Proceedings of Umeå's 23rd Student Conference in Computing Science

USCCS 2019

S. Bensch, T. Hellström (editors)

UMINF 19.02
ISSN-0348-0542

Department of Computing Science
Umeå University

Preface

The Umeå Student Conference in Computing Science (USCCS) is organized annually as part of a course given by the Computing Science department at Umeå University. The objective of the course is to give the students a practical introduction to independent research, scientific writing, and oral presentation.

A student who participates in the course first selects a topic and a research question that he or she is interested in. If the topic is accepted, the student outlines a paper and composes an annotated bibliography to give a survey of the research topic. The main work consists of conducting the actual research that answers the question asked, and convincingly and clearly reporting the results in a scientific paper. Another major part of the course is multiple internal peer review meetings in which groups of students read each others' papers and give feedback to the author. This process gives valuable training in both giving and receiving criticism in a constructive manner. Altogether, the students learn to formulate and develop their own ideas in a scientific manner, in a process involving internal peer reviewing of each other's work and under supervision of the teachers, and incremental development and refinement of a scientific paper.

Each scientific paper is submitted to USCCS through an on-line submission system, and receives reviews written by members of the Computing Science department. Based on the review, the editors of the conference proceedings (the teachers of the course) issue a decision of preliminary acceptance of the paper to each author. If, after final revision, a paper is accepted, the student is given the opportunity to present the work at the conference. The review process and the conference format aims at mimicking realistic settings for publishing and participation at scientific conferences.

USCCS is the highlight of the course, and this year the conference received 10 submissions (out of a possible 21), which were carefully reviewed by the reviewers listed on the following page.

We are very grateful to the reviewers who did an excellent job despite the very tight time frame and busy schedule. As a result of the reviewing process, 7 submissions were accepted for presentation at the conference. We would like to thank and congratulate all authors for their hard work and excellent final results that are presented during the conference.

We wish all participants of USCCS interesting exchange of ideas and stimulating discussions during the conference.

Umeå, 8 January 2019

Suna Bensch
Thomas Hellström

Organizing Committee

Suna Bensch
Thomas Hellström

With special thanks to the reviewers

Neha Baranwal
Suna Bensch
Johanna Björklund
Patrik Eklund
Jerry Eriksson
Thomas Hellström
Timotheus Kampik
Lars Karlsson
Ewnetu Bayuh Lakew
Michele Persiani
Kai-Florian Richter

Table of Contents

Performance comparison of deep learning with traditional time series weather forecasting techniques 1
Chaitanya Ganesh

Automated Drowsiness Detection while Driving using Depth Camera 11
Omar Haggag

Wait - Responses to loading behaviour inching toward completion 23
Hanna Konradsson

Measuring how colors and geometrical shapes influence reaction times . . . 41
Ivan Lyxzen

The effect of digital gamification when learning a third language 51
Anna Nystedt

Physical Approach with Varying Degrees of Intention Expression with the Pepper Robot 63
Georgios Tsiatsios

Author Index 73

Performance comparison of deep learning with traditional time series weather forecasting techniques

Chaitanya Ganesh

Department of Computing Science
Umeå University, Sweden
mrc17cka@cs.umu.se

Abstract. Temperature forecasts are important for electrical load forecasting and other applications in industry, agriculture, and the environment. Traditional methods of forecasts use time series analysis to make predictions on future values of temperatures. In this paper temperature from weather data is forecasted by fitting the existing data into LSTM neural networks and predicting the near future conditions. The performance of this neural network is then compared to the traditional time series models which have been long used to understand the behaviour of time series data. This paper proves that the use of neural networks dramatically improves the chances of more accurate predictions compared to traditional models. The predictions exhibited a root mean square error of 12.49 °C in time analysis and 2.77 °C while using LSTM networks.

1 Introduction

Forecasting the weather is important since it determines the expectations for the day-to-day human activity. Through common knowledge we can assume the climate of a region depends on the latitude. However, in order to get the correct weather estimates in a particular place over a period of time in the future needs more data analysis. These predictions can be used for various developmental activities like pinpointing the best location for exploiting renewable energy resources with maximum efficiency, plan transportation in case of any areas highly prone to hazards, help to predict natural disasters to reduce loss of human lives, plan daily activities for businesses which heavily depend on outside weather etc. We analyze the change in performance in terms of accuracy of forecasts using LSTMs against traditional methods.

Until the early 1990s, weather forecasts were believed to be an intrinsically deterministic endeavour. For a given set of input data which aptly defines the climate of the region, it was expected that a finite number of fairly accurate forecasts could be generated. This was possible through huge computational costs to run the models which could adapt their behavioural parameters to the climate data provided and produce a deterministic forecast of future atmospheric states. [6] The weather data collected for training a model is usually of time series nature. A time series is a series of data points indexed (or listed or graphed)

S. Bensch, T. Hellström (Eds.): Umeå's 23rd Student Conference in Computing Science USCCS 2019, pp. 1–9, January 2019.

in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data [7]. This paper describes the implementation of a Long Short-Term Memory Network (LSTM) which was trained on a given data set to understand the past behaviour and predict the future. The performance of this neural network is compared to traditional time series analysis models like Autoregression (AR), Moving Averages (MA), Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) [14].

2 Related Work

There has been a considerable amount of research in the field of time series forecasting. Forecasting models were used to fit the previous data to predict the future state of the weather. Various parameters from weather conditions were collected at equal intervals and time series datasets were produced. Abhishek [1] has investigated time series analysis on weather data using artificial neural network (ANN). In this system, the minimum and maximum temperature of the days are forecasted. The accuracy of the results were calculated using the Mean Square Error method. From this study, we gather that the prediction errors are very low if the neural network is tuned with the right parameters.

Nury [16] has conducted research by using time series models like ARMA and ARIMA for data fitting. In his paper, the best order of coefficients was found for the data fitted into the model. Box Jenkins methodology [3] was used to arrive at the best order of coefficients for the ARMA and ARIMA models. This is a better approach compared to trial and error method to decide the coefficients in time series models.

A survey on time series data [12], various time series modelling techniques suitable for different types of data sets are described. Optimization techniques have also been presented along with appropriate evaluation strategies. This paper was effective in narrowing down the evaluation strategy in this research to Root Mean Square Error calculation.

2.1 Background

Traditional methods Traditional methods like AR, MA, ARMA are very simple method by design, but still powerful enough to fit previous weather data and forecast the near future. They are based on an approach that several values from the past generate a forecast of the next point with the addition of some random variable, which is usually white noise. As you can imagine, forecasted values in the future will generate new values and so on.

AutoRegression(AR) In this model the forecasting of the current value Y_t is done based on a combination of the past values Y_{t-1}, Y_{t-2}, Y_{t-3} etc. This means

that the current values may be dependent on one or more past values such as [14]

$$Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, ...\epsilon_t) \tag{1}$$

Hence a common representation of the AR model considering p of its past values also known as the AR(p) model would be:-

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \epsilon_t \tag{2}$$

where β is the coefficient of linear combination to be calculated based on the number of latent variables in the model.

Moving Average(MA) In this model the forecasting of the current value Y_t is done based on a combination of the past random error terms ϵ_{t-1}, ϵ_{t-2}, ϵ_{t-3} etc. This means that the current values may be dependent on one or more past errors such as [14]

$$Y_t = f(\epsilon_{t-1}, \epsilon_{t-2}, \epsilon_{t-3}, \epsilon_{t-4}, ...) \tag{3}$$

Hence a common representation of the MA model considering q of its past random errors also known as the MA(q) model would be :-

$$Y_t = \beta_0 + \epsilon_t + \psi_1 \epsilon_{t-1} + \psi_2 \epsilon_{t-2} + ... + \psi_q \epsilon_{t-q} \tag{4}$$

where the error terms are ϵ_t assumed to be white noise processes with mean $\mu = 0$ and variance $\sigma^2 = constant$. The value of ψ depends on the number of latent variables considered in the model.

AutoRegressive Moving Average(ARMA) In this model we represent the current value Y_t in the time series model as a mix of AR and MA models described above. Hence a common representation of the AR model considering p of its past values and q of its past random errors also known as the ARMA(p,q) model would be

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \epsilon_t + \psi_1 \epsilon_{t-1} + \psi_2 \epsilon_{t-2} + ... + \psi_q \epsilon_{t-q} \tag{5}$$

Auto-Regressive Integrated Moving Average(ARIMA) In this method, the time series data is made invariant and stationary. We need to ensure that the relation between the input and the output is constant with respect to time. Also, the mean and variance of the data should remain constant. This can be done through differencing. Lags of the stationarized series in the forecasting equation are called "autoregressive" terms, lags of the forecast errors are called "moving average" terms, and a time series which needs to be differences to be made stationary is said to be an "integrated" version of a stationary series. In terms of

y, the general forecasting equation considering p past values and q past errors is [10, 15]

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} - \theta_1 e_{t-1} - \theta_2 e_{t-2} - .... - \theta_q e_{t-q} \quad (6)$$

Recurrent Neural Networks In this part, we describe a plan to predict the data points on the time series data using recurrent neural networks consisting of Long Short Term Memory units. Suppose we have a 2-D graph with the x-axis indicating time and the y-axis indicating the data points. Typically predictions work in a way where we assume a few data points in the beginning are absolutely true. Therefore, we take a small window of values as the input and try to predict the next value. We can formulate the problem of predicting future values based on past values as a supervised learning problem using the LSTM.

LSTM Long Short Term Memory networks [9, 17, 18] (usually just called LSTMs) are a special kind of RNN, capable of learning long-term dependencies. The structure of a simple RNN is as shown in Figure 1 [8] and the LSTM is as shown in Figure 3 [9].

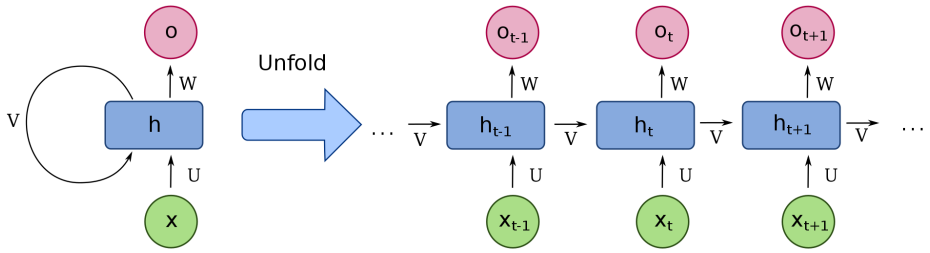

Fig. 1. RNN schematic diagram [5], where, x_t: Input Vector, h_t: Hidden Layer Vector, y_t: Output Vector, W, U: Parameter Matrices, V : Vector, σ_h and σ_y: Activation functions

In the RNN network, we supply the input to the hidden layer. In the structure towards the left, we have one hidden layer which represents several layers with the same weights and biases. This can also be interpreted as a chain structure towards the right. The unfolded structure shows the successive layers with their own individual inputs. We observe that the interaction of neurons inside an LSTM is a bit different as compared to the RNN. The LSTM has the ability to remove or add information to the cell state, carefully regulated by structures called gates. These gates have a sigmoid layer which can control the amount of information they can let in and out of the cell as the outputs range from 0 to 1.

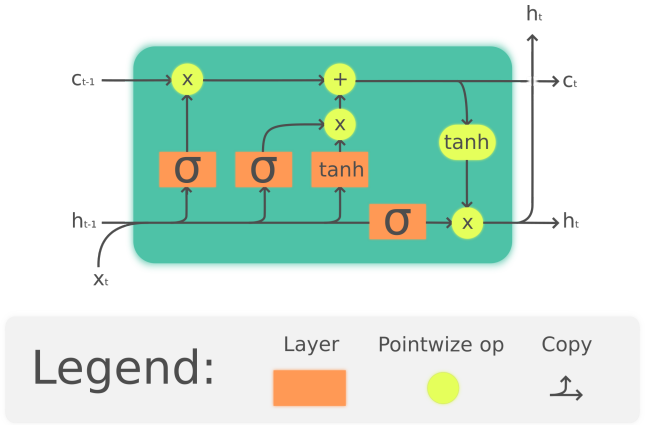

Fig. 2. Single LSTM cell schematic diagram [4], where, x_t: Input vector to the LSTM unit, f_t: Forget gate's activation vector, i_t: Input gate's activation vector, o_t: Output gate's activation vector, h_t: Hidden state vector also known as output vector of the LSTM unit, c_t: Cell state vector, W, U and b: Weight matrices and bias vector parameters which need to be learned during training

3 Method

3.1 Procedure

The data was extracted from the website weather underground using a provided api. The data consists of multiple timestamps per day and the values of temperature, pressure, humidity, rain, precipitation,etc. at that time period. From this multivariate data we extract the temperature readings. This data can be called as a univariate time series since there is only one variable that changes according to time. Data analysis may be done with traditional time series AR, MA and ARMA. The second method is through a neural network can be achieved through Supervised Learning. In this method we assume that future data is somehow related to a few past observations. Then the results of these techniques were compared with an appropriate evaluation strategy at the end.

Preparing Data In this section we explain how the data was prepared in order to feed into the two systems. For time series analysis, we need to have stationary data as an input. Therefore, the given data is checked using the Dickey-Fuller test for stationarity.

It was observed that the test statistics were not less than the critical values, therefore we concluded that the series was non-stationary. Therefore, we use differencing method due to ease of implementation to convert our data in to stationery data. We again perform the Dickey-Fuller test for stationarity on the processed data to validate our result. The observation indicate that the processed data was stationary.

Fitting Data and Predictions This data is then fitted into our time series models to check for the best order of coefficients using the Box Jenkins methodology [13]. This is one of the ways to eliminate trial and error technque for choosing the best order of coeficients while using time series analysis. The Akaike information criterion (AIC) [2] is calculated for different orders while fitting the processed time series data which gives the quality of the time series model with this data in comaprison to the other models. The best order produces the lowest score. This order is then used for future predictions. However, this is one of the bottlenecks while doing predictions in this particular case. We train each time series model for a given amount of train data and then try to predict a finite number of future values based on the estimated models.

LSTM Structure The network has a visible layer with 1 input, a hidden layer with 400 LSTM blocks or neurons, and an output layer that makes a single value prediction. The default sigmoid activation function is used for the LSTM blocks. The network is trained for 100 epochs and a batch size of 100 is used. The mean absolute error is used for calculating the loss and the adam optimizer [11] is used for configure the model. These parameters were observed the best possible combination to fit the time series weather data.

3.2 Tools

The experiment was conducted in python. Libraries such as pandas, numpy, statsmodels, scipy, matplotlib, keras, sklearn, tensorflow are used for rapid prototyping the models. The evaluation for future predictions was done by calculating the Root Mean Square Errors (RMSE) of the predictions compared to the actual values.

4 Result

4.1 AR

Prediction Accuracy In the estimation of the most accurate AR model, we observe that there is a need for at least 24 latent variables which is a very high number and it is safe to assume that the data in hand is more complex than an AR model can handle.

4.2 MA

Prediction Accuracy A moving average model with 4 latent variables was found to be the best fit according to the Akaike Information Criterion (AIC) and the RMSE of 1706 temperature forecasts is 12.49.

