enable support agents to provide more personalized support. We propose a method which groups situations according to their psychological characteristics, and in turn determines which personal values of the user would be promoted or demoted in each group of situations. To do this, we conduct a user study to gather data from participants about situations that they encounter in their daily lives. Results show that the created groups of situations significantly promote or demote certain personal values. This approach can allow support agents to help the user in a way which is in line with their personal values.

1 INTRODUCTION

Kurt Lewin, already 80 years ago, proposed that human behaviour is a function of both the personality of the person, as well as the situation in which they are in [18]. This is now a widely accepted idea in social psychology, after multiple debates in the field [24]. However, applications of support agents (e.g. [13, 20, 31]) focus mostly on modelling internal aspects of the user. Personal values are one of these aspects. They represent what is important to people [9], and because of that, they guide behavior. However, how important a certain value is for the user is not the only factor that guides behaviour. Whether that value is actually relevant in a given situation also plays an important role. For example, the fact that having an exciting life is important to someone plays a role in deciding the next holiday destination, but most probably does not affect the decision whether to have pizza or salad for dinner. On the other hand, the fact that someone values health would affect that decision, since having salad is supposed to promote the value health (i.e., help you fulfill it), whereas having pizza can demote it (i.e., prevent you from fulfilling it). This means that apart from personal values, it is important to also consider how the situation in which someone is in affects those values. This information can be used by a support agent in combination with information about the value preferences of the user in order to offer support on how to handle daily life situations. Continuing the previous example, the agent would suggest having a salad to a user that finds health important.

In this work we will explore the relationship between the situation in which a user is in, and the personal values that are affected by the situation. To achieve this, first of all we explore ways how to group similar situations together. To do so, we will extend the work on Context Space Theory [21], which refers to a group of similar situations as a *subspace*. A situation subspace is a group of situations which have the same range of numerical values on certain attributes (Section 3.1). In this work, we use psychological characteristics of situations as attributes. Psychological characteristics are seen as dimensions that can be used to describe situations, similar to the manner in which people can be described with traits, attributes, or qualities [7]. Examples of these characteristics are positivity, duty, intellect, mating etc [24] (Section 3.2). This leads to the following research question:

- What methods can we use to group situations according to their psychological characteristics as context attributes?

Then, we investigate whether the identified subspaces significantly promote or demote personal values. Our research hypothesis is:

- Situations of the same subspace significantly promote or demote the same personal values, in comparison to a general population of situations.

While the research question and hypothesis guide the work presented in this paper, we do not aim to provide definitive answers here. Rather, as this is a novel research direction, our aim is to assess the feasibility of the approach as a basis for future work, as we proposed in previous work [15]. Our results indicate that it is possible to group situations into subspaces by using domain knowledge and insights from the data, and that situations from the same subspace tend to promote and demote the same personal values.

The rest of this paper is structured as follows: In Section 2 we present a high level architecture of our approach, and compare it to related work. In Section 3 we motivate our research choices for the use of psychological characteristics to group situations into subspaces, and we provide a short introduction to the concept of personal values. In Section 4 we present the user study in which we gather data in order to build the method which we described in the architecture. We present and discuss the results in Section 5, showing that situation subspaces can promote or demote specific personal values. Section 6 concludes this paper.

2 AGENT ARCHITECTURE

We propose an architecture which explains how a support agent can use information about the psychological characteristics of situations

[1] TU Delft, The Netherlands, email: i.kola@tudelft.nl

[2] TU Delft and Leiden Institute of Advanced Computer Science, The Netherlands, email: c.m.jonker@tudelft.nl

[3] TU Delft, The Netherlands, email: m.l.tielman@tudelft.nl

[4] University of Twente, The Netherlands, email: m.b.vanriemsdijk@utwente.nl

in order to determine the promoted or demoted personal values, and in turn offer support to the user. The architecture (Figure 1[1]) depicts two main components: a learning component in which we use data gathered from people to identify situation subspaces, and a support agent which uses this information to provide support to the user.

In the first component, participants of a user study describe situations from their lives and provide us with the psychological characteristics as well as the promoted and demoted values of these situations (Section 4). We use these psychological characteristics together with domain knowledge in order to form situation subspaces (Section 5.2). Then, we determine the promoted or demoted values for each situation subspace (Section 5.3). When the support agent is interacting with the user, once presented with a new situation, the agent uses the subspace rules to classify the situation to a subspace, as done in Context Space Theory [21]. By knowing the subspace values, the agent can reason about the promoted or demoted values of that specific situation. This information, in combination with the value preferences of the user, can be used in order to reason about support. This last part is not tackled in this work, but is displayed in the architecture in order to make the bigger picture clear.

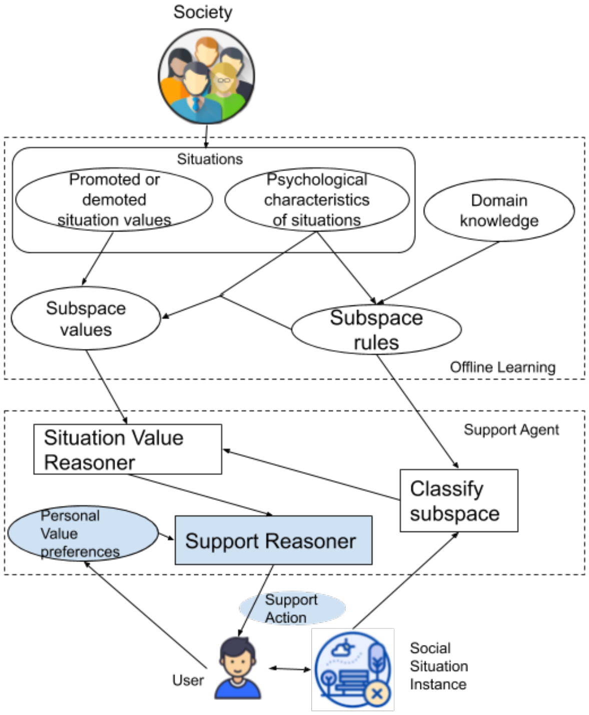


Figure 1. High-level architecture of the approach. Concepts shaded in blue represent aspects which we do not directly tackle in this work (i.e. Modeling the user preferences, extracting psychological characteristics of situations, and reasoning about the type of support). Circles represent knowledge elements (e.g. personal value scores, subspace rules), whereas squares represent reasoning steps. Arrows indicate the workflow of the approach.

This approach would allow support agents to align their suggestions with the personal values of their users. Let us consider an agent that recommends free time activities to the user, and the options are going to a party and attending a workshop to learn a new skill. Following the architecture depicted in Figure 1, the agent might infer that the first would promote pleasure, and the second would promote capability. This way, the agent would suggest going to a party to a user who prefers the value pleasure, and attending the workshop to a user who prefers the value capability.

Related work Other work also focuses on using concepts such as personal values and context in socio-technical systems, in order to enable them to understand and adapt to human motivations. We introduce some of these approaches in order to position our work. Tielman et al [32] propose an approach to derive norms from a combination of values, context and actions. Context is used as a modifier to determine how much a value is promoted or demoted when performing a certain action, and this information is elicited from the user. Context is not modelled explicitly, and can be represented by a list of variables, depending on the situation. Similarly, Cranefield et al. [6] propose an approach on how to use values in order to help users with moral decisions. The work focuses on the reasoning about aligning the values of the user to the values that are promoted or demoted by different actions. Similarly, the values and context are assumed to be predetermined. Our work focuses on the other point of view: how to actually infer what values are promoted or demoted in a given situation? In a way, our work can be considered an extension of these approaches, since the output of our work can be used as an input for these reasoning frameworks. Kayal et al [13] also take a step in this direction. In their work, they ask participants about their personal values and about the promoted and demoted values of different social commitments. They then use this information to break ties when different commitments overlap. Our work aims at taking this a step further, since we present a procedure that automatically reasons about the promoted or demoted values of a situation, rather than always having to ask the user. Other work (e.g. [17, 14]) describes the relation between the environment and the people in terms of contextual affordances, which represent potential actions that the environment (or parts of it) allows people to perform. For instance, a chair allows the action “sit”. This is in principle similar to what we are doing, since personal values can be seen as affordances of a situation, since some situations allow people fulfill specific personal values. For example, a situation in which a person is exercising would help them fulfill the personal value of health.

3 SITUATIONS AND PERSONAL VALUES

3.1 Situation Subspaces

Research in computer science uses terms such as situation awareness (e.g. [8]) and context awareness (e.g. [1]) to describe attempts to enable artificial agents to better understand their surrounding environment. According to Barwise [3], these concepts refer to the same thing, and situations represent a way of modelling contexts. Other researchers (e.g. [2]) see context as a lower level of abstraction, and situations can be seen as “logically aggregated pieces of context”. In Endsley’s situation awareness framework [8], the aforementioned interpretation of context would refer to the situation cues in the perception level of situation awareness. There is vast research on modelling and reasoning about context and situations, and describing this research in depth is beyond the purview of this paper. For a detailed account, readers can refer to [4, 33]. In this section, we will introduce possible approaches on how to use context elements in order to determine the promoted and demoted values of a situation.

[1] Icons used in the architecture were made by Freepik and retrieved from www.flaticon.com

Our proposed approach is to first group similar situations into so-called situation subspaces, and then to determine the promoted and demoted values of that subspace. This is inspired by work on Context Space Theory [21]. In their approach, context is represented as an object in a multidimensional Euclidean space, called situation subspace. A context state is represented in terms of attributes, and each dimension of the situation subspace represents an accepted region for a specific context attribute. This way, when given a set of attributes that define a context state, we can infer whether the state is or is not part of the situation subspace. For example, a situation subspace can be "Person is healthy" and its attributes are "Body temperature" with an accepted region of values between 36.0 and 37.5 and "Resting heart rate" with an accepted region of values between 60 and 100. In our approach, we consider a situation subspace to be the set of situations having similar psychological characteristics (Section 3.2). For instance, a subspace can consist of situations where the characteristics Duty and Intellect have a value between 4 and 7.

Using situation subspaces facilitates the process of explaining the suggestions of the support agent to the user, since each subspace is defined by a set of attributes. The reasoning is explicit: for instance, situation X is in subspace A because of attributes B and C, and situations in subspace A promote values Y and Z. These steps can be available to the user. Furthermore, this way of approaching situations is also in line with work on social psychology on how people actually deal with situations. Gigerenzer [11] suggests that people have different modules of interaction, and when presented with a new situation they "classify" it as part of one of the modules, and then follow the "interaction script" of that module.

One other option for reasoning about the values of a situation would be to look at the correlation of each individual psychological characteristic of the situation with specific personal values (e.g. as done by [24]). However, this approach does not take into account the possibility that the ways in which characteristics are combined in a situation also affect the values that are promoted or demoted in it. We will explore this possible connection in Section 5.3. In the current section we will simply give an intuition. For instance, situations with a high level of mating can in general affect the value pleasure, however it is the combination with high positivity or high negativity that affects whether the value is promoted or demoted. Furthermore, if we consider each psychological characteristic individually, it is not clear whether the low score of a characteristic indicates that a value is demoted or not affected. For instance, knowing that situations with high intellect promote capability is not enough to determine whether situations with low intellect demote this value or do not affect it. Our approach takes the potential effect that the combination of psychological characteristics have on personal values into account, but does not rely on it: if that effect does not hold, our approach would simply take into account the correlation between individual psychological characteristics and personal values.

Lastly, we can reason about personal values by training a model that takes as input the situation's psychological characteristics, and predicts the score for each value. This way, the model would actually take into account all the psychological characteristics of the situation and their potential interactions. Putting aside the requirement for high amounts of data and the non-trivial task of building such a model, our primary reason for not following this approach is its black box nature. We believe one of the key features of a support agent is its ability to explain its suggestions to the user. Such a comparison, and the potential trade-off between accuracy and explainability, is something that we plan to explore in future work.

3.2 Psychological Characteristics of Situations

Research in social psychology has explored ways in which situations can be systematically described. Rauthmann et al. [24] discuss three ways in which situational information can be taxonomized: Cues (e.g. persons, places, objects etc.); (psychological) Characteristics (which attributes can be used to describe situations - e.g. positivity, intellect, duty etc.); Classes (which kind of situations are there - e.g. social situations, work situations etc.).

In this work we focus on the use of psychological characteristics. There are several taxonomies of situations on the psychological characteristics level. We choose the DIAMONDS taxonomy since it covers a wide variety of daily life activities and it provides a validated 24-items survey which allows the measurement of the psychological characteristics of situations through online surveys. Horstmann et al [12] suggest that the dimensions of the existing taxonomies have a high level of similarity when compared across taxonomies, so our choice should not influence the outcome of the work. The DIAMONDS taxonomy describes situations in terms of the following dimensions:

- **Duty** - situations where a job has to be done, minor details are important, and rational thinking is called for;
- **Intellect** - situations that afford an opportunity to demonstrate intellectual capacity;
- **Adversity** - situations where you or someone else are (potentially) being criticized, blamed, or under threat;
- **Mating** - situations where potential romantic partners are present, and physical attractiveness is relevant;
- **Positivity** - playful and enjoyable situations, which are simple and clear-cut;
- **Negativity** - stressful, frustrating, and anxiety-inducing situations;
- **Deception** - situations where someone might be deceitful. These situations may cause feelings of hostility;
- **Sociality** - situations where social interaction is possible, and close personal relationships are present or have the potential to develop.

There are different reasons for using the psychological characteristics of situations in order to group them. First of all, psychological characteristics allow us to assess similarities between situations beyond their physical cues (e.g. where is the situation taking place, how many people are involved). Social psychology (e.g. [5, 7, 23, 30]) suggests that people think about situations by using their psychological characteristics. They create impressions of situations as if they were real, coherent entities. These impressions allow people to better navigate through the world by being able to predict what will happen and coordinate behaviour accordingly. This inherent psychological component of situations makes them difficult to interpret only in terms of physical context. For instance, let us consider a scenario where our user, Alice, is going out with friends. The relevant physical attributes would be the activity (i.e. going out), the location, time etc. A support agent might determine that such situation promotes pleasure. On the other hand, it is also possible that at some point Alice is going out and some people that she dislikes join. In that case, the situation could actually demote the value pleasure. However, from the point of view of physical cues, everything would remain the same and this difference would not be captured. Kola et al. [16] propose a set of social cues that can be used to capture such differences, for example the quality of the relationship with the other person or the level of trust. However, despite capturing the psychological component of situations, these social cues remain a low-level description.

Another advantage of focusing on the psychological characteristics is easiness of explainability. This means the support agent can explain its suggestions to the user in a way that is understandable and intuitive to people. To continue the previous scenario, we assume our support agent wants to propose an activity which promotes the value of pleasure to Alice, since this value is important to her. It would be more intuitive for Alice to understand that the situation “going out with friends” promotes pleasure because it has high positivity and low adversity, rather than because it is an activity that takes place after 8pm, at a bar, and a certain amount of people are present.

Focusing on the psychological characteristics of situations allows us to identify similarities in situations that look very different. For instance, a situation in which a parent is helping their child with a school project and a situation in which that same parent has an important work meeting do not have anything in common when it comes to physical cues, however they both potentially involve a high level of duty and intellect, and promote values such as helpfulness and capability. This also brings forward practical considerations from a technical point of view: there can be a very high number of physical cues that can be measured, and what is actually relevant differs from situation to situation. Furthermore, highly general concepts such as “activity” are difficult to model in a way which actually makes them comparable from a situation to another. For these reasons, deciding which elements to model and how to do it is both crucial and non-trivial. Our approach allows us to abstract from the physical context, which results in a low dimensionality of characteristics that are proven to be relevant across daily situations [24].

There is some difference in terminology when comparing Context Space Theory with DIAMONDS. A context state from Context State Theory is simply referred to as a situation in DIAMONDS, and context attributes would be represented by the situation dimensions. In this work, we will use the DIAMONDS terminology.

3.3 Personal Values

Values represent key drivers of human decision making (e.g. [26, 27]). Friedman and colleagues [9] define values as “what a person or group of people consider important in life”. People hold various values (e.g. wealth, health, independence) with different degrees of importance. The main features of personal values which make them relevant to our work have been explicitly described by Schwartz [29], but are also implicitly present in other work on values. First of all, values refer to desirable goals that motivate action, and they serve as standards to guide the selection of actions, people, or events. This means that (unconsciously) people’s decisions are influenced by values. Secondly, values transcend specific actions and situations. For instance, values such as honesty are important to someone regardless of the activity they are doing or who they are with. Lastly, what puts this all together is the fact that in order for values to influence action not only should they be important to the actor, but they should also be relevant in that specific context. This suggests that if we know which values are likely to be activated in a certain context (or situation) and have information about the value preferences of a user, we can use that information to evaluate how much does a situation promote or demote personal values that are important to the user. It is also important to notice that in this work, we talk about personal values on three different levels:

- Personal values are important to an individual - e.g. Alice values achievement;
- A specific situation can promote or demote personal values to someone in the situation - e.g. Being a speaker at a conference promotes the value achievement for Alice;
- A situation from a certain subspace usually enables promoting or demoting a personal value to someone in the situation - e.g. Being part of situations with high intellect and high duty usually promotes the value achievement for people.

The most prominent models of human values were proposed by Rokeach [26] and Schwartz [27]. These models are universal and domain-independent, making them suitable for our purpose, in which we will deal with a wide range of every day situations. This is different from other approaches where the first step was to find a subset of values which are more applicable to a certain domain, for instance mobile location sharing [13] or music recommendations [19]. In our work we use the model proposed by Schwartz since it offers validated measurement instruments with fewer items than Rokeach, which makes them more applicable to online surveys. Furthermore, it is to be noted that Schwartz builds on the work of Rokeach and other researchers, so there is overlap in their proposed value lists. The Schwartz theory of basic human values [27] recognizes 10 universal value groups, namely: Self-direction, Stimulation, Hedonism, Achievement, Power, Security, Conformity, Tradition, Benevolence and Universalism. Each of these value groups includes more “specific” values, as depicted in Table 2.

4 USER STUDY

In this user study we gather data[2] for constructing and evaluating our methods. The study consists of three parts: first, participants were asked to describe situations from their daily lives (part 1), then they had to answer questions about the psychological characteristics of the situations (part 2) and finally they had to answer questions about how much the situations promote or demote certain personal values (part 3). The study was approved by the ethics committee of TU Delft.

Participants We collected answers from 150 participants recruited in the crowd-sourcing platform Prolific Academic[3]. Using a crowd-sourcing platform allowed us to efficiently obtain a large sample size in a short amount of time. Respondents received a monetary compensation for the time they spent, as per the platform policies. The average age of participants was 32.38 (SD=12.1). 51.3% were female, 44% male, and 4.7% selected the option “other” when asked about their gender.

Procedure [4] In order to have enough data to evaluate whether clustering situations is useful, it is important that we use a method that generates a diverse sample of situations. To this end, we use a method applied in other research that asks participants to describe a situation in their daily lives (e.g. [10, 25]). This retrospective procedure was shown to encourage participants to report on a wide range of situations. We asked participants to think about two situations which occurred during the past weeks which involved one other person, since our focus is on social situations. We specifically asked for situations involving only one other person, since if needed it is possible to control the effect of the relationship with the other person on the situation. However, the approach would work the same way for situations involving multiple other people. We instructed participants to think

[2] The data can be accessed under: https://doi.org/10.4121/12867041

[3] https://www.prolific.co/

[4] The survey questions can be found in Appendix A

of situations where a concrete activity took place, and not situations such as “I saw someone in the street and said hello”. A positive example was not given in order to avoid priming the participants towards certain situation types. Participants were asked to describe the situations in 3-4 sentences and to focus on describing the activity, their relation to the other person, as well as how each person behaved in the situation. Furthermore, we instructed participants to try to think of diverse situations, which involved different people and where different activities took place. To check for consistency, participants had to answer four open questions about the situation they just described: when did the situation take place, what was the main activity, where did the situation occur, and what is the role of the other person.

In the second part of the study, participants were presented with a set of statements to measure psychological characteristics of situations, and they were asked how much each statement applies to each of the situations that they had just described. Examples of statements were “A job needs to be done”, “Task-oriented thinking is required” etc. The statements were taken from the S8* scale proposed by Rauthmann and Sherman [25]. This is a validated instrument which can be used to measure the DIAMONDS dimensions of a situation. Each dimension is represented by three statements, for an overall total of 24 statements. Participants could indicate their answers on a scale ranging from 1 (not at all) to 7 (totally).

In the last part, participants were presented with a list of personal values, and they were asked on a slider with values from -10 (fully demote) to 10 (fully promote), how much is each value promoted or demoted in each of their two situation. Participants were presented with 21 personal values, which are based on a version of the Schwartz Value Survey [27] which was used on the European Social Survey [28]. Each of the universal value groups is represented by two values, apart from Universalism which is represented by three. In the original survey, each item of the list describes a feature that a person might exhibit (e.g. “She seeks every chance she can to have fun. It is important to her to do things that give her pleasure.”), which correspond to a personal value (e.g. “pleasure”). This was done because the aim of the European Social Survey was to explore personal values that people find important, and for that purpose framing values as features of a person was useful. In this study, we want to know how much a value is promoted or demoted in a certain situation, therefore framing values as qualities of a person would not work. For this reason, we presented participants with the underlying value of each item on the list. The only change that was made to the list was to replace the value “National security” with the value “Health”, which is also a value from the Security value group. The reason for this is that we believe it is common for people to commonly encounter situations that can affect their health (e.g. sports, choice of food), but we do not expect them to encounter situations that affect national security.

5 RESULTS AND DISCUSSION

5.1 Variety of Situations

Participants reported situations involving a wide range of other people, including a friend (24%), a family member (20%), a co-worker or supervisor (17%), a romantic partner (12%), an acquaintance (3%) or other (24%, mostly consisting of strangers). These situations comprised a high variety of activities, ranging from work meetings to dinner dates, from sport activities to discussions with other drivers, and everything in between. This is also shown by the high variety of the ratings that participants gave to the psychological characteristics of these situations. The rating for each dimension was calculated as the average score that the participant gave to the three statements representing that dimension, following the guidelines of the S8* measurement scale that we are adopting [25]. As seen in Figure 2, most of the dimensions have ratings across the whole range of possible alternatives, apart from Adversity and Mating which tend to have a more confined distribution and less variety in general. The score for each dimension is calculated as the average score across the three statements of the questionnaire that define that dimension. We provide a detailed distribution of answers for each psychological characteristic in Figure 2, since this insight will be used to form the subspaces in Section 5.2.

When it comes to personal values that are afforded in these situations according to the participants, the scores also have high variety, as depicted in the distribution presented in Figure 3. This distribution suggests that that values were differently promoted or demoted across situations. However, it also holds that most values were slightly promoted on average (overall mean=1.24, SD=4.68). This is in line with research on personal values [26] which views them as positive concepts.

5.2 Forming Situation Subspaces

In this subsection, we will group situations according to their psychological characteristics into situation subspaces. We will try an automatic approach, as well as one based on domain knowledge and insights from the data.

5.2.1 Automatic Clustering

The most straightforward way to form the situation subspaces is by using a clustering algorithm. We tried state of the art algorithms such as K-Means, Affinity Propagation and Agglomerative Clustering using different parameters. The algorithm would receive as an input the psychological characteristics scores of each situation, and return the cluster to which that situation should belong. We evaluated them with standard metrics used in cases where there is no ground truth when it comes to cluster memberships, such as the Silhouette coefficient and the Davies-Bouldin Index. We used the implementations from the scikit-learn package [22] in Python. The best configuration was achieved by the K-Means algorithm with two clusters, which achieved a Silhouette score of 2.4, and a Davies-Bouldin index of 1.59. These metrics suggest that the data is not well separable when we use all the dimensions in order to perform the clustering. This was to some extent to be expected, considering the high variety of situations, and the fact that there are 8 dimensions and only 300 situations in total. In future work we will collect more situations and explore whether that leads to a higher number of similar situations in the dataset, which could potentially lead to better defined clusters.

While exploring the scores of the dimensions in these two clusters, we notice that in the first cluster Positivity and Mating have a higher score than the average and the other six dimensions have a lower score. In the second cluster this trend is inverted. However, we also notice that each cluster contains situations with scores across the full range of scores for each of the dimensions. First of all, this suggests that these clusters are difficult to interpret/explain since they do not have clear distinguishing features. Secondly, in order to be able to use the Context Space Theory framework, attributes need to have a defined range, which means for at least some of the dimensions we need to have a cutting threshold. This is not the case for the formed clusters, and when faced with a new situation, it is not trivial

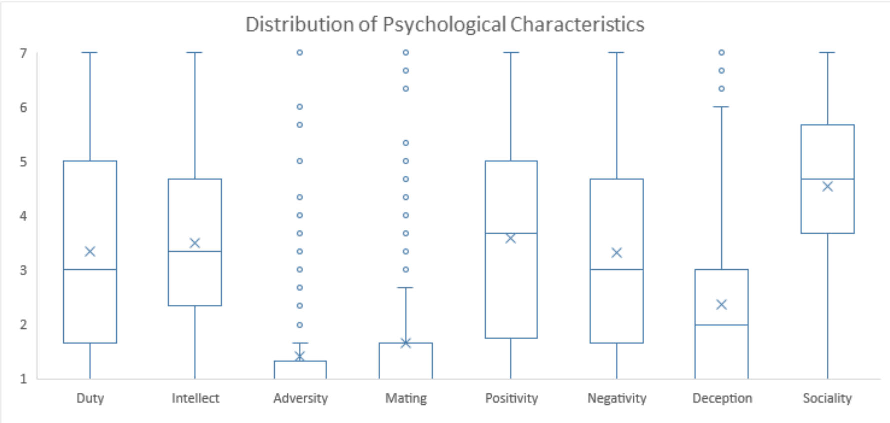


Figure 2. Distribution of scores across situations for each dimension, expressing the variety of situations from a point of view of their psychological characteristics. For each boxplot, the middle line represents the median, the sides of the boxes represent the first and third quartiles, and the whiskers represent the minimum and maximum values without considering outliers (which are represented by round points). The x represents the mean scores of the dimensions.

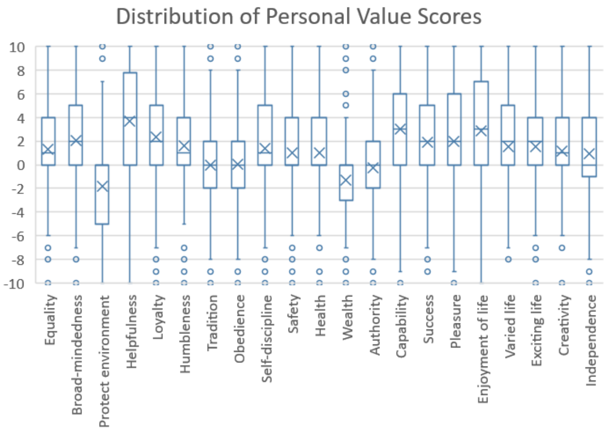


Figure 3. Distribution of scores for each personal value across situations.

to determine to which cluster it belongs. Overall, we notice that performing automatic clustering on our data leads to clusters consisting of situations which share some similarity in terms of psychological characteristics, but the division is not granular enough.