4.3 ARMA

Prediction Accuracy In order to make the ARMA model, we try to choose the order of (β,ψ) according to the best combination based on which model produces the lowest Akaike Information Criterion (AIC), which provides the quality of models when compared against each other. This was predicted as (6,4). The RMSE of the predicted temperatures was 12.49.

4.4 ARIMA

Prediction Accuracy We try to estimate an additional parameter in this model where we try to find the number of times we are differencing the series along with the AR and MA parameters. The order is found to be (6,0,4) which is the same as the ARMA model detected previously. The RMSE of the predicted temperatures was 12.49.

4.5 LSTM

The LSTM network is also given the same amount of data as the time series models mentioned above.

Prediction Accuracy The mean square values for training the data through one iteration was 5.13. The error gradually decreases to 2.77 as the number of iterations increases to 10. The results have been presented in Figure The prediction accuracy can be improved on this network through adding additional layers of neural networks. There is a significant scope of improvement in terms of speed and accuracy of the network when processing over a GPU is taken into consideration.

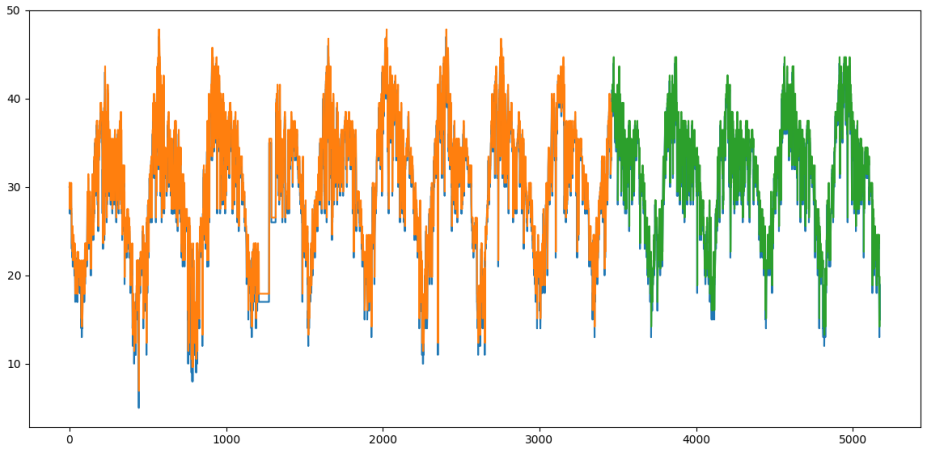

Fig. 3. Result of LSTM predictions

In this figure the timesteps are taken on the X-axis and the values of temperature are taken on the Y-axis. The original values of the temperatures are plotted in blue. The performance of the LSTM on the trainset is plotted in orange and the predictions have been plotted in green.

5 Conclusion

We observe that in time series models, the RMSE values are constant for all the models. This shows that the model needs to be iteratively updated as we get newer results from predictions and new errors are supposed to be accommodated continuously. Therefore, it is safe to conclude that the time series models cannot capture minor details in the train data such as seasonality, trend, irregularity etc without necessary additional modifications.

5.1 Limitations

All the time series models considered in this experiment had very intense bottlenecks while estimating the best order for each model. The all the data was fit into each model multiple times to estimate the best possible order for future predictions and this took a lot of time. Therefore, a severe hardware limitation was faced.

6 Future work

A sliding window approach maybe used in order to estimate the best order and adapt the model to lesser number of data points at a time. The error of the next few predictions would be taken into account while modifying the existing model. This seems to be a better approach which could not be implemented due to hardware and time constraints.

Ensemble learning maybe performed on the available data with base learners such as LSTM, HMM and ARIMA (from time series modelling). This may significantly improve accuracy of the neural network when a proper evaluation technique like Mean Absolute Percent Error or Symmetric Mean Absolute Percent Error [12] are used.

7 Acknowledgement

I would like to thank Prof. Thomas Hellstrom and Prof. Suna Bensch for their support and encouragement throughout this project.

References

1. Ashish Pandey Imran Khan Abhishek Agrawal, Vikas Kumar. An application of time series analysis for weather forecasting. *International Journal of Engineering Research and Applications*, 2:974–980.

2. Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1998.
3. George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Inc., San Francisco, CA, USA, 1990.
4. Guillaume Chevalier. The lstm cell.png, 2018. `https://creativecommons.org/licenses/by/4.0/`.
5. François Deloche. Recurrent neural network unfold.svg, 2017. `https://creativecommons.org/licenses/by/4.0/`.
6. Tilmann Gneiting and Adrian E. Raftery. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005.
7. C.W.J. Granger. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16(1):121 – 130, 1981.
8. A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connections system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, May 2009.
9. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.
10. Rob J. Hyndman, George Athanasopoulos, and OTexts.com. *Forecasting : principles and practice / Rob J Hyndman and George Athanasopoulos*. OTexts.com [Heathmont, Victoria], print edition. edition, 2014 2014.
11. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
12. G. Mahalakshmi, S. Sridevi, and S. Rajaram. A survey on forecasting of time series data. In *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, pages 1–8, Jan 2016.
13. Spyros Makridakis and Michele Hibon. Arma models and the box-jenkins methodology. *Journal of Forecasting*, 16(3):147–163.
14. Terence C. Mills. *Time series techniques for economists*. Cambridge University Press, Cambridge, 1990.
15. Robert Nau. Introduction to arima models. `https://people.duke.edu/~rnau/411arim.htm`, accessed 2018-10-05.
16. A. H. Nury, Manfred Koch, and Muhammad Jawaherul Alam. Time series analysis and forecasting of temperatures in the sylhet division of bangladesh. 2013.
17. Christopher Olah. Understanding lstm networks, 2015. `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`, accessed 2018-10-05.
18. Andy Thomas. Keras lstm tutorial - how to easily build a powerful deep learning language model, 2018. `http://adventuresinmachinelearning.com/keras-lstm-tutorial/`, accessed 2018-10-05.

Automated Drowsiness Detection while Driving using Depth Camera

Omar Haggag

Department of Computing Science
Umeå University, Sweden
mcs17ohg@cs.umu.se

Abstract. Drowsiness during driving has been a serious issue for decades causing thousands of deaths and billions of dollars in loss every year. Some research has already been done in this area, however none of them is practical, real-time and can be applied in real life. In this paper, we will introduce a new framework for drowsiness detection which uses depth cameras for tracking and processing the dynamic motion of the upper body of the driver in real time. In our experiments and results, we were able to detect drowsiness of different drivers on the fly without the need of placing markers on their body or to have initial preparation or setup before they start driving.

1 Introduction

Every year, thousands of drivers die or get severely injured due to drowsiness while driving. A drowsy driving can result from several conditions of the driver, these conditions could be lack of sleeping, high alcohol overdose, sleep disorders, fatigue, medications, etc. [12]. In [19], it is mentioned that the main huge danger comes from the fact that drivers do not know the exact moment they will become too drowsy and fall asleep over the steering wheel. According to the National Highway Traffic Safety Administration (NHTSA) in the United States[1], drowsy driving is responsible for 21% of total accidents, causing at least 1,550 death, 40,000 injuries and 100,000 accidents every year. Moreover, according to the National Sleep Foundation (NSF)[2], at least half of the American drivers admitted that they have experienced a drowsy driving incident in the last year and one in five drivers actually fell asleep which is 20% of them. According to NHTSA, the accidents that are caused by driver fatigue result in $12.5 billion losses only in the United States every year.

In the last few years, technology allowed drivers to get distracted while driving and to drive too fast without developing real and effective safety guarantees for human mistakes. Drivers believe that the latest technologies in their cars such as driving assistance, auto-brake, lane detection, anti-lock braking (ABS), anti-skid or air bags systems can allow them to prevent accidents [10] and that

[1] Access by: https://www.nhtsa.gov
[2] Access by: https://www.sleepfoundation.org

S. Bensch, T. Hellström (Eds.): Umeå's 23rd Student Conference in Computing Science USCCS 2019, pp. 11–22, January 2019.

is the reason why drivers feel safer and get encouraged to take higher risks. However, none of the previous safety systems could detect drowsy or sleepy drivers in order to warn them in real time [16]. By warning and alerting them, we could potentially save thousands of lives every year and billions of dollars resulting from physical destruction in roads, human injuries and sadly death incidents.

The goal of our research in this paper is to take the advantage of using depth cameras and its accuracy and high sampling rate to monitor and alert drivers if their current position is indicative for drowsiness or that they may be falling asleep. We have seen a huge expansion and interest of using depth cameras in recent years [1]. Nowadays, depth cameras became embedded into many devices such as game consoles, home appliances and even new smart phones. Many smart phones are just unlocked using depth cameras such as iPhone and Samsung's new models. Depth cameras can allow us to handle some problems in an efficient, real-time and inexpensive way. Depth cameras such as Microsoft's Kinect give us the opportunity to add depth to video processing without the need to have markers attached to our bodies [7] or even to have special light conditions [11]. The challenge in our solution is tracking the dynamic motion of drivers in real-time without having a marker on their bodies as it is extremely difficult since the body is flexible and human bodies have different heights, sizes and even different type of clothes [18]. Our research is done by tracking the dynamic motion of body of the driver in real-time, estimating the driver's pose and detecting whether the driver is still on their position or their upper body is falling to the front. In this case, an automated warning in the form of sound alert occurs which warns the driver and allow the driver whether to concentrate or stop the car. Our paper is divided into the following sections. In Section 2, we present a literature review on the Kinect depth camera and its usage in the body motion tracking field. We also present and discuss the related work and its limitations. In Section 3, we present our proposed method for detecting drowsiness or sleeping during driving through depth camera (Kinect). Finally, in Section 4, we present our results and conclusions and the directions for our future research.

2 Earlier Work

Kinect depth camera was invented and patented by PrimeSense company that was bought by Microsoft later. Kinect detects Xbox players postures and movements by having the depth map of the world in-front of the camera without the need of putting markers on their bodies. Kinect technology can be very useful for the research done in motion tracking and image processing fields. The Kinect consists of five components, an infrared (IR) emitter, depth camera, RGB camera, tilt motor and four microphones as shown in Figure 1.

Kinect's IR emitter sends out a pattern of infrared light in which when it hits a surface it is distorted and read by depth camera. Then the depth camera analyze it to construct a 3D map for the environment in-front of the Kinect. The RGB camera is a normal color camera that captures a real video image for the environment in-front of the Kinect. The tilt motor adjusts the angle of

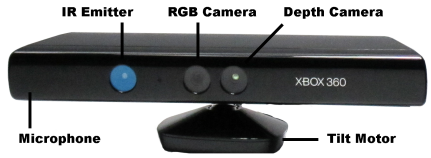

Fig. 1: Kinect depth camera components

the Kinect towards the player or the environment in-front of it with a range of -27 degrees (down) to +27 degrees (up). Generally, it is very challenging to develop a real-time, accurate, marker-less, 3D motion capture especially with a low-cost solution, however, this challenge can be simplified by using Kinect depth camera [22].

The accidents data and statistics indicated in the introduction resulted from drowsiness and sleeping during driving revealed the importance of finding a solution to that problem. Some research has been already done in this area to detect the problem of drowsiness and sleeping during driving [17]. One way was by processing electroencephalogram (EEG) brain signals. EEG is the measurement of the electrical activity initiated by the brain and processed on the scalp surface through a cap containing electrodes [4]. When the driver starts to feel drowsiness or sleeping occurs, it will be detected by EEG brain signals. EEG brain signals research was not only tested on car drivers but also on plane pilots. That solution is very expensive, non-practical or applicable in real life since the drivers will need to wear caps and put special gel on their head every time they need to use that system [14].

Other researchers used normal cameras to detect, process and analyze the movements and actions of the drivers such as eye's blinking. The eye blinking is considered to be a suitable occular indicator for fatigue or drowsiness diagnostics [5]. However, the problem is that those cameras will perform very bad and inaccurate at night or dark conditions [8]. Also, if the driver wear any type of glasses that are considered as main problems for all eye detection systems designed so far [21].

Other research was done on detecting the drowsiness of the drivers by using motion sensors, however, motion sensors cannot detect skeleton hierarchy [3]. Motions sensors can detect if the body of the driver is moving, but it will not detect accurately if the driver's neck fall to front, which can lead to false outcomes and very low accuracy. That problem can be solved by using marker-based sensors, however, it still unpractical, as it is not automated and the drivers will always need to place the markers on their body on the right place so that the system can work as it should be. The three previous methods of drowsiness detection during driving are illustrated in Figure 2.

In [13, 9], the proposed systems and techniques runs with good accuracy. However, in some conditions, they totally fail and will not work. So, if the driver is driving after sunset, which happens all the time as the system mainly depends on the lighting conditions. Also, the videos sometimes have motion blur which stops the system from working as expected. Other systems, which are tacking

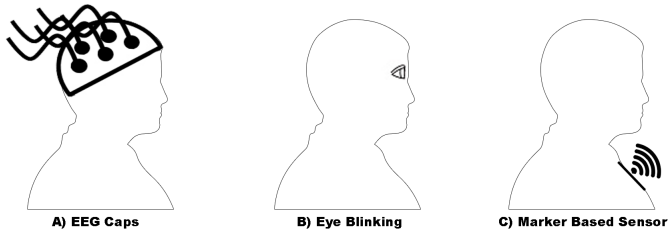

Fig. 2: Different types of driving drowsiness detection

the eye totally fails when the driver is wearing glasses as it reflects lights in case of eyeglasses or cover the entire eye if they are wearing sunglasses. reflections. These issues has been fixed using our system which uses depth cameras instead of ordinary ones.

Some industrial car manufactures companies such as Volkswagen (VW) or Volvo are currently developing and testing new systems for driver fatigue detection. For example, VW[3] fatigue detection work mainly on long journeys and in the dark at speeds above 65 km/h. VW fatigue detection function detect and process the drivers behavior such as steering behavior, pedal usage and lateral acceleration. When the system detects that the driver concentration is decreasing, it provides a visual warning and sound alert. If the driver does not take a break within the next 15 minutes the warning is repeated to force the driver to take a break. VW neither revealed how their system actually work, nor released cars with this functionality to public. In Volvo, they use a different technique for drowsiness detection than VW called driver alert control (DAC)[4] and already developed and available in many of Volvo car models such as S90, V90 and XC90. DAC uses a camera installed between the windshield and the interior rear view mirror. In addition to installing a number of sensors and a processor in order to constantly monitor the distance between the car and the road markings. If the driver is continuously not driving in the middle of the lane and is shifting to right or left, the alert control determines that this driving behavior is caused by tired or inattentive driver. The system displays a warning sound and the message time for a break. The main drawback of this technique is that it will not work with snowy roads since the camera will not be able to detect the the road markings. Also, if the road is not marked, DAC technique will not work at all.

3 Methodology

In our proposed solution, we take the advantage of depth cameras to detect the upper body posture of the driver without the need of the driver to have special markers attached to their body which needs a special setup. Then, we detect the joints and construct the bones (the neck and shoulder bones) as shown in

[3] Access by: http://inside.volkswagen.com/Take-a-break.html
[4] Access by: https://www.volvocars.com/intl

Figure 3. Following that, we capture and calibrate the pose of the skeleton of every driver in a few seconds just when the driver is seated.