5.2.2 *Using Data Insight and Domain Knowledge*

The next approach will be to use insights from the data as well as domain knowledge in order to manually group situations into situation subspaces. It is important to notice that by "data insights" we only refer to the scores given to the situation dimensions, and not the scores assigned to personal values. From the previous subsection, we learn that trying to cluster over all dimensions is not effective because of the low amount of data and its high variety. For this reason, we use less dimensions in order to define each situation subspace. In order to identify these dimensions, first of all we explore the data. In Figure 2 we notice that the dimensions which bring the highest variety to the data are positivity, negativity, intellect and duty. This makes combinations of these dimensions suitable for defining the situation subspaces, since their scores have a high range, and the combinations would lead to subspaces with similar numbers of situations in them. Another insight from the data is that adversity has a very low variety, which makes the situations with a high adversity to form a particular group when compared to the rest. The same applies to mating, but adversity serves the purpose more since it contains outliers. Domain knowledge about the nature of these dimensions can also inform the process of selecting dimensions used to define subspaces. Positivity and negativity, despite being independent concepts, have an inherently opposite flavor. On the other hand, negativity has similar connotations with deception. This is also confirmed by the Pearson correlation coefficients between the data (positivity-negativity: -0.56, negativity-deception: 0.37). This information was used to define six situation subspaces:

- Subspace 1 - High Duty, High Intellect, Low Adversity;
- Subspace 2 - High Positivity, Low Duty, Low Intellect;
- Subspace 3 - High Duty, Low Intellect;
- Subspace 4 - High Adversity;
- Subspace 5 - High Negativity, Low Positivity, Low Duty, Low Intellect, Low Adversity;
- Subspace 6 - High Intellect, Low Duty.

The description "High" refers to scores between 4-7, while the description "Low" refers to scores between 1-3.99 (non-integer scores are possible since each dimension is calculated as the mean of three items from the survey). That means the dimension is highly or lowly characteristic of situations in that subspace. These subspaces allow us to classify 262 out of the 300 situations in our data set. When exploring the remaining situations, we notice that all dimensions other than sociality have a low score. For this reason, we use sociality as a dimension to define the final split, thus forming the last two subspaces:

- Subspace 7 - Low Sociality, and all other dimensions also Low;
- Subspace 8 - High Sociality, and all other dimensions Low.

These subspaces are designed to work well with the Context Space Theory framework, since each of them is defined by a set of attributes with specific values. This allows for a straightforward way for classifying a new situation to a subspace. Figure 4 provides a visualisation of this, by depicting four of the subspaces projected onto their defining dimensions, for illustration purposes. These defining dimensions enable the subspaces to be more interpretable and explainable in terms of the psychological characteristics that apply to their situations, when compared to the automatic clusters that were created.

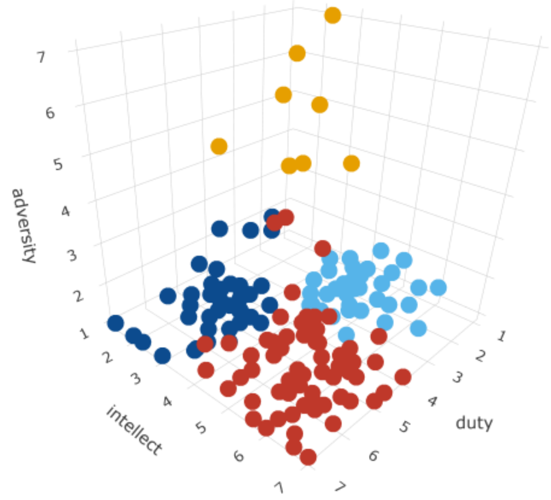


Figure 4. Visualisation of four situation subspaces defined by Adversity, Intellect and Duty. Red dots represent situations from Subsp. 1, dark blue dots represent situations from Subs. 3, orange dots represent situations from Subsp. 4, and light blue dots represent situations from Subsp. 6.

We notice that the subspaces are not strictly disjoint. However, this is not a restriction from Context Space Theory, where our approach is based. This also works on an intuitive level, since situations are fluid concepts which can be "in between" two different subspaces. In future work, we will work on strategies on how to break possible ties. Padovitz et al. [21] propose using optional attributes which would increase the probability of a situation being in a subspace.

Using intrinsic metrics for evaluating clusters like we did for the automatic clusters (Silhouette score, Davis-Bouldin Index) would heavily penalize the manual subspaces, since these scores apply to all eight dimensions, whereas the subspaces were defined on a smaller subset of dimensions. For example, in Figure 4 we see that the subspaces would be well separated if we only consider the dimensions on which they were defined. In future work it will be important to define evaluation metrics for manually created subspaces.

We notice a high diversity of activities taking place in the situations of each subspace. For example, Subspace 1 (defined by high duty, high intellect and low adversity), comprises, apart from work situations, also activities such as going to a suture course with a friend, or discussing the family finances with the partner. Similarly, Subspace 4 (defined by high adversity) includes situations ranging from someone being accused of cheating in a card game, to someone being lectured from the CEO of the company. This supports our initial premise that analysing the psychological characteristics of situations can point out to similarities between situations that seem very different at first sight. A similar variety is also present when it comes to the role of the other person in the situation. In our setup, roles are mutually exclusive. The distributions are depicted in Table 1. As we can see, in each subspace there are people from almost all the roles present. As expected, Subspace 1 (situations with high intellect and duty) include more colleagues, and Subspace 2 (situations with high positivity, low duty and low intellect) include more family and friends, and less colleagues. This aspect will be analysed further in future work.

	Fam	Rom	Fr	Coll	Gr	Other
Subspace 1 (n=74)	12.5	9.72	16.67	37.5	4.17	19.44
Subspace 2 (n=77)	23.08	14.1	34.62	8.97	1.28	17.95
Subspace 3 (n=44)	20.45	9.09	20.45	11.36	4.55	34.09
Subspace 4 (n=10)	12.5	0	12.5	25	25	25
Subspace 5 (n=19)	45	15	15	5	5	15
Subspace 6 (n=40)	12.5	12.5	35	15	0	25
Subspace 7 (n=24)	18.52	18.52	11.11	7.41	3.7	40.74
Subspace 8 (n=12)	36.36	9.09	27.27	9.09	0	18.18
All situations (n=300)	20	12	24	17	3.33	23.67

Table 1. Distribution of the other person's roles in the situations of each subspace (in percentage). n represents the number of situations (and therefore, the number of people, since situations involve the user and one other person) in each subspace. Fam = Family Member, Rom = Romantic Partner, Fr = Friend, Coll = Colleague, Gr = Group Member

5.3 Promoted and Demoted Personal Values

In this section, we explore whether specific values tend to be more promoted or demoted across situation subspaces. We look at this from two points of view. First of all, we take into consideration statistical significance. For this, we perform the Wilcoxon rank-sum test to check whether the scores of each value in the situations of a subspace are significantly different from the ones in the rest of situations. Secondly, we look at the mean scores. We consider that a subspace strongly promotes a value when the mean score of the values in its situations is higher than 3.5, and it strongly demotes a value when the mean score is lower than -2.5. Demoting has a lower threshold since we notice that participants tend to give slightly more positive scores overall (the overall mean is 1.24). Despite the distributions not being strictly normal, we believe the mean can be informative since the scale is limited between -10 and 10 so there are no values that can greatly skew it. We also calculated the median, and there is a very high overlap in the values that fulfill the criteria (22 out of 26). We do not report the medians for space purposes. We perform this analysis for the automatically created clusters, as well as for our manually formed subspaces.

When it comes to the automatically created clusters, we notice that the first one significantly promotes the values pleasure (3.87) and enjoyment of life (4.87), whereas the second cluster significantly promotes the value capability (4.08). No values are significantly demoted in either cluster. We do not report all values for space purposes. When comparing these results to the interpretation of the clusters using the psychological characteristics of situations, it seems intuitive that the cluster with higher positivity and mating promotes pleasure and enjoyment of life, whereas the cluster with higher duty and intellect promotes capability. The divisions are not granular enough to help us determine a larger number of promoted and demoted values, since we have only two clusters which consist of diverse situations. However, this analysis hints towards the idea that subsets of the data which share similar psychological characteristics do tend to promote certain values more than others, when compared to the overall data.

Next, we perform the same analysis for our manually crafted situation subspaces (Table 2). We notice that 5 of the subspaces significantly promote or demote some personal values, thus supporting our initial hypothesis. By analysing these results further, we notice that they are also aligned with the common sense understanding of these concepts: values such as pleasure and enjoyment of life are promoted in situations defined by high positivity (Subspace 2) and demoted in situations defined by high adversity (Subspace 4). Moreover, situations defined by high intellect and duty promote values such as help-

Table 2. Average score for each value in each cluster as well as the full data set. Scores in bold mean that the value is promoted or demoted in that cluster, with boundaries at $<$-2.5 for demoting and $>$3.5 for promoting. Scores marked with * suggest statistical significance with $p<0.05$ when performing the unpaired Wilcoxon rank-sum test for the cluster vs. the rest of the data.

Value (value group)	Subsp 1	Subsp 2	Subs 3	Subs 4	Subsp 5	Subsp 6	Subsp 7	Subsp 8	All Situations
Equality (Universalism)	2.2	1.72	1.11	-1.5*	**-2.63***	2.03	0.96	1.82	1.32
Broad-mindedness (Universalism)	**3.5***	1.74	1.07*	-0.5*	1	**3.98***	-0.37*	1.36	2.07
Protect environment (Universalism)	-2.04	-2.7	-0.95	-1.88	-2.37	-1.25	-0.52	-1.36	-1.79
Helpfulness (Benevolence)	**5.58***	2.5*	**4.41**	-1.5*	0.63*	3.48	2.89	**6.18**	**3.66**
Loyalty (Benevolence)	3.07	3.26	1.8	-2.38*	0.37*	2.78	0.33*	3.45	2.33
Humbleness (Tradition)	2.47	2.05	1.34	-0.75*	-0.63	1.68	1.07	1.09	1.64
Tradition (Tradition)	0.45	-0.09	0.25	-0.88	**-3.05***	0.85	-0.93	1.45	-0.04
Obedience (Conformity)	1.49*	-0.79	0.52	**-2.63**	2	-1.15	-1.11	0.55	0.05
Self-discipline (Conformity)	**3.68***	-1.18*	2.82*	1.5	1.68	1.18	1.33	1	1.39
Safety (Security)	1.95	-0.21*	2.3	**-3.88***	0.11	1.4	0.78	2.36	1.02
Health (Security)	1.18	0.2*	1.89	-1	-0.32	1.8	1.33	**3.91**	1.01
Wealth (Power)	-0.89	-1.55	-1.32	-0.88	-1.26	-1.48	-1.63	-0.09	-1.28
Authority (Power)	1.27*	-1.86*	1.34*	-1	-1.47	-0.48	-1.3	1	-0.24
Capability (Achievement)	**5.45***	1.78*	**3.86**	1	0.74*	3.15	1.11*	2.09	2.99
Success (Achievement)	**4.04***	1.29	2.55	0.63	-1.63*	1.83	1.19	0.82	1.93
Pleasure (Hedonism)	1.15	**5.76***	-0.77*	**-3.5***	**-3.63***	**4.55***	0.3	0.18	1.94
Enjoyment of life (Hedonism)	1.93	**6.82***	0.02*	**-3.25***	-0.63*	**4.73***	1.15*	2.45	2.9
A varied life (Stimulation)	1.7	2.62*	1.5	-0.63	-0.05	2.33	-1.04*	1.82	1.56
An exciting life (Stimulation)	0.85	**4.01***	0*	-1.38*	-0.05	2.58	-0.26*	-0.18*	1.5
Creativity (Self-direction)	2.68*	1.54	0.39	-1.13	**-2.74***	2.15	-0.96	1.64	1.18
Independence (Self-direction)	2.39*	0.33	1.39	-0.88	-1.53	0.65	0.37	1.64	0.91

fulness, capability and success. These intuitive connections suggest that a support agent that uses this method would have the possibility to explain its suggestions to the users in an understandable way. Furthermore, it seems like the promoted or demoted values are affected by the combination of dimensions, rather than by each dimension individually. For instance, situations defined by both high intellect and duty (Subspace 1) significantly promote success and helpfulness, whereas situations defined by high duty and low intellect (Subspace 3) or low duty and high intellect (Subspace 6) do not promote these values.

6 CONCLUSION

6.1 Contributions

In this work we present an approach in which we group situations into subspaces by using their psychological characteristics as attributes, and show that these subspaces can be used to determine which personal values are promoted or demoted in these situations. In order to explore our research question, we use automatic clustering, as well as insights from the data combined with domain knowledge, in order to group situations according to their psychological characteristics. We notice that automatic methods lead to clusters which are not well defined, while the manual method allowed us to form groups that fit the requirements of Context Space Theory.

Secondly, we show that certain personal values are significantly more promoted or demoted in specific situation subspaces, thus confirming our research hypothesis. This can be used as a method to automatically determine how the situation that a user faces affects the personal values of the user. This would be a useful extension for current support agents [6, 32] that rely only on information from the users to know the effect it has on personal values.

An advantage of this approach is its potential for providing explainable support to the user. Our methods are inherently more explainable than black box approaches, and we borrow the attributes that form the basis of our approach from social psychology. Concepts such as the psychological characteristics or personal values are potentially more understandable for users.

6.2 Limitations and Future Work

Considering that the work is still in its early stage, there are limitations which we aim to tackle in the future. First of all, we assume that we already know the psychological characteristics of a situation. This is not a trivial task, and in order to have a supportive agent that can help in real life cases, these characteristics will have to be inferred from situation cues. Work from Kola et al. [16] provides initial evidence that they can be used to infer concepts such as the priority of situations. In the future, we will explore whether that approach can be applied to the psychological characteristics of situations.

Secondly, we detect more affected values in the manually defined situation subspaces. While this approach is not necessarily weaker than an automatic approach, it has to be tested with a wider range of situations. The reason for this is that it was crafted particularly for this set of situations, so its effectiveness for another set of situations is to be determined. In the future, we will work on having a well-defined formal procedure on how to form situation subspaces by using the psychological characteristics of situations as context attributes. Another option will be to explore forming automatic clusters by considering a subset of the dimensions.

Next, the promoted and demoted values need to be analysed further. We notice three of the subspaces do not promote or demote any personal values, and some personal values are neither promoted nor demoted in any subspace. In future work, we will explore using a more specific list of values which are salient to daily life situations. Lastly, we will explore whether situation subspaces can help determine concepts other than personal values, such as expected behaviour.

ACKNOWLEDGEMENTS

This work is part of the research programme CoreSAEP, with project number 639.022.416, which is financed by the Netherlands Organisation for Scientific Research (NWO).

REFERENCES

[1] Varol Akman and Mehmet Surav, 'Steps toward formalizing context', *AI magazine*, **17**(3), 55, (1996).

[2] Christos B Anagnostopoulos, Yiorgos Ntarladimas, and Stathes Hadjiefthymiades, 'Situational computing: An innovative architecture with imprecise reasoning', *Journal of Systems and Software*, **80**(12), 1993–2014, (2007).

[3] Jon Barwise, 'Situations and Small Worlds', *Handbook of Semantics*, (1987).

[4] Claudio Bettini, Oliver Brdiczka, Karen Henricksen, Jadwiga Indulska, Daniela Nicklas, Anand Ranganathan, and Daniele Riboni, 'A survey of context modelling and reasoning techniques', *Pervasive and Mobile Computing*, **6**(2), 161–180, (2010).

[5] Nancy Cantor, Walter Mischel, and Judith C Schwartz, 'A prototype analysis of psychological situations', *Cognitive psychology*, **14**(1), 45–77, (1982).

[6] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum, 'No pizza for you: Value-based plan selection in bdi agents.', in *IJCAI*, pp. 178–184, (2017).

[7] John A Edwards and Angela Templeton, 'The structure of perceived qualities of situations', *European journal of social psychology*, **35**(6), 705–723, (2005).

[8] Mica R Endsley, 'Toward a theory of situation awareness in dynamic systems', *Human Factors*, **37**(1), 32–64, (1995).

[9] Batya Friedman, Peter H Kahn, and Alan Borning, 'Value sensitive design and information systems', *The handbook of information and computer ethics*, 69–101, (2008).

[10] Fabiola H Gerpott, Daniel Balliet, Simon Columbus, Catherine Molho, and Reinout E de Vries, 'How do people think about interdependence? a multidimensional model of subjective outcome interdependence.', *Journal of Personality and Social Psychology*, **115**(4), 716, (2018).

[11] Gerd Gigerenzer, '10 the modularity of social intelligence', *Machiavellian intelligence II: Extensions and evaluations*, **2**(264), 264–288, (1997).

[12] Kai T Horstmann, John F Rauthmann, and Ryne A Sherman, 'The measurement of situational influences', *The SAGE handbook of personality and individual differences*, (2017).

[13] Alex Kayal, Willem-Paul Brinkman, Rianne Gouman, Mark A Neerincx, and M Birna Van Riemsdijk, 'A value-centric model to ground norms and requirements for epartners of children', in *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*, pp. 329–345. Springer, (2013).

[14] Franziska Klügl, 'Using the affordance concept for model design in agent-based simulation', *Annals of Mathematics and Artificial Intelligence*, **78**(1), 21–44, (2016).

[15] Ilir Kola, Catholijn M Jonker, and M Birna van Riemsdijk, 'What does it take to create social awareness for support agents?', in *International Workshop on Engineering Multi-Agent Systems*. Springer, (2019).

[16] Ilir Kola, Catholijn M Jonker, and M Birna van Riemsdijk, 'Who's that?-social situation awareness for behaviour support agents', in *International Workshop on Engineering Multi-Agent Systems*, pp. 127–151. Springer, (2019).

[17] John Bruntse Larsen, Virginia Dignum, Jørgen Villadsen, and Frank Dignum, 'Querying social practices in hospital context', in *10th International Conference on Agents and Artificial Intelligence*, pp. 405–412. SCITEPRESS Digital Library, (2018).

[18] Kurt Lewin, *Principles of topological psychology*, New York, NY: Mc-Graw Hill, 1936.

[19] Sandy Manolios, Alan Hanjalic, and Cynthia CS Liem, 'The influence of personal values on music taste: towards value-based music recommendations', in *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 501–505, (2019).

[20] Karen L Myers and Neil Yorke-Smith, 'Proactivity in an intentionally helpful personal assistive agent.', in *AAAI Spring Symposium: Intentions in Intelligent Systems*, pp. 34–37, (2007).

[21] Amir Padovitz, Seng Wai Loke, and Arkady Zaslavsky, 'Towards a theory of context spaces', in *IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second*, pp. 38–42. IEEE, (2004).

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, **12**, 2825–2830, (2011).

[23] Lawrence A Pervin, 'A free-response description approach to the analysis of person-situation interaction', *ETS Research Bulletin Series*, **1975**(2), i–26, (1975).

[24] John F Rauthmann, David Gallardo-Pujol, Esther M Guillaume, Elysia Todd, Christopher S Nave, Ryne A Sherman, Matthias Ziegler, Ashley Bell Jones, and David C Funder, 'The situational eight diamonds: A taxonomy of major dimensions of situation characteristics.', *Journal of Personality and Social Psychology*, **107**(4), 677, (2014).

[25] John F Rauthmann and Ryne A Sherman, 'Measuring the situational eight diamonds characteristics of situations: An optimization of the rsq-8 to the s8*.', *European Journal of Psychological Assessment*, **32**(2), 155, (2016).

[26] Milton Rokeach, *The nature of human values.*, Free press, 1973.

[27] Shalom H Schwartz, 'Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries', *Advances in experimental social psychology*, **25**(1), 1–65, (1992).

[28] Shalom H Schwartz, 'Human values', *European Social Survey Education Net*, (2005).

[29] Shalom H Schwartz, 'An overview of the schwartz theory of basic values', *Online readings in Psychology and Culture*, **2**(1), 2307–0919, (2012).

[30] Ryne A Sherman, Christopher S Nave, and David C Funder, 'Properties of persons and situations related to overall and distinctive personality-behavior congruence', *Journal of Research in Personality*, **46**(1), 87–101, (2012).

[31] Myrthe Tielman, Willem-Paul Brinkman, and Mark A Neerincx, 'Design guidelines for a virtual coach for post-traumatic stress disorder patients', in *International Conference on Intelligent Virtual Agents*, pp. 434–437. Springer, (2014).

[32] Myrthe L Tielman, Catholijn M Jonker, and M Birna van Riemsdijk, 'What should i do? deriving norms from actions, values and context', in *10th International Workshop on Modelling and Reasoning in Context*, (2018).

[33] Juan Ye, Simon Dobson, and Susan McKeever, 'Situation identification techniques in pervasive computing: A review', *Pervasive and mobile computing*, **8**(1), 36–66, (2012).

A User Study Survey

The survey can be accessed in the following link: `https://tudelft.fra1.qualtrics.com/jfe/form/SV_bsdYhzLjbJH64zX`

A.1 Part 1 - Collecting Situations

Introductory text: In this part, you will be asked to describe two situations involving you and one other person that occurred in your life during the previous weeks. Try to think of situations in which a concrete activity took place (e.g., not a situation such as ”I saw someone in the street and said hello”). Describe the situation in 3-4 sentences, and focus on describing the activity, your relation to the other person, as well as how each of you behaved in the situation. Think of concrete and specific situations that actually took place, and not of ”situation types”. Please, think of two diverse situations (i.e., they involved different people, and different activities took place). After describing the situations, you will be asked some general questions about them.

For each situation, the following questions were asked:

- Please describe the situation.

- When did the situation that you just described take place, approximately? (day and time)
- What was the main activity that took place in that situation?
- Where did the situation occur? Please do not give the exact address/name of the place, the type of place suffices (e.g. at a bar, in my office, etc.).
- What's the role of the other person that is present in the situation? (e.g. ”child” would suggest that that person is your child). options: {*partner, parent, sibling, child, friend, extended family member, neighbor, coworker, supervisor, member of the same group (e.g., sports team), other*}

A.2 Part 2 - Psychological Characteristics of Situations

For each situation, participants were presented with the text of their described situation, and for each situation they were asked:

“How much does each of these statements apply to the situation that you just described?”. options: {*Not at all, Very little, A little, Moderately, A lot, Very much, Totally*}

- A job needs to be done.
- I have to fulfill my duties.
- Task-oriented thinking is required.
- The situation contains intellectual stimuli.
- There is the opportunity to demonstrate intellectual capacities.
- Information needs to be deeply processed.
- I am being blamed for something.
- I am being criticized.
- I am being threatened by something or someone.
- The situation is sexually charged.
- Potential sexual or romantic partners are present.
- Physical attractiveness is relevant.
- The situation is joyous and exuberant.
- The situation is pleasant.
- The situation is playful.
- The situation could entail frustration.
- The situation could elicit stress.
- The situation could elicit feelings of tension.
- It is possible to deceive someone.
- Someone in this situation could be deceived.
- Not dealing with others in an honest way is possible.
- Communication with other people is important or desired.
- Close personal relationships are important or can develop.
- Others show many communicative signals.

A.3 Part 3 - Personal Values

Introductory text: Personal values represent things that can be important to you in life. Different situations can promote or demote some specific values. For example, skiing can promote values such as pleasure or having an exciting life, but on the other hand it can demote values such as safety, since there's always the chance of getting hurt. In the last part of this survey you will be presented with a list of values, and for each of them you will be asked to answer to what extent they would be promoted/demoted in the situations that you described in the first part of the survey.

For each situation, participants were presented with the text of their described situation, and for each situation they were asked:

To what extent does this situation promote/demote each of these values? options: slider from -10 (fully demote) to 10 (fully promote), where 0 is marked as ‘neither promote nor demote’.

- Equality;
- Broad-mindedness;
- Protecting the environment;
- Helpfulness;
- Loyalty;
- Humbleness;
- Respect for tradition;
- Obedience;
- Self-discipline;
- Safety;
- Health;
- Wealth;
- Authority;
- Capability;
- Success;
- Pleasure;
- Enjoyment of life;
- A varied life;
- An exciting life;
- Creativity;
- Independence.

Linking actions to value categories - a first step in categorization for easier value elicitation

Djoshua D. M. Moonen and **Myrthe L. Tielman** [1]

Abstract. Computer systems are increasingly involved in making decisions. Therefore, it is increasingly important that they understand our values. To make values usable, context is important, both of the individual and the actions they underlie. This work aims to study if it is possible to make it easier to elicit an individual's values by using the context of the action. Practically, we first held an expert survey (n = 7) to see if some values are more likely to underlie some actions than others. The results were positive on this score, so a second study (user, (n = 135)) was done showing that restricting the number of values made it easier to elicit values from users while not unnecessarily limiting their expression. This work shows that when linking actions to values, it is possible to make the elicitation easier by only showing the applicable options. This is an important step in being able to incorporate values in computerized decision making.

1 Introduction

Computer systems are increasingly helping us to make and stick to important decisions in life. Reminder systems, health apps and social-media blockers all function to help us change behavior in some way [5, 7]. However, such systems often blindly stick to a single goal, and do not truly understand the motivations behind our actions, nor the context in which we make our decisions. To help technology understand these motivations, values have been proposed [1]. Values represent the things we find important in life, and which guide our decisions [8]. Therefore, they have long been taken into account in system design [3]. However, to flexibly adapt to individual values, systems require values in the reasoning as well in the design. In recent years, a number of systems have attempted to model this reasoning by linking values to our choices in some way [2, 10]. Ideally, such work will lead to systems that can more flexibly adapt their decision making and take into account values in their reasoning [1].

Values are general, abstract concepts. However, for a system to use them, they need to be made concrete. They need to be linked to actions [10], or to choices [2]. Often, this is also done by transforming values into norms [3]. This concretization of values means that information needs to be added about the context in which they are applied. We identify two main types of context. Firstly, the individual needs to be taken into account, as people have different values, as well as different views on what a value means for them. Secondly, what type of choices or actions the value is applied to is relevant, values will take on different meanings in different domains.

The first type of context is the individual, which means that information about values should ideally come from them. The most obvious source for this information are the users themselves, but people have often not explicitly thought about values, or do not even fully understand the concept. Moreover, the conversational capabilities of many automated systems are not yet capable of this type of conversation. So this information is difficult for a system to obtain [6]. Therefore, most existing value-elicitation methods are based in human-human interaction [11], or are aimed at what values are important in general [9].

In order to make this elicitation of an individual's values easier, it is helpful to consider the second form of context, namely the action. Most systems have attempted to elicit values in general. But values can take on different meanings in different domains. For instance, safety might mean something different for choosing a car than for choosing a doctor. Similarly, the choice to go to work is motivated by different types of values than the choice to go to a party. This also means that we could use this type of context to narrow the conversation about values between a system and human.