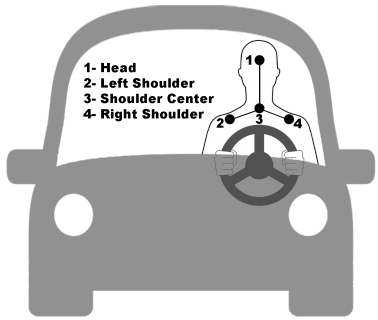

Fig. 3: Driver body hierarchy

After detecting and calibrating the upper body of the driver, we start pose estimation of every single movement the driver does, process it and then update the new driver's posture in real time by displaying a warning and playing a warning sound. Through online monitoring, we calculate the new angle between the head of the driver and the shoulders and compare it to the driver's initial calibrated pose. If the head started to fall to front, we start to notify the driver in real time. Also, in our implementation and graphical user interface (GUI), we developed and constructed a smooth color range interpolation system ranging between green, yellow and red depending on the head angle [6] as shown in Figure 4.

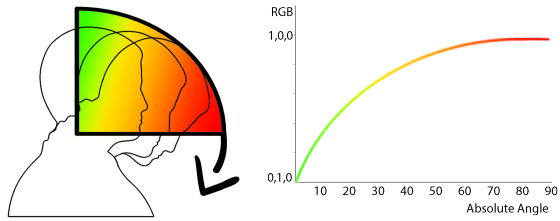

Fig. 4: A smooth color range interpolation system ranging between green, yellow and red depending on the head angle.

The green color indicates that the current posture of the driver is similar or close to the calibrated ones. The yellow color degradation indicates that driver's head is starting to fall to front, so we can start to warn the driver. Finally, the red color degradation indicates that the head fell to the front and the driver is notified. By tracking the body postures using depth cameras, we are able to track any driver regardless of his size, height, weight and the way of seating of the driver as shown in Figure 5. Also, our motion tracking model is shown in Figure 6.

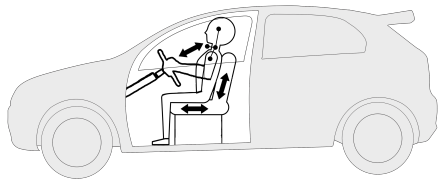

Fig. 5: Different driver's positions inside cars

In model initialization, we create two drawing layers (portions) for both the depth and colour streams. Then, we start an online streaming of both the the depth stream (captured depth frames) and the colour stream (captured RGB frames). Then we start detecting the needed joints we are interested in, in our case, head (neck), shoulder center, right and left shoulders using the depth stream and frames. Afterwards, we start creating and drawing the needed bones which are the Head (connecting top head-neck), right shoulder (connecting neck with right shoulder), left shoulder (connecting neck with left shoulder), then we map the location of the detected joints and constructed bones to the drawing depth image canvas. That is done by measuring the distance between any two body joints $\mathbf{p}_1 = (x_1, y_1, z_1)$ and $\mathbf{p}_2 = (x_2, y_2, z_2)$ using the following steps [15, 2], where the distance between two points in 3-D is shown in Figure 7.

The distance between the two points (x_1, y_1, z_1) and (x_2, y_2, z_2) in the plane $z = z_1$ is calculated using Pythagoras' theorem. The distance between (x_1, y_1, z_1) and (x_2, y_2, z_2) is $\sqrt{\Delta x^2 + \Delta y^2}$ where $\Delta x = x_2 - x_1$ and $\Delta y = y_2 - y_1$.

Then we draw a line from (x_1, y_1, z_1) to (x_2, y_2, z_2) representing Δz, where $\Delta z = z_2 - z_1$ in which a right angled triangle is formed using the vertices (x_1, y_1, z_1), (x_2, y_2, z_1) and (x_2, y_2, z_2).

Then we use Pythagoras to calculate the distance between two points (x_1, y_1, z_1) and (x_2, y_2, z_2) using the following equation:

$$\sqrt{(\sqrt{\Delta x^2 + \Delta y^2})^2 + \Delta z^2} = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}$$

After model initialization, we update the drawing image canvas with online feeded input depth stream from the depth sensor. Later on, we calculate the hierarchical orientation (angle) of the neck joint relative to the connected bones (shoulders). This angle is calculated by getting the dot product of the two vectors of the two connected bones, in which the first bone is between the neck joint and right shoulder joint, the second bone is between the neck joint and left shoulder joint [20, 2].

That can be calculated as follows: $\mathbf{l}_1 = (l_1^x, l_1^y, l_1^z)$ and $\mathbf{l}_2 = (l_2^x, l_2^y, l_2^z)$ and calculate it using this formula:

$$\theta = \cos^{-1} \frac{\mathbf{l}_1 . \mathbf{l}_2}{\|\mathbf{l}_1\| . \|\mathbf{l}_2\|}$$

where the dot product is computed by multiplying the components of each vector along each axis. This is followed by the addition of three multiplication products

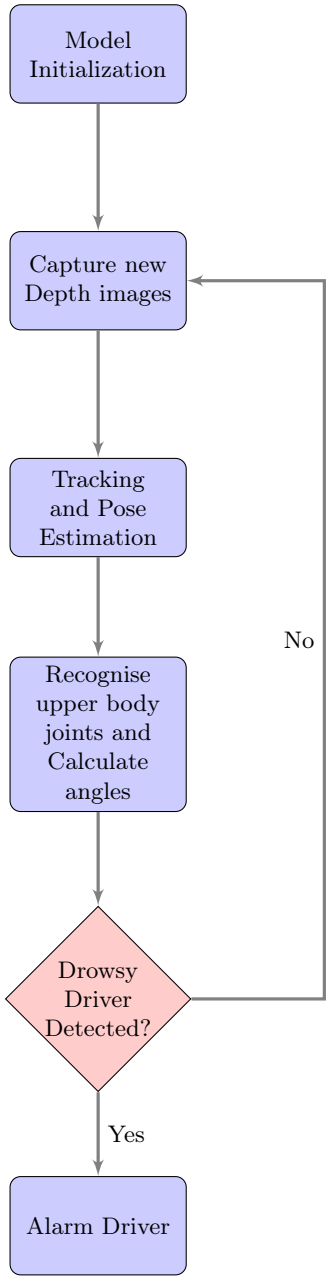

Fig. 6: Motion tracking model flowchart.

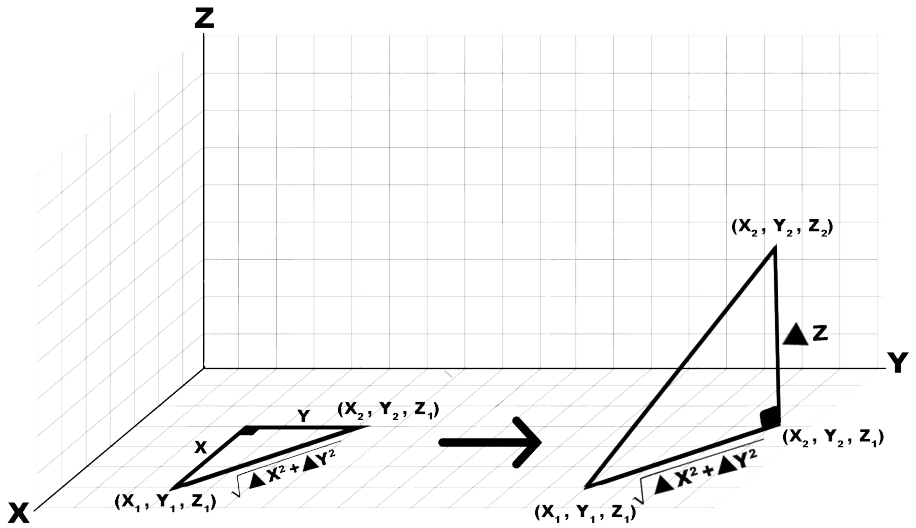

Fig. 7: 3D line formulation

$$\mathbf{l}_1.\mathbf{l}_2 = \mathbf{l}_1\mathbf{l}_2{}^T = (l_1^x * l_2^x) + (l_1^y * l_2^y) + (l_1^z * l_2^z)$$

where the magnitudes of each vector is calculated using the formula

$$\|\mathbf{l}_1\| = \sqrt{\mathbf{l}_1 \cdot \mathbf{l}_1} \text{ and } \|\mathbf{l}_2\| = \sqrt{\mathbf{l}_2 \cdot \mathbf{l}_2}$$

After that, the results are substituted in this formula, where the angle theta is calculated by getting the inverse cos of this formula.

$$\cos\theta = \frac{\mathbf{l}_1.\mathbf{l}_2}{\|\mathbf{l}_1\|.\|\mathbf{l}_2\|}$$

Next step, we calibrate the posture of the driver before starting driving, we perform an online monitoring of the neck orientation from the feeded input depth stream and we process the calculated orientation and continuously compare the new captured posture and orientation with that of the calibrated one. We constructed a colour palette to display the difference in orientation using very smooth colour degradation system as shown in Figure 4. Depending on the orientation difference a given colour is given to the bone, where green is normal which close to the calibrated orientation and red is dangerous where a state of sleepiness is detected. Finally, once a state of drowsiness is detected and the head of driver fall more than 70 degrees to the initial calibrated position, a warning system is activated where a message is displayed on the screen along with a loud notification sound to alert the driver and prevent him from sleeping while driving.

4 Results

We tested our program on eight drivers. The drivers pretend to be drowsy (simulated drowsiness). They had different body shapes and sizes, also they were

wearing different type of clothes since we did not have a prerequisite of what they can wear. We also tested our program in both good light conditions and darkness. Our results showed that both worked the same since the depth camera is totally independent on light conditions. Each experiment took a couple of minutes. The details of our participants are shown in Table 1.

Participant	**Height**	**Weight**	**Age**	**Light Condition**
Male 1	171	76	29	Normal
Male 2	174	76	43	Dark
Female 1	184	64	23	Normal
Male 3	186	95	31	Dark
Male 4	172	68	22	Normal
Male 5	167	71	25	Normal
Male 6	186	79	24	Normal
Female 2	169	53	26	Dark

Table 1: Details of our participants.

The driver sat in front of the kinect and all what they had to do was to pretend they were driving and then to pretend drowsiness by falling forward. First, the depth camera captures and calibrates the driver's initial position, and display the neck posture in clear green color shown in Figure 8a. Afterwards, our framework calculates the angle between the neck and the shoulders. Depending on the angle shown in Figure 4, our framework starts to color the neck bone by color degradation ranging from green to red, depending on the risk of drowsiness as shown in Figure 8b, c and d. As shown in Figure 8a, once the driver gets inside the car and get ready for driving, our framework detects their body and start calibrating their position including their head and neck positions. Then the driver started to feel sleeping and their head started to fall forward as shown in Figure 8b and the framework starting to detect that their is something wrong with their head position and the related bar started to turn yellow. In figure 8c the driver became more sleepy and their head was falling more and our framework detected that and shows an orange color. Finally the driver's head moved down in a dangerous position as shown in Figure 8d and our framework detected that by showing the red color. Moreover, the relation between the time and the absolute difference of calibrated pose and real time pose is shown in Figure 9. The detection outcome is based on these angles and as we can see in the Figure the angles are changed by the framework every second and the detection was done accuracy based on those angles.

5 Limitations and Future work

Our system and technique will not work if the driver fell a sleep without moving their head to front at all. That will create a false negative drowsiness detection. Moreover, during our testing, we found out that if the driver is listening to music and started to interact by extremely nodding their head to up and down more

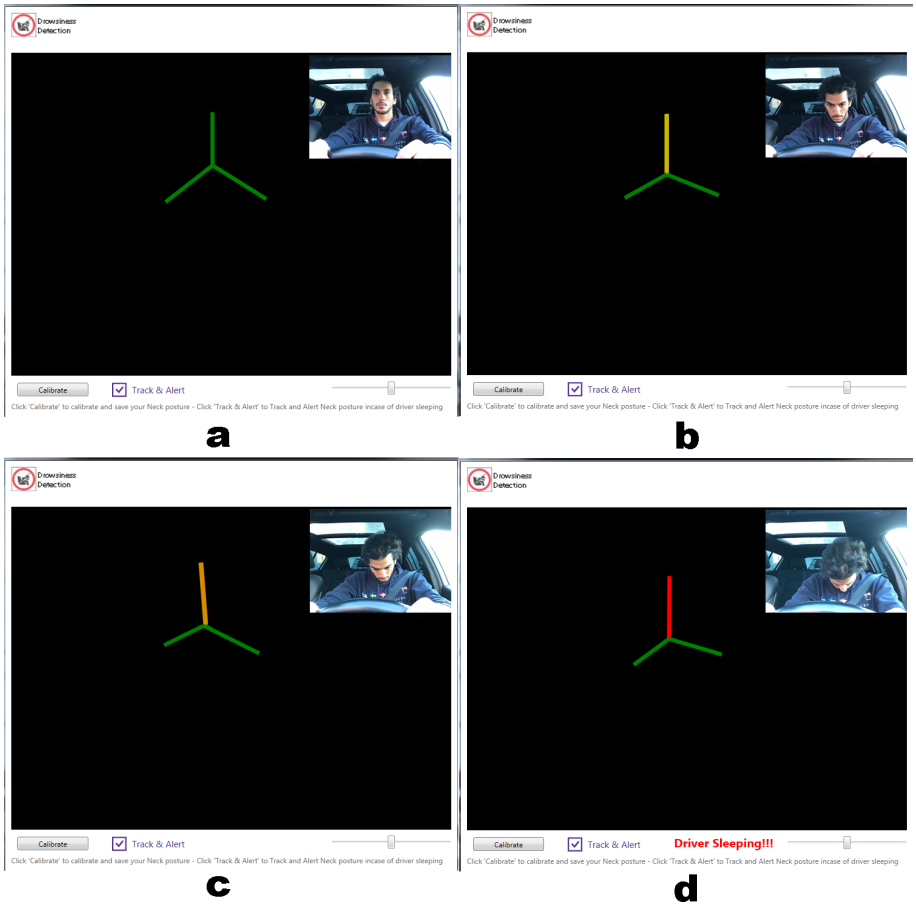

Fig. 8: Different driver's position and risks.

than 70 degrees from the initial calibrated position, our program will generate a false positive drowsiness detection, however, the driver is just nodding his head. That occurs since our technique only process the driver's upper body positions from depth camera, so it could not differentiate if it is true or false drowsiness in this case. In addition to, it will not be able to detect if the driver slept without moving their head forward. These problems can be solved by checking the pulse of the driver using a smart fitness watch and integrating this to our system wirelessly to monitor the driver's pulse. Also, we will need to adjust our experiments in order to enable the determination of accuracy values. In order to do that we will need to make sure that the software attempts to determine whether a person in a car's driver seat is in fact sleeping or awake.