If we want to know what value underlies a certain action for a specific individual, we could pose this as a question in which the user can pick from all possible values. However, this would mean a very large answer space. And as mentioned, the action probably also limits what values are most likely to underlie that choice. So it might be possible to use this context to limit the amount of possible values an individual has to pick from, for instance in the form of a pre-selection of the list of values. However, as we are interested in the individual's values, not just the most likely ones underlying a general action, it is also important to not limit the individual too much in what they can express by making this pre-selection too small. In this paper, we wish to explore whether it is possible to make elicitation easier in this way without limiting expression.

Thus, in this work we explore two things. Firstly, whether it is possible to make a pre-selection of values which are more likely to underlie a choice for a specific action. And secondly, whether a pre-selection like this makes it easier for users to select a value from a list while not limiting them in their expressive ability. In section 2 the first question is explored by means of an expert study. Section 3 explores the second question by means of a user study and 4 presents the results. A discussion and conclusion based on the findings can be found in section 5.

2 Value Categorization

In order to make value-selection easier, we propose to make a pre-selection based on the type of activity the value promotes. Our hypothesis is that different actions have different value types which often underlie them. For instance, the values which underlie people's choice to go to work are probably different from the one to watch a movie. In order to study whether such a pre-selection can be made and what it would be, an expert-study was performed. The goal of

[1] Delft University of Technology

this study was two-fold. Firstly, to see if there is agreement amongst experts in what categories of values are most likely to underlie the choice to perform a specific action. And secondly, if there is such agreement, what categories of values are most likely for what actions.

2.1 Participants

The study was conducted with 7 participants (71.4% male), recruited from research staff and PhD students of Delft University of Technology. All participants were familiar with or have worked on value-based topics. Average age was 33.4 (sd 7.2) and they had an average of 3.83 years (sd 4.41) of experience with value-based research.

2.2 Procedure

The participants were sent a survey along with instructions. The instructions defined value as used by Schwarz (1992) including a detailed description of each of the 10 value categories [8]. Participants were asked to consider 40 actions, and for each indicate which top three of value categories would be most likely to underlie a person's choice to perform those actions. The full list of actions can be seen in Table 1. These actions were selected in such a way that the list represented a diverse set of daily activities, and the authors felt all value categories were most likely to be covered at least once.

2.3 Measures

After the surveys were filled in, the anonymized data was aggregated. This was done by counting the frequency of each value category in the 1st, 2nd and 3rd places for each action. Then, first place was awarded a score of 4, second place a score of 2 and third place a score of 1 for each time it appeared in said place. The scores were summed up such that every value category received an overall score per action. This formula was chosen such that a first place was worth a little more than a third and second place combined, and the same as two second places combined. After this score was created, a threshold of 9 was chosen in order to determine which categories were most relevant for each action. All categories scoring 9 or over were marked as relevant. This threshold was chosen such that each action had at least has one value category above the threshold.

2.4 Results

Table 1 shows the full results, marking each value category's score for each of the included actions. The rightmost column shows the difference between the mean score and the maximum score per action. This number indicates how much agreement existed between experts, with higher numbers indicating more agreement. Furthermore, it shows which value categories were marked by the experts as being relevant (above the threshold of 9) in red/bold.

From Table 1 the average distance from the highest score to the mean was computed, which is 11.4 on average. This indicates that for many actions a value category exists which scores visibly better than the rest. After all, to get an overall score of 11, at least 3 of the 7 participants needed to have scored one particular category in at least 2nd place. To get this number as difference from the mean score, this means the majority of the 7 experts agreed on the highest scoring category. This consensus indicates that we might, indeed, use value categories that are in Table 1 to pre-select what values a user can choose from. However, more work is necessary to study if this pre-selection truly does not limit users in the expression of their values, as well as to know if it actually achieves its goal of making value selection easier.

3 User Study

The results from the expert study show the potential of using a pre-selection of possible values based on the action. The goal of this pre-selection would be to make it easier for users to indicate what values underlie decisions to perform actions. However, it is important that people do not feel this pre-selection limits their freedom of expression, as the pre-selection is not meant to push users into giving certain answers. To study these two aspects, an online between-subject user study was performed. Participants were asked what value would most likely underlie an action. Half were only shown the pre-selection to pick from, while participants in the other condition were shown the full list of values from Schwarz [8].

3.1 Participants

For this study, participants were recruited via Amazon Mechanical Turk. 297 started the survey, and 231 completed it. Of these 231, 64 did not answer the control question correctly and were, therefore, excluded. Of the 167 remaining, 8 filled in the survey twice, and the data of their second time was deleted, leaving 159. One final participant was excluded because they did not collect their payment, leaving us with 158 participants included in the initial analysis.

When looking at this initial data, we noticed that some of the participants had only clicked once on the pages with the questions, namely for going to the next page. This can be taken as evidence that they did not look at the full drop down list of values, just leaving the first, default answer in place. In some cases, this might just indicate that the default answer seemed correct, but some participants also did this for every question. In the end, it was decided to remove participants that had answered 10 or more questions within a second of seeing the page, as it would've been nearly impossible for them to have fully read a question in that time. The threshold of 10 was chose due to it being over half of the questions. This way 23 participants were removed. This made the final number of participants included in the analysis 135.

3.2 Procedure

The participants were asked to fill in a survey. The survey starting with some general information, followed by asking for informed consent of the participants. After obtaining consent the participants were placed in 1 of 2 conditions after which 19 questions were asked where the amount of answers was dependant on the condition the participant were in. The 19 questions were asked in random order where on each question the answers were also in random order. The survey concluded by asking the participants 5 questions on their experience completing the survey.

3.3 Measures

We measured the total time spent to complete the survey and the first click, last click, the total amount of clicks and time at which the questions was submitted. The difference between time of the first and last click was used to measure the time actually spent on each of the questions. This metric proved to be useful as some of the participants had

Table 1. Weighted numerical representation of action per value category. Achievement(AC), Benevolence(BE), Conformity(CO),F Hedonism(HE), Power(PO), Security(SE), Self-Direction(SD), Stimulation(ST), Tradition(TR), Universalism(UN). First place is worth 4 points, second 2 and third 1. The mean to highest represents the difference from the highest to the average score. Highlighted in red/bold are the value categories higher or equal then 9, so marked as relevant for that action

Promoted activity	AC	BE	CO	HE	PO	SE	SD	ST	TR	UN	Mean to highest
Act politely	2	**12**	**11**	0	1	1	4	0	4	7	7,8
Buy something	8	4	4	**11**	1	4	5	4	1	0	6,8
Care for someone	2	**20**	1	2	2	4	2	4	0	6	15,7
Celebrate holiday	0	0	2	**11**	4	5	2	5	**13**	0	8,8
Communicate	5	4	3	0	4	2	3	8	1	**12**	7,8
Compete	**10**	0	2	1	7	1	2	8	4	0	6,5
Cook	**9**	0	0	**13**	0	**10**	4	2	3	1	8,8
Create something (e.g. painting)	**10**	0	0	6	1	0	**11**	**12**	1	1	7,8
Decide what to do	4	0	0	1	**11**	0	**20**	4	0	2	15,8
Do something exciting	6	0	2	**16**	1	1	2	**14**	0	0	11,8
Drink	4	0	2	**20**	0	4	3	4	4	1	15,8
Eat	0	2	6	**12**	0	**14**	3	0	5	0	9,8
Enjoy art	0	0	0	**15**	2	4	3	**10**	2	6	10,8
Exercise	**13**	0	0	2	2	**11**	**9**	5	0	0	8,8
Exercise influence	4	7	0	1	**18**	0	4	8	0	0	13,8
Follow a ceremony	0	0	**11**	4	0	3	1	1	**20**	2	15,8
Follow the law	0	1	**22**	4	0	7	0	1	4	3	17,8
Help someone	1	**18**	2	0	4	2	4	1	0	**10**	13,8
Learn	8	0	2	1	2	2	**16**	6	0	5	11,8
Make decisions for others	8	5	0	1	**18**	0	5	0	3	2	13,8
Make money	**13**	0	0	**11**	8	8	5	0	0	0	8,5
Meditate	2	7	1	5	1	2	**16**	4	4	0	11,8
Perform (e.g. a play)	**11**	4	1	2	2	0	6	**15**	0	1	10,8
Plan your day	**10**	0	2	4	3	1	**18**	0	2	0	14
Play games	2	0	3	**11**	0	0	6	**14**	5	1	9,8
Pray	2	2	1	0	0	8	6	4	**17**	2	12,8
Protect others	0	**18**	4	0	5	8	0	0	1	6	13,8
Protect your belongings	1	0	4	2	5	**24**	2	0	1	2	19,9
Protect yourself	2	1	2	0	7	**20**	4	0	5	1	15,8
Read	0	4	1	4	2	1	**9**	**11**	0	**10**	6,8
Relax	0	6	1	**18**	0	8	6	1	0	2	13,8
Repair something (e.g. car)	**18**	2	0	5	4	1	1	**9**	0	2	13,8
Sleep	0	1	0	**11**	3	**16**	8	2	0	0	11,9
Spend time with family	0	6	4	8	1	3	0	6	**10**	4	5,8
Spend time with friends	0	5	2	**9**	4	5	6	**9**	1	1	4,8
Study	**10**	0	0	5	2	0	**12**	6	1	6	7,8
Take responsibility	2	**11**	0	2	**16**	2	5	0	1	3	11,8
Travel	1	0	2	**12**	0	0	**11**	**15**	0	1	10,8
Watch movies	2	0	1	**18**	0	0	2	**13**	4	2	13,8
Work	**11**	0	1	0	4	8	**11**	6	1	0	6,8

taken breaks over 10 minutes long before the first click on a question was made, so we could not look at total time spent on the page. The first 19 questions were regarding values, there the last 5 questions were about the participants' experience taking the survey. These 5 consisted of 4 questions about the difficulty of the survey, followed by 1 question asking if the participant was missing the option for the answer they wanted to give. The first 4 questions regarding difficulty of the survey used a 5-point Likert scale ranging from -2 (Extremely difficult) via 0 (Neither easy nor difficult) to 2 (Extremely easy). The last question regarding missing answer options used a 4-point Likert scale ranging from 1 (Only some of the questions) to 4 (All of the questions).

4 Results

The data was analyzed with R version 3.6.1 and the analysis was split into 3 parts. The first part is analyzing the time spent on questions about values. The second part is on the questions regarding difficulty of the survey. And the third and last part is on the perceived lack of answers to the questions of the survey.

The time spent on the questions on values was analysed by using the mean time spent per question. The Shapiro-Wilk normality test was used, indicating that the data was not normally distributed (W = 0.77, $p < 0.01$). Therefore the Wilcoxon rank sum test with continuity correction was used, indicating that a significant difference exists between conditions in the amount it took for people to answer what value was most relevant (W = 3068, $p < 0.01$).

Difficulty was tested with four questions. In order to create a single difficulty score, the questions had their internal cohesion tested using Cronbach's alpha and were found to be internally cohesive ($\alpha .83$). The Shapiro-Wilk normality test shows the data was not normally distributed (W = 0.95, $p < 0.01$). Therefore the Wilcoxon rank sum test with continuity correction was used, showing significant difference in the answers on questions regarding the difficulty of the survey between the two conditions (W = 1394.5, $p < 0.01$).

The question regarding freedom of answers was analysed separately. On average, people indicated that they could answer as they wished for 'most of the answers' (3) for both conditions (all answers: M=2.95, SD=0.73, pre-selection: M=3.01, SD=0.86). As the data was not normally distributed (Following Shapiro-Wilk W = 0.79, $p < 0.01$), the Wilcoxon rank sum test with continuity correction was used, showing no significant difference between the two groups regarding their experience of missing answers (W = 2112.5, $p = 0.455$).

5 Discussion and Conclusion

The results show that participants that received the pre-selection spent significantly less time on average per value question, implying that it was easier to select an answer from the pre-selection. This was probably partly because there are less answers to consider, but could also be because people already had had an answer in mind and it would take less time to find their answer. Overall this means that the survey with pre-selected answers was less of a time investment, and that it was potentially easier to complete. This implication is supported by the results from the questionnaire, which also show that the participants that received the pre-selection found the survey significantly easier to complete. One concern with only presenting people with a pre-selection would be that it limits people's freedom of expression. However, our results show no significant difference in the amount of times people wanted to pick a value which was missing from the list. Note that the average score of both conditions indicated that they were able to find their value for 'most of the actions'. Therefore, we found no evidence that making a pre-selection lead to people feeling restricted in their expression.

5.1 Contributions

Values are abstract concepts, but when a system needs to use them, they need to be seen in the context of both the individual and what actions they are applied to. In this work, we use the context of these actions to inform us about what values are most likely, in order to more easily elicit values from an individual. More specifically, this study shows that it is possible to present a pre-selected list of values to participants based on the context of the action it is applied to. This pre-selected list makes the process of picking underlying values faster and easier to perform, without it affecting the freedom of expression perceived by participants. This is important as this technique can be used by systems to learn what values underlay an individual's choice to perform an action. In this way, values can be used by system's to adjust their advice and decision making processes, and to align better with their users. Values form a large part of the moral context in which people make decisions, so it is important that we take steps to allow systems to understand these better [1].

5.2 Limitations

Firstly, our pre-selection was based on a limited number of expert participants. Although our results indicate that this was a good pre-selection, we do not assume full consensus on what this should look like. For a fully validated pre-selection of what value type corresponds to what action, more work would need to be done. However, our main intention was to study whether such a pre-selection was even possible in the first place and we believe this smaller sample was enough to show that this is indeed the case. Secondly, the questions about difficulty and perceived amount of missing answers used self-reported data for the analysis. We do not fully know to what extent people truly found it more difficult because of the long list, or because the selection made values easier to think about. Moreover, the results with respect to freedom of answers were all relatively high, which might indicate a ceiling effect. Although we did not find that a pre-selection limited people's perceived freedom in choice, this might be because they simply could not think of anything else. However, when presented with a full list some people might still pick things which were not in the pre-selection. As we did not show the same people both the full and the pre-selection lists, a direct comparison like this was not possible.

5.3 Future Work

Firstly, this paper focused on a pre-selection on values for ease of use. At the moment, you need to have the pre-selection for each specific action. To be able to scale up to any arbitrary set of actions it would be worthwhile to explore the existence of a groupings of actions that share the same values. The possibility exists that values can be extrapolated, making it easier for the system to scale in the amount of actions. Secondly, this paper only looks at actions to narrow down a pre-selection of possible underlying values. However, in indicating what value underlies an action, more contextual factors might play a role. Things like time of day, weather and surrounding actions might be relevant. But a good starting point for taking into account more context might also be social situation. Social norms are highly dependent on our values, so whether we perform an action with friends or with colleagues might change what value underlies it [4]. More work is necessary to see whether such additional context factors would allow for better pre-selections of values. Finally, this paper assumes that the answers filled in by the participants in the surveys are representative of their beliefs. However, talking about values is difficult, and so is verifying whether what people say about their values matches with what they actually value in practise. Therefore, it would be interesting to see to what extent the answers given in the survey coincide with the values that the participants actually hold.

5.4 Conclusion

Values are increasingly being incorporated in technology, but their elicitation remains difficult. In this work, we explore whether it is possible to make value elicitation for specific actions easier by presenting people with a pre-selection containing only those values most relevant to that action context. In an expert study, we found that there is indeed some consensus on what value categories are most likely to correspond to an action. This indicates that it is indeed possible to make a pre-selection of most relevant values based on the actions that are looked into. Additionally, in a user study with such a pre-selection we found that it made it easier for people to choose the most likely underlying value for an action, without diminishing their perceived freedom of choice. These results are important for the process of value elicitation and through that of value-based reasoning, which is becoming more important in today's society where we increasingly interact with technology on a personal level.

Acknowledgement: This work is part of the research programme CoreSAEP, with project number 639.022.416, which is financed by the Netherlands Organisation for Scientific Research (NWO).

REFERENCES

[1] *Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2*, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems., 2017.

[2] S. Cranefield, M. Winikoff, V. Dignum, and F. Dignum, 'No pizza for you: Value-based plan selection in BDI agents', in *International Joint Conference on Artificial Intelligence*, (2017).

[3] Batya Friedman, Peter H. Kahn Jr., and Alan Borning, *Human-Computer Interaction and Management Information Systems: Foundations Advances in Management Information Systems, Volume 5 (Advances in Management Information Systems),*, chapter Value Sensitive Design and Information Systems, 348–372, M.E. Sharpe, 2006.

[4] Ilir Kola, Catholijn M. Jonker, and M. Birna van Riemsdijk, 'Mode-model the social environment: Towards socially adaptive electronic partners', in *International Workshop Modelling and Reasoning in Context (MRC), Held at FAIM*, (2018). AAMAS/IJCAI Workshop on Modeling and Reasoning in Context.

[5] Eleonora Milić, Dragan Janković, and Aleksandar Milenković, 'Health care domain mobile reminder for taking prescribed medications', in *ICT Innovations 2016*, eds., Georgi Stojanov and Andrea Kulakov, pp. 173–181, Cham, (2018). Springer International Publishing.

[6] Alina Pommeranz, *Designing Human-Centered Systems for Reflective Decision Making*, Ph.D. dissertation, Delft University of Technology, 2012.

[7] Danielle E. Schoffman, Gabrielle Turner-McGrievy, Sonya J. Jones, and Sara Wilcox, 'Mobile apps for pediatric obesity prevention and treatment, healthy eating, and physical activity promotion: just fun and games?', *Translational Behavioral Medicine*, **3**(3), 320–325, (2013).

[8] Shalom H Schwartz, 'Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries', in *Advances in experimental social psychology*, volume 25, 1–65, Elsevier, (1992).

[9] Shalom M. Schwarz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens, 'Extending the cross-cultural validity of the theory of basic human values with a different method of measurement', *Journal of Cross-Cultural Psychology*, (2001).

[10] M.L. Tielman, C.M. Jonker, and M.B. van Riemsdijk, 'What should I do? Deriving norms from actions, values and context', in *International Workshop Modelling and Reasoning in Context (MRC), Held at FAIM*, (2018). Under revision at the AAMAS/IJCAI Workshop on Modeling and Reasoning in Context.

[11] Ibo van de Poel, *Translating Values into Design Requirements*, chapter Philosophy and Engineering: Reflections on Practice, Principles and Process, Springer, 2013.

A Transparent Framework towards the Context-Sensitive Recognition of Conversational Engagement

Alexander Heimerl, Tobias Baur, Elisabeth André [1]

Abstract.

Modelling and recognising affective and mental user states is an urging topic in multiple research fields. This work suggests an approach towards adequate recognition of such states by combining state-of-the-art behaviour recognition classifiers in a transparent and explainable modelling framework that also allows to consider contextual aspects in the inference process. More precisely, in this paper we exemplify the idea of our framework with the recognition of conversational engagement in bi-directional conversations. We introduce a multi-modal annotation scheme for conversational engagement. We further introduce our hybrid approach that combines the accuracy of state-of-the art machine learning techniques, such as deep learning, with the capabilities of Bayesian Networks that are inherently interpretable and feature an important aspect that modern approaches are lacking - causal inference. In an evaluation on a large multi-modal corpus of bi-directional conversations, we show that this hybrid approach can even outperform state-of-the-art black-box approaches by considering context information and causal relations.

1 Introduction

Nowadays, machine learning approaches are most often purely data-driven as they use so-called ”black-box” approaches that map low-level features or decisions of previous classifiers onto abstract labels following statistical methods. Here we usually have no transparent concept of how the model is internally represented, e.g. how and why weights on the nodes of artificial neural networks are related.

In most research areas (e.g., in psychology, behaviour analysis, but also physics), the goal of creating a model is to reason about observations in the world, while creating and validating theories that aim to find causation and explanations. Then, such models are often validated in simulations, or collated with real-world observations. That means on the one hand, we have data-driven models in machine learning that do a decent job in creating predictions for a huge amount of recognition problems, but deliver no transparent way to understand their decisions and don't necessarily have a theory behind them. On the other hand, we have models that aim to explain interrelations of observations of the world and/or of their inner states. Such models are also called ”white-box” approaches.

In this paper, we suggest a hybrid approach that combines state-of-the-art ”black-box” recognition models with a transparent causal inference model. Lately, the focus of research tends towards deep end-to-end learning with artificial neural networks. While such approaches deliver promising results on audio-visual data, they only give little insight on how and why they predict behaviours the way they do. In this work, we investigate the recognition of ”conversational engagement”. Especially in scenarios where it is essential to know *why* a person's behaviour is interpreted as, e.g., ”strongly disengaged”, the idea is often to identify cues that led to this interpretation, providing an additional abstraction layer. Here, the relevance of a comprehensible model becomes very clear. Imagine a system that gives feedback on how engaged a person appeared in a social coaching scenario. A model should be able to give feedback on *why* it decided a person appeared to be strongly engaged or disengaged, so that a human can learn from the feedback. In order to infer complex social signals with a transparent model, we combine predictions of multiple high-precision classifiers with dynamic Bayesian networks (DBN) [32]. DBNs are probabilistic models that allow expressing causal relationships between nodes in a network, while at the same time considering previous observations. Even tough the parameters for such nodes and even the overall network structure may be learned with machine learning techniques, DBNs allow retracing the decisions they are making for each node or layer of nodes visually and are therefore inherently interpretable. While the structure of a DBN may be modelled based on a theory and grounded in social sciences, our framework allows to consider parallel observations, so it can learn correlations between concurrent behaviours, context and the complex phenomena of interest.

2 Related work

2.1 Engagement in psychology

Engagement is a complex social attitude. This becomes apparent when being confronted by the mass of available definitions. In fact Glas et. al [17] gave an overview of many different engagement definitions, with some of them being very context specific. The definition of Poggi coincides best with a general understanding of engagement. She describes it as: “The value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction.” [36]. As complex as it is to acquire a fitting definition, equally complex is the manifestation of engagement in conversations. There are multiple behaviours that are strongly connected to it.

In general, body language is an elemental part in expressing conversational engagement. To be more precise, the alignment of the body and the limbs play an important role on broadcasting the state of engagement [31]. Interlocutors, that are engaged during a conversation, align their bodies to each other, as described in [22], “to create a frame of engagement”.

However not only the body position and body movement relative to each other is an important criteria, also the individual body behaviour is of great interest. Lots of body movement may indicate some kind

[1] Augsburg University, Germany, Universitätstr. 6a, email: heimerl@hcm-lab.de

of restlessness. This was found to be connected to boredom, which is a manifestation of low engagement [14]. Also depending on the level of engagement the body reacts with more subtle signals. Heart rate, blood pressure, EEG and galvanic skin response are all potential candidates to draw conclusions about engagement [51].

Moreover specific gestures may allow to draw conclusions about the level of engagement. Lausberg [25] investigated, among other things, the origin of self-touch gestures. She describes, that self-touch gestures occur when people are emotionally engaged. Alongside self-touch gestures there are also more complex gestures, that reflect different affective states [28].

Another crucial part in human interaction are “Feedback / Backchannels”. It describes a high-level behaviour that is related to engagement. Backchannels are a kind of feedback. They occur between interlocutors and are typically in the form of non-intrusive acoustic or visual signals, e.g. a simple “Yes” or a headnod. Backchannels are a tool, to not only signal the success of communication, but also provide information about the level of engagement [17].

A strong form of engagement manifestation is mirroring of behaviours, be it acoustic or visual, from one interlocutor by the other. Those go by the terms “Synchrony”, “Mimicry” or “Alignment”. All of those represent a connection or bonding between interlocutors [17].

2.2 Recognition of engagement

Engagement has been investigated from various research angles, e.g. how to define engagement, how to annotate engagement or how to automatically predict engagement. Therefore it is no surprise that there are many different systems available to automatically predict engagement.

Rich et al. [39] introduced a reusable module for the recognition of engagement in human-robot interaction. They identified four connection events that they found to be tools for the maintenance of engagement. The four events were, directed gaze, mutual facial gaze, adjacency pairs, verbal and non-verbal backchannels. Those concepts built the theoretical foundation for their engagement recognition module.

Sanghvi et al. [45] predicted engagement based on body posture features. All their features have been extracted from video signals. They identified following important posture features: “Body lean angle”, “Slouch factor”, “Quantity of motion” and “Contraction index”. For the classification they used Weka [16] and evaluated 63 different classifiers. The best ones achieved a prediction accuracy of 82% on the two classes “engaged” and “not engaged”.

Roman Bednarik et al. [7] focused on recognising conversational engagement with gaze data. Further, they introduced an annotation scheme for the different levels of conversational engagement. They defined a total of six levels. In ascending order, the first being the lowest level of engagement and the last being the highest level of engagement: “No interest”, “Following”, “Responding”, “Conversing”, “Influencing discussion discourse/topic” and “Governing/managing discussion”. To ease down the classification task the authors decided to reduce the six classes of engagement to a two-classes problem - low and high engagement. For the automatic estimation they computed a total of 26 features from the raw eye gaze data, e.g. number of fixations, number of saccades, minimal and maximal fixation duration, minimal and maximal saccade amplitude, quantity of fixation at the speakers’ face. Those features have been used to train a SVM. Following this approach they achieved a prediction accuracy of 74%.

Yun et al. [56] proposed a convolutional neural network(CNN) to automatically predict engagement of children. For training their CNN they relied solely on facial images. However due to limited training data they used CNNs that have been pre-trained on face recognition tasks. Their network architecture includes a new layer combination to model temporal dynamics in order to extract high-level features from low-level features. For predicting engagement they distinguished between four levels of engagement, high engagement, low engagement, low disengagement and high disengagement. On the given task their network architecture achieved a balanced accuracy of 0.7807.

There is already plenty of research available that targets recognising engagement. However most of the systems focus solely on finding feasible features, either handcrafted or extracted from convolutional layers to optimise prediction accuracy. Little attention is payed to context, which is important when it comes to recognising engagement in everyday scenarios. Depending on the environment individuals are in it can affect how people behave and also what kind of cues they are using during a conversation. Imagine a student talking to his friend during a break in comparison to a student attending an oral exam. However not only external factors can influence the broadcasting of engagement. Also the very unique psychological traits every person has can influence their behaviour. An extrovert person in comparison to an introvert person can appear totally different during a conversation. Those examples illustrate potential context information that should be considered when recognising engagement.

2.3 Bayesian networks

Bayesian networks have been successfully applied in earlier work in the area of high-level interpretation of social signals. One of the pioneer studies is the work by Conati et al. [11]. They have incorporated bio-feedback sensors into a complex emotion model, that was based on a subset of the emotions proposed by OCC theory [34]. They employed a dynamic decision network (a generalisation of a dynamic Bayesian network) to capture many of the complex phenomena associated with appraisal theories. In particular, their model estimated student goals based on personality traits and events which represent changes in the environment (e.g., progress in the system) as well as evidence from physical feedback channels to support the model’s prediction.