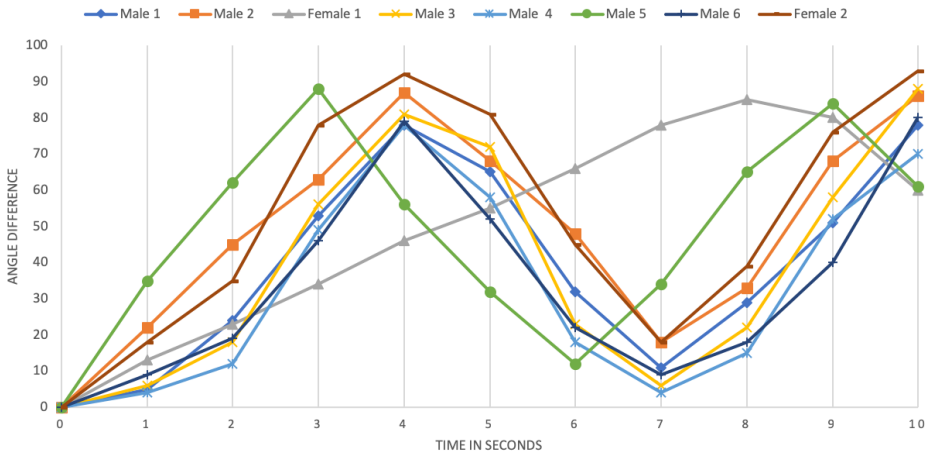

Fig. 9: Relation between time and angle difference for our participants.

6 Conclusion

Drowsy driving causes thousands of deaths and billions of dollars loss every year. In this paper, we have discussed the importance of having a framework to detect drowsiness while driving and showed by statistics how safety modern technology in a car still do not prevent accidents from happening, which shows how serious this issue is. Some research was already developed in this area and we have shown the issues that we face in case we use them in real life in the literature review section. We proposed and developed a framework for drowsiness detection using depth cameras by tracking and processing the dynamic motion of the driver in real time. As shown in the methodology and results sections, our experiments and results were able to perform a simulated drowsiness detection of different drivers and all the work was done on the fly and in real time without the need of placing markers on their body or to have initial preparation or setup before they drive.

References

1. Jake K Aggarwal and Quin Cai. Human motion analysis: A review. *Computer vision and image understanding*, 73(3):428–440, 1999.
2. Rutherford Aris. *Mathematical modelling techniques*. Courier Corporation, 2012.
3. Luis Miguel Bergasa, Jesús Nuevo, Miguel A Sotelo, Rafael Barea, and María Elena Lopez. Real-time system for monitoring driver vigilance. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):63–77, 2006.
4. Gianluca Borghini, Laura Astolfi, Giovanni Vecchiato, Donatella Mattia, and Fabio Babiloni. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44:58–75, 2014.

5. Philipp P Caffier, Udo Erdmann, and Peter Ullsperger. Experimental evaluation of eye-blink parameters as a drowsiness measure. *European journal of applied physiology*, 89(3-4):319–325, 2003.
6. In-Ho Choi, Chan-Hee Jeong, and Yong-Guk Kim. Tracking a driverâs face against extreme head poses and inference of drowsiness using a hidden markov model. *Applied Sciences*, 6(5):137, 2016.
7. Guanglong Du, Ping Zhang, Jianhua Mai, and Zeling Li. Markerless kinect-based hand tracking for robot teleoperation. *International Journal of Advanced Robotic Systems*, 9(2):36, 2012.
8. Pia M Forsman, Bryan J Vila, Robert A Short, Christopher G Mott, and Hans PA Van Dongen. Efficient driver drowsiness detection at moderate levels of drowsiness. *Accident Analysis & Prevention*, 50:341–350, 2013.
9. Lex Fridman, Philipp Langhans, Joonbum Lee, and Bryan Reimer. Driver gaze region estimation without use of eye movement. *IEEE Intelligent Systems*, 31(3):49–56, 2016.
10. Jeffrey K Gurney. Sue my car not me: Products liability and accidents involving autonomous vehicles. *U. Ill. JL Tech. & Pol'y*, page 247, 2013.
11. Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE transactions on cybernetics*, 43(5):1318–1334, 2013.
12. Melinda L Jackson, Rodney J Croft, GA Kennedy, Katherine Owens, and Mark E Howard. Cognitive components of simulated driving performance: sleep loss effects and predictors. *Accident Analysis & Prevention*, 50:438–444, 2013.
13. Li Li, Klaudius Werber, Carlos F Calvillo, Khac Dong Dinh, Ander Guarde, and Andreas König. Multi-sensor soft-computing system for driver drowsiness detection. In *Soft computing in industrial applications*, pages 129–140. Springer, 2014.
14. Chin-Teng Lin, Ruei-Cheng Wu, Sheng-Fu Liang, Wen-Hung Chao, Yu-Jie Chen, and Tzyy-Ping Jung. EEG-based drowsiness estimation for safety driving using independent component analysis. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 52(12):2726–2738, 2005.
15. Eli Maor. *The Pythagorean theorem: a 4,000-year history*. Princeton University Press, 2007.
16. Paul Stephen Rau. Drowsy driver detection and warning system for commercial vehicle drivers: field operational test design, data analyses, and progress. In *19th International Conference on Enhanced Safety of Vehicles*, pages 6–9. Citeseer, 2005.
17. Arun Sahayadhas, Kenneth Sundaraj, and Murugappan Murugappan. Detecting driver drowsiness based on sensors: a review. *Sensors*, 12(12):16937–16953, 2012.
18. Jan Smisek, Michal Jancosek, and Tomas Pajdla. 3d with kinect. In *Consumer depth cameras for computer vision*, pages 3–25. Springer, 2013.
19. Heikki Summala and Timo Mikkola. Fatal accidents among car and truck drivers: effects of fatigue, age, and alcohol consumption. *Human factors*, 36(2):315–326, 1994.
20. PJG Teunissen. *Adjustment theory*. VSSD, 2003.
21. Bao-Cai Yin, Xiao Fan, and Yan-Feng Sun. Multiscale dynamic features based driver fatigue detection. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(03):575–589, 2009.
22. Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.

Wait - Responses to loading behaviour inching toward completion

Hanna Konradsson

Department of Computing Science
Umeå University, Sweden
Email: tfy14hkn@cs.umu.se
Website: hannakon.se

Abstract. The purpose of this study was to find if loading bar behaviour significantly affect user response and time estimation of their duration. A test was conducted where test users estimated the duration of loaders between menial tasks, and scored their behaviours on seven subjective attributes. The results of which lead to conclusions on what type of behaviour may be favourable and a discussion on why.

1 Introduction

The progress indicator is a visual representation of ongoing work, such as loading assets, and is used in many digital applications. It is more commonly known as a loader, see Figure 1. The progress indicator is typically used when there is a significant amount of waiting time for the user, and has several purposes. One purpose is to show users how much has already been loaded and how much is left. Another purpose is to indicate to users that the loading application has not malfunctioned or stopped. Examples of indicators are a bar filling up or a spinning hourglass. This study investigates different behaviours of a specific class of progress indicator; the determinate indicator. More on this in Section 2.1. The behaviour of the progress indicator depends on the type of indicator, and refers to how the indicator changes from one instant to another. The specific behaviours investigated here are detailed in Section 2.2.

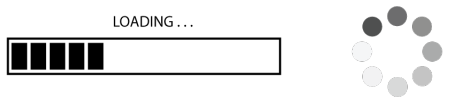

Fig. 1. Two examples of progress indicators, or "loaders". To the left is a loading bar and to the right is a spinning loader.

The aim is to analyze how some different progression rate behaviours affect both user appeal and user speed perception. The underlying hypothesis was: *Showing users a progression indicator which increases to completion at a steady rate has user experience benefits with regards to the perceived elapsed time and*

S. Bensch, T. Hellström (Eds.): Umeå's 23rd Student Conference in Computing Science USCCS 2019, pp. 23–40, January 2019.

emotional response. This hypothesis was chosen in order to examine a stance of this research subject, even thought the reverse hypothesis may be an equally legitimate based on earlier work.

Gaining insight into how users respond to different progress indicator behaviours is important to developers who try to create appealing applications. Waiting time is often a point of annoyance for users as it interrupts their activity. Investigations has been made regarding design, amount of feedback and time perception management of progress indicators. However, user response to progression behaviour has not been as common of a subject.

In order to investigate this, three specific behaviours were applied to the same loader design. An online mock application interspersed these loaders and menial word association tasks. Twelve test users were recruited and asked to complete the set of tasks in the application and to evaluate for each loader both perceived elapsed time and seven subjective attributes related to satisfaction. In order to remove bias, the users were not made aware which aspect of the mock application was being researched until the end of the test session. They were also asked some personal background details such as computer experience and perceived stress in order to discuss the influence of these aspects.

While significant differences were not found between time estimations, there were notable results in regards to the users' experiences and evaluations of the selected attributes.

1.1 Background

The work in [9] covers both progress indicator classification and how to design for processes to appear faster without actually shortening the completion time, i.e. perception management. Techniques include non-linear progression and placing heavy processes first, both of which result in the process appearing to speed up as it progresses. This has two benefits. Firstly, the user is surprised by the completion time which appears faster than they first approximated. Secondly, the last section of the progress, where users typically focus on the indicator as normal activity can soon resume, goes by faster. Another aspect of time perception is evaluated in [8]. This paper investigates how users apply the phrase "time flies when you're having fun" to retroactively evaluate their enjoyment of mundane tasks based on the time it seemingly took them. In their study it was shown that users who knew of the "time flies" phrase would rate a task as more enjoyable when it felt shorter than it was. While both [9] and [8] investigate how users respond to a perceived acceleration of progression, moment to moment behaviour is not considered. However, looking at [8] one might suggest that irregularities in progress indicator rates might garner a positive response if they act a certain way. Additionally, since [9] confirms that perception management can manipulate time perception, they same may very well be true for moment to moment behaviour. Both thoughts relate to the hypothesis formed in Section 1, which examines the latter and rejects the former.

[3] supplies a philosophical and psychological background to human time perception, and explains how humans perceive time by observing environmental

changes and repetitive actions. Estimating time when affected by colour is investigated by [2]. This paper concludes that the color blue, when compared to a set of other hues, has a calming effect on users, causing time to be perceived as passing quicker. One may assume that this indicates that both mood and interpretation of aesthetic design elements may influence user response.

The nature of feedback and positive user responses has been connected by several studies. The paper [1] studied the correlation between user attitude toward applications and how informative the applications feedback and labels are. The authors hypothesize from their conclusions that system transparency is equated with trustworthiness by the users. [4] investigates user attitude to an application with and without a progress indicator using both variable and constant waiting times. The test results show that a constant waiting time does not have a significant effect on user time perception. Increased feedback is also argued to improve user response by [7]. The authors evaluated both classes of progress indicators in mobile apps. They conclude that an added textual percentage indicator increases satisfaction but not necessarily speed perception. The hypothesis formed in Section 1 assumes that behaviour may be informative and act as feedback. Thus, by the logic of [1], it should be able to invoke positive subjective responses from the user, such as trust.

The author of [6] suggests some time spans for users' waiting tolerance and argues which type of progress indicator is suitable for each span. The suggestion is that any loading that takes over 10 seconds ought to use a percent-done indicator, while a waiting time of 2 to 10 seconds may do better to use simpler indicators. When the expected waiting time is even shorter, the indicator is usually excluded to avoid flashing visual elements at the user. [5] also investigated how waiting time affects user tolerance. Their paper aims to find the point at which users give up when presented with infinite website loading. The authors show that the user notes a disturbance at around 2 seconds, but are significantly more willing to wait when a percentage indicator is present compared to when indeterminate indicators are used.

Of the aforementioned papers, only [5] and [7] did user tests aimed at comparing affect on user attitude to different progress indicators. However, both studied the differences between indicator classes, rather than focusing on behaviours of the determinate class, which is the focus in this study. Comparing moment to moment behaviour of determinate indicators seem like a generally unexplored subject, with regards to studying the effect on time perception and subjective response.

2 Preliminary notions

2.1 Classification

There are two distinct classifications of progress indicators [9]. The first is indeterminate indicators which do not indicate to the user how far the loading has progressed. The spinning hourglass falls into this class. Indeterminate indicators

are typically used for short loading times of maximum 10 seconds [6]. Determinate indicators on the other hand approximate the remaining time and display the progression towards completion [9]. Typically by incrementally enlarging a visual representation of loaded assets. The loading bar belongs to this class. See examples of both classes in Figure 2.

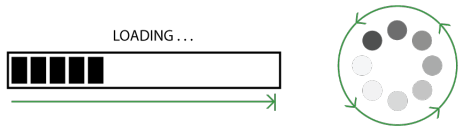

Fig. 2. A determinate progress indicator is shown to the left: A classic loading bar. An indeterminate progress indicator is shown to the right: Dots taking on colours to indicate a repeating spinning motion. The arrows indicate the progression change of the indicators.

Determinate indicators often use the number of assets to be loaded to calculate total and remaining workload, disregarding asset size. This leads to indicators which progress at an irregular rate, presenting an inaccurate time approximation. Another method is to, when possible, also take the size of individual assets into account. This produces a steadier progression rate but requires more work from developers since predicting the workload on the hardware for any specific asset may be complicated.

2.2 Behaviours

Some aspects of progression rate such as certain non-linear progression has already been shown to have a positive effect on user satisfaction and time perception [9]. This by achieving a perceived acceleration in progress. However, momentary behaviour such as pausing, increasing in larger chunks, and steady progression rates are not commonly compared. These are the three behaviours that will be defined and compared in this research. Other behaviours, such as the percentage counter and surrounding implementation, is kept constant and not considered to significantly influence results. The word increment is used in this study to refer to any visual step of progression in a loader.

Regular The Regular behaviour uses a constant progression rate. It starts at 0%, and increases in constant intervals with an increment size of 1%, until the completion of 100% is reached. The result is a smooth loader progression at constant speed and small step sizes. See Figure 3.

Pause The behaviour defined here as Pause has a progression which pauses shortly at several points. In order to reach the completion of 100% in the same amount of time as Regular behaviour, Pause needs to increase at a higher average rate than Regular. Between each pause the behaviour increases with a new

interval. Like Regular behaviour, the increment size is always 1%, going from 0% to 100%. The result is a loader progression which has small step sizes and varying speeds between pauses. See Figure 3.

Chunks The Chunks behaviour increases from 0% to 100%, but with larger increment sizes. Each increment has a different size and the time it takes to progress is relative its size. Thus, an increment chunk of size 23% will take the same amount of time to be added to the progression as Regular behaviour will take to increase 1% 23 times. No constant interval is used in chunks. The result is a loader progression which "jumps" directly to percentages between longer pauses than the Pause behaviour. See Figure 3.