Sabourin et al. [43] focused, similar to Conati et al., on learners’ emotions, and employed multiple variations of Bayesian networks. More specifically, they investigated the benefits of using cognitive models of learner emotions, to guide the development of Bayesian networks for prediction of student affect. Predictive models were empirically trained on data, acquired from 260 students interacting with a game-based learning environment. As a dynamic Bayesian network turned out to be the most successful model, they emphasised the importance of temporal information in predicting learner emotions. They concluded that predictive models may be used to validate theoretical models of emotion.

Wöllmer et al. [55] combined a hierarchical dynamic Bayesian network to detect linguistic keyword features together with long short-term memory (LSTM) neural networks [19] which model phoneme context and emotional history to predict the affective state of the user. This way, they are combining acoustic, linguistic, and long-term context information to continuously predict the current valence and activation in a two-dimensional emotion space.

Lugrin et al. [26] used Bayesian networks to incorporate culture into intelligent systems by combining theory-based and data-driven approaches. Their network aims to generate non-verbal culture-dependent behaviours. While the model is structured based on cultural theories and theoretical knowledge of their influence on prototypical behaviour, the parameters of the model are learned from a multi-modal corpus recorded in the German and Japanese cultures. In their work, they aim to generate adequate behaviours for an agent to show, based on its simulated culture.

Finally, one could conclude that (dynamic) Bayesian networks have been successfully employed for some predefined contexts and applications. Especially when considering context, as it is essential in e.g. appraisal emotion models, or in specific applications, DBNs turn out to be a promising approach. In contrast to most other fusion mechanisms their structure may be actively modelled, based on existing theories, so that the structure contains valuable information implicitly, allowing to include existing knowledge in the model. This is especially useful when it is required to make assumptions *why* the model predicted one outcome and not another. It is worth mentioning that context information has only rarely been taken into account - or in most cases, limited to aspects like temporal context in previous research. Yet, in human communication multiple aspects of context [6] continuously influence our behaviours.

2.4 Explainable AI Approaches

The current trend in machine learning tends towards deep learning and neural network architectures that in contrast to Bayesian networks aren't inherently interpretable. Therefore efforts are made to provide explanations for such "black-box" approaches. In general we can distinguish between two kinds of systems providing explanations: model-agnostic or model-specific. Model-agnostic systems are capable of generating explanations independent of the underlying model. Ribeiro et al. introduce in [38] LIME, a model-agnostic approach for the generation of explanations. LIME is able to provide explanations for any given model by approximating an interpretable model around the passed model.

Alber et al. [3] introduced a library named iNNvestigate that provides implementations of common analysis methods for neural networks, e.g. PatternNet and LRP. The generated explanations come in the form of highlighted regions, that have been important for the classification. The supported methods are in contrast to Lime model-specific.

Same goes for SHAP developed by Lundberg et al. [27]. Their framework generates explanations by assigning each feature a value, that describes its importance in regard to the prediction.

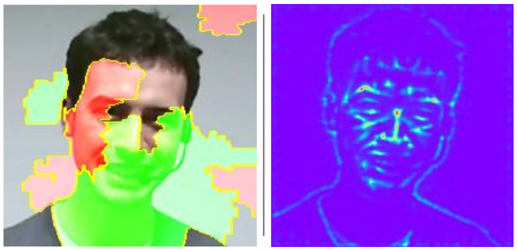


Figure 1. The left image shows an explanation generated with LIME. The right image displays an explanation generated with the iNNvestigate Library using Guided Backpropagation. The neural network to be explained was trained on raw image data from the NoXi corpus (see section 4) to predict different emotions, in this particular case the network predicted happiness as the subject's emotional state.

Figure 1 displays what visual explanations generated by LIME and iNNvestigate could possibly look like. The images have been generated within the scope of the presented work. While such visual explanation systems are of great value in helping to better understand which part of the input data was relevant for a decision, they don't provide causal explanations. The explanation generated by LIME highlights areas that are important for predicting a specific class in green colour, whereas the red coloured shapes describe areas that speak against the predicted class. In the example provided in Figure 1 it is evident that a large part of the face including the smile of the person is important for classifying happiness. However the other half of the face is coloured red and even some areas in the background are coloured green. With this information alone it is not easily comprehensible what the exact reasoning to predict a particular class has been. The explanations generated with iNNvestigate are even harder to correctly interpret. In the provided examples several edges outlining the facial features of the subject are marked being relevant for predicting. Those explanations often leave the user guessing and applying self made causal coherencies to further explain the prediction. Rather these approaches help to get better insight on the decisions of a network on a feature level. A big advantage of Bayesian networks is that the structure of a network can be modelled to have intrinsic meaning. Those causal coherencies might be used as a foundation for generating human-interpretable textual explanations.

3 The Role of Context

In current systems for recognising human behaviours only little attention is given to *context* (e.g. context that is represented by surrounding frames when training a model). Yet there are behaviours that are difficult to analyse and interpret correctly without further information about the *context* of a situation. *Context* is a wide-ranging term that has different meanings depending on the paradigm of research, application and scenario. Duranti et al. [15] noted that it seems impossible to present a single, precise and technical definition of context. Context information might appear as a single impact factor on the interaction or as a combination of multiple types of information. In addition to that, various challenges occur when it comes to context in multimodal communication [50]. In this section we approach different aspects of context:

Temporal context: In classical linguistics, context is "a frame that surrounds the event and provides resources for its appropriate interpretation" [15]. Wöllmer et al. [54] considered context as the temporal surroundings of an observation. In their work they successfully applied bidirectional long-short-term memory (BLSTM) neural networks to consider contextual long-range observations for the prediction of emotions. They further investigated algorithms such as multidimensional dynamic time wrapping (DTW) and asynchronous hidden-markov models to fuse mutual information from multiple modalities, while considering their temporal alignment [53]. An overview on algorithmic approaches, such as dynamic and canonical time wrapping in the context of facial expression analysis is given in [35].

When analysing complex social signals and emotions, the temporal order of behaviours is of vast importance. As an example, Keltner [21] describes a typical time series of behaviours in multiple modalities, that represent a typical instance for the complex emotion "embarrassment" in a social situation - a similar times series of events as we consider here for recognising engagement. Typically, the gaze shifts towards the bottom, the lips make slight

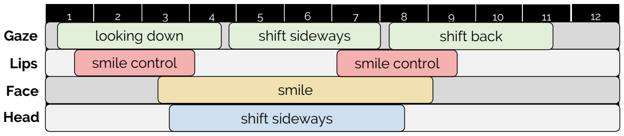


Figure 2. A typical time series of social cues that are performed when a person is feeling ”embarrassed”

movements that often turn into a smile followed by the gaze and head shifting to the side and back. Considering such sequences of social signals adds valuable information to the interpretation, compared to the analysis of isolated single cues.

Interaction dynamics context Analysing the dynamics in human communication includes being able to investigate both, the individual multi-modal dynamics (see temporal context) as well as the interpersonal dynamics. Researchers consider interpersonal dynamics on multiple abstractions. For example, Delaherche et al. and Varni et al. [13, 46] consider the synchronicity of people in dyadic interactions on a signal level. Therefore, they developed a set of synchronicity measurements. Rich et al. [40] defined state machines to automatically recognise the four interpersonal cues ”mutual gaze”, ”directed gaze”, ”adjacency pairs” and ”backchannels”. In their work they counted the appearance of such bi-directional cues and considered their appearance as an indicator of a person’s engagement. Another aspect is the current role in a conversation. Depending on whether the user is in the role of a listener or a speaker, the same kind of behaviour might be interpreted in a completely different way. The influence of the interaction role is illustrated by the following example. Let us assume we observe a person showing a high amount of gestural activity. If the person is in the role of a listener, the observed activity could be interpreted as restlessness. On the opposite, if the person is in the role of a speaker, we might conclude that the person is actively engaged in the conversation. Salam et al. [44] classify multiple aspects of context as parts of the relationship of a social robot and a human during an interaction. More precisely, the interaction context in their definition describes how a scenario relates multiple interlocutors.

Semantic context: The interpretation of detected social cues can be entirely altered through the semantics of accompanying verbal utterances. For example, a laughter in combination with an utterance commenting a negative event would no longer be interpreted as a sign of happiness, but rather be taken as sarcasm. By considering the semantics of accompanying spoken content, detected social cues could be interpreted more accurately. Studies further indicate that humans use semantic context for the interpretation of facial expressions [8, 37, 48].

Environmental context: The location and environmental surroundings may also influence the way we behave during an interaction. As an example, Zimmermann et al. [57] argues that the environmental surroundings directly influence our behaviours e.g. in the way we breathe or speak. In human-computer interaction and especially in ubiquitous computing, a system is called context-aware when it understands the circumstances and conditions surrounding the user. Abowd et al. [1], define context as ”any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves”. They further state that context is highly dependable on the current perspective.

Social context: Another aspect of context is the so called ”social context”. Riek et al. [41] stress the importance of considering social context when creating automated behaviour analysis systems. In their definition, social context is the ”environment where a particular person is situated with four factors that may influence (their) behaviour: situational context, cultural context, the person’s social role context, and the environmental social norms”. Such aspects may be addressed by the following questions: In what kind of situation does the conversation happen? What is the setting of the interaction? (situational context), How well do the interlocutors know each other? Do they share common knowledge? What culture or gender do they have? What is their personality like? (cultural context). How is their relationship? How is their social status? (the person’s social role). What are the social norms in the location of the interaction? What are the social norms in the community of the interlocutors? (environmental social norms). Questions like these play an important role, especially when interpreting non-verbal behaviour. Some of these aspects might be difficult to retrieve in an automated manner during the interaction between multiple interlocutors. However, if it is not possible to automatically gather such context information, it could be collected upfront.

When humans interpret behaviours of other people, they consciously or unconsciously include these and similar considerations in their reasoning process. Machines that aim to correctly interpret human behaviours should therefore consider contextual aspects in their interpretation models as well. Yet, besides temporal context (e.g. [54]), only little attention has been put to contextual aspects in current social signal processing research.

4 NoXi Database

The data for the upcoming evaluation tasks has been gathered from the NoXi Database [9]. NoXi provides dyadic novice-expert conversations. One participant took the role of the expert and the other one the role of the novice. Experts were free to chose the topic they wanted to talk about. Furthermore, the novices were evenly free in choosing what to listen to. This resulted in conversations covering a broad scope of different topics ranging from photography to dementia. Both participants were placed in separate rooms during the recording. They interacted remotely through TV screens and microphones. An example for the setup can be seen in Figure 3.

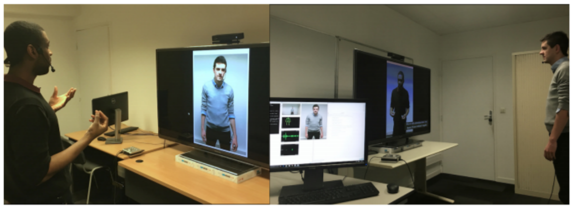


Figure 3. Recording of a novice-expert conversation in the NoXi database [9].

The database covers multiple languages and ethnicities, e.g. English, French, German, Indonesian, Arabic, Spanish, Italian. However, English, German and French have been the languages that occurred the most. A total of 84 sessions have been recorded, providing

25 hours and 18 minutes of conversational data. Additionally, demographic information of the participants have been collected, which include gender, cultural identity, age and level of education. The range of age has been from 21 to 50 years. We decided for the NoXi corpus due to the fact that it contains multi-modal multi-person interaction data and its transferability to social coaching scenarios. Moreover the setup of the corpus allowed for both, engaging, as well as non-engaging interactions.

A total of 19 sessions of the NoXi corpus have been annotated regarding conversational engagement. The annotators followed the engagement definition of Poggi, which we introduced in subsection 2.1. For most of the sessions novice and expert annotations have been created. Of the 19 sessions twelve are associated with French, four with English and three with German. For annotating, a continuous scheme has been chosen. The engagement annotations were created on the ratings of 4-7 different annotators. To measure the quality of the created annotations, from every annotator, they are validated against each other using the Pearson Correlation Coefficient (PCC). Based on the PCC, a gold standard for the annotations has been created. Whenever different annotators have scored a PCC value greater than 0.5 they have been merged to a gold standard annotation. Depending on the definition a value greater than 0.5 is considered a strong uphill (positive) linear relationship. However, at least two annotators have to score higher than 0.5, otherwise no gold standard has been created for the specific session and the session has been discarded. The gold standard itself is calculated by averaging the corresponding annotations.

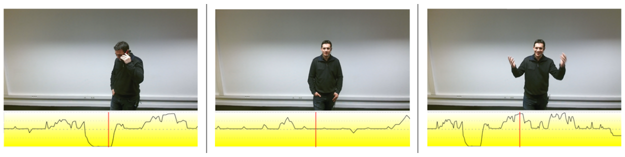


Figure 4. Examples for very low (left), medium (middle) and very high (right) engagement. In addition, the corresponding gold standard annotation is provided.

Figure 4 displays examples for very low, medium and very high engagement, with the corresponding gold standard annotation. The first image has been interpreted by the annotators as very low engagement. This scene occurred, as the novice decided to answer his phone, during the conversation (there were planned interruptions in the NoXi corpus, e.g. by calls from the experimenters or walk-ins). Answering the phone can be considered as a strong signal of the individual not willing to maintain the interaction. The alignment of head and body, away from the interlocutor, go along with a very low level of engagement. The next picture displays a neutral body position of the novice. This behaviour has been associated with a medium level of engagement. He aligned his body towards the other participant and is focusing the TV-screen. The last image represents very high engagement. The novice is smiling and shows a very open body posture, with the arms wide spread using a large gesture space. Again his body is aligned towards his interlocutor.

Engagement comes in various facets and sometimes the determination of its degree is distinct, like the just presented examples for very low and very high engagement. However, sometimes things are less obvious and leave room for a different interpretation. During the continuous annotation of conversational engagement we faced similar problems, as the ones mentioned by Whitehill et al. in [51]. They faced the issue, that an annotator tends to classify the level of engagement in the context of the currently annotated individual. Furthermore, they argue this could lead to annotations that are not comparable between different sessions. In fact, during the process of annotating, we often caught ourselves with statements like, “For their type of character, this should be considered as low/medium/high engagement”. However, we figured out that this causal chain is not wrong. It shows, that the way the level of engagement of an individual is perceived, also depends on the psychological traits the annotator attributes to the individual. Those traits can be considered as context information, which could be modelled inside the Bayesian network.

5 Engagement Model

Based on the evidences presented in subsection 2.1 we developed an annotation scheme that has been used to train our Bayesian networks. We considered different modalities besides context information.

Audio: First of all we considered the general voice activity of the interlocutors as valuable information. Even though it is very basic in its nature it allows to draw a conclusion about the overall involvement of the individuals regarding the conversation. An overall low voice activity may imply a conversation with low engaged interlocutors. On top of that we distinguished between different types of voice activity. We considered speech, filler and silence. The fillers are a particularly interesting type of voice activity as they also cover audio backchannels. In subsection 2.1 we mentioned that backchannels are a very common tool during conversation and provide information about the level of engagement [17].
Further Knapp et al. [23] argue that emotions are reliably transported by the voice. Therefore we trained a support vector machine (SVM) to predict the arousal of the voice [5]. The output of both SVM models (arousal, speech/filler/silence) is used to train the Bayesian network.

Face/Head: During conversations the face usually occupies most of the interlocutors attention. A lot of important information regarding the level of engagement can be extracted from the face respectively the head. Therefore we aimed in our annotation scheme to cover a general impression of the region, as well as looking for specific behaviour that is strongly connected to engagement. We defined features that represent the overall movement of the head in regard to X,Y and Z-Axis. Those features were mainly inspired by the research of Ryota Ooko et al. [33]. They found that a moderate positive correlation of head movement regarding the level of conversational engagement is present. Further we considered the individual gaze behaviour of the participants. There are multiple studies present about the recognition of engagement solely based on gaze data, with good recognition scores [20] [7]. Finally we trained a neural network on the facial action units (FACS) extracted with Openface [4] to predict valence [5]. We used the output of the neural network to train our Bayesian network.

Body: We mentioned earlier in subsection 2.1 that the alignment and movement of the body play an important role in the recognition of engagement. We followed an approach that has been similar to the head features. We tried to cover the general behaviour of the body, as well as specific gestures or poses that are connected to engagement. Therefore we defined a group of features, called body properties. They are mainly inspired by the coding system introduced in [12]. It contains values for the distance between the arms and the hips for X and Z-Axis. Moreover, the alignment of the arms is covered, by calculating the rotation of the elbow joints. Those values are supposed to describe a general level of openness.

Also the distances of each arm to the hip allow interpretation of the symmetry of the arms. In addition to that, the standard deviation of the distance travelled by the head during a frame and the rotation of the head is calculated. Those values have been chosen based on [29] [12].

In subsection 2.1 we identified restlessness to be connected to low levels of engagement. This is the reason we decided to calculate the continuous movement of the interlocutors. Continuous movement is a cumulative value, which describes the overall body movement. Lots of movement may indicate restlessness. In addition to that we wanted to cover the amount of gesticulation an individual performs. Gesticulation is mentioned in [29] and [12] as a crucial nonverbal queue in communication. Therefore we mapped the amount of movement done by both hands onto a real number value, which represents a numeric value for gesticulation.

Furthermore we considered the crossed arms and head touch gestures. The crossing of the arms is a common and often observed gesture. In research it is often interpreted as the expression of a negative emotional attitude by individuals [18] [49]. Based on this we argue that a negative emotional state is bonded to low engagement. In subsection 2.1 we mentioned self touches as a possible signal of being emotionally engaged. Moreover, Gunes et al. [18] were able to achieve good recognition rates for emotions, based on face and body features. Their system associated the emotions of fear, sadness and surprise mostly with gestures of the hands touching the head.

We believe that context plays an important role when it comes to correctly identifying social behaviour. The same applies to recognising conversational engagement. Depending on the context a specific gesture or behaviour may have a different meaning. Recall the example of the very actively moving engaged expert. His continuous movement is not a sign of restlessness. Given the fact that he is talking and gesticulating he should be considered as actively engaged in the conversation. Based on the different types of context we defined in section 3 we considered following context to predict conversational engagement.

Turn hold: During a conversation the interlocutors usually alternate their speaking turns. Therefore we determine the interlocutor that is currently holding the turn. Turn taking and vocal cues play an important part during conversations [23]. This kind of information can be considered as interaction dynamics context.

Role: In the used corpus two roles have been present: novice and expert. The novice has been the one with little to no knowledge about the topic presented by the expert. Accordingly, the expert has been the one introducing and providing information about the topic to the novice. Furthermore, it is in the nature of the expert to be more talkative than the novice, therefore a rather silent expert tends to be in a state of lower engagement, when compared to a similar silent novice, who might be just interestedly listening. In terms of context the information about the role covers multiple aspects. As we just elaborated, most of the time novices and experts operate differently during conversations. Therefore this can be seen as interaction dynamics context. Besides that, the role also covers social context. This is due to the fact, that specific expectations are raised towards the expert. By putting themselves in the role of an expert they signal the novice that they have sophisticated knowledge about their topic. This may result in novices being rather reserved regarding their interactions and comments. Moreover, it is common for the expert to take the lead during the conversation, which automatically results in more speaking time.

Gender: There are differences in the behaviour during conversations depending on the gender of the interlocutors [29]. For example, in same-gender conversation pairs females tend to have more eye contact with each other then males do. Also, males are more prone to decrease eye contact over time, while females have a tendency to increase it [29]. That is only one of many examples where the different genders behave differently. Due to that we think that not only gender itself, but also the constellation of interlocutor pairs, e.g. male-male, male-female, female-female, will be beneficial to the recognition of engagement. By considering the gender we aim to cover another aspect of social context.

Temporal context: In section 3 we argued that the temporal order of behaviours is important when it comes to analysing complex social signals, such as engagement. That means, time series and patterns of behaviours have different meaning when performed differently.

Coming up with a suitable architecture for the Bayesian network has been an incremental approach. This process included systematically adding, removing and exchanging classifiers, because even though specific characteristics for engagement are suggested in the literature, it does not necessarily mean they will work for any given context.

To provide more insight about the actual architecture Figure 5 displays an excerpt of the multi person dynamic Bayesian network. Basically the network is a graphical representation of the just presented annotation scheme. However, a big advantage of Bayesian networks is that the structure has intrinsic meaning compared to other models (e.g. artificial neural networks). This way, we were able to take knowledge about causal coherencies into account. Context nodes such as the gender or role are represented by conditional nodes, so that engagement is predicted ”given” the context information, while social cues are ”symptoms” shown by the observed person. In other words, social cues can be observed, given that a person has a certain level of engagement. Most of the context information we considered important is focused on a single interlocutor. However we also identified interaction dynamics context as a key element in correctly interpreting conversational engagement. Therefore we chose to model a multi person Bayesian network that also takes the interaction context and the interaction dynamics of the different interlocutors into account when estimating conversational engagement. For the NoXi Database this resulted in a network considering two persons - expert and novice. Moreover, we modelled our network as a dynamic Bayesian network. This way we were able to take temporal context into account.

6 Transparency

Bayesian networks not only allow us to easily model context and other causal coherencies, but also provide transparency by default [52]. In subsection 2.4 we mentioned that machine learning models, in the context of explainable AI, can be distinguished between inherently interpretable models and black-box models. Bayesian networks are inherently interpretable. This is due to the fact that for a given set of variables a Bayesian network is a representation of the joint probability distribution [30]. Usually we want a trained Bayesian network - given a set of observation - to predict what the most likely class of our target node is. In our use case we want to know how engaged one of the interlocutors is. However in a Bayesian network we are not only able to find out how engaged a person is but also what are the most important features for a specific class and what characteristics

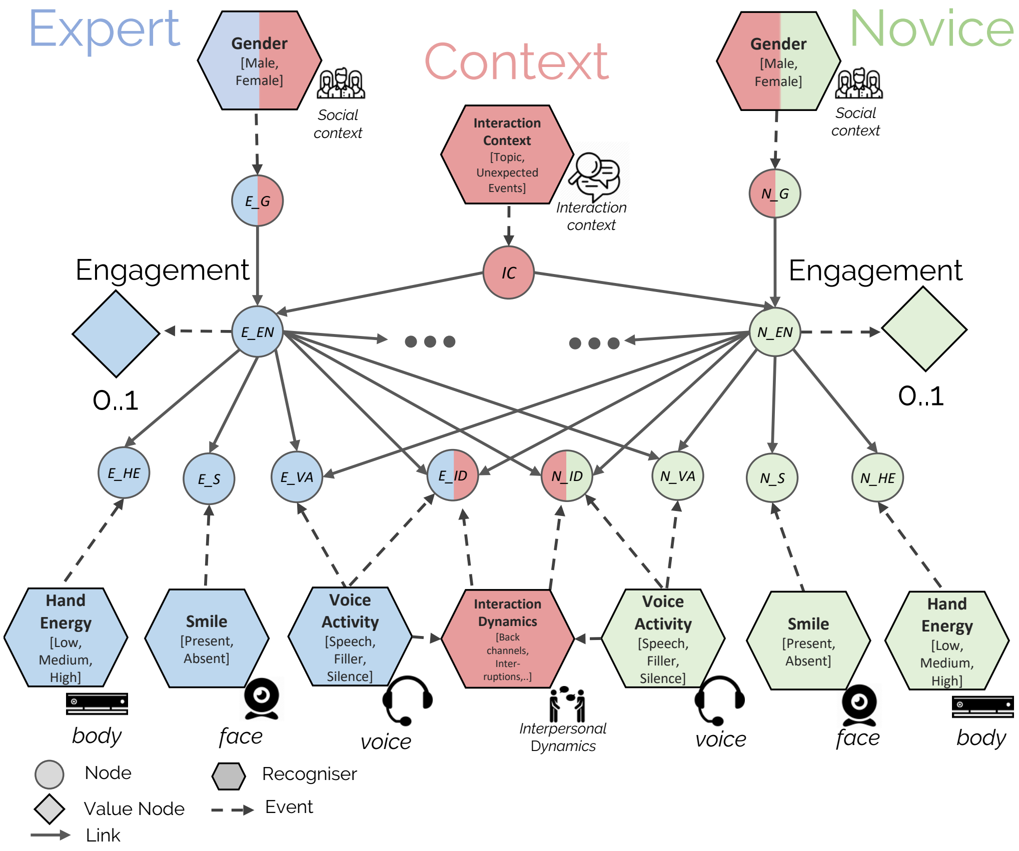


Figure 5. Schematic of a single time slide in a dynamic Bayesian network for two persons.

does the feature have. In Figure 6 a schematic of a reduced Bayesian network for the recognition of engagement is presented. The network contains the features Hand Energy and Voice Activity, which can take the characteristics low, medium and high. Moreover we have our target node Engagement, which also can be low, medium and high. Finally we considered some social context by adding the Role of the interlocutors. The schematic displays the probability distribution of the nodes given the person is highly engaged. This information tells us that when a person is highly engaged they are most likely in the role of the expert (70%) and show most likely high levels of Hand Energy and Voice Activity. We could now apply the same approach to find out more about low and medium engagement and get extensive insight about the learnt representations of our network.

7 Evaluation

Even though transparency is important in the context of machine learning, there is little use for a transparent model that isn't able to accurately predict the task at hand. That is why we investigate in the following the performance of the introduced architectures compared to other state-of-the-art machine learning approaches.

We split the acquired data into dedicated sets for training and evaluation. The training set included 13 sessions and had a size of 616374 samples. The evaluation set consisted out of six sessions, with a total of 328385 samples. So we ended up with the evaluation set having roughly half the samples of the training set.

To evaluate the different models, the Pearson correlation coefficient has been calculated between the model's prediction and the gold standard annotation.

[ht]

Table 1. Average PCCs on multimodal inputs

Method	Modalities	PCC
LSVM	Face, Body, Voice	.6253
Keras RNN	Face, Body, Voice	.6034
BN	Face, Body, Voice, Context	.7373
DBN (10 timesteps)	Face, Body, Voice, Context	.7443
MDBN (10 timesteps)	Face, Body, Voice, Context	.7680

As described earlier, developing a suitable Bayesian network has been an incremental approach by adjusting the classifier composition. An early Bayesian network (BN) based on multiple modalities including some context information achieved promising results with a PCC of 0.7373. By extending this network with temporal context for selected nodes that are related to body and face movement as well as voice activity we were able to further improve the correlation score to 0.7443. During our tests the network that performed best has been a multi-person dynamic Bayesian network (MDBN). It incorporates interpersonal dynamics, like mutual gaze and turn transitions between the novice and expert. The network achieved a PCC of 0.768 which is significantly better ($p < 0.001$) than the best single-user DBN (0.7443).