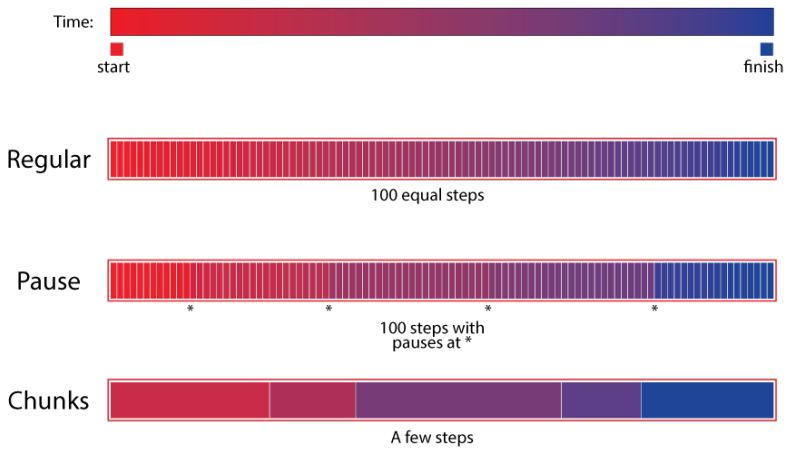

Fig. 3. Figure of all three progress indicator behaviours over time. Time is represented by a gradient from red to blue. The colour of each increment shows at which point in time it is added to the bar. Note that the figure does not represent the final design in terms of colour, and only aims to describe the general behaviour.

3 Methodology

The aim of the user tests conducted was to find if user perception of speed and satisfaction related attributes is affected by the progression rate behaviour. Inspiration for evaluating satisfaction as a key point came from [7] as subjective opinion and experience is at least equally important to speed perception.

In order to evaluate with less bias, users were not told that the loading rate experience was the subject of the research. Instead, they were told that spontaneous user perception and attention were tested. This way they were more likely to pay attention to each aspect of the test, while not overly focused on the loading time. The loaders had two possible completion times. If only one constant

completion time was used, the user might have instinctively approximated the same amount of time for all loaders in the evaluation forms. Using noticeably different completion times, the users would be less inclined to simplify evaluation by answering the same time approximation for every loader. Additionally, according to [4] this variance itself should not affect time perception, allowing the behaviour itself to be the reason for any difference found.

Consent and anonymity of the test user were established before the test begun. Additionally, the way that the results of the study would be presented was explained. After the user completed the set of tasks in the application and answered the task specific set of questions, they were debriefed on the true purpose of the test. They were asked if they figured out the true purpose of the test on their own, and if they felt that this influenced their answers. Thus, possibly biased results could be identified and removed. After the debrief they were presented with the progress indicator behaviours and asked to evaluate which one they felt held their attention and which one they favoured overall.

Test users were also asked about their professional background, computer experience and if they were currently feeling stressed. These are aspects that could affect the user's evaluation and perception during the test. Thus they were noted in order to discuss whether or not they might have affected the results. The test was conducted on twelve participants. Their backgrounds included technician, librarian, economy student, and interaction designers among others. Their ages varied from 20's to 60's. Only one user did not consider themselves experienced with computers. Half of the users reported that they felt stressed, some in the moment and some generally in life. Only one participant noted that they figured out the true purpose of the test. They assured, however, that they made a conscious effort not to let this affect their answers. Thus, their results were marked by the supervisor to be reviewed later. As their results did not significantly differ from other participants' results upon investigation, they were considered valid and kept in the statistics.

3.1 Test setup

The test had to be conducted in an environment without visual or auditory distractions. There the user was presented with the test application. The user was informed that they could ask the supervisor for help regarding the word tasks or if other complications arose.

In the application the user was first introduced to a start screen only containing a title and some instructions on how the tasks would be structured as well as which attitude to have toward the tasks. From there they could press a button to start the test. See Figure 4.

Once start was clicked, a progress indicator loaded the first task which was then presented. In the task the user was asked to select any number of words which they felt related to the given theme, from a list of 30 words. The task view can be seen in Figure 5. The word amount of 30 was found adequate after pilot testing and user feedback. 30 words require some cognitive effort for the

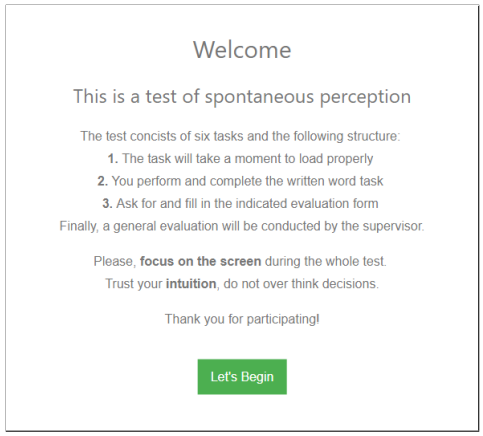

Fig. 4. Image of the test application starting view. Note that the browser and some empty space has been cropped out.

task evaluation and does not bore the user into not caring about their selection. The choices the user did in the word task were not recorded in any way.

Once done with a task, the user could click a submit button and a prompt instructed them to ask for the evaluation form corresponding to the id shown on screen. The id consisted of a number and a letter, see Figure 6. The letter corresponded to a certain task and its evaluation form. The id number corresponded to which indicator behaviour preceded the task and was noted on the evaluation form before the supervisor handed it over to the test user. The evaluation content is elaborated on in Section 3.3. When finished with the survey the user could move onto the next task. This was repeated for all tasks and then the application part of the test was complete. The order in which the user completed the tasks was noted by the supervisor.

Each behaviour appeared as a progression indicator twice. There were six tasks in total. Task and indicator behaviour order were random in order to minimize potential bias in users. The tasks were designed to be as equal in stimulus as possible in order to not change user experience which might have affected time perception and tolerance. The choice of words for each word task were therefore randomized as well.

When the user was finished they were shown all three loader behaviours and asked to do a short evaluation on satisfaction, comparing the behaviours.

3.2 Implementation details

The mock web application used to present tasks and loaders was built with HTML, CSS and Javascript, and stored locally on a laptop.

Each loader was programmed to act exactly the same for each user. The completion times for the indicator were 10 and 14 seconds. Each behaviour was presented twice, once using each completion time. There was also slight variations between the two presentations of Chunks and Pause so as to make their

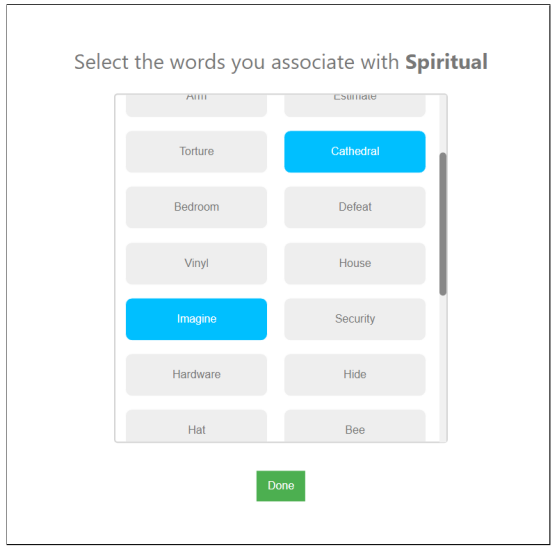

Fig. 5. Image of test application task view. All word tasks lists 30 words for the user to freely select from. Users had to use the scroller to access all words. Note that the browser and some empty space has been cropped out.

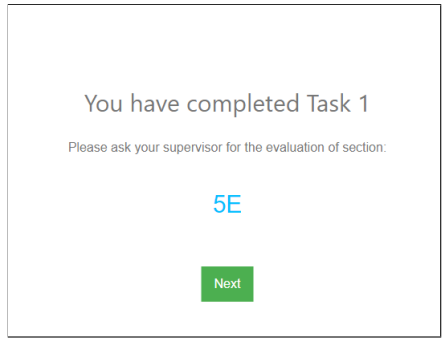

Fig. 6. Image of test application survey view. The id number and letter corresponds to the indicator behaviour and word task respectively. Note that the browser and some empty space has been cropped out.

irregularity less predictable. Which variation that had which completion time was always the same due to programming limitations. The Pause behaviour had two different pause patterns: pausing at 10%, 58%, 63% and 96%, and pausing at 4%, 44%, 67% and 90%. In order to achieve the correct completion times the length of pauses and progression speed between pauses varied as well. The points at which pauses occurred were distributed with the intention to imitate a random pattern, yet still be relatively evenly distributed in conjunction with the varying progression speed. The Chunk behaviour incremented according to the two following patterns: 15%, 35%, 70% and 90%, and 10%, 30%, 65% and 85%. The increment sizes were mirrored: 15, 20, 35, 20 and 10, in order to avoid non-linear progression bias [9].

The design was chosen to be as minimalist as possible, while still using a percentage indicator to add another feedback aspect. The percentage adds satisfaction to the design as argued by [7]. Attempting to add satisfaction in the design lowers the risk of a biased low evaluation of satisfaction, regardless of behaviour. The added feedback also acted as another visual indicator that can hold user attention. Keeping the user focused on the application is important for evaluation of the loader. Yellow, orange and red are colours that are attention grabbing[1], but have therefore traditionally been used to indicate caution, error, and warning within technological contexts. Therefore they were ill suited to indicate loader progression. While blue is a color associated with technology and security, it has also been shown to have an effect on time perception [2]. Green is another culturally common color and is associated with action and success, which communicates the purpose of the loader. Colour is culturally dependent and very much up to individual taste, however in the end an arbitrary green hue was seen as the most suitable. The final indicator design can be seen in Figure 7.

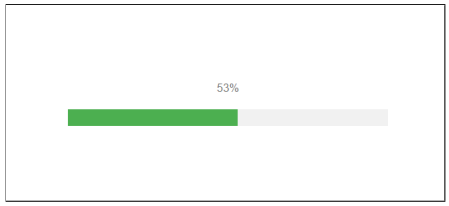

Fig. 7. Image of test application loader view, specifically of the Pause behaviour. Note that the browser and some empty space has been cropped out.

The completion times 10 and 14 seconds were chosen as they have a mean completion time of 12 seconds, just over the 10 second limit suggested for percentage done indicators [6]. The difference of 4 seconds was chosen because it creates a noticeable change, so that users were more likely to pay attention to the change and the time in general.

[1] https://99designs.com/blog/tips/color-meanings/

3.3 Evaluation format

The evaluation was separated into task specific answering sheets, and a final general answering sheet once the debrief of the test had been done.

The task evaluation forms consisted of ten questions of which seven related to the words in the task. Some of these questions appeared for several tasks, some were unique for each task, and some compared different tasks. Two of the remaining questions examined other aspects such as sound, color, and button labels. While these questions encouraged attention and distracted from the test's purpose, they were not relevant to the study and their answers not documented. The only exception being the tenth question: "How many seconds did the loading take?", which was always present and documented. Memory related questions such as "What was the first question in the list?" and "How many of your selected words began with B?" served to engage the user in the tasks. Questions like "What did the button in the previous view say?" on the other hand encouraged users to pay attention to visual aspects. All questions not relevant to the study were slight variations of the ones noted above.

The final evaluation began with the debrief of the true test purpose and questioning whether or not the user deducted this themselves. Then, users were asked to grade each of the three progress indicator behaviours on several attributes using a Likert scale. These included: Efficient, Engaging, Frustrating, Fun, Informative, Interesting and Predictable. A 10-point scale was chosen, using labels only for anchor points, namely "Barely" and "Very". The even number of points removes neutral answers and forces users to take a stance toward each attribute. This served to encourage more thoughtfulness in evaluation[2]. The relatively high amount of points on the scale also helped users express small subjective differences between behaviours. Finally, the final evaluation form was concluded with two elaborating questions on the behaviours: "Which behaviour would you generally prefer?" "Why?" and "Which behaviour does you eyes feel drawn to?" "Why do you think that is?".

The first attribute, Efficient, referred to the impression the user got regarding the imagined process behind the loader. In essence, if they felt that the process was going well. Informative was present to see how well the feedback was received. Predictability is key in time management and planning, and therefore was included as well. Engaging, Interesting and Fun were chosen to cover different aspects of how distracted the user would get from the passage of time. Engaging refers to how invested the user felt in the progression. Interesting aimed to gauge how likely the user was to pay attention. Fun let the user show if they found any enjoyment from the loader. Frustrating was used to let the user express and compare annoyance of any kind, hoping that the reasoning would be caught by other evaluations. These attributes are all abstract concepts which partly overlap. This was intentional in order to cover a lot of ground for discussion.

[2] J. Losby and A. Wetmore. CDC Coffe Break: Using Likert Scales In Evaluation Survey Work. Centers for Disease Control and Prevention. `https://www.cdc.gov/dhdsp/pubs/docs/CB_February_14_2012.pdf`

4 Results

All results of the collected time estimations are summarized in Table 1. Using an f-test, the variances were determined unequal and assumed as such in the following t-tests. None of the results significantly differ between behaviours when investigated using two-tailed t-tests of a 5% significance level.

User	Regular 14s	Regular 10s	Chunks 14s	Chunks 10s	Pause 14s	Pause 10s
1*	25	10	22	25	20	13
2*	7	7	20	10	20	15
3	3	3	5	4	5	3
4	10	10	15	5	15	10
5*	10	10	15	12	10	10
6	15	15	10	10	15	8
7*	8	8	15	8	10	10
8	10	10	13	8	12	7
9**	30	10	15	10	6	8
10**	7	5	7	5	10	7
11	15	10	10	7	8	12
12	5	N/A	15	10	5	N/A
Mean	12,083	8,909	13,5	9,5	11,333	9,363
VAR	65,356	9,891	24,090	29,909	27,515	10,854
STD	8,084	3,144	4,908	5,468	5,245	3,294
Count	12	11	12	12	12	11
Conf.	5,136	2,112	3,118	3,474	3,332	2,213

Table 1. Table containing all user time estimation data for each behaviour and true time. * notes a user that reported feeling stressed in their day to day life. ** notes a user that reported feeling stressed during the test. VAR is variances and STD is the standard deviation. Conf. is the confidence interval at 95%.

A visual representation of the time estimations along with 95% confidence intervals are shown in Figure 8.

One user answered "don't know" for the estimated time of the loader for the behaviours Regular and Pause, of true duration 10 seconds. Thus, the mean value for those two results are based on the remaining eleven test users. Another test user figured out the focus on progression indicators during the tasks. When asked, they assured that they had not been counting the seconds during loading in order to estimate the elapsed time more accurately.

The mean values of the subjective attribute evaluations and variances are shown in Table 2. The variances were investigated using f-tests and very few could be considered equal. Those were the variances for Engaging, as well as Interesting when comparing Regular and Chunks, and Fun when comparing Chunks and Pause. T-tests were later conducted with regards to which variances could be considered equal.

Confidence intervals for the attribute evaluation for each behaviour and attribute are shown in Figures 9, 10 and 11.

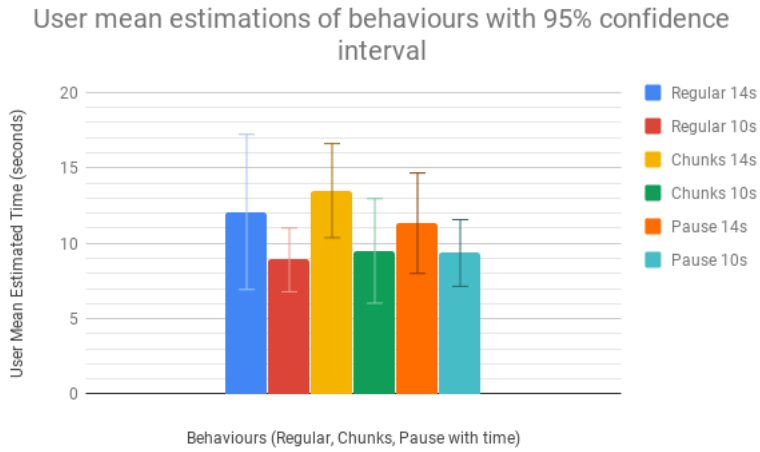

Fig. 8. Diagram of the mean estimated times as well as confidence intervals of 95% for each behaviour and time.