The (D)BNs we applied are created using a hybrid approach where classification results for sub-recognition tasks, as well as threshold

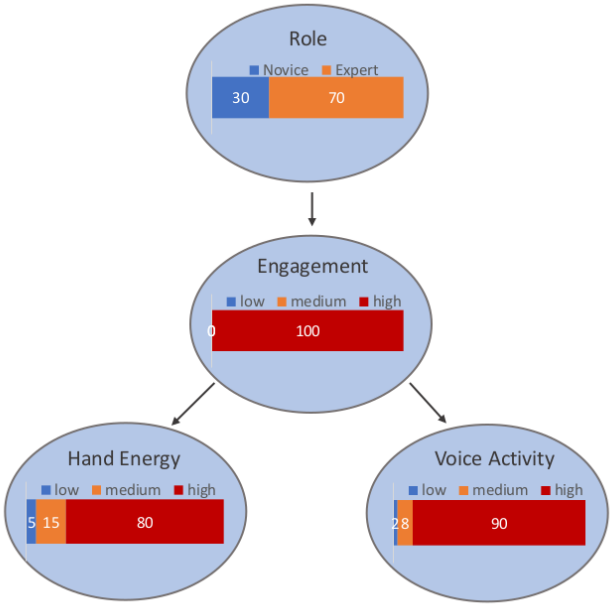


Figure 6. Schematic of a simplified Bayesian network displaying the probability distribution given the observation of high engagement.

based features are used to update the evidences in the network. This makes it difficult to compare the multi-modal model with other classification models that rely on low level features. In order to have a baseline to evaluate our approach, we created an engagement feature set that is heavily influenced by the previously introduced engagement annotation scheme. It contains features on body movement, body posture, head movement, facial expression and audio. We trained a linear support vector machine (LSVM) on this feature set and achieved a PCC of 0.6253. Moreover, we tested several neural networks implemented in Keras. The best one has been a fully connected deep recurrent neural network (RNN) and was able to score a PCC of 0.6034 on the engagement feature set. Those results are significantly (p <0.001) worse than our introduced hybrid model.

8 Discussion

We were able to show that our hybrid approach using a theory-modelled DBN can deliver comparable results to purely statistical black-box approaches. This is in compliance with the research of Rudin [42]. On our corpus it even slightly outperformed the other classification methods. With the introduction of a multi person dynamic Bayesian network architecture we were able to further increase the prediction accuracy. We explain this with several aspects: by employing the transparent DBN we could intuitively refine our first assumptions on what influences engagement, which allowed us to incrementally add classifiers, until the network achieved satisfying correlations with our gold standard annotation. Further, through the update mechanism on annotation/event abstraction we aimed to simulate a decision making and reasoning process that's similar to the one of humans. To our understanding, humans will consciously or unconsciously map abstractions of behaviours (e.g. smiles) on their perception of the other person (e.g. happiness). Further, we conclude that for our particular use-case of recognising conversational engagement, considering different types of context information leads to improvements in terms of the correct and adequate interpretation. In fact the more context information we added the better our model performed.

9 Conclusion

Deep learning can be considered as the current gold standard in machine learning. Deep neural networks proved themselves on various problem domains by performing exceptionally well [24] [10] [2]. However their biggest weakness is their lack of interpretability. That is why efforts are made to provide additional insight to otherwise ”black-boxes” (see subsection 2.4). Even though there are approaches present that help in gaining additional insight on the decision-making of neural network architectures, they rather provide additional information on a feature-level basis. In contrast to that there are models, like Bayesian networks that are inherently interpreteable and can be modelled to have intrinsic meaning. This enables a user to gather causal coherencies on why a model made a specific prediction. Often this seems to come down to a trade-off between prediction performance and transparency. However, we showed for the use case of multi-modal engagement recognition that by applying a hybrid approach that fuses abstractions of multiple social cues in a causal recognition model, accuracy and transparency do not necessarily need to exclude each other. Moreover we were able to improve the recognition rates of our model by incorporating social, temporal and interaction dynamics context. The significant impact of context on recognition scores stresses the importance of context in correctly and adequately interpreting conversational engagement. The proposed system has been implemented within the SSI Framework [47], so that all social cue classification models, as well as the overall BN inference step can be performed in a real-time system. This allows to apply this approach in a variety of applications, such as human-agent or human-robot scenarios.

ACKNOWLEDGEMENTS

This work has received funding from the DFG under project number 392401413, DEEP.
Further this work presents and discusses results in the context of the research project ForDigitHealth. The project is part of the Bavarian Research Association on Healthy Use of Digital Technologies and Media (ForDigitHealth), funded by the Bavarian Ministry of Science and Arts.

REFERENCES

[1] Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles, ‘Towards a better understanding of context and context-awareness’, in *Handheld and Ubiquitous Computing, First International Symposium, HUC’99, Karlsruhe, Germany, September 27-29, 1999, Proceedings*, ed., Hans-Werner Gellersen, volume 1707 of *Lecture Notes in Computer Science*, pp. 304–307. Springer, (1999).

[2] Igor Aizenberg and Gonzalez Alexander, ‘Image recognition using mlmvn and frequency domain features’, *International Joint Conference on Neural Networks*, (2018).

[3] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans, ‘innvestigate neural networks!’, *CoRR*, **abs/1808.04260**, (2018).

[4] T. Baltrušaitis, P. Robinson, and L. P. Morency, ‘Openface: An open source facial behavior analysis toolkit’, in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, (March 2016).

[5] S. Basu, N. Jana, A. Bag, Mahadevappa M, J. Mukherjee, S. Kumar, and R. Guha, ‘Emotion recognition based on physiological signals using valence-arousal model’, in *2015 Third International Conference on Image Information Processing (ICIIP)*, pp. 50–55, (2015).

[6] Tobias Baur, Dominik Schiller, and Elisabeth André, ‘Modeling user’s social attitude in a conversational system’, in *Emotions and Personality in Personalized Services*, 181–199, Springer, (2016).

[7] Roman Bednarik, Shahram Eivazi, and Michal Hradis, ‘Gaze and conversational engagement in multiparty video conversation: An annotation scheme and classification of high and low levels of engagement’, in *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, Gaze-In ’12, pp. 10:1–10:6, New York, NY, USA, (2012). ACM.

[8] Vicki Bruce and Andy Young, *In the eye of the beholder: the science of face perception.*, Oxford University Press, 1998.

[9] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar, ‘The noxi database: Multimodal recordings of mediated novice-expert interactions’, *ICMI’17*, (November 2017).

[10] Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber, ‘Multi-column deep neural networks for image classification’, *arXiv preprint arXiv:1202.2745*, (2012).

[11] Cristina Conati and Heather Maclaren, ‘Modeling user affect from causes and effects’, in *User Modeling, Adaptation, and Personalization, 17th International Conference, UMAP 2009, formerly UM and AH, Trento, Italy, June 22-26, 2009. Proceedings*, pp. 4–15, (2009).

[12] Nele Dael, Marcello Mortillaro, and Klaus R. Scherer, ‘The body action and posture coding system (bap): Development and reliability’, *Journal of Nonverbal Behavior*, **36**(2), 97–121, (Jun 2012).

[13] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen, ‘Interpersonal synchrony: A survey of evaluation methods across disciplines’, *IEEE Trans. Affective Computing*, **3**(3), 349–365, (2012).

[14] Sidney S D’Mello, Patrick Chipman, and Art Graesser, ‘Posture as a predictor of learner’s affective engagement’, in *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, (2007).

[15] Alessandro Duranti and Charles Goodwin, *Rethinking context: Language as an interactive phenomenon*, number 11 in Studies in the Social and Cultural Foundations of Language, Cambridge University Press, 1992.

[16] Stephen R Garner et al., ‘Weka: The waikato environment for knowledge analysis’, in *Proceedings of the New Zealand computer science research students conference*, pp. 57–64, (1995).

[17] N. Glas and C. Pelachaud, ‘Definitions of engagement in human-agent interaction’, in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 944–949, (Sept 2015).

[18] Hatice Gunes and Massimo Piccardi, ‘Affect recognition from face and body: early fusion vs. late fusion’, in *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 4, pp. 3437–3443. IEEE, (2005).

[19] Sepp Hochreiter and Jürgen Schmidhuber, ‘Long short-term memory’, *Neural Computation*, **9**(8), 1735–1780, (1997).

[20] Ryo Ishii and Yukiko I. Nakano, ‘An empirical study of eye-gaze behaviors: Towards the estimation of conversational engagement in human-agent communication’, in *Proceedings of the 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction*, EGIHMI ’10, pp. 33–40, New York, NY, USA, (2010). ACM.

[21] Dacher Keltner, ‘Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame’, *Journal of personality and social psychology*, **68**(3), 441, (1995).

[22] Mardi Kidwell, ‘Framing, grounding, and coordinating conversational interaction: Posture, gaze, facial expression, and movement in space’, in *Body - Language - Communication. An International Handbook on Multimodality in Human Interaction*, 100 – 113, De Gruyter Mouton, (2013).

[23] Mark Knapp, L. and Judith Hall, A., *Nonverbal Communication in Human Interaction*, Harcourt Brace, 1997.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, ‘Imagenet classification with deep convolutional neural networks’, in *Advances in Neural Information Processing Systems 25*, eds., F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 1097–1105, Curran Associates, Inc., (2012).

[25] Hedda Lausberg, ‘Neuropsychology of gesture production’, in *Body - Language - Communication. An International Handbook on Multimodality in Human Interaction*, 168 – 182, De Gruyter Mouton, (2013).

[26] Birgit Lugrin, Julian Frommel, and Elisabeth André, ‘Combining a data-driven and a theory-based approach to generate culture-dependent behaviours for virtual characters’, in *Advances in Culturally-Aware Intelligent Systems and in Cross-Cultural Psychological Studies*, 111–142, Springer, (2018).

[27] Scott M Lundberg and Su-In Lee, ‘A unified approach to interpreting model predictions’, in *Advances in Neural Information Processing Systems 30*, eds., I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–4774, Curran Associates, Inc., (2017).

[28] Marwa Mahmoud and Peter Robinson, ‘Interpreting hand-over-face gestures’, in *Affective Computing and Intelligent Interaction*, 248–255, Springer, (2011).

[29] Albert Mehrabian, *Nonverbal Communication*, AldineTransaction, 2007.

[30] Michael Mitchell, Tom, *Machine Learning*, 177–197, MacGraw-Hill, 1997.

[31] Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill, and Sedinha Tessendorf, *Body - Language - Communication. An International Handbook on Multimodality in Human Interaction*, De Gruyter Mouton, 2013.

[32] Kevin Patrick Murphy and Stuart Russell, ‘Dynamic bayesian networks: representation, inference and learning’, *Ph.D Thesis*, (2002).

[33] Ryota Ooko, Ryo Ishii, and Yukiko I. Nakano, ‘Estimating a user’s conversational engagement based on head pose information’, in *Intelligent Virtual Agents*, eds., Hannes Högni Vilhjálmsson, Stefan Kopp, Stacy Marsella, and Kristinn R. Thórisson, pp. 262–268, Berlin, Heidelberg, (2011). Springer Berlin Heidelberg.

[34] Andrew Ortony, Gerald L Clore, and Allan Collins, *The cognitive structure of emotions*, Cambridge university press, 1990.

[35] Yiannis Panagakis, Ognjen Rudovic, and Maja Pantic, ‘Learning for multi-modal and context-sensitive interfaces’, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*, **2**, in press, (2018).

[36] Isabella Poggi, *Mind, hands, face and body: a goal and belief view of multimodal communication*, Weidler, 2007.

[37] Carl Ratner, ‘Back to dr. ratner’s home page journal of mind and behavior, 1989, 10, 211-230 a social constructionist critique of naturalistic theories of emotion’, *Journal of Mind and Behavior*, **10**, 211–230, (1989).

[38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ‘”why should I trust you?”: Explaining the predictions of any classifier’, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144, (2016).

[39] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, ‘Recognizing engagement in human-robot interaction’, in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 375–382, (March 2010).

[40] Charles Rich, Brett Ponsleur, Aaron Holroyd, and Candace L. Sidner, ‘Recognizing engagement in human-robot interaction’, in *Proceedings of the 5th ACM/IEEE International Conference on Human Robot Interaction, HRI 2010, Osaka, Japan, March 2-5, 2010*, eds., Pamela J. Hinds, Hiroshi Ishiguro, Takayuki Kanda, and Peter H. Kahn Jr., pp. 375–382. ACM, (2010).

[41] Laurel D. Riek and Peter Robinson, ‘Challenges and opportunities in building socially intelligent machines [social sciences]’, *IEEE Signal Process. Mag.*, **28**(3), 146–149, (2011).

[42] Cynthia Rudin, ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, *Nature Machine Intelligence*, **1**(5), 206–215, (2019).

[43] Jennifer Sabourin, Bradford W. Mott, and James C. Lester, ‘Modeling learner affect with theoretically grounded dynamic bayesian networks’, in *Affective Computing and Intelligent Interaction - 4th International Conference, ACII 2011, Memphis, TN, USA, October 9-12, 2011, Proceedings, Part I*, pp. 286–295, (2011).

[44] Hanan Salam and Mohamed Chetouani, ‘A multi-level context-based modeling of engagement in human-robot interaction’, in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, May 4-8, 2015*, pp. 1–6. IEEE Computer Society, (2015).

[45] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira,

Peter W. McOwan, and Ana Paiva, 'Automatic analysis of affective postures and body motion to detect engagement with a game companion', in *Proceedings of the 6th International Conference on Human-robot Interaction*, HRI '11, pp. 305–312, New York, NY, USA, (2011). ACM.

[46] Giovanna Varni, Marie Avril, Adem Usta, and Mohamed Chetouani, 'Syncpy: a unified open-source analytic library for synchrony', in *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence, INTERPERSONAL@ICMI 2015, Seattle, Washington, USA, November 13, 2015*, pp. 41–47, (2015).

[47] Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André, 'The social signal interpretation (ssi) framework: Multimodal signal processing and recognition in real-time', in *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, p. 831–834, New York, NY, USA, (2013). Association for Computing Machinery.

[48] Harald G Wallbott, 'In and out of context: Influences of facial expression and context information on emotion attributions', *British Journal of Social Psychology*, **27**(4), 357–369, (1988).

[49] Harald G Wallbott, 'Bodily expression of emotion', *European journal of social psychology*, **28**(6), 879–896, (1998).

[50] Rebekah Wegener, *Studying Language in Society and Society through Language: Context and Multimodal Communication*, 227–248, Palgrave Macmillan UK, London, 2016.

[51] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan, 'The faces of engagement: Automatic recognition of student engagementfrom facial expressions', *IEEE Transactions on Affective Computing*, **5**(1), 86–98, (2014).

[52] Wim Wiegerinck, Willem Burgers, and Bert Kappen, *Bayesian Networks, Introduction and Practical Applications*, 401–431, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[53] Martin Wöllmer, Marc Al-Hames, Florian Eyben, Björn W. Schuller, and Gerhard Rigoll, 'A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams', *Neurocomputing*, **73**(1-3), 366–380, (2009).

[54] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn W. Schuller, and Shrikanth S. Narayanan, 'Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling', in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pp. 2362–2365, (2010).

[55] Martin Wöllmer, Björn W. Schuller, Florian Eyben, and Gerhard Rigoll, 'Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening', *J. Sel. Topics Signal Processing*, **4**(5), 867–881, (2010).

[56] W. Yun, D. Lee, C. Park, J. Kim, and J. Kim, 'Automatic recognition of children engagement from facial video using convolutional neural networks', *IEEE Transactions on Affective Computing*, 1–1, (2018).

[57] Heinz Zimmermann, *Speaking, listening, understanding*, SteinerBooks, 1996.

Multi-Modal Subjective Context Modelling and Recognition

Qiang Shen [1,2] and **Stefano Teso**[2] and **Wanyi Zhang**[2] and **Hao Xu** [1] and **Fausto Giunchiglia** [1,2]

Abstract. Applications like personal assistants need to be aware of the user's context, e.g., where they are, what they are doing, and with whom. Context information is usually inferred from sensor data, like GPS sensors and accelerometers on the user' smartphone. This prediction task is known as context recognition. A well-defined context model is fundamental for successful recognition. Existing models, however, have two major limitations. First, they focus on few aspects, like location or activity, meaning that recognition methods based on them can only compute and leverage few inter-aspect correlations. Second, existing models typically assume that context is objective, whereas in most applications context is best viewed from the user's perspective. Neglecting these factors limits the usefulness of the context model and hinders recognition. We present a novel ontological context model that captures *four dimensions*, namely time, location, activity, and social relations. Moreover, our model defines *three levels of description* (objective context, machine context, and subjective context) that naturally support subjective annotations and reasoning. An initial context recognition experiment on real-world data hints at the promise of our model.

1 INTRODUCTION

The term "context" refers to any kind of information necessary to describe the situation that an individual is in [2]. Automatic recognition of personal context is the key in applications like personal assistants, smart environments, and health monitoring apps, because it enables intelligent agents to respond proactively and appropriately based on (an estimate of) their user's context. For instance, a personal assistant aware that its user is at home, alone, doing housework, could suggest him or her to order a take-away lunch. Since context information is usually not available, the machine has to infer it from sensor data, like GPS coordinates, acceleration, and nearby Bluetooth devices measured by the user's smartphone. The standard approach to *context recognition* is to train a machine learning model on a large set of sensor readings and corresponding context annotations to predict the latter from the former. Existing implementations are quite diverse, and range from shallow models like logistic regression [14] to deep neural networks like feed-forward networks [15], LSTMs [7], and CNNs [12].

A context model defines how context data are structured. A good context model should capture all kinds of situational information relevant to the application at hand [2] and use the right level of abstraction [1]. Ontology is a widely accepted tool for formalizing context information [10], and several context ontologies have been proposed. Typical examples include CONON [16] and CaCONT [17]. CONON focuses on modeling locations by providing an upper ontology and lower domain-specific ontologies organized into a hierarchy. CaCONT defines several types of entities, and provides different levels of abstraction for specifying location of entities, e.g., GPS and location hierarchies. Focusing on semantic information of place, the work in [18] proposed a place-oriented ontology model representing different levels of place and related activities and improve the performance of place recognition. In [9], they proposed an ontology model involving social situation and the interaction between people.

These models, however, suffer from two main limitations. First, in order to support context recognition, the model should account for subjectivity of context descriptions. For instance, the *objective* location "hospital" plays different roles for different people: for patients it is a "place for recovering", while for nurses it is a "work place". This makes all the difference for personal assistants because the services that a user needs strongly depend on his or her subjective viewpoint. Most context models ignore this fact, with few exceptions, cf. [8]. Second, arguably answers to four basic questions – "what time is it?", "where are you?", "what are you doing?", and "who are you with?" – are necessary to define human contexts. Correlations between these aspects are also fundamental in recognition and reasoning: if the user is in her room, a personal assistant should be more likely to guess that she is "studying" or "resting", rather than "swimming". In stark contrast, most models are restricted to one or few of the above four aspects and therefore fail to capture important correlations, like those between activity and location or between time and social context.

As a remedy, we introduce a novel ontological context model that supports both reasoning and recognition from a subjective perspective, that captures time, location, activity, and social relations, and and that enables downstream context recognition tools to leverage correlations between these four fundamental dimensions. Our model also incorporates three levels of description for each aspect, namely objective, machine-level, and subjective, which naturally support different kinds of annotations. We apply and test our approach by collaborating with sociology experts within the SmartUnitn-One project [6]. We validate empirically our model by evaluating context recognition performance on the SmartUnitn-One context and sensor annotation data set [6], which was annotated consistently with our context model. Our initial results shows that handling correlations across aspects substantially improves recognition performance and makes it possible to predict activities that are otherwise very hard to recognize.

[1] College of Computer Science and Technology, Jilin University, Changchun, China, email: shenqiang19@mails.jlu.edu.cn, xuhao@jlu.edu.cn

[2] University of Trento, Italy, email: {stefano.teso, wanyi.zhang, fausto.giunchiglia }@unitn.it

2 CONTEXT MODELLING

Context is a theory of the world that encodes an individual' subjective perspective about it [3]. Individuals have a limited and partial view of the world at all times in their everyday life. For instance, consider a classroom with a teacher and a few students. Despite all the commonalities, each person in the room has a different context because they focus on different elements of their personal experience (the students focus on the teacher while the teacher focuses on the students) and ignore others (like the sound of the projector, the weather outside, and so on.) Given the diversity and complexity of individual experiences, formalizing the notion of context in its entirety is essentially impossible. For this reason, simpler but useful application-specific solutions are necessary.

Previous work has observed that reasoning in terms of questions like "what time is it?", "where are you?", "what are you doing?", "who are you with?", "what are you with?" is fundamental for describing and collecting the behavior of individuals [3]. Motivated by this observation and our previous work [4, 5, 11] , we designed an ontology-based context model organized according to the aforementioned dimensions of the world: time, location, activity, social relations and object. Formally, context is defined as a tuple:

$$\text{Context} = \langle \text{TIME}, \text{WE}, \text{WA}, \text{WO}, \text{WI} \rangle$$

where:

TIME captures the exact time of context, e.g., "morning". We refer to it as the *temporal context*. Informally, it answers the question "When did this context occur?".

WE captures the exact location of context, e.g., "classroom". We refer to it as the *endurant context*. Informally, it answers the question "Where are you?".

WA captures the activity of context, e.g., "studying". We refer to it as the *perdurant context*. Informally, it answers the question "What are you doing?".

WO captures the social relations of context, e.g., "friend". We refer to it as the *social context*. Informally, it answers the question "Who are you with?".

WI captures the materiality of context, e.g., "smartphone". We refer to it as the *object context*. Informally, it answers the question "What are you with?".

Figure 1 shows a scenario as a knowledge graph representing the personal context of an individual in the class. For instance, attributes of WO are "Class", "Name", and "Role", and their values are "Person", "Shen", and "PhD student", respectively. Edges represent relations between entities, e.g., "Shen" is in relation "Attend" with "Lesson".

The example in Figure 1 is presented in objective terms, that is, facts are stated as if they were independent of personal conscious experiences. However, each person interprets the world and her surroundings from her personal privileged point of view, which accounts for her personal knowledge, mental characteristics, states, etc. For instance, while in Figure 1 "Shen" has an objective role of Ph.D student, for other people "Shen" plays the roles of a "friend" or a "classmate" subjectively. The subjective context which is related to personal consciousness, knowledge, etc. can provide more information for applications such as personal assistant in order to give more intelligent services.

Notice that a person's view of her context is radically different from what her handheld personal assistant observes. In fact, machines interpret the world via sensors, while humans do not only interpret the world via their perceptions but with their knowledge as well. For instance, while a machine views location (e.g., a building) as a set of coordinates, humans interpret it based on its *function* (e.g., whether the building is their home or office).

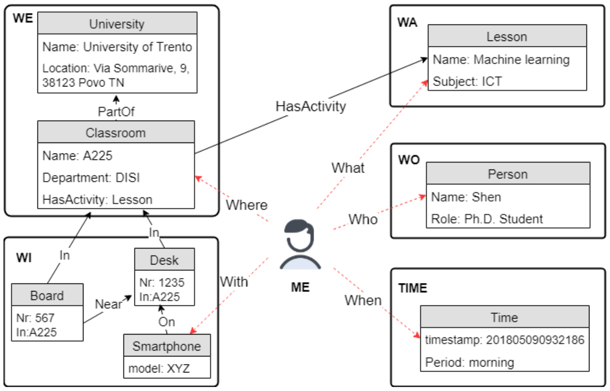
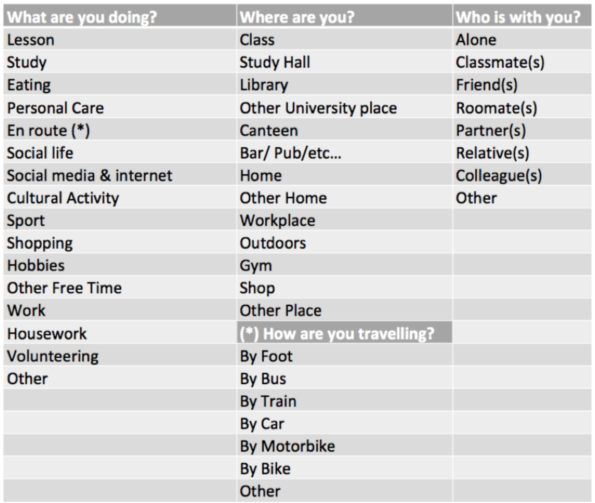


Figure 1. Illustration of our context model.



What are you doing?	Where are you?	Who is with you?
Lesson	Class	Alone
Study	Study Hall	Classmate(s)
Eating	Library	Friend(s)
Personal Care	Other University place	Roomate(s)
En route (*)	Canteen	Partner(s)
Social life	Bar/ Pub/etc...	Relative(s)
Social media & internet	Home	Colleague(s)
Cultural Activity	Other Home	Other
Sport	Workplace	
Shopping	Outdoors	
Hobbies	Gym	
Other Free Time	Shop	
Work	Other Place	
Housework	(*) How are you travelling?	
Volunteering	By Foot	
Other	By Bus	
	By Train	
	By Car	
	By Motorbike	
	By Bike	
	Other	

Figure 2. Questions and answers in the SmartUnitn-One questionnaire.

To model context precisely and completely, in addition to considering four dimensions, as discussed above, we also model three perspectives: objective context, subjective context and machine context. Table 1 shows the above example viewed through three types of perspective. The objective context captures the fact that at the University of Trento, Italy, at 11:00 AM, a person is attending a class together with Shen. When moving from objective to subjective, things change dramatically. From the perspective of the machine, the temporal context "11:00 AM" is viewed as a timestamp timestamp "1581938718026", and in subjective terms it becomes "morning"; similarly, "University of Trento" becomes coordinates "46°04'N,11°09'E" for the machine and "classroom" from a subjective perspective. For the perdurant context, the activity of taking lesson can be subjectively annotated as "study" by user, but it can be described as "connecting WIFI of classroom, sensors such as gy-

Level	TIME	WE	WA	WO
Objective Context	2020-02-17 11am	Via Sommarive, 9, 38123 Povo TN	Lesson	Shen
Machine Context	1581938718026	46°04'01.9"N 11°09'02.4"E	Accelerometer: 0g,0g,0g	"Shen" is in contact list
Subjective Context	Morning	Classroom	Studying	Friend

Table 1. An example of our three-partitioned context model. Each row gives a different description of the same underlying situation from the perspective of the world (top), the machine (middle), and the user (bottom).