Subjective value	**Regular (variance)**	**Chunks (variance)**	**Pause (variance)**
Efficient	8.6 (0.97)	4.5 (2.45)	6.6 (3.54)
Engaging	6.2 (5.24)	4.3 (5.11)	6.3 (5.39)
Frustrating	3.0 (1.64)	7.9 (1.72)	6.7 (4.75)
Fun	4.4 (4.45)	3.5 (5.72)	4.0 (4.45)
Informative	8.3 (3.52)	4.8 (4.39)	6.8 (2.52)
Interesting	4.4 (5.17)	4.8 (5.06)	5.8 (3.97)
Predictable	9.6 (0.99)	4.0 (4.00)	4.3 (2.20)

Table 2. Table containing the mean user evaluation for subjective attributes with regards to the different progress indicator behaviours. The attributes had a minimum score 1 and maximum score 10, as the evaluation was conducted with a 10-point Likert scale. The variance is shown in parentheses.

Fig. 9. Diagram of the mean evaluation of each attribute for the Regular behaviour as well as confidence interval of 95% for each attribute.

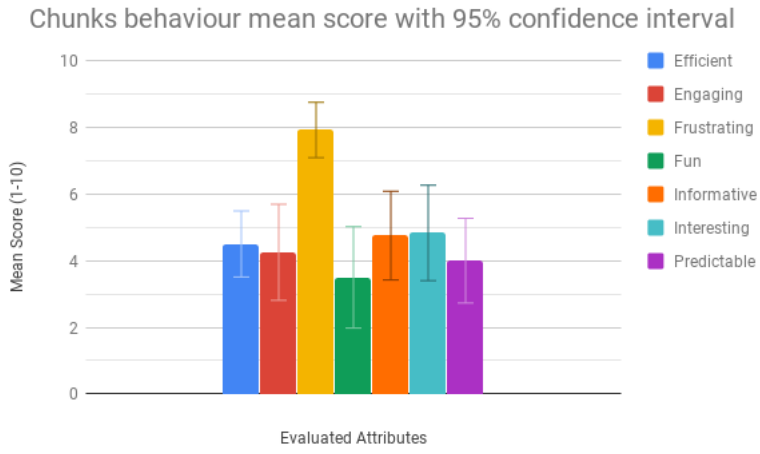

Fig. 10. Diagram of the mean evaluation of each attribute for the Chunks behaviour as well as confidence interval of 95% for each attribute.

Fig. 11. Diagram of the mean evaluation of each attribute for the Pause behaviour as well as confidence interval of 95% for each attribute.

When doing t-tests on the collected data with a 5% significance level, the following statements can be made about the attribute evaluation: The Regular behaviour is shown to be significantly more efficient, predictable, informative and less frustrating than both Chunks and Pause behaviours. Pause is significantly more efficient, engaging and informative than Chunks.

All test users answered that out of the three behaviours, they generally preferred the Regular behaviour. Noted motives behind their choice include ease of time estimation, feeling of progression, feeling of security and trust, informativeness, predictability and a sense of it being the least frustrating.

Upon answering which behaviour drew their attention the most when viewing all three side by side, responses where more diverse. Three chose Regular, stating that since it always moved their eyes were constantly drawn back to it. Four chose Chunks and argued that its sudden large changes and "movement" drew attention and increased the desire to see it continue. Five test users chose Pause, which was then mentioned to be more exciting and fun due to it's unpredictability and seemingly random behaviour. One user also noted that it drew attention since, at the start of loading, it progressed the fastest out of the three.

The stressed users approximated the elapsed time of loaders as slightly longer than non-stressed users. Calculating separate mean values for stressed and non-stressed users results in a 4 second difference for the 14 second duration. For the 10 second duration, this difference dropped to 2 seconds longer.

5 Conclusion

The hypothesis presented in Section 1 can not be completely confirmed using the results of this study. There are no significant differences in user time approx-

imation for the duration of loaders of different behaviours, so there can be no assumption made about these specific behaviours affecting time perception in users.

This study looked only at a waiting duration of 10 to 14 seconds, the shortest time span in which it is advisable to use determinate progression indicators [6] like the loading bar. Thus, this study is relevant when the waiting time increases to a point were consideration is put into changing an indeterminate progression indicator for a determinate one. In that case, there is not enough evidence to argue that development resources need to be spent on changing the progression behaviour in order to affect user time perception. If the waiting time is longer, it may still have a significant effect.

Significant results however, were found when looking at emotional responses to the progression indicator behaviours. As user attitude towards applications can be influenced by the user experience in the design of the system [1], it is important to consider these results when designing for user experience. The Regular behaviour was evaluated significantly less frustrating than both Chunks and Pause, which suggests that it is beneficial to strive for this behaviour if one wants to gain user favour in this respect.

Predictability is an attribute that was highly ranked by test users upon answering why they preferred one behaviour over the others. The Regular behaviour was also evaluated significantly higher than the other behaviours with regards to this attribute, and the fact that it was unanimously chosen as the favourite by users reflect the importance of predictability to users.

Generally, a behaviour close to Regular should be striven for when designing applications that rely on determinate progress indicators of this type and in similar contexts. If a Regular-type behaviour cannot be achieved, the next best thing is the Pause behaviour which was favourable to Chunks with regards to efficiency, engagement and informativeness. As the test user base consisted of only twelve test participants, further tests must be conducted to confirm and support the conclusions drawn from this test.

6 Discussion

In the results gathered in this study, there is no foundation for irregular behaviour having positive effect on the experience of progress indicators, as might have been suggested from the findings of [8]. Instead, regular, predictable behaviour seem favoured which may in turn have improved perception on informativeness and efficiency.

The irregular aspect of the behaviour did gain some attention, which can be explained by human natural attention to change. Changes in our environment are a stimuli, distracting us from the passage of time, which may speed up our perception of time. However, changes are also a way for us to measure time [3] and if they are not stimulating enough, they simply aid us in noticing the passage of time instead. While users evaluated the irregular behaviours Chunks and Pause as less predictable than Regular, they were not evaluated as significantly

more stimulating by the attributes engaging, fun or interesting. That seems to be the reason that the irregular behaviours lost user favour. Still, one user argued that the Pause behaviour did keep their attention due to being more random and fun. The unpredictability of the behaviour may therefore act as a stimuli. Whether individuals keep or shift attention upon waiting may depend on how experienced they consider themselves to be. As humans gain experience through repeated events, the event itself loses its stimulative effect over time [3]. Computer experience could therefore play a role in this. However, in this study only one participant noted that they were not experienced with computers, and this user's results did not stick out in any particular way.

Stress seemed to have an impact on time perception as well, although little can be said without a larger study. It could also have caused more homogenized results for the behaviours, as it might be that we tend to generalize under stress and discomfort.

Varying professional backgrounds and ages were sought and found for the test participants in order to eliminate bias of interests in technology and generational shifts. Regardless of background, a majority of the test users were keen to expand on their answers regarding their favourite behaviour in order to talk about other frustrating behaviours they had encountered in real life. From this generally eager attitude one can argue that user opinion on this subject is not negligible. Users care, speak of and compare their experiences with progress indicators.

As the ages of test participants ranged from 20s to 60s, the results could have shown significant differences between users of different ages. This since members of different generations generally have different experiences due to the available technology growing up. Thus they may react and evaluate the loading behaviours differently. However, no notable differences were found between the users of vastly different ages.

Much of the test was designed to eliminate any possible influence on loader opinion from the setup. However, equal care was not taken with regards to the environment. While the environment was kept calm and other people were not permitted in the room, other distracting elements such as colors, distant sounds, windows and other background stimuli were not considered influential enough to affect the results. This could have been avoided entirely if the same location had been used for every test.

Upon having conducted all tests, a suggested adjustment to the evaluation setup is to change the way that the users approximate the time it took each loader to load each task. In the method used, the user could freely write their own guess. As a result, users tended to limit themselves to certain values, such as 5, 10, 15 and 20, which eliminated some of the potential nuance to the evaluation. If users had been given explicit options instead, such as a scale, they might have gone for more specific answers such as 9 or 13 seconds.

6.1 Limitations

As this study aimed to gather time approximations of loader duration, the number of user tests were crucial. As only twelve tests were conducted, a lot of

accuracy is left to be desired. Each test took from 30 to 50 minutes to conduct which made planning and recruiting test users of varying backgrounds difficult within the time frame of student conference course.

Since this study had loader duration of 10 to 14 seconds, nothing can be said about user responses to loaders when waiting much longer than that. Processes such as system updates and booting up complex applications can take several minutes and may yield completely different responses. Longer unpredictable processes, such as studied by [5], require another study and a different test setup. The time span inspected in this study is suited mainly to web and mobile applications which generally avoid such long processes entirely in order to keep the user invested.

6.2 Future work

For future work within this field, conducting the same type of test on a larger scale would garner more accurate and reliable results. An expanded study could also include longer time spans to find if the effect on time perception from different behaviours may take significant hold at some point, or if it is simply non-existent when compared to time management techniques such as non-linear progression [9]. In order to gain deeper insight into the correlation between user opinion on progress indicator behaviour and on application opinion, one could make use of extensive interviews and a psychological perspective.

7 Acknowledgements

We would like to thank the study supervisors, Suna Bensch and Thomas Hellström, who provided many insights into test setup and report structure. Thanks to those who participated in the test and to the members of the peer review group who carefully studied and corrected faults in logic, grammar and methodology several times. Finally, some recognition to the inspiration for this paper, the study of [7], which brings attention to how loaders affect our sense of time and satisfaction.

References

1. Abe Crystal and Sririm Kalyanaraman. Usability, cognition, and affect in web interfaces: The role of informative feedback and descriptive labeling. In *Proceedings of the 54th annual conference of the International Communication Association*, 2004.
2. Gerald Gorn, Amitava Chattopadhyay, Jaideep Sengupta, and Shashank Tripathi. Waiting for the web: How screen color affects time perception. *Journal of Marketing Research*, 41:215–225, 2004.
3. Robin Le Poidevin. The experience and perception of time. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2015 edition, 2015.
4. Brad Myers. The importance of percent-done progress indicators for computer-human interfaces. In *ACM SIGCHI Bulletin*, pages 11–17, 1985.

5. Fiona Fui-Hoon Nah. A study on tolerable waiting time: How long are web users willing to wait? In *9th Americas Conference on Information Systems, AMCIS 2003, Tampa, FL, USA, August 4-6, 2003*, page 285, 2003.
6. Jakob Nielsen. *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
7. Hannah Pehrson and Nikola Pantic. From zero to hundred in eight seconds. Master's thesis, University West, Trollhättan, Sweden, 20015.
8. Aaron M. Sackett, Tom Meyvis, Leif D. Nelson, Benjamin A. Converse, and Anna L. Sackett. You're having fun when time flies: The hedonic consequences of subjective time progression. *Psychological Science*, 21(1):111–117, 2010. PMID: 20424031.
9. Steven Seow. *Designing and Engineering Time: The Psychology of Time Perception in Software*. Addison Wesley Professional, Glenview, IL, 2008.

Measuring how colors and geometrical shapes influence reaction times

Ivan Lyxzen

Department of Computing Science
Umeå University, Sweden
lyxzen.ivan@gmail.com

Abstract. In this study we measured the effect on reaction time when users were presented with different types of data: colors, shapes and the combination of both. Specifically this study aimed at looking into how using colors and shapes at the same time would alter reaction time compared to only using color or only shapes. Based on earlier research there were two main hypotheses that could be found to be the main factor in reaction time tests. The first is that using both colors and shapes would add unnecessary information and increase the users cognitive load, thus decreasing reaction time. The other is that using colors and shapes increases contrasts and thus increase reaction time. The study found that using colors and shapes in combination is comparable to, but slightly worse than, using only colors and that using only shapes is by far worse with regards to reaction time. There is still a need for further research into this effect and whether it scales to larger sample groups.

1 Introduction

This study aims at measuring how color and shapes influence reaction time with the purpose of finding out how using both at the same time affects the results. It measured reaction time with a simple game programmed in HTML and JavaScript. In the game five squares are positioned horizontally in a line, each square has a unique color except for two. The two squares that have a matching color make up the only color pair. The goal of the game is to click on either of the squares that are a part of the pair and then five new squares present themselves. This continues for one minute in which the user attempts to get as many correct clicks as possible. Since colors are simple pieces of data they can transfer information very quickly [1]. Compared to the colors used the shapes used, see Figure 3, are more similar to one another and therefore should require more processing. Colors are distinct and contain only one piece of relevant information, the color itself, while shapes are more complex and each line needs to be processed for the person to understand what shape they represents. Being presented with both colors and shapes at the same time could either lead to an excess of data and therefore slow down reaction time, or it could lead to a more clear distinction between the squares and thus improve reaction time. Not knowing which of the two previously stated hypotheses are correct can lead to a

S. Bensch, T. Hellström (Eds.): Umeå's 23rd Student Conference in Computing Science USCCS 2019, pp. 41–50, January 2019.

designer implementing a sub optimal solution for their user interface. Knowing how using both color and shapes at the same time affect reaction time might prove useful for system designers when deciding how to code their user interfaces, especially if the UI is time dependant. After the tests were conducted we can see that simply using colors improves reaction time slightly but also increases the number of misses by a larger amount, more on this in the result section.

2 Theory

This study is based on the matching, in other words coding, of sets of data, along with reaction time. The theory section is therefore split into two parts, the first of which is related to color coding and the second to reaction time. The previous studies that were most relevant for this one were mainly [2] which discussed reaction time using color coding and [3] which looked at reaction time, visual filtering and interpreting shapes. The coming two subsections will go over earlier studies and look at their findings. While this paper focuses on the difference between using colors and using both colors and shapes in reaction time testing, the subsection color coding will only look at color coding and its affect on many aspects, such as learning, navigation and reaction time. The reaction time subsection will only look at using reaction time in general with studies about age, male and female difference in reaction time and how humans filter information. Studies looking specifically at the use of colors and shapes in combination with regards to reaction times are generally hard to find.

2.1 Color Coding

Color coding is the use of structuring information in such a way that items or data pertaining to a particular type or class is presented in the same color.

An area where color coding is useful is finding an item from a large set of similar items. When looking for a specific object in a set one can limit the number of objects one needs to focus on by specifically targeting a certain color. The amount of time needed to find the object is linearly correlated to the number of objects needed to search through [2][1]. For instance if a test person is presented with a set of random digits and told to find a specific digit the search time is correlated with the total number of digits. However if half the digits are green and the other half are red and the same user is told to find a specific digit that is red they only need to search through half of the total number of digits. Therefore if color coding reduces the number of object needed to search through by half the search time is twice as fast [2]. A color code that eliminates a majority of unwanted digits would be an example of a good color code while one that only eliminates a few digits would be a poor color code.