Sensor	Frequency	Unit
Acceleration	20 Hz	m/s^2
Linear Acceleration	20 Hz	m/s^2
Gyroscope	20 Hz	rad$/s$
Gravity	20 Hz	m/s^2
Rotation Vector	20 Hz	Unitless
Magnetic Field	20 Hz	μT
Orientation	20 Hz	Degrees
Temperature	20 Hz	°C
Atmospheric Pressure	20 Hz	hPa
Humidity	20 Hz	%
Proximity	On change	0/1
Position	Every minute	Lat./Lon.
WIFI Network Connected	On change	Unitless
WIFI Networks Available	Every minute	Unitless
Running Application	Every 5 seconds	Unitless
Battery Level	On change	%
Audio from the internal mic	10 seconds per minute	Unitless
Notifications received	On change	Unitless
Touch event	On change	0/1
Cellular network info	Once per minute	Unitless
Screen Status, Flight Mode, Battery Charge, Doze Mode, Headset Plugged in, Audio Mode, Music Playback	On change	0/1

Table 2. List of sensors. Proximity triggers when the phone detects very close objects, e.g., the user's ear during a phone call.

roscope, accelerometer are sensed as static". For the social context, "Shen" is described as friend subjectively by the user and the machine senses "Shen" is in the contact list of the user.

3 EMPIRICAL EVALUATION

In order to evaluate the proposed context model, we carried out a context recognition experiment using the SmartUnitn-One data set [6], and studied whether recognition of subjective context is feasible and whether taking inter-aspect correlations into account helps recognition performance.

Data Collection. The SmartUnitn-One data set consists of sensor readings and context annotations obtained from 72 volunteers (university students) for a period of two weeks. All participants were required to install the i-Log app [19], which simultaneously records sensor data from several sensors (cf. Table 2) and context annotations. During the first week, students were asked to report their own context every 30 minutes by administering them questionnaires comprising three questions about location, activity, and social relations. The i-Log app collected sensor data at the same time. During the second week, the participants were only required to have the application running for the sensor data collection. All records were timestamped automatically. The questions were designed according to our context model and possible answers were modelled following the America Time Use Survey (ATUS) [13], leading to an ontology with over 80 candidate labels, see Figure 2 for the full list. Object context (WI) information was not collected as it is too hard to track without disrupting the volunteer's routines. All records were processed as in [20]. This resulted in 23309 records, each comprising 122 sensor readings (henceforth, features) and self-reported annotations about location, activity, and social context.

Experimental Setup. For every aspect in $\{\mathrm{WA}, \mathrm{WE}, \mathrm{WO}\}$, we trained a random forest to predict that aspect from sensor measurements. We randomly split the dataset into training (75% of the records) and validation (25% of the records) subsets and then selected the maximum depth of the forest using the validation set only. The classifier performance was evaluated using a rigorous 5-fold cross validation procedure. The data set was randomly partitioned into 5 folds. We hold out the selected fold as the test set to train a classifier on the remaining folds and compute the performance on the held out (test) fold. Then, we compared this model to another random forests (with the same maximum depth) that was supplied both sensor data and annotations for (a subset of) the other aspects as inputs. In order to account for label skew (e.g., some locations and activities are much more frequent than others), performance was measured using the *micro-average* F_1 score to account for class imbalance.

Results and Discussion. The average F_1 score across users are reported in Figure 3. The plots show very clearly that knowledge of other aspects substantially improves recognition performance regardless of the aspect being predicted: supplying the other aspects as inputs increases the F_1 score of predicting WA and WE by more than 10% and for WO by more than 5%. A breakdown of performance increase can be viewed in Table 3. The table shows that all aspects are correlated, as expected, especially activity and location, and that providing more aspects as inputs increases F_1 almost additively.

Inputs	WA	WE	WO
Sensors + WA	–	+8.80%	+2.36%
Sensors + WE	+8.27%	–	+3.09%
Sensors + WO	+3.34%	+3.27%	–
Sensors + Other Aspects	+11.25%	+11.57%	+5.31%

Table 3. Improvement in F_1 score when using other aspects as inputs to the recognition model. Columns indicate the aspect being predicted.

Figure 4 shows F_1 scores (again, averaged across users) for each label. For **WO**, some labels are clearly easier to predict than others. The performance improvement is usually in the 5–10% range, with the notable exception of "other", which improves by about 20%. It seems that location information always facilitates recognition of WO, while activity does not. Their combination, however, is always beneficial. For **WE**, looking at either WO and WA helps recognition performance in all cases, and providing both WO and WA gives a

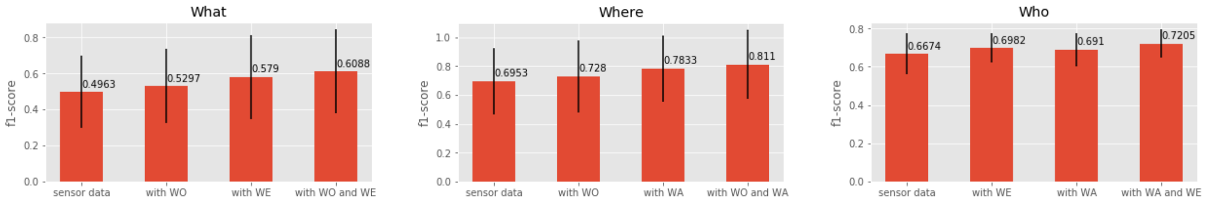


Figure 3. F_1 of our context recognition model. From left to right: perdurant (WA), endurant (WE), and social context (WO), respectively. The leftmost column refers to a predictor that uses sensor data only, while the other columns to predictors that in addition have access to context annotations.

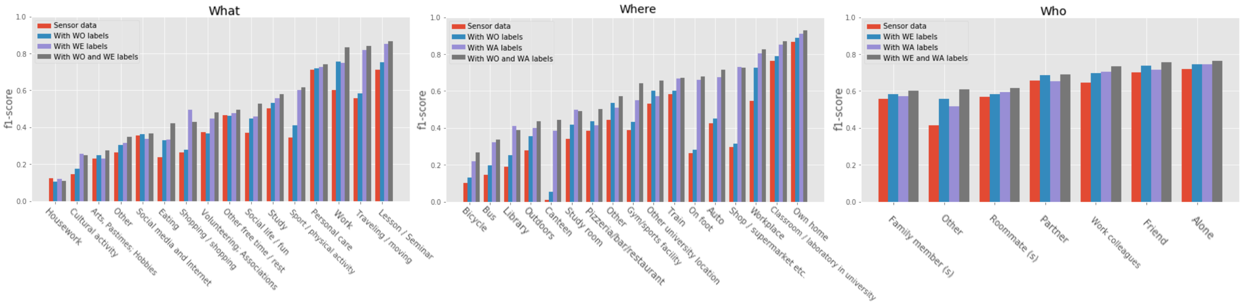


Figure 4. F_1 of individual labels (averaged over users). From left to right: perdurant, endurant, and social context, respectively.

larger improvement than than providing them separately. The exceptions are "library", "study room", and "shop", for which knowing WA improves more than knowing both WO and WA. This is somewhat surprising, as we expect social context to be moderately indicative of location, and deserves further investigation. Some locations ("canteen", "on foot", "auto", "shop", and "workplace") receive a major increase in recognition performance, from 25% to 40% approximately. This is partly due to the rarity of these classes in the data set, which shows that inter-aspect correlations supply to the lack of supervision. Finally for **WA**, some activities (like "housework", "cultural activities", and "hobbies") are very hard to predict, as their F_1 score is below 30%, while others ("work", "moving", and "lesson") are much easier to predict, with more than 80% F_1 score. This mostly shows that rare activities are harder to predict, understandably, although other factors might play a role. Using the full context (with WE and WO) always improves performance, except for "housework". For all the other activities, the improvement is from 5% to 20%, and even larger for "Shopping", "Sport" and "Traveling", for which the improvement is up to 30%.

This analysis provides ample support for our context model: correlations between different aspects improve context recognition performance for most users and, even more importantly, some values (like "Canteen") that are essentially impossible to recognize suddenly become much easier when full context information is provided.

4 CONCLUSION

We designed a novel context model that captures situational information about time, location, activity, and social relations of individuals using subjective—rather than objective—terms. An initial context recognition experiments on real-world data showed that machine learning models built using our context model produce higher quality predictions than models based on less complete context models. As for future work, we plan to study the effects of subjectivity more in detail, to migrate our architecture to more refined learning approaches (e.g., deep neural nets), and to carry out an extensive comparison against the state-of-the-art in context recognition.

5 ACKNOWLEDGEMENT

The research of FG has received funding from the European Union's Horizon 2020 FET Proactive project "WeNet – The Internet of us", grant agreement No 823783. The research of ST and WZ have received funding from the "DELPhi - DiscovEring Life Patterns" project funded by the MIUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) 2017 – DD n. 1062 del 31.05.2019.

REFERENCES

[1] Claudio Bettini et al., 'A survey of context modelling and reasoning techniques', *Pervasive and Mobile Computing*, (2010).

[2] Anind K Dey, 'Understanding and using context', *Personal and ubiquitous computing*, (2001).

[3] Fausto Giunchiglia, 'Contextual reasoning', *Epistemologia, special issue on I Linguaggi e le Macchine*, (1993).

[4] Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni, 'Human-like context sensing for robot surveillance', *International Journal of Semantic Computing*, **12**(01), 129–148, (2017).

[5] Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni, 'Personal context modelling and annotation', in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, (2017).

[6] Fausto Giunchiglia et al., 'Mobile social media usage and academic performance', *Computers in Human Behavior*, (2018).

[7] Nils Y Hammerla et al., 'Deep, convolutional, and recurrent models for human activity recognition using wearables', *arXiv preprint arXiv:1604.08880*, (2016).

[8] Mieczyslaw M Kokar, Christopher J Matheus, and Kenneth Baclawski, 'Ontology-based situation awareness', *Information fusion*, **10**(1), 83–98, (2009).

[9] Ilir Kola, Catholijn M Jonker, and M Birna van Riemsdijk, 'Who's that?-social situation awareness for behaviour support agents', in *International Workshop on Engineering Multi-Agent Systems*, pp. 127–151. Springer, (2019).

[10] Reto Krummenacher and Thomas Strang, 'Ontology-based context modeling', in *Proceedings*, (2007).

[11] Nardine Osman, Carles Sierra, Ronald Chenu-Abente, Qiang Shen, and Fausto Giunchiglia, 'Open social systems', in *17th European Conference on Multi-Agent Systems (EUMAS)*, Thessaloniki, Greece, (2020).

[12] Aaqib Saeed et al., 'Learning behavioral context recognition with multi-stream temporal convolutional networks', *arXiv preprint arXiv:1808.08766*, (2018).

[13] Kristina J Shelley, 'Developing the american time use survey activity classification system', *Monthly Lab. Rev.*, (2005).

[14] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet, 'Recognizing detailed human context in the wild from smartphones and smartwatches', *IEEE Pervasive Computing*, (2017).

[15] Yonatan Vaizman et al., 'Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification', *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, (2018).

[16] Xiaohang Wang et al., 'Ontology based context modeling and reasoning using owl.', in *Percom workshops*, (2004).

[17] Nan Xu et al., 'CACOnt: A ontology-based model for context modeling and reasoning', in *Applied Mechanics and Materials*, (2013).

[18] Laura Zavala, Pradeep K Murukannaiah, Nithyananthan Poosamani, Tim Finin, Anupam Joshi, Injong Rhee, and Munindar P Singh, 'Platys: From position to place-oriented mobile computing', *Ai Magazine*, **36**(2), 50–62, (2015).

[19] Mattia Zeni et al., 'Multi-device activity logging', in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pp. 299–302, (2014).

[20] Mattia Zeni et al., 'Fixing mislabeling by human annotators leveraging conflict resolution and prior knowledge', *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, (2019).

Curriculum Learning with Diversity for Supervised Computer Vision Tasks

Petru Soviany[1]

Abstract. Curriculum learning techniques are a viable solution for improving the accuracy of automatic models, by replacing the traditional random training with an easy-to-hard strategy. However, the standard curriculum methodology does not automatically provide improved results, but it is constrained by multiple elements like the data distribution or the proposed model. In this paper, we introduce a novel curriculum sampling strategy which takes into consideration the diversity of the training data together with the difficulty of the inputs. We determine the difficulty using a state-of-the-art estimator based on the human time required for solving a visual search task. We consider this kind of difficulty metric to be better suited for solving general problems, as it is not based on certain task-dependent elements, but more on the context of each image. We ensure the diversity during training, giving higher priority to elements from less visited classes. We conduct object detection and instance segmentation experiments on Pascal VOC 2007 and Cityscapes data sets, surpassing both the randomly-trained baseline and the standard curriculum approach. We prove that our strategy is very efficient for unbalanced data sets, leading to faster convergence and more accurate results, when other curriculum-based strategies fail.

1 Introduction

Although the accuracy of automatic models highly increased with the development of deep and very deep neural networks, an important and less studied key element for the overall performance is the training strategy. In this regard, Bengio et al. [2] introduced curriculum learning (CL), a set of learning strategies inspired by the way in which humans teach and learn. People learn the easiest concepts at first, followed by more and more complex elements. Similarly, CL uses the difficulty context, feeding the automatic model with easier samples at the beginning of the training, and gradually adding more difficult data as the training proceeds.

The idea is straightforward, but an important question is how to determine whether a sample is easy or hard. CL requires the existence of a predefined metric which can compute the difficulty of the input examples. Still, the difficulty of an image is strongly related to the context: a big car in the middle of an empty street should be easier to detect than a small car, parked in the corner of an alley full of pedestrians. Instead of building hand-crafted models for retrieving contextual information, in this paper, we use the image difficulty estimator from [12] which is based on the amount of time required by human annotators to assess if a class is present or not in a certain image. We consider that people can understand the full context very accurately, and that a difficulty measure trained on this information can be useful in our setting.

The next challenge is building the curriculum schedule, or the rate at which we can augment the training set with more complex information. To address this problem, we follow a sampling strategy similar to the one introduced in [28]. Based on the difficulty score, we sample according to a probability function, which favors easier samples in the first iterations, but converges to give the same weight to all the examples in the later phases of the training. Still, the probability of sampling a harder example in the first iterations is not null, and the more difficult samples which are occasionally picked increase the diversity of the data and help training.

The above-mentioned methodology should work well for balanced data sets, as various curriculum sampling strategies have been successfully employed in literature [19, 28, 34, 37], but it can fail when the data is unbalanced. Ionescu et al. [12] show that some classes may be more difficult than others. A simple motivation for this may be the context in which each class appears. For example, a potted plant or a bottle are rarely the focus of attention, usually being placed somewhere in the background. Other classes of objects, such as tables, are usually occluded, with the pictures focusing on the objects on the table rather than on the piece of furniture itself. This can make a standard curriculum sampling strategy neglect examples from certain classes and slow down training. The problem becomes even more serious in a context where the data is biased towards the easier classes. To solve these issues, we add a new term to our sampling function which takes into consideration the classes of the elements already sampled, in order to emphasize on images from less-visited classes and ensure the diversity of the selected examples.

The importance of diversity can be easily explained when comparing our machine learning approach to actual real-life examples. For instance, when creating a new vaccine, researchers need to experiment on multiple variants of the virus, then test it on a diverse group of people. As a rule, in all sciences, before making any assumptions, researchers have to examine a diverse set of examples which are relevant to the actual data distribution. Similar to the vaccines, which must be efficient for as many people as possible, we want our curriculum model to work well on all object classes. We argue that this is not possible in unbalanced curriculum scenarios, and it is slower in the traditional random training setup.

Since it is a sampling procedure, our CL approach can be applied to any supervised task in machine learning. In this paper, we focus on object detection and instance segmentation, two of the main tasks in computer vision, which require the model to identify the class and the location of objects in images. To test the validity of our approach, we experiment on two data sets: Pascal VOC 2007 [4] and Cityscapes [3], and compare our curriculum with diversity strategy

[1] University of Bucharest, Department of Computer Science, Romania, email: petru.soviany@yahoo.com

against the standard random training method, a curriculum sampling (without diversity) procedure and an inverse-curriculum approach, which selects images from hard to easy. We employ a state-of-the-art Faster R-CNN [24] detector with a Resnet-101 [11] backbone for the object detection experiments, and a Mask R-CNN [10] model based on Resnet-50 for instance segmentation.

Our main contributions can be summarized as follows:

1. We illustrate the necessity of adding diversity when using CL in unbalanced data sets;
2. We introduce a novel curriculum sampling function, which takes into consideration the class-diversity of the training samples and improves results when traditional curriculum approaches fail;
3. We prove our strategy by experimenting on two computer vision tasks: object detection and instance segmentation, using two data sets of high interest.

We organize the rest of this paper as follows: in Section 2, we present the most relevant related works and compare them with our approach. In Section 3, we explain in detail the methodology we follow. We present our results in Section 4, and draw our conclusion and discuss possible future work in the last section.

2 Related Work

Curriculum learning. Bengio et al. [2] introduced the idea of curriculum learning (CL) to train artificial intelligence, proving that the standard learning paradigm used in human educational systems could also be applied to automatic models. CL represents a class of easy-to-hard approaches, which have successfully been employed in a wide range of machine learning applications, from natural language processing [8, 16, 19, 21, 31], to computer vision [6, 7, 9, 15, 18, 27, 35], or audio processing [1, 22].

One of the main limitations of CL is that it assumes the existence of a predefined metric which can rank the samples from easy to hard. These metrics are usually task-dependent with various solutions being proposed for each. For example, in text processing, the length of the sentence can be used to estimate the difficulty of the input (shorter sentences are easier) [21, 30], while the number and the size of objects in a certain sample can provide enough insights about difficulty in image processing tasks (images with few large objects are easier) [27, 29]. In our paper, we employ the image difficulty estimator of Ionescu et al. [12] which was trained considering the time required by human annotators to identify the presence of certain classes in images.

To alleviate the challenge of finding a predefined difficulty metric, Kumar et al. [17] introduce self-paced learning (SPL), a set of approaches in which the model ranks the samples from easy to hard during training, based on its current progress. For example, the inputs with the smaller loss at a certain time during training are easier than the samples with higher loss. Many papers apply SPL successfully [26, 32, 33], and some methods combine prior knowledge with live training information, creating self-paced with curriculum techniques [14, 36]. Even so, SPL still has some limitations, requiring a methodology on how to select the samples and how much to emphasize easier examples. Our approach is on the borderline between CL and SPL, but we consider it to be pure curriculum, although we use training information to advantage less visited classes. During training, we only count the labels of the training samples, which is a priori information, and not the learning progress. A similar system could iteratively select examples from every class, but this would force our model to process the same number of examples from each class. Instead, by using the class-diversity as a term in our difficulty-based sampling probability function, we impose the selection of easy-to-hard diverse examples, without massively altering the actual class distribution of the data set.

The easy-to-hard idea behind CL can be implemented in multiple ways. One option is to start training on the easiest set of images, while gradually adding more difficult batches [2, 7, 16, 27, 30, 37]. Although most of the models keep the visited examples in the training set, Kocmi et al. [16] suggest reducing the size of each bin until combining it with the following one, in order to use each example only once during an epoch. In [19, 28] the authors propose a sampling strategy according to some probability function, which favors easier examples in the first iterations. As the authors show, the easiness score from [28] could also be added as a new term to the loss function to emphasize the easier examples in the beginning of the training. In this paper, we enhance their sampling strategy by adding a new diversity term to the probability function used to select training examples.

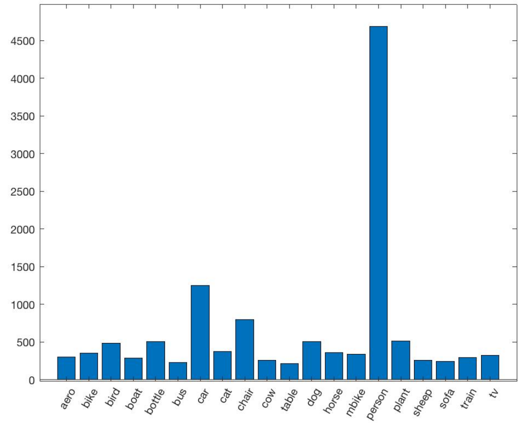


Figure 1. Number of instances from each class in the trainval split of the Pascal VOC 2007 data set.

Despite leading to good results in many related papers, the standard CL procedure is highly influenced by the task and the data distribution. Simple tasks may not gain much from using curriculum approaches, while employing CL in unbalanced data sets can lead to slower convergence. To address the second problem, Wang et al. [34] introduce a CL framework which adaptively adjusts the sampling strategy and loss weight in each batch, while other papers [13, 25] argue that a key element is diversity. Jiang et al. [13] introduce a SPL with diversity technique in which they regularize the model using both difficulty information and the variety of the samples. They suggest using clustering algorithms to split the data into diverse groups. Sachan et al. [25] measure diversity using the angle between the hyperplanes the samples induce in the feature space. They choose the examples that optimize a convex combination of the curriculum learning objective and the sum of angles between the candidate samples and the examples selected in previous steps. In our model, we define diversity based on the classes of our data. We combine our predefined difficulty metric with a score which favors images from less visited classes, in order to sample easy and diverse examples at the beginning of the training, then gradually add more complex elements. Our idea works well for supervised tasks, but it can be extended to unsupervised learning by replacing the ground-truth labels with a clustering model, as suggested in [13]. Figure 1 presents the class distribution on Pascal VOC 2007 data set [4] which is heavily biased towards class *person*.

Object detection is the task of predicting the location and the class of objects in certain images. As noted in [29], the state-of-the-art object detectors can be split into two main categories: two-stage and single stage models. The two-stage object detectors [10, 24] use a Region Proposal Network to generate regions of interest which are then fed to another network for object localization and classification. The single stage approaches [20, 23] take the whole image as input and solve the problem like a regular regression task. These methods are usually faster, but less accurate than the two-stage designs. **Instance segmentation** is similar to object detection, but more complex, requiring the generation of a mask instead of a bounding box for the objects in the test image. Our strategy can be implemented using any detection and segmentation models, but, in order to increase the relevance of our results, we experiment with high quality Faster R-CNN [24] and Mask R-CNN [10] baselines.

3 Methodology

Training artificial intelligence using curriculum approaches, from easy to hard, can lead to improved results in a wide range of tasks [1, 6, 7, 8, 9, 15, 16, 18, 19, 21, 22, 27, 31, 35]. Still, it is not simple to determine which samples are easy or hard, and the available metrics are usually task-dependent. Another challenge of CL is finding the right curriculum schedule, i.e. how fast to add more difficult examples to training, and how to introduce the right amount of harder samples at the right time to positively influence convergence. In this section, we present our approach for estimating difficulty and our curriculum sampling strategies.

3.1 Difficulty estimation

To estimate the difficulty of our training examples, we employ the method of Ionescu et al. [12] who defined image difficulty as the human time required for solving a visual search task. They collected annotations for the Pascal VOC 2012 [5] data set, by asking annotators whether a class was present or not in a certain image. They collected the time people required for answering these questions, which they normalized and fed as training data for a regression model. Their results correlate fine with other difficulty metrics which take into consideration the number of objects, the size of the objects, or the occlusions. Because it is based on human annotations, this method takes into account the whole image context, not only certain features relevant for one problem (the number of objects, for example). This makes the model task independent, and, as a result, it was successfully employed in multiple vision problems [12, 29, 28]. To further prove the efficiency of the estimator for our task, we show that automatic models have a lower accuracy in difficult examples. We split the Pascal VOC 2007 [4] test set in three equal batches: easy, medium and hard, and run the baseline model on each of them. The results in Table 1 confirm that the AP lowers as the difficulty increases.

We follow the strategy of Ionescu et al. as described in the original paper [12] to determine the difficulty scores of the images in our data sets. These scores have values ≈ 3, with a larger score defining a more difficult sample. We translate the values between $[-1, 1]$ using Equation 1 to simplify the usage of the score in the next steps. Figure 2 shows some examples of easy and difficult images.

$$Scale_{min-max}(x) = \frac{2 \cdot (x - min(x))}{max(x) - min(x)} - 1 \qquad (1)$$

Table 1. Average Precision scores for object detection using the baseline Faster R-CNN, on easy, medium and hard splits of Pascal VOC 2007 test set, as estimated using our approach.

DIfficulty	mAP (in %)
Easy	72.93
Medium	72.16
Hard	67.03

3.2 Curriculum sampling

Soviany et al. [28] introduce a curriculum sampling strategy, which favors easier examples in the first iterations and converges as the training progresses. It has the advantage of being a continuous method, removing the necessity of a curriculum schedule for enhancing the difficulty-based batches. Furthermore, the fact that it is a probabilistic sampling method does not constrain the model to only select easy examples in the first iterations, as batching does, but adds more diversity in data selection. We follow their approach in building our curriculum sampling strategy with only a small change in the position of parameter k in order to better emphasize the difficulty of the examples. We use the following function to assign weights to the input images during training:

$$w(x_i, t) = (1 - diff(x_i) \cdot e^{-\gamma \cdot t})^k, \forall x_i \in X, \qquad (2)$$

where x_i is the training example from the data set X, t is the current iteration, and $diff(x_i)$ is the difficulty score associated with the selected sample. γ is a parameter which sets how fast the function converges to 1, while k sets how much to emphasize the easier examples. Our function varies from the one proposed in [28] by changing the position of the k parameter. We consider that we can take advantage of the properties of the power function which increases faster for numbers greater than the unit. Since $1 - s_i \cdot e^{-\gamma \cdot t} \in [0, 2]$, and the result is > 1 for easier examples, our function will focus more on the easier samples in the first iterations. As the training advances, the function converges to 1, so all examples will have the same probability to be selected in the later phases of the training. We transform the weights into probabilities and we sample accordingly.

3.3 Curriculum with diversity sampling

As [13, 25] note, applying a CL strategy does not guarantee improved quality, the diversity of the selected samples having a great impact on the final results. A simple example is the case in which the data set is biased, having fewer samples of certain classes. Since some classes are more difficult than others [12], if the data set is not well-balanced, the model will not visit the harder classes until the later stages of the training. Thus, the model will not perform well on classes it did not visit. This fact is generally valid in all kind of applications, even in real life reasoning: without seeing examples which match the whole data distribution, it is impossible to find the solution suited for all scenarios. Because of this, we enhance our sampling method, by adding a new term, which is based on the diversity of the examples.

Our diversity scoring algorithm is simple, taking into consideration the classes of the selected samples. During training, we count the number of visited objects from each class ($num_{objects}(c)$). We subtract the mean of the values to determine how often each class was visited. This is formally presented in Equation 3. We scale and translate the results between $[-1, 1]$ using Equation 1 to get the score

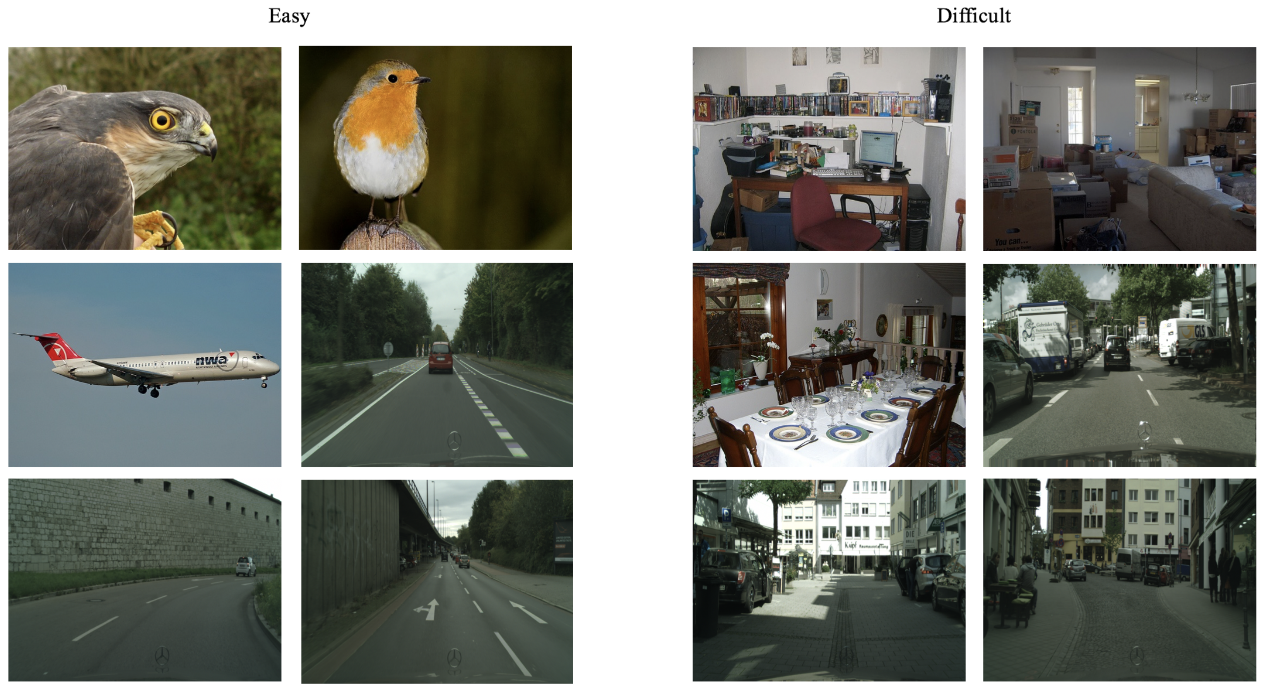


Figure 2. Easy and difficult images from Pascal VOC 2007 and Cityscapes according to our estimation.

of each class, then, for every image, we compute the image-level diversity by averaging the class score for each object in its ground-truth labels (Equation 4).

$$visited(c_i) = num_{objects}(c_i) - \frac{\sum_{c_j \in C} num_{objects}(c_j)}{|C|} \quad \forall c_i \in C. \tag{3}$$

$$imgVisited(x_i) = \frac{\sum_{obj \in objects(x_i)} visited(class(obj))}{|objects(x_i)|} \quad \forall x_i \in X. \tag{4}$$

In our diversity algorithm we want to emphasize the images containing objects from less visited classes, i.e. with a small $imgVisited$ value, closer to -1. We compute a scoring function similar to Equation 2, which also takes into consideration how often a class was visited, in order to add diversity:

$$\begin{aligned} w(x_i, t) = [1 - \alpha \cdot (diff(x_i) \cdot e^{-\gamma \cdot t}) \\ - (1 - \alpha) \cdot (imgVisited(x_i) \cdot e^{-\gamma \cdot t})]^k, \end{aligned} \tag{5}$$

where α controls the impact of each component, the difficulty and the diversity, while the rest of the notation follows Equation 2. We transform the weights into probabilities by dividing them by their sum, and we sample accordingly.

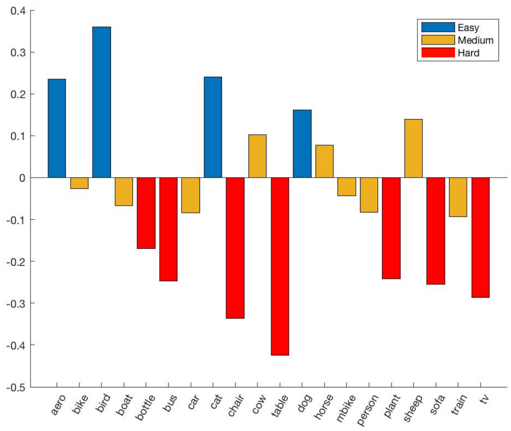


Figure 3. Difficulty of classes in Pascal VOC 2007 according to our estimation. Best viewed in color.

4 Experiments

4.1 Data sets

In order to test the validity of our method, we experiment on two data sets: Pascal VOC 2007 [4] and Cityscapes [3]. We conduct detection experiments on 20 classes, training on the 5011 images from the Pascal VOC 2007 trainval split. We perform evaluation on the test split which contains 4952 images. For our instance segmentation experiments, we use the Cityscapes data set which contains eight labeled object classes: person, rider, car, truck, bus, train, motorcycle, bicy-

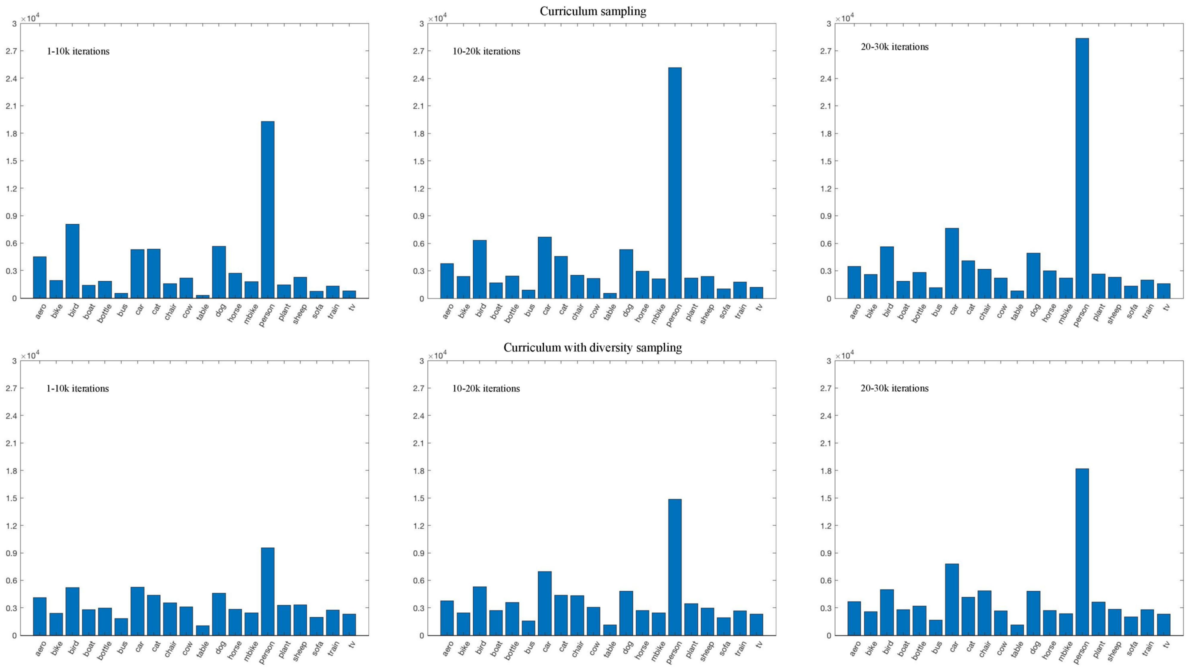


Figure 4. Number of objects from each class sampled during our training on Pascal VOC 2007. On the first row it is the curriculum sampling method and on the second row it is the curriculum with diversity approach. We present the first 30000 iterations for each case, with histograms generated from 10k to 10k steps.

cle. We train on the training set of 2975 images and we evaluate on the validation split of 500 images.

4.2 Baselines and configuration

We build our method on top of the Faster R-CNN [24] and Mask R-CNN [10] implementations available at: https://github.com/facebookresearch/maskrcnn-benchmark. For our detection experiments, we use Faster R-CNN with Resnet-101 [11] backbone, while for segmentation we employ the Resnet-50 backbone on the Mask R-CNN model. We use the configurations available on the web site, with the learning rate adjusted for a training with a batch size of 4. In our sampling procedure (Equation 5) we set $\alpha = 0.5$, $\gamma = 6 \cdot 10^{-5}$, and $k = 5$. We do not compare with other models, because the goal of our paper is not surpassing the state of the art, but improving the quality of our baseline model. We also present the results of a hard-to-easy sampling, in order to prove the efficiency of the easy-to-hard curriculum approaches inspired by human learning.

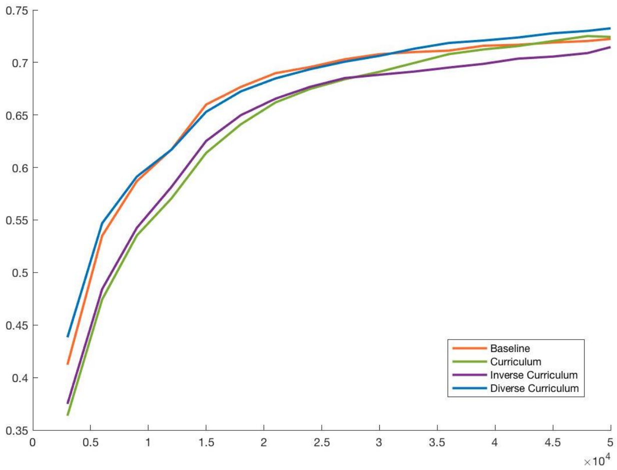


Figure 5. Evolution of mAP during training on Pascal VOC 2007 for object detection. Best viewed in color.

4.3 Evaluation metrics

We evaluate our results using the mean Average Precision (AP). The AP score is given by the area under the precision-recall curve for the detected objects. The Pascal VOC 2007 [4] metric is the mean of precision values at a set of 11 equally spaced recall levels, from 0 to 1, at a step size of 0.1. The Cityscapes [3] metric computes the average precision on the region level for each class and averages it across 10 different overlaps ranging from 0.5 to 0.95 in steps of 0.05. We also report results on Cityscapes using AP50% and AP75%, which correspond to overlap values of 50% and 75%, respectively. Since the exact evaluation protocol has some differences for each data set, we use the Pascal VOC 2007 [4] metric for the detection experiments and the Cityscapes [3] metric for the instance segmentation results. We use the evaluation code available at https://github.com/facebookresearch/maskrcnn-benchmark. More details about the evaluation metrics can be found in the original papers [3, 4].

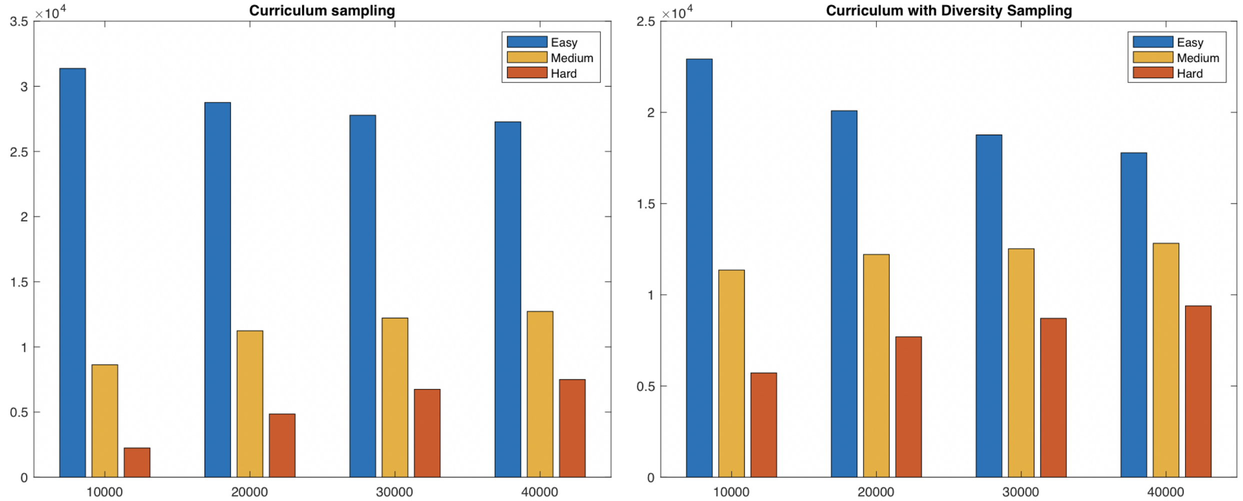


Figure 6. Difficulty of the images samples during our training on Pascal VOC 2007. On the left it is presented the curriculum sampling method and on the right the curriculum with diversity approach. We present the first 40000 iterations for each case, with histograms generated from 10k to 10k steps. Best viewed in color.

4.4 Results and discussion

The class distribution of the objects in Pascal VOC 2007 clearly favors class *person*, with 4690 instances, while classes *dinningtable* and *bus* only contain 215 and 229 instances, respectively. This would not be a problem if the difficulty of the classes was similar, because we can assume the test data set has a matching distribution, but this is not the case, as it is shown in Figure 3.

Figure 4 presents how the two sampling methods behave during training on the Pascal VOC 2007 data set. In the first 10k iterations, curriculum sampling selects images with almost 20k objects from class *person* and only 283 instances from class *diningtable*. By adding diversity, we lower the gap between classes, reaching 10k objects of persons and 1000 instances of tables. This behaviour continues as the training progresses, with the differences between classes being smaller when adding diversity. It is important to note that we do not want to sample the exact number of objects from each class, but to keep the class distribution of the actual data set, while feeding the model with enough details about every class. Figure 6 shows the difficulty of the examples sampled according to our strategies. We observe that by adding diversity we do not break our curriculum learning schedule, the examples still being selected from easy to hard.

To further prove the efficiency of our method, we compute the AP on both object detection and instance segmentation tasks. The results are presented in Tables 2 and 3.

We repeat our object detection experiments five times and average the results, in order to ensure their relevance. The sampling with diversity approach provides an improvement of 0.69% over the standard curriculum method, and of 0.79% over the randomly-trained baseline. Although the improvement is not large, we can observe that by adding diversity we boost the accuracy where the standard method would fail, without much effort. Our experiments, with an inverse curriculum approach, from hard to easy, lead to the worst results, showing the utility of presenting the training samples in a meaningful order, similar to the way people learn.

Moreover, Figure 5 illustrates the evolution of the AP during training. The curriculum with diversity approach has superior results over the baseline from the beginning to the end of the training. As the figure shows, the difference between the two methods increases in the later stages of the training. A simple reason for this behaviour is the fact that the curriculum strategy is fed with new, more difficult, examples as the training progresses, continuously improving the accuracy of the model. On the other hand, the standard random procedure receives all information from the beginning, reaching a plateau early during training. The standard CL method starts from lower scores, exactly because it does not visit enough samples from more difficult classes in the early stages of the training. For instance, after 5000 iterations, the AP of the standard CL approach on class *dinningtable* was 0. Thus, by adding diversity, our model converges faster than the traditional methods.

Table 2. Average Precision scores for object detection on Pascal VOC 2007 data set.

Model	mAP (in %)
Faster R-CNN (Baseline)	72.28 ± 0.34
Faster R-CNN with curriculum sampling	72.38 ± 0.32
Faster R-CNN with inverse curriculum sampling	70.89 ± 0.53
Faster R-CNN with diverse curriculum sampling	$\mathbf{73.07 \pm 0.28}$

Table 3. Average Precision scores for instance segmentation on Cityscapes data set.

Model	AP	AP50%	AP75%
Faster R-CNN (baseline)	38.72	69.15	34.95
Curriculum sampling	38.47	**69.88**	35.01
Inverse curriculum sampling	37.40	68.17	34.22
Diverse curriculum sampling	**39.12**	69.86	**35.4**

The instance segmentation results on the Cityscapes data set confirm the conclusion from our previous experiments. As Table 3 shows, the curriculum with diversity is again the optimal method, surpassing the baseline with 0.4% using AP, 0.71% using AP50%, and 0.45% using AP75%. It is interesting to point out that, although the diverse curriculum approach has a better AP and AP75% than the standard CL method, the former technique surpasses our method with 0.02% when evaluated using AP50%. The inverse curriculum approach has the worst scores again, strengthening our statements on the utility of curriculum learning and the importance of providing training examples in a meaningful order.

5 Conclusion and future work

In this paper, we presented a simple method of optimizing the curriculum learning approaches on unbalanced data sets. We consider that the diversity of the selected examples is just as important as their difficulty, and neglecting this fact may slow down training for more difficult classes. We introduced a novel sampling function, which uses the classes of the visited examples together with a difficulty score to ensure the curriculum schedule and the diversity of the selection. Our object detection and instance segmentation experiments conducted on two data sets of high interest prove the superiority of our method over the randomly-trained baseline and over the standard CL approach. A benefit of our methodology is that it can be used on top of any deep learning model, for any supervised task. Diversity can be a key element for overcoming one of the shortcomings of CL which can lead to the replacement of the traditional random training and a larger adoption of meaningful sample selection. For the future work, we plan on studying more difficulty measures to build an extensive view on how the chosen metric affects the performance of our system. Furthermore, we aim to create an ablation study on the parameter choice and find better ways to detect the right parameter values. Another important aspect we are considering is extending the framework to unsupervised tasks, by introducing a novel method of computing the diversity of the examples.

REFERENCES

[1] Dario et al. Amodei, 'Deep speech 2: End-to-end speech recognition in english and mandarin', in *Proceedings of ICML*, p. 173–182, (2016).

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, 'Curriculum learning', in *Proceedings of ICML*, pp. 41–48, (2009).

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, 'The cityscapes dataset for semantic urban scene understanding', in *Proceedings of CVPR*, (2016).

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[6] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang, 'Multi-modal curriculum learning for semi-supervised image classification', *IEEE Transactions on Image Processing*, **25**(7), 3249–3260, (2016).

[7] L. Gui, T. Baltrušaitis, and L. Morency, 'Curriculum learning for facial expression recognition', in *Proceedings of FG*, pp. 505–511, (2017).

[8] Junliang Guo, Xu Tan, Linli Xu, Tao Qin, Enhong Chen, and Tie-Yan Liu, 'Fine-tuning by curriculum learning for non-autoregressive neural machine translation', *arXiv preprint arXiv:1911.08717*, (2019).

[9] Guy Hacohen and Daphna Weinshall, 'On the power of curriculum learning in training deep networks', in *Proceedings of ICML*, (2019).

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, 'Mask r-cnn', in *Proceedings of ICCV*, pp. 2961–2969, (2017).

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of CVPR*, pp. 770–778, (2016).

[12] Radu Tudor Ionescu, Bogdan Alexe, Marius Leordeanu, Marius Popescu, Dim P. Papadopoulos, and Vittorio Ferrari, 'How hard can it be? estimating the difficulty of visual search in an image', in *Proceedings of CVPR*, pp. 2157–2166, (2016).

[13] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander Hauptmann, 'Self-paced learning with diversity', in *Proceedings of NIPS*, pp. 2078–2086, (2014).

[14] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann, 'Self-paced curriculum learning', in *Proceedings of AAAI*, (2015).

[15] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei, 'Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels', in *Proceedings of ICML*, pp. 2304–2313, (2018).

[16] Tom Kocmi and Ondřej Bojar, 'Curriculum learning and minibatch bucketing in neural machine translation', in *Proceedings of RANLP*, pp. 379–386, (2017).

[17] M Pawan Kumar, Benjamin Packer, and Daphne Koller, 'Self-paced learning for latent variable models', in *Proceedings of NIPS*, pp. 1189–1197, (2010).

[18] Siyang Li, Xiangxin Zhu, Qin Huang, Hao Xu, and C.-C. Jay Kuo, 'Multiple instance curriculum learning for weakly supervised object detection', in *Proceedings of BMVC*. BMVA Press, (2017).

[19] Cao Liu, Shizhu He, Kang Liu, and Jun Zhao, 'Curriculum learning for natural answer generation.', in *Proceedings of IJCAI*, pp. 4223–4229, (2018).

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, 'Ssd: Single shot multibox detector', in *Proceedings of ECCV*, pp. 21–37. Springer, (2016).

[21] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell, 'Competence-based curriculum learning for neural machine translation', in *Proceedings of NAACL*, pp. 1162–1172, (2019).

[22] Shivesh Ranjan and John HL Hansen, 'Curriculum learning based approaches for noise robust speaker recognition', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(1), 197–210, (2017).

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, 'You only look once: Unified, real-time object detection', in *Proceedings of CVPR*, pp. 779–788, (2016).

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, 'Faster r-cnn: Towards real-time object detection with region proposal networks', in *Proceedings of NIPS*, pp. 91–99, (2015).

[25] Mrinmaya Sachan and Eric Xing, 'Easy questions first? a case study on curriculum learning for question answering', in *Proceedings of ACL*, pp. 453–463, (2016).

[26] Enver Sangineto, Moin Nabi, Dubravko Culibrk, and Nicu Sebe, 'Self paced deep learning for weakly supervised object detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**(3), 712–725, (2018).

[27] Miaojing Shi and Vittorio Ferrari, 'Weakly supervised object localization using size estimates', in *Proceedings of ECCV*, pp. 105–121. Springer, (2016).

[28] Petru Soviany, Claudiu Ardei, Radu Tudor Ionescu, and Marius Leordeanu, 'Image difficulty curriculum for generative adversarial networks (cugan)', in *Proceedings of WACV*, (2020).

[29] Petru Soviany and Radu Tudor Ionescu, 'Frustratingly Easy Trade-off Optimization between Single-Stage and Two-Stage Deep Object Detectors', in *Proceedings of CEFRL Workshop of ECCV*, pp. 366–378, (2018).

[30] Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky, 'Baby Steps: How "Less is More" in unsupervised dependency parsing', in *Proceedings of NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*, (2009).

[31] Sandeep Subramanian, Sai Rajeswar, Francis Dutil, Christopher Pal, and Aaron Courville, 'Adversarial generation of natural language', in *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 241–251, (2017).

[32] James S Supancic and Deva Ramanan, 'Self-paced learning for long-term tracking', in *Proceedings of CVPR*, pp. 2379–2386, (2013).

[33] Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller,

‘Shifting weights: Adapting object detectors from image to video’, in *Proceedings of NIPS*, pp. 638–646, (2012).

[34] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan, ‘Dynamic curriculum learning for imbalanced data classification’, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (October 2019).

[35] Daphna Weinshall and Gad Cohen, ‘Curriculum learning by transfer learning: Theory and experiments with deep networks’, in *Proceedings of ICML*, (2018).

[36] Dingwen Zhang, Junwei Han, Long Zhao, and Deyu Meng, ‘Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework’, *International Journal of Computer Vision*, **127**(4), 363–380, (2019).

[37] Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat, ‘An empirical exploration of curriculum learning for neural machine translation’, *arXiv preprint arXiv:1811.00739*, (2018).

Observation of Communicative Behaviour when Learning a Movement Sequence: Prequel to a Case Study

Julian Blohm[1] and **Jörg Cassens**[2] and **Rebekah Wegener**[3]

Abstract. When trying to improve human-machine communication it can be helpful to better understand human thinking and behaviour. In some cases, it is not only feasible, but also helpful to transfer recognised communicative patterns to machine interaction. The benefits of multimodal interfaces have been explored for quite some time, arguably starting with the famous "put that there!" demonstration system [4], leading to a variety of theoretical works and application systems [14]. However, there is still a lot of work to be done before non-verbal elements of communication can challenge the predominant paradigms for human-computer interaction [18, 35]. We have previously worked on multimodal behaviour in specific contexts of interaction [17] and on explanation-aware systems [16] as well as a combination thereof [8]. In order to better understand which aspects of human-to-human communicative behaviour can (at least) be mimicked by computational systems, we perform empirical research with humans in this area. In this paper, we present a pre-study for an experimental setup that looks at human-to-human communicative behaviour during movement sequence learning. This will enable us to better understand the role of different features in explanatory behaviour. In the end, a better understanding of this behaviour will hopefully enable us to optimize human-machine interaction as it pertains to explainable AI and might aid the development of better training systems for learning complex motor skills in high risk environments such as surgery or emergency medicine.

1 Research Questions

An increasing number of tasks in all walks of life are being taken on or supported at least partially by technology e.g. learning in high risk environments like surgery, where learning a new complex motor skill can be essential to saving life, but where learning by doing is life threatening [21, 40]. An important aspect here is the notion of cooperative systems, mixed-initiative systems or, more general, the notion of "human-in-the-loop" [34, 42].

For the often implied sharing of tasks between humans and machines to be effective, it is necessary that the exchange of information between human and machine runs smoothly. While it has been (and in some cases still is) common to model humans as information processing systems [6, 19], which means that they perceive signals from stimuli through the sensory perception system, process that information through the cognitive processing system and finally act on that information; human information processing is quite distinct from machine data processing. Despite the different capabilities and potentials, however, a better understanding of human communicative behaviour will perhaps enable us to build systems for better communication between humans and machines [39].

Communication is so much more than spoken or written language. Natural language is inherently multimodal in nature [36]. Because of this, the classic transmitter-receiver model of information processing that is often used in computer science is typically extended to include other modalities depending on the needs of the research [27, 32]. In natural interaction, the progression and the success or failure of the interaction can be shaped by many different factors including behavioural or contextual elements [27, 29].

The underlying research program of which this paper is a part aims to investigate whether the consideration of behavioural and contextual elements can provide insights that can be used for the optimization of future explanation-aware systems. To this end, an experimental setup was developed in a pre-structured explanatory situation. In this experimental domain, the test subjects' goal is to learn a behavioural sequence that is indicative of complex motor skill learning, in this instance a Judo technique. The aim is to design the instructional material in such a way that it is relatively self explanatory, making verbal communication superfluous. Non-verbal behaviours such as gestures, facial expressions and body movement are observed and the communicative behaviour is recorded as accurately and unobtrusively as possible. This allows for the analysis of the timing and potential motivation for additional communication and the consideration of how this might relate to contextual and individual factors.

2 Human-Machine Communication

Human-machine communication (HMC) refers to the mutual Information exchange between human and machine [41]. This means the "intuitive"[4], natural, and therefore multimodal interaction between people and information processing systems. Early textual chatbots such as Eliza [38] mainly responded to keywords or phrases and answered with canned responses. This was then amended using template-based systems [5].

By now, systems using spoken natural language and learned models have become mainstream. Technologies such as Google Duplex, Alexa (Amazon) and Siri (Apple) respond to questions and answer them appropriately, even mimicking non-task oriented aspects of human communication. For example, Google Duplex uses typical human behaviours like a short pause for reflection between sentences or uttering "hm" [22].

Turning to other modalities than spoken (or written) language, modern sensor technology in principle opens up the potential for

[1] University of Hildesheim, Germany, email: blohmj@uni-hildesheim.de
[2] University of Hildesheim, Germany, email: cassens@cs.uni-hildesheim.de
[3] University of Salzburg, Austria, email: rebekah.wegener@sbg.ac.at

[4] Intuitive is used here in a cultural-historic sense and is not referring to an assumed inherent property.

simulation of communication that is comparatively close to human-to-human communication [5]. Despite these improvements, communication does not always running smoothly.