In a monochromatic environment colored object are more salient, this is one of the reasons why color coding increases searching efficiency [4]. Another reason

[1] BF Green, WJ McGill, HM Jenkins - 1953 - MIT Lincoln Laboratory, The time required to search for numbers on large visual displays

for the effectiveness of color coding when it comes to learning new material is that users can more easily relate different objects to one another via the color [4].

There are areas where color coding is useless, such as in spatial navigation [5]. Users who are presented with a color code designed to help them navigate a virtual reality space did not improve their time nor their choices of paths [5]. This indicates that color coding is not suited for navigation in a three dimensional real world, instead its application might be useful solely in two dimensions. More research on color coding in 3D environments is required to draw any final conclusions.

In certain situations color coding is not only inefficient it can also be damaging if used in an inconsistent ways [6]. As a result the learning process can slow down. If the user has to spend a lot of time getting familiar with a complex and illogical color code when the underlying data is simple this can take more time than using non coded material [2]. Therefore one should be careful when using color coding and conduct testing on its efficiency before publishing material in which learning is time sensitive, such as the use of military equipment [2].

When users are presented with new data they use cognitive abilities to process it [6]. Combining visual information along with text improves learning [4, 7, 8] since two different channels of cognitive load can be used simultaneously without much interference [4]. An everyday example of using two different cognitive channels is driving a car while listening to the radio. In this case the music takes very little of our attention away from driving, this is not the case if the driver is texting. While texting they need to focus their visual attention on their phone and their driving, using the same cognitive space for both leads to one or both being impaired. The same principle extends to color coding since the processing of color requires very little cognitive load but helps make important information salient, easier to recall and to connect to similar data [4].

2.2 Reaction time

Studies have shown that reaction time varies with age. For simple reaction time tests the reaction time slowed at a rate of 0,5 msec per year [9]. This will not play a big part in this study since all participants will be of roughly the same age. Reaction time may vary for many different reasons other than age. One of which is gender [10]. Women tend to be more accurate when doing visually related reaction time tasks but as a consequence have slower reaction times. The gender difference is of moderate size, though [10] argues that with practice doing visually related tasks the gender difference may change and further research is needed. The reasons for overall slower reaction times for women is contributed to a general speed factor but also more specifically in their study:

- Slower mental rotation
- Slower mental comparison
- Possible bias towards higher accuracy

Of these factors only the mental comparison and the possible bias is relevant for this study since no mental rotation will take place. Since the tasks that will

be tested in this study are of a generally lower difficulty we should be able to remove some of the gender difference. However, to maintain accuracy an equal amount of men and women will be tested. If the bias towards accuracy is relevant we should be able to measure its effect in the percentage of misses between men and women in this study.

When it comes to reaction time in a visual search task humans are very adaptable. Humans can quite quickly filter what kind of information they are looking for and ignore unwanted data [3]. This means that when looking through data that visually looks similar to one another it will be harder to filter out the unwanted information. Since people cannot be sure it is unwanted until they take a closer look. This means that in this study we should more easily be able to filter the colors since they are more unalike compared to the shapes. Another point of interest might be that a circle will be easy to discern unlike the other geometrical shapes since it contains no edges or straight lines, making it easier to filter out when looking for a triangle. In [3] they also found that with practice it is possible to compare multiple items at the same time almost as quickly as a just two items. Subjects learned what to look for at a glance and could scan multiple shapes at once. For instance the test person might get familiar with looking for a particularly acute angle when scanning for a triangle.

3 Method

3.1 Test setup

The study was conducted at Umeå University on students around the age of twenty. This group was chosen since it was important to test on users of roughly the same age and this group was widely available. All students were informed that the test was anonymous and that only their numerical results would be used, their verbal consent was given. In total 21 students participated in the study. This number was deemed a minimal requirement to lower the effect of variance between each individual's reaction time. The participants were brought to an environment with as little background noise and disruption as possible to partake in the testing. Each participant did, in random order, the three tests. One in which they tested their reaction time using color coding, another with shape coding and finally one with both. These three tests will be noted as *colors, shapes* and *both* from now on. The tests were done on a laptop using a mouse to move the cursor. To avoid any influence on the result due to the users learning the test and thereby getting a quicker reaction time the tests were done in a random order.

Before the testing began, the process was described in detail and the users were allowed to ask any questions. When they were ready they did a short trial run to get familiar with the movements of the mouse and how the game worked. The trial run was conducted on the test that they were randomly assigned to do first. After that the tests began properly, each test taking one minute. The participants total score was calculated and also their total number of misses. The game itself will be described in detail in the next subsection.

After the three tests were done a few questions followed. Questions pertaining to which test felt most difficult and why, and which felt easiest and why. This might help explain the difference in scores between the tests and also give insight into how the subject perceived the tests. Meaning, that even though the score might indicate no difference between *colors* and *both* there might still be a perceived difference.

3.2 The game

The game is programmed in HTML and JavaScript to be as simple as possible. The game begins by clicking on the start button, this also functions as a reset button if one should be needed. Once the game starts five squares are presented in a horizontal line, see Figure 1. In the five squares there are four colors spread out randomly. This means that the patterns are not set beforehand and also that there will always be exactly one color pair. The total score will also be updated live as shown in Figure 1 and presented at the top of the page, the same is true for the total number of misses.

The objective of the game is to get as many correct clicks as possible, the number of misses will be irrelevant in the results. Though, the users are not told this since it might make them more eager to simply guess the correct answer instead of actually scanning the squares. If the user clicks on any square five new squares will appear, no matter if they miss or not. It's the job of the test leader to measure the time elapsed. The test will be conducted a total of three times, one for each type of coding. The order of the tests is random. *Colors* is shown in Figure 1, *shapes*, Figure 2, and *both*, Figure 3.

The reason for the choice of shapes, see Figure 2, is to make no one shape completely unique. The aim is that there should be only minor differences when trying to scan for a square compared to a triangle. Therefore there are straight lines, sharp angles and roundness represented in multiple shapes to keep them similar and to not make one or more of them easier to scan for than the others.

The colors were chosen from the primary colors red, blue and yellow with the addition of green. Since the primary colors are most distinct this should give the biggest contrast between the colors but since there are only three primary colors green was chosen at random, other choices for this fourth color could have been orange or purple who also fit between two primary colors.

A possible bias in the test will be the different levels of contrast in *both*, see Figure 3 center square and center right square. The circle in the yellow stands out more since black has higher contrast to yellow than to blue. This is hard to remedy when striving to be consistent in the colors of the shapes from Figure 2(the shapes have black outline) when moving them to the final test, an option for future testing would have been to alter the colors of the shapes to maximize the contrast for each background. How this would affect the results is currently unknown but is a subject that might be interesting in pursuing further.

Fig. 1. The game with only colors, also showing the total score and total misses, top left

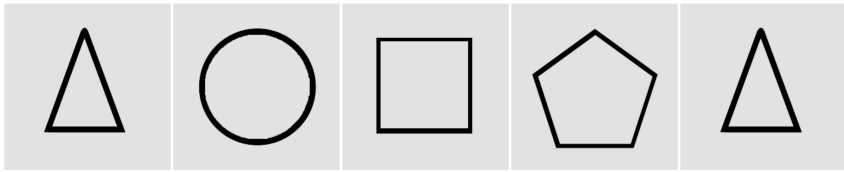

Fig. 2. The game with only shapes

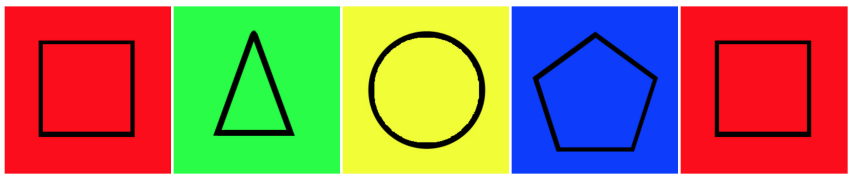

Fig. 3. The game with colors and shapes. Note that the square will always be red and the triangle always green, in other words the colors and shapes stay static in relation to one another.

4 Results

A total of 21 students have participated in the trial, 10 females and 11 males, all participants were in their twenties. The results for for each individual test can be seen in Table 1.

	Colors	*Shapes*	*Both*
Total score	1305	937	1288
Average score	62,14	44,62	61,33
Average score/Second	1,04	0,74	1,02

Table 1. Total and average scores

The test with the best performance was colors but only with a slim margin not statistically significant. Both was not far behind when looking at averages or score per second. As previous research indicated [3] shapes was quite far worse regarding general reaction time with a p-value of less than 0.0001 in a two tailed T-test.

The results of misses in each individual test can be seen in Table 2.

	Colors	Shapes	Both
Total Misses	16	5	6
Average misses	0,76	0,24	0,28

Table 2. Total and average misses

Here the test with the most misses was also the test with the highest overall score, colors. Among the other tests the number of misses remained overall very low with most participants having no misses at all.

The results of which test was perceived as most difficult and easiest can be seen in Table 3.

	Only Colors	Only shapes	Both at once
Perceived as Easiest	13	0	8
Perceived as most Difficult	1	19	1

Table 3. Perceived difficulty

In this study colors was frequently thought of as easier than both, and shapes is vastly perceived as the most difficult. When looking closer at the comparisons between colors and both we can see that all results from Tables 1 through 3 are leaning in favor of colors, though for the most part only with a slim margin. With the one exception being that the number of misses in colors is a lot higher. More on this the Future works and Discussion.

5 Conclusion

The results between *colors* and *both* was not statistically significant and a larger sample size is required to draw final conclusions. The results might indicate

that *colors* maximizes reaction time but decreases accuracy. Using *both* lead to high reaction time, though slightly lower than *colors*, and better accuracy than *colors*. Using *shapes* results in a much lower reaction time but similar accuracy to *both*. *Shapes* is perceived as the most difficult test and *colors* as the easiest. This indicates that colors are easier to compare against other colors while shapes require more strenuous comparisons. In design aspects this means that when speed is a larger factor than accuracy we should focus on using colors to separate objects or sets of data, and when accuracy is more important we use shapes instead. More research is needed to be able to conclusively say that colors and shapes together increase accuracy without slowing down reaction time.

6 Discussion

As [2, 3, 3] found in their studies, we too can conclude that color increase reaction time. Though it appears that perceiving color happens in a different way compared to shapes. Shapes require much more thorough scanning and analyzing, which angles are present, how are the lines placed, is the symbol round and similar questions must be answered. However, when looking at colors the only question that needs to be answered is, is this color the same as another color. Many users reported that they could scan all colors at once, by letting their focus go and just hazily gazing over the entire screen they could grasp which squares were a match. This was not possible when looking at the shapes. It was interesting to find that a lot of users reported that when doing the test with both they simply ignored the shapes and did the same as with the colors, focusing on the entire screen. Since they did not have the same number of misses it suggests that even though their perception of how they solved the problem remained the same their outcome changed, meaning the shapes did in fact have an effect on them. As to what exactly caused the effect warrants further research.

During the testing very few people improved their technique and their method. There are several ways for the user to make each test easier, one of which is to always strive to keep your mouse in the center of the screen, since most of the pairs will have one or more squares in the center three squares. Some test users did report that they got more familiar with the mouse which might have had improved their performance. However, since the tests were conducted in a random order each time the effect of this should be spread out and can therefore be ignored.

There was also very little difference between genders as was suggested by [9]. In total 14 of the misses were generated by men and the remaining 13 where generated women. This means that there appears to be no clear correlation between women and a bias towards higher accuracy in this sample group. This runs contrary to the results found in [9], however a larger sample is required to refute their results. There were also 11 men and 10 women in this study and only the total number of misses were measured, not the frequency of misses compared to hits. Another future study could measure this more specifically. Something to note is that some of the students, mostly males, had experience

playing video games which does not only increase reaction time and accuracy [2] but often means they are familiar with handling a mouse. In this study the effect of video games might play a bigger role than gender or age. Since it is hard to tell exactly which factors will play a part in reaction time there are only two real solutions. The first is to have a much larger sample group where people are selected at random without any commonality that is apparent. The second is to test only on an extremely homogeneous group, say only males ages 20-23 who frequently play video games, or some other group with common traits. This of course depends on the exact hypothesis of the study being conducted. In some cases the efficiency of the participants does not matter at all, as was the case for this study. These are the cases where the only thing being measured is the difference between the tests, not between people.

7 Future works

Since the sample group was quite small for a study about reaction time none of the results will be entirely conclusive. However, based on earlier research and the results of this study we will be able to make adjustments and improvements to future works.

Since the results when comparing colors only and shapes only is quite clear from this study and other similar previous work the symbol only test could be excluded from a future version of this study. Instead one could focus on only measuring the difference between the colors and the one with both.

The game itself functioned quite well with only two exceptions.

Not everyone was familiar with using a mouse, especially when it was not their own mouse. A future version of this test could work with touchscreen or another mechanism that is less reliant on the users previous experience with it.

Green, yellow and red are closer to each other in the color wheel and therefore are more similar than for example red and blue. A future version of this test could work with only four squares instead of five so that only primary colors, with high contrast, can be used. The background can also be altered to increase contrast since yellow tends to blend into the white background more than the darker colors red and blue. To be consistent with color contrast a hypothetical game that uses ten colors have to select them from a color wheel at a constant rate, where the top of the wheel is the first color and then the next color would be the color located at $1/10$ of the way along the the axis, and so on.

Another avenue that one might embark upon in the future is trying to find out why the color only test resulted in a higher number of misses compared to the color and symbol test, even though both had roughly the same score. This could be applicable in works that want the benefits of color coding to increase reaction time but also want to minimize the number of errors, even it if comes at the cost of a slightly lower reaction time. An hypothesis could be that since colors require less processing it might lead to higher reaction time but since

[2] `https://journals.sagepub.com/doi/abs/10.1111/j.1467-8721.2009.01660.x`

there was less contrast between each square that also lead to more errors. A future study could try and prove this by testing only the four primary colors for maximum contrast and compare that with colors and shapes, then do the same for a different set of colors with less contrast. If the level of contrast corresponds to the number of misses it would indicate that the hypothesis was correct.

When it came to the perceived difficulty it was in favor of colors only but there was still a substantial amount of people who chose the test with both as the easiest. This might indicate that different people perceive these things differently. Finding out why might also be another possible future study. Based on the answers given by the test users the general feeling was that those who favored the *colors* simply ignored the shapes when they did *both* and those who favored *both* did so since it increased the contrast, especially between green and yellow.

8 Acknowledgement

I would like to thank all the people who helped make this paper into what it is today through their feedback and peer review. I also want to thank all of those who participated in the study.