3 Human-Human Communication

Interpersonal communication can be described by various linguistic, semiotic or communication models [20]. In the pre-study described here, we focused on the characteristics of communication approach [33] as well as an integrative model of communication [27]. Communication, therefore, is here understood as a process that arises through interactions. Verbal and non-verbal elements such as gestures, facial expressions and body language are used.

Decisive for the unfolding of the communication process is the respective context, especially personal and situational context. Besides observable elements, non-visible activities determine communication behaviour (communication rules, sympathy, tenor, prejudices). Basic prerequisite for successful communication is the use of a common repertoire of signs and symbols by the communication partners. Nevertheless, misunderstandings and errors can occur when coding and decoding a message. The overall course of events is influenced by contextual and psychological factors. The objectives of a communication, the response and feedback behaviour, and the mutual perception also influence the course [27]. These factors should be taken into account when planning the empirical study.

According to Watzlawick, humans will always communicate even if they don’t intend to communicate [37]. Thus every behaviour has a communicative character. Part of the non-verbal side of communication pertains to affect. Body language is related to individual variation and the situation. However, it is not possible to draw conclusions about the emotions of the communication partner by interpreting a single body language expression. Not only are they not unique in themselves, but we will always only see the expression of affect, and not the underlying emotion. Facial expressions vary individually, contextually and culturally, therefore other elements are usefully included [3], for instance, our voice contains important and surprisingly reliable information about our emotional state [25].

4 Planning

In a random sample, test subjects are to learn a Judo technique, i.e. a complex motor skill in the form of a motion sequence. Instruction on how to perform this motion sequence is given via text, video, and photo sequences. The respective learning steps are evaluated when the motion sequence is enacted afterwards. The study consisted of two phases, a small pilot to test the experimental protocol, and the case study itself. For the remainder of this article, we will focus on the small pilot phase and the process leading up to the experiment.

In the preparatory phase, the focus lies on reflections on the method, the context of situation, the explanatory materials, and the evaluation strategy. Influencing factors and barriers which may complicate the course of communication are to be considered. These preliminary considerations are then evaluated in test runs checked and corrected. The trainer is part of the communication process and since the test subject and the trainer together determine the course of communication, the behaviour of the trainer has to be taken into consideration as well.

4.1 Method

4.1.1 Participatory observation

Participatory observation was chosen as the method of data collection as this is a standard method of field research and thus offers a point of comparison [11, 23, 28]. During the procedure, two observers recorded the behaviour of the participants and the trainer. They used pre-formulated observation sheets with the option to note down individual remarks. In addition, the trainer wrote down their observations after the exercise using a memory protocol.

The multi-perspective data collection (trainer, respondent, observers) served to achieve comprehensive observations by relating data points to each other and allowing them to be corrected if necessary. In the run-up to the project, the aim was to take into account (and where possible control) all factors that could plausibly have an influence on the result and thus on the reliability of the data to be collected. For example, the context (place, time, atmosphere), the behaviour of the trainer and the observers, and the structuring of the execution was precisely defined. By pre-structuring the observation sheets, the focus of the observers was specifically directed to essential aspects in contrast to free wording (validity). Elements of movement, verbal expressions and observations on the execution of the Judo technique were recorded.

All observations were made with the same observers and in the same room. Those carrying out the observations kept an unobtrusive external appearance. The test persons were addressed randomly and did not have any personal relationship to the persons performing the observations. In order to achieve reliable results, the test persons had no prior knowledge or reservations. To ensure this, a preparatory questionnaire was used. Using teaching material that was produced specifically for the task, the observation can be repeated reliably. Since the trainer was also part of the exercise, various safeguards were put it place to ensure consistency over the course of the experiment. The behaviour of the trainer was precisely defined and was also checked by an observer. With the help of the reflection sheet, observations made in different runs could then be compared.

4.1.2 Selection of the object of explanation: Learning a Judo technique

For the analysis of the non-verbal communication elements, learning of a movement sequence was chosen. In contrast to a purely cognitive learning situation, it can be clearly seen whether the respondent has understood the given explanations by following the exercise in action. The fact that understanding and learning has taken place can be demonstrated by the action itself [26].

While the guidelines for the correct execution of Judo technique by the German Judo Association (DJB) [10] were taken into account, they were applied in a modified form because participants in this study were complete novices. The use of the DJB guidelines however gave a consistent and detailed measure for evaluating. The didactic structure of a training unit is familiar to the first author of this article who takes part in the experiment as a trainer. He has been active in Judo himself for about 20 years and has been active as a trainer for 5 years. In his role as a Judo trainer, he has to be able to teach the Judo techniques in an understandable way.

His personal experience that the exclusive use of simple statements (verbalizations), pictures (visualizations of throwing techniques), or even throwing descriptions in text form are not sufficient is consistent with the literature [13, 21, 24] and translates to complex motor skill learning in other disciplines than sports [9, 40]. Often, a combination of different explanation strategies are used and Judo instructors generally teach the technique using the following steps:

- Verbal explanation,
- Demonstration
- In sequences with explanations
- Clarifying demands
- Practice phase with individual help

Even if the underlying mechanics and movements are understood in principle, when learning a new complex motor skill it is not unusual to initially have difficulties in performing it correctly. If necessary, the technique should be explained again or shown repeatedly. The motor skill chosen for this study was the “joint lock” because it does not require any previous knowledge or additional equipment. With an arm joint lock it is important to fix the elbow joint of the partner and then overstretch it.

4.1.3 Multimodality, sequential explanations and action

The instructional material was presented to the participants in digital form using the keynote presentation software. It consisted of 10 pages, 5 photos and 5 videos. It was designed to be self-explanatory so that the verbal communication components were reduced. The focus of the observation was on non-verbal behaviour and movement elements.

The sequence of movements to be learned was broken down into individual learning sequences, which are modelled on the normal training situation in Judo practice session. The acquisition phase was followed by an action phase, in which participants act and practise what they have learned. The training texts were written such that the participants were directly addressed and could identify more easily with their role. The texts were kept simple and were developed as an instruction manual. The written description of the movement sequences is supplemented with photos and videos. The photos show the current state or the initial situation and details. The video provides the movement sequence. All pages are structured identically to provide the participants with an easier orientation of the learning path.

4.2 Setting

4.2.1 Place and Time

The location of the study has an influence on the mood and motivation of the participants [15]. For this reason, a room on the premises of the University of Applied Sciences and Arts Hanover was chosen as it is a simple, small working room that is located on the fifth floor with little disturbance from outside noise. Students of the university are familiar with this type of room and the choice of a workspace as opposed to a private room or training facility provided the experiment with a quiet, neutral space. A clock was not visible so that no time pressure was built up and sessions were scheduled in the early evening or on weekends, so that the participants arrived relatively rested.

4.2.2 Atmosphere

The explanatory situation tested was a learning situation with clearly defined roles. The trainer is the instructor, the participant has the role of a student. Stress elements contained in this situation were alleviated by the surrounding atmosphere. The behaviour of the trainer played an important role in creating a pleasant and open atmosphere. In order to enable the participants to act as relaxed and pressure-free as possible, the appearance, clothing, language style, posture etc. of both the trainer and the observers were prescribed before running the experiments [2, 12].

These considerations were confirmed in the test runs where all participants noted that they felt comfortable in the situation and even enjoyed it.

4.2.3 Observation

As the experiment aimed for a relaxed atmosphere as close to everyday life as possible, observers were used for both external and self-observation. They went directly into the setting, actively participating and writing notes which are then evaluated. They were briefed and trained beforehand [30]. The use of cameras was deliberately avoided because the awareness of being under observation can lead to changes in behaviour (Hawthorne Effect) [1].

In practice runs before the small pre-test, the observers were trained in the handling of the different observation sheets. It turned out that the observers were able to follow the practice runs well and that the pre-defined structure of the observation sheets was helpful. The overall impression and individual peculiarities could be easily recognized and recorded.

Nevertheless, some details were missed. In contrast to the planned setup, a recording device was deemed necessary in order to record verbal utterances instead of transcribing them on-the-fly. This was done by using a mobile phone during the later runs. According to the test subjects, this small, inconspicuous camera was not noticeable or even perceived as disturbing.

4.2.4 Selection of participants

The test subjects were recruited directly and invited to participate after a short eligibility interview. The following selection criteria were established:

- Age group 18 and above (legal adult).
- Body height, approx. between 1.70-1.90m.
- The potential participants should have an average physical fitness.
  - A movement exercise is carried out with the persons addressed in order to test their coordination and movement skills (opposite windmill arm movement).
- Good German language skills are necessary, as texts must be read and understood.
- No previous knowledge of Judo or wrestling, determined by means of a questionnaire.
- Persons who do not wish to be touched or who do not agree with the general conditions of the experiment are also excluded.

For organisational and technical reasons, the participants were recruited at the university campus in Hanover. The total of 10 participants were young adults.

4.3 Execution

4.3.1 Procedure

The trainer invited the participants and gave an initial overview of the goals and progress of the experiment. The participants filled in questionnaires I (personal) before, and II (feedback) after the experiment. The explanatory material alternates between acquisition and action phases. The participant could scroll forward or backward and repeat individual parts at any time. The trainer was available as a contact person for questions and interaction during the entire process and operated the PC. Two observers filled out observation sheets of the communication partners A1 and A2 (participants) and B1 and B2 (trainer). In parallel, video recordings were made using a mobile phone. Immediately after completion of the experiment, the trainer completed a memory protocol C on their own perceptions.

4.3.2 *Questionnaires*

The questionnaire I (personal questions), was handed out to the participants before the movement task was performed. The exclusion criteria for the selection of participants were checked and personal data was queried. Following to the integrative communication model [27], potential influencing factors such as previous knowledge, motives, age, gender, etc. were taken into account. Volunteers were asked about their ability to understand instruction manuals because the judo technique is essentially developed step by step, similar to an instruction manual.

Questionnaire II (feedback) was given to the participants immediately after the practice task had been executed in order to record the immediate experience. Questions were asked about the Judo technique, the instructional material and the general conditions. When filling out the questionnaire, the test persons had the materials at their disposal. The feedback was intended to point out possible restructuring necessities for later follow-up studies. For example, the test runs performed showed that some changes in the design had to be made in order to achieve clarity. Also the detailed demand for previous knowledge of certain martial arts was reformulated into a more general question.

In addition, Questionnaire II asked for a self-assessment and inquires whether additional help was necessary both in terms of understanding the material and performing the movement. Implicit in the answers given is a distinction between whether the respondent asked for help of their own accord or whether the trainer intervened proactively. Since the trainer is an essential part of the exercise, their behaviour was described from the test person‘s perspective. In Questionnaire II, the test person also gave a self-assessment of the degree of difficulty and whether they needed help with the exercise.

4.3.3 *Observation sheets*

The behaviour of the participant and of the trainer was recorded in separate observation sheets A and B. There is one observation sheet each for the acquisition phase (A1 and B1) and a second for the action phase (A2 and B2).

The observation sheets were pre-formulated according to selected criteria (verbal language, gestures, facial expressions, movement) and serve as an aid for the observer. They follow the chronological sequence and repeat the fields of observation for the individual sections in the same way. The pre-formulated fields of observation should enable the observer to note many aspects in as short a time as possible. There is room for individual remarks so that the observers can record unforeseen events. Nevertheless, the pre-formulated aspects ensure a structured approach, especially for later evaluation.

The action part was mainly recorded using observation sheets A2 and B2. In A2 the observer recorded descriptions in general form for implementation of the movement. In addition, aspects about the transition from acquisition to practice were recorded. The focus here was on the manner of implementation, i.e. whether the test subject starts hesitantly or actively. The trainer describes the non-verbal or verbal communication behaviour during the action phase.

The study uses a semi-standardised procedure, since it works with pre-formulated criteria, but it also leaves room for the recording of new aspects. In addition, the video recordings were available for comparison.

4.3.4 *Reflection sheet*

The reflection sheet C was filled in by the trainer directly after the execution. The questions were answered spontaneously and reflect the first impression. The first questions refer to the execution of the judo technique. From the perspective of the experienced judo trainer, the extent to which the technique is executed correctly was assessed and the process of learning was also examined. Afterwards, the relationship between subject and trainer was described, especially its subjective impact. Attention was paid to the application of additional help, when and why was this necessary, how help was given and whether it was successful.

4.4 Evaluation

The evaluation was derived from the observers‘ notes, the video recordings and the trainer‘s reflection sheet C. The questionnaires filled in by the test persons supplemented the data collected. Similarities and differences in the observations were interpreted and analysed. In this way the observations on handling of material, the method and the course of communication can be viewed from different perspectives. This is intended to achieve the highest possible degree of coverage.

When describing the course of communication, the verbal and non-verbal remarks were recorded. The focus here is on the questions of *when*, *what* and *how* it was communicated. The verbal comments are clearly recognizable and can be written down. The non-verbal communication results from the context and the behavioural elements. Every “additional communication” was recorded. First of all, a time stamp is noted, i.e. when the communication took place. In a second step the cause was examined. This resulted in the following areas for the evaluation:

1. general personal data for the classification of the test subject
2. recording of personality and behavioural characteristics (situational and context-related)
3. situation/atmosphere
4. time, an average value is calculated
5. information part: handling of the materials/method
6. linguistic comments
7. body language
8. action part

Each test subject was described individually. The self-reported aspects and the observed behaviour were related to observed non-verbal communication behaviour. Hypotheses could then be formed as to whether the non-verbal additional communication was due to the inter-personal differences, the material, or the situation. The self-assessments were always related to the observed data.

5 First observations from trial runs

Two trial runs were carried out and these test runs were intended to familiarise the observers and the trainer with the use of the observation sheets and with the flow of the test.

Overall, it was found that the planned procedure was reasonable and practicable. Materials offered proved to be sufficient for the participants. The test subjects were able to understand them and implement the motions correctly. In the end, the participants were able to successfully perform the judo technique. They were satisfied with their results and considered this learning path an acceptable alternative to classical Judo training.

Additionally, the multimodal explanation strategy, the decomposition of the movement sequence to be learned into sequences, and the alternation of acquisition and action phases, has proven to be useful.

Participants confirmed that they felt comfortable and enjoyed it. This indicates an overall relaxed atmosphere. After the introduction in the first action part, participants wanted to perform the whole movement sequence immediately.

The trainer had to intervene at this point and point out that only the sequences shown should be practised. Here, the instructions by the trainer had to be optimized so that the sequence would be clearer. Participants had to be encouraged to switch to the first action phase. Hesitation was signalled by eye contact and by waiting, indicating that the test subjects needed some form of interactive response. This despite the fact that transition from acquisition to action is signalled in the training material in such a way that execution could in principle take place without any intervention by the trainer.

The need for interactive response could indicate that there is a specific need for communication and information, especially in the initial phase of becoming familiar with the learning path. Although the explanatory material and the trainer‘s presentations contain a lot of relevant information, this did not seem to be sufficient for the participants during the acquisition phase.

In contrast, the need for eye contact during the action phase is likely a result of the setup, as the technique is a partner task and it is necessary to respond to each other. So the search for eye contact can here be interpreted as coordination during the execution. The trainer reported that participants tend to react affirmatively to the search for eye contact. Furthermore, eye contact was a frequently occurring behavioural signal.

The evaluation of the different body signals, which of course are to be understood contextually, already suggest that the trainer should respond adaptively to different test subjects. For example, test subjects that are very cautious and reserved in the execution phase would need encouragement in the action phase for a more courageous and powerful execution. It is important to note that this encouragement can be shared using non-verbal cues.

6 Conclusions and Further Work

The goal of this experiment was two fold: to test the experimental protocol for the larger study and to see what aspects of human-to-human interaction might be useful for designing and developing for human-to-machine interaction, particularly for explainable AI and training systems for high risk environments. In terms of testing the protocol, a number of aspects of the pre-study are being revised for the larger study. It should be noted that the pre-study showed deficiencies in our experiment protocol that will be rectified. Individual behaviour of the participants could not be fully recorded and transcribed. The observers also made individual judgements and set priorities themselves and this added discrepancies in the evaluations. It is difficult to counter this effect, but it may be necessary to improve the training cycle for the observers. After all, ethnographic recording is a skill that itself requires a lot of practice. While the observation can not be considered representative due to the small number of test subjects (10), the pre-study provided crucial learning for the revision of the larger study and it was also possible to obtain results that were useful and indicative in nature.

In the test runs of the designed study, non-verbal behaviour of the test subjects was transcribed in addition to verbal comments. Test subjects showed different communication needs, which could be read from behavioural cues. The non-verbal behaviour could also be seen as expressions of inter-personal difference. Thus, for example, uncertainties that are shown through behaviour could be reacted to accordingly.

Even although participants worked independently with the training material and did not verbally ask for help, situations were identified in which they signalled a need to communicate, e.g. by eye contact or waiting [31]. It is helpful for explanatory systems, be they human or technological, to react to this behaviour.

The prevalence and diversity of situations where eye contact played a crucial role in the interaction is indicative that a richer model of gaze might be beneficial for upcoming studies, particularly since gaze is a feature that can readily be captured by existing sensors.

In the initial development of a situation where cooperation of multiple entities is central (collaborative or team work situations), the need for additional communication is higher, so that the process and the procedure are understood and mutual trust is created.

In failure situations where corrective action and explanations are necessary, an appropriate communication strategy that includes multimodal feedback should be developed so that users do not give up in frustration or fail to recognize the error at all. An incidental finding is that it appears from this experiment that impending frustration and possible abandonment of learning can be predicted from the behaviour before it occurs so that an intervention might be possible. This is consistent with findings in other work we have been done on multimodal markers of importance [7].

It was very clear from the experiment that explanations should be offered multimodally and, depending on the subject, also sequentially. Repetitions and some redundancy, if necessary also in variations, help participants to habituate to working methods and provide security and ultimately build a trust relationship.

Within a human-centred approach to intelligent systems development, the better a system knows its user, the better it can potentially respond to them and their individual needs. The experiment showed that by taking behavioural elements into account, it is possible to get to know the user or participant better. From the recognition of the individual needs for assistance, appropriate communication strategies can be designed.

References

[1] Renee L. Allen and Andrew S. Davis, *Hawthorne Effect*, 731–732, Springer US, Boston, MA, 2011.

[2] Michael Argyle, Adrian Furnham, Jean Ann Graham, et al., *Social situations*, Cambridge University Press, 1981.

[3] Thomas Bøgevald Bjørnsten and Mette-Marie Zacher Sørensen, ‘Uncertainties of facial emotion recognition technologies and the automation of emotional labour’, *Digital Creativity*, **28**(4), 297–307, (2017).

[4] Richard A. Bolt, ‘“put-that-there”: Voice and gesture at the graphics interface’, in *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH ’80, p. 262–270, New York, NY, USA, (1980). Association for Computing Machinery.

[5] Luka Bradeško and Dunja Mladenić, 'A survey of chatbot systems through a loebner prize competition', in *Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies*, pp. 34–37, (2012).

[6] Stuart Card, Thomas P. Moran, and Allen Newell, 'The model human processor - an engineering model of human performance', *Handbook of perception and human performance.*, **2**(45–1), (1986).

[7] Jörg Cassens and Rebekah Wegener, 'Supporting students through notifications about importance in academic lectures', in *Proceedings of AmI 2018 – International Joint Conference on Ambient Intelligence*, eds., Achilleas Kameas and Kostas Stathis, volume LNCS, pp. 227–232, Larnaca, Cyprus, (November 2018). Springer. ISBN: 978-3-030-03061-2.

[8] Jörg Cassens and Rebekah Wegener, 'Ambient explanations: Ambient intelligence and explainable ai', in *Proceedings of AmI 2019 – European Conference on Ambient Intelligence*, eds., Ioannis Chatzigiannakis, Boris De Ruyter, and Irene Mavrommati, volume LNCS, Rome, Italy, (November 2019). Springer.

[9] Ligia Cordovani and Daniel Cordovani, 'A literature review on observational learning for medical motor skills and anesthesia teaching', *Advances in Health Sciences Education*, **21**(5), 1113–1121, (2016).

[10] Hannes Daxbacher, Klaus Hanelt, Roman Jäger, Klaus Keßler, Ulrich Klocke, Ralf Lippmann, Rudi Mieth, Hans Müller-Deck, Jan Schröder, Mario Staller, Hans-Jürgen Ulbricht, and Franz Zeiser. Begleitmaterial zum dan-prüfungsprogramm - ein nachschlagewerk zu verschiedenen themen der dan-prüfungsordnung im deutschen judo bund e.v., 4. überarbeitete auflage, 2011. [Online; Stand 28.06.2020].

[11] Uwe Flick, 'Qualitative sozialforschung: Eine einführung (4. aufl., vollst. überarb. und erw. neuausg.)', *Rororo Rowohlts Enzyklopädie*, **55654**, (2006).

[12] Owen Hargie, 'Interpersonal communication: A theoretical framework.', in *The handbook of communication skills*, 29–63, Psychology Press, (1997).

[13] Avi Karni, Gundela Meyer, Christine Rey-Hipolito, Peter Jezzard, Michelle M. Adams, Robert Turner, and Leslie G. Ungerleider, 'The acquisition of skilled motor performance: Fast and slow experience-driven changes in primary motor cortex', *Proceedings of the National Academy of Sciences*, **95**(3), 861–868, (1998).

[14] Athanasios Katsamanis, Vassilis Pitsikalis, Stavros Theodorakis, and Petros Maragos, *Multimodal Gesture Recognition*, 449–487, Association for Computing Machinery and Morgan & Claypool, 2017.

[15] Naz Kaya and Feyzan Erkip, 'Invasion of personal space under the condition of short-term crowding: A case study on an automatic teller machine', *Journal of Environmental Psychology*, **19**(2), 183–189, (1999).

[16] Anders Kofod-Petersen and Jörg Cassens, 'Explanations and context in ambient intelligent systems', in *Modeling and Using Context – CONTEXT 2007*, eds., Boicho Kokinov, Daniel C. Richardson, Thomas R. Roth-Berghofer, and Laure Vieu, volume 4635 of *LNCS*, pp. 303–316, Roskilde, Denmark, (2007). Springer.

[17] Anders Kofod-Petersen, Rebekah Wegener, and Jörg Cassens, 'Closed doors – modelling intention in behavioural interfaces', in *Proceedings of the Norwegian Artificial Intelligence Society Symposium (NAIS 2009)*, eds., Anders Kofod-Petersen, Helge Langseth, and Odd Erik Gundersen, pp. 93–102, Trondheim, Norway, (November 2009). Tapir Akademiske Forlag. ISBN: 978-8-2519-2519-8.

[18] Stefan Kopp, 'Giving interaction a hand: Deep models of co-speech gesture in multimodal systems', in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, p. 245–246, New York, NY, USA, (2013). Association for Computing Machinery.

[19] Karl-Friedrich Kraiss, 'Der mensch als kognitive systemkomponente', in *Fahrzeug-und Prozeßführung*, 26–70, Springer, (1985).

[20] Robert M Krauss and Susan R Fussell, 'Social psychological models of interpersonal communication', *Social psychology: Handbook of basic principles*, 655–701, (1996).

[21] Danielle E Levac, Meghan E Huber, and Dagmar Sternad, 'Learning and transfer of complex motor skills in virtual reality: a perspective review', *Journal of NeuroEngineering and Rehabilitation*, **16**(1), 121, (2019).

[22] Yaniv Leviathan and Yossi Matias, 'Google duplex: an ai system for accomplishing real-world tasks over the phone', *Google AI blog*, **8**, (2018).

[23] Philipp Mayring, *Einführung in die qualitative Sozialforschung*, Beltz Verlag, 2002.

[24] Karl M Newell, 'Motor skill acquisition', *Annual review of psychology*, **42**(1), 213–237, (1991).

[25] Virginia P Richmond, James C McCroskey, and Mark Hickson, *Nonverbal behavior in interpersonal relations*, Allyn & Bacon, 2008.

[26] Ulrike Rockmann, 'Bewegungen verstehen und beherrschen', *Vom Lernen zum Lehren: Lern-und Lehrforschung für die Weiterbildung*, 159, (2006).

[27] Jessica Röhner and Astrid Schütz, *Psychologie der Kommunikation*, Springer-Verlag, 2015.

[28] Gabriele Rosenthal. Interpretative sozialforschung. eine einführung. juventa, 2005.

[29] N. Santiano, L. Young, L.S. Baramy, R. Cabrera, E. May, Rebekah Wegener, David Butt, M. Parr, and Clinical Analysis Group, 'The impact of the medical emergency team on the resuscitation practice of critical care nurses', *BMJ Quality and Safety*, **20**(2), 115–120, (2011).

[30] Helmar Schöne, 'Participant observation in political science: Methodological reflection and field report', in *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, volume 4, (2003).

[31] Beatrix Schönherr, *Syntax-Prosodie-nonverbale Kommunikation: empirische Untersuchungen zur Interaktion sprachlicher und parasprachlicher Ausdrucksmittel im Gespräch*, volume 182, Walter de Gruyter, 2013.

[32] Claude Elwood Shannon, *The Mathematical Theory of Communication, by CE Shannon (and Recent Contributions to the Mathematical Theory of Communication), W. Weaver*, University of illinois Press, 1949.

[33] *Kommunikationspsychologie—Medienpsychologie*, eds., Ulrike Six, Uli Gleich, and Roland Gimmler, Beltz, 2007.

[34] Norbert Streitz, 'Reconciling humans and technology: The role of ambient intelligence', in *Ambient Intelligence*, eds., Andreas Braun, Reiner Wichert, and Antonio Maña, pp. 1–16, Cham, (2017). Springer International Publishing.

[35] Matthew Turk, 'Multimodal interaction: A review', *Pattern Recognition Letters*, **36**, 189–195, (2014).

[36] Gabriella Vigliocco, Pamela Perniss, and David Vinson, 'Language as a multimodal phenomenon: Implications for language learning, processing and evolution', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **369**, (09 2014).

[37] Paul Watzlawick, 'Die axiome von paul watzlawick', *Unter: http://www. paulwatzlawick. de/axiome. html [26.06. 2011] S*, **38**, 60, (2016).

[38] Joseph Weizenbaum, 'Eliza—a computer program for the study of natural language communication between man and machine', *Communications of the ACM*, **9**(1), 36–45, (1966).

[39] *Dorsch Lexikon der Psychologie*, ed., Markus Antonius Wirtz, Verlag Hans Huber, 2020. [Online; Stand 13.06.2020].

[40] Gabriele Wulf, Charles Shea, and Rebecca Lewthwaite, 'Motor skill learning and performance: a review of influential factors', *Medical education*, **44**(1), 75–84, (2010).

[41] Tang Yuan Yan, Wang Patrick SP, and Yuen Pong Chi, *Multimodal interface for human-machine communication*, volume 48, World Scientific, 2002.

[42] Fabio Massimo Zanzotto, 'Human-in-the-loop artificial intelligence', *Journal of Artificial Intelligence Research*, **64**, 243–252, (2019).