References

[1] Tanja, K., Peter, G., Katharina, S., Bärbel, G.: Information visualizations for knowledge acquisition: The impact of dimensionality and color coding. Computers in Human Behavior (2006)

[2] Bert K, G., Lois K, A.: Color coding in a visual search task. Journal of experimental psychology (1956)

[3] Neisser, U.: Decision-time without reaction-time: Experiments in visual scanning. The American Journal of Psychology **76**(3) (1963) 376–385

[4] Sonja, F., Helge, R., Lorenz, S.: Processing and integrating multimodal material — the influence of color-coding. Proceedings of the Annual Meeting of the Cognitive Science Society, 27(27) (2005)

[5] Ranjbar, A., Fasllija, E., Aksel, E.: The effects of color coding on users wayfinding performance in a virtual shopping mall environment. (2016)

[6] Dzulkifli, M., Mustafar, M.: The influence of colour on memory performance: A review. he Malaysian Journal of Medical Sciences (2013)

[7] Erol, O., Turkan, K., Engin, K., Kursat, C.: An eye-tracking study of how color coding affects multimedia learning. Computers & Education (2009)

[8] Gutierrez, K.: The psychology of color: How do colors influence learning? (2016)

[9] Fozard, J.L., Vercruyssen, M., Reynolds, S.L., Hancock, P.A., Quilter, R.E.: Age differences and changes in reaction time: The baltimore longitudinal study of aging. Journal of Gerontology **49**(4) (1994) P179–P189

[10] Blough, P., Slavin, K.: Reaction time assessments of gender differences in visual-spatial performance. (1987)

The effect of digital gamification when learning a third language

Anna Nystedt

Department of Computing Science
Umeå University, Sweden
id14ant@cs.umu.se

Abstract. Gamification has been a popular tool within the area of education during the last decade. This is mainly because of its ability to provide rich learning experiences and create motivation and engagement among users. In this study we have examined what affect gamification has on a user's performance when learning a third language. The study compared the results of two groups using two different approaches for learning a previously unknown language, one that were based on gamification and one not including any gaming elements. The results showed no difference between the two methods regarding the long-term memory, but the group that used the method not including gamification performed significantly better when testing the short-term memory.

1 Introduction

Recently there has been an expanding interest for applying game-design elements and game principles in a non-game context. The term most commonly used for this type of application is gamification. Education is one of the areas where there exists a great interest for applying gamification [7], as games can be used to promote active learning, critical thinking, collaboration, give appropriate feedback and encourage progress [9].

Several studies show that gamification can be used as a tool to enhance engagement and motivation when learning new knowledge. Motivation is a crucial part of learning, as it usually resembles a power or an energy that drives people to do a task or achieve a goal [13]. One aspect that has a positive effect on engagement, immersion and the perceived learning, is when games provides a challenge to the user [12]. Individual gamification elements can also be used as a way to create engagement and increase motivation. A summary of eight studies [24] conducted within the area shows that experience points, badges, scoreboard and feedback are the elements with the highest impact on engagement, motivation and enjoyment. The elements that had poor or negative impact on these aspects were avatars, time track and competition.

Motivation has a positive effect on educational performance [23], and gamification can therefore be a useful tool for improving a user's learning capacity. However, when learning a third language the heightened motivation usually

S. Bensch, T. Hellström (Eds.): Umeå's 23rd Student Conference in Computing Science USCCS 2019, pp. 51–61, January 2019.

makes the user willing to spend more time on the learning activity [8], which improves the prerequisite for learning. What impact gamification has on the user's ability to learn has rarely been addressed in earlier research.

The aim of this study is to examine how the use of gamification affects the learning process of a third language. The study will look at broadening of vocabulary, and whether gamification can be used as a tool to improve the performance. The relations between gamification and learning are important to establish, since the ability to acquire new knowledge will always be relevant.

To examine the relations, a study was conducted. The results of two different learning methods practiced by two test groups were conmpared. The task for each test participant was to learn 18 different Hungarian words during a limited time. The first test group learned through an application based on gamification, while the other group learned without any game elements. The study showed no differences between the two methods when it comes to the long-term memory. However, it was shown however that the method not including gamification had a positive effect on the short-term memory.

2 Background

In order to understand the results from the study it is necessary to have a basic understanding of gamification and learning, both individually and in combination. This section therefore describes the foundation of the two fields with references to previous studies where these two areas were combined.

2.1 Gamification

The term gamification originate from the digital media industry [6] and its first documented use was in 2008. The term got a wide spread in the latter part of 2010 [21] as a result of the topic's popularity at conference presentations and the industry's adoption of the concept. Since then the application of gamification has been a thriving trend among many different fields, due to its ability to motivate, engage, increase user activity and retain consumers [6].

Gamification is achieved by adding a number of game elements and techniques in a non-game context. Each element has a specific purpose and can be adapted to suit several different types of environments [10]. A brief definition of game elements commonly used has been compiled in the list below:

Experience Points (XP) Used to reward users for activities.

Badges Visual representation of achievements.

Leaderboards Used to compete with others

Progress bars/Progression Shows the progress of the user.

Quests Missions the user should accomplish.

Levels Used to section the game.

Rewards/reward system Used to motivate the user to accomplish a task.

2.2 Learning

Learning is always present wherever people are situated and it is therefore a part of life that is both essential and inevitable [1]. A person's learning capacity can be streamlined by a variety of external factors. By using different memory techniques, a person's memory can be enhanced [16], which is a central part of the learning process. The techniques include actions like dividing the memory work over several sessions, recite material out loud, organize the material into meaningful patterns and to test yourself. Even a positive attitude can be a powerful tool to increase the memory - by consciously deciding to remember, the memory usually follows.

Learning style is one commonly used term within the area of learning. The term refers to a person's preferred way of absorbing and processing new information [20]. There are several different learning styles, but the most common are visual, auditory and kinesthetic [22]. The general perception is that information is easier to address if it is presented in a way that matches the individual's preferences [19]. However, when expanding the vocabulary, the capacity can be further increased by integrating a mixed learning style [22], compared to only using a person's preferred style.

Learning a language is similar to other learning areas. The learning capacity of a language can therefore also be affected by utilizing the processes, styles and strategies available [3]. Other factors that potentially have an influence on the acquisition of a language are a person's willingness to communicate, taking risks and their motivation to learn.

2.3 Applying gamification in a learning situation

There are numerous reasons why the combination of gamification and education has been a growing trend in recent years. Not only does it enhance the engagement, it also provides rich learning experiences by supplying the user with a controlled environment [14]. The controlled environment allows the user to explore, think outside the box and try things out, with the permission to fail.

Gamifiaction is also a way to improve the learning process due to its ability to incorporate learning principles [11]. One example of this is to provide the user with information just in time and not out of context from his or her purpose or goal. This can be used to enhance learning, since people have difficulties understanding or remembering information gained out of context or too long before it becomes useful [2]. Another principle that can be integrated into games is to place the difficulty on the edge of the user's competence. A task that is challenging, but achievable, is often perceived as pleasantly frustrating, which is a highly motivating state for people [11].

Another reason why gamification is a popular tool within the area of education, is its abilities to include feedback. Feedback is indispensable in a learning situation [15] and it is most efficient when it is accurate, timely, specific, substantive and constructive [5]. Games have the potential to create feedback that

possesses all of these characteristics, which can be used to guide the user through the learning process and serve as a motivator to learn more [4].

Individual game elements can, in addition to produce motivation, also be used to affect performance. A study has shown that points, levels and leaderboards can help people improve quantity performance [17], meaning produce more results during the same time. However, the study could not identify a connection between these elements and an increased quality of the performance. Another study has shown that game elements can be used to improve the performance of a person's working memory [18]. By utilizing game elements such as progress bar, levels, background story and visual feedback, a person can be brought closer to his or hers maximum memory performance. These two studies [17, 18] indicates that there exists a relation between gamification and learning, but neither of them examines how gamification affect learning and memory over time.

3 Methodology

The study was conducted with two test groups who studied a previously unknown language using different approaches. The language all participants studied was Hungarian, as it is a language not widely known in Sweden. The focus of the study was to examine acquirement of vocabulary more closely. The first test group learned Hungarian through a learning activity based on gamification, while the other learned the same language using an approach not including any game elements. This section describes the details of the study.

3.1 Assignment description

The task assigned to all test participants was to learn 18 different Hungarian words. The same language was used for all tests to avoid the difficulty of language affecting the result. A set of Hungarian words of different complexity were selected in advance so the test groups could train on the same set.

Both test groups trained on the chosen words during two training sessions each, using their respective method. Each training session was six minutes long and performed sequentially with a 15 minute interval between. In order to compare the two groups with each other, the training was followed by three test sessions held at different times, so-called posttests. The first posttest was made immediately after the two training sessions, while the second and the third were held one and two weeks later. By performing several tests at different times, it has been possible to investigate the effects of gamification on both short-term and long-term memory.

The test sessions consisted of a questionnaire that contained 18 questions representing the words the participants had practiced on. Each question contained a Hungarian word and four alternatives to the English equivalent. The task of the participants was to choose the correct definition among the alternatives. The test result was equal to the number of correct answers in the questionnaire. The

order of the questions and the corresponding alternatives were randomized before each test session. After the test the participants did not get to know the results of their completed test.

3.2 Test participants

The study included 16 test participants with eight participants in respective group. The selected test participants all had the same mother tongue, Swedish in this case. This ensured that all participants had the same prerequisites for learning Hungarian. The participants had also not studied Hungarian before, but it was considered okay if they ever had visited Hungary during a vacation.

None of the test groups included any bilingual persons, since these persons may have it easier to acquire yet another language. The two groups have, on the other hand, included Swedes that have extensive knowledge in both Swedish and English, as this is the case for most Swedes.

Before the test each test participant were informed about the scope of the study, how the results would be used and the procedure of the test. Each participant also gave their verbal consent to participate in the study before starting.

3.3 Methods for learning

In order for the two test groups to perform the test in environments that were as equal as possible, both learning methods were performed on a computer. The learning methods, on the other hand, differed in structure. The first test group trained on the words using an application based on gamification. The application included game elements such as progress bar, experience points (XP), a quest and feedback. These elements were chosen since studies have shown that they enhance motivation [24] and also improve the performance of the working memory [18].

The game contained two different practice modes, a quiz-like exercise and one exercise where words should be matched. At the start, the quiz was the only mode available to the participant. To be able to gain access to the matching exercise the participant was forced to succeed with the quest of collecting 50XP, which could be collected by completing training exercises.

The quiz exercise worked as a regular quiz, containing the 18 Hungarian words. To avoid that learning occurred as a result of the words being placed in a specific sequence, the order of the questions were randomized at the beginning of each started exercise. The order of the four alternatives associated to a question were randomized for the same reason. During the exercise the participants were presented with one Hungarian word and told to choose the English equivalent among four alternatives, see Figure 1. If the selected word was correct it was considered as a known word and was removed from the current exercise. If the given answer was incorrect the word got placed at the end of the word list so it could reappear later in the quiz for repetition. The two training sessions of six minutes guaranteed that all words were shown at least twice for each participant, but most participants had time to redo the exercise multiple times and therefore saw all words repeatedly.

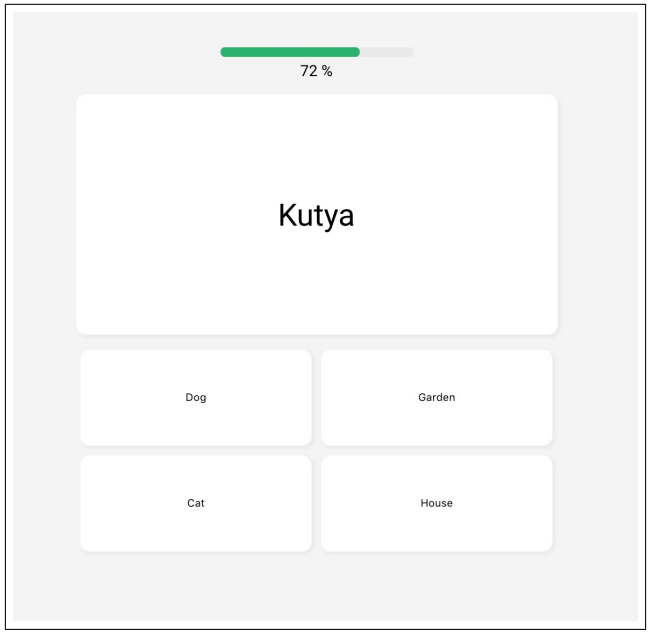

Fig. 1. The layout of the quiz exercise.

When the participants chose a word, they immediately got feedback on the given answer. The feedback was presented with corresponding sound, color and text. If the question was answered correctly, the participant got positive feedback and was presented with the Hungarian word and its English translation. This worked as a way to create further repetition of the word. If the given answer was incorrect the participant got encouraging feedback, to help keep up the motivation. In addition to the Hungarian word and its translation, the participants also got presented with the incorrect answer he or she had given. This was meant as a way to facilitate the participant's reply the next time the word returned, as it would help the participant to remember the incorrect answered word.

The progress was visible to the participant during the whole exercise. The progress was presented through a progress bar and the percentage of the completed words. When a word was answered correctly the progression increased. This way the participant could always be aware of the percentage of the completed words, and the part that remained.

When an exercise was completed, the participants were rewarded with XP. When 50XP had been collected the user unlocked the exercise for matching words. In this exercise the user got presented with all Hungarian words and their English equivalent. The participant's task was to create pairs by dragging and dropping matching words on each other, see Figure 2. When a correct pair was found, the sound of a correct answer was played and the two matching words got transparent and were no longer movable. If the participant tried to create an incorrect pair the two words flashed red and the sound for an incorrect answer

was played, to notify him or her that the words did not match. Just like in the quiz the participants progression was visible at all times.

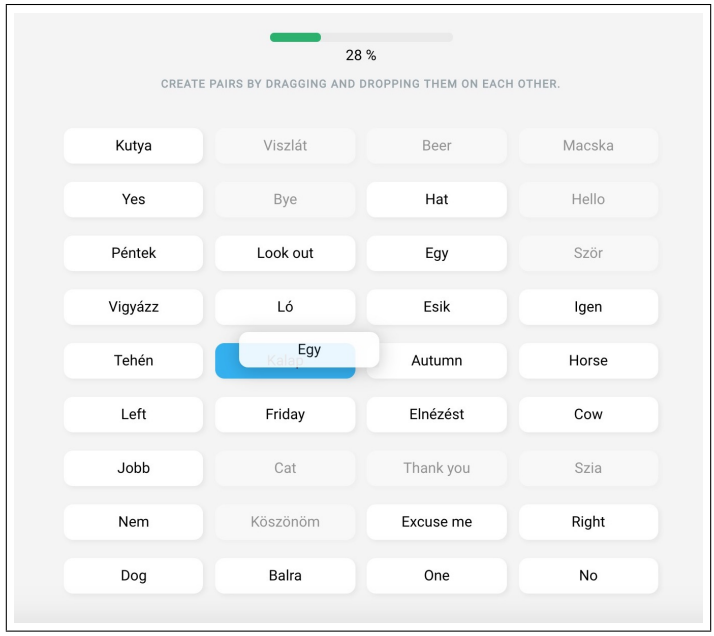

Fig. 2. The layout of the exercise for matching words.

The second test group used a digital version of flashcards to practice the Hungarian words, see Figure 3 . The cards were designed with the Hungarian word on one side, and the English equivalent on the other. The test participants had free access to the all of the cards, and could move between them as they desired. By using either the keyboard or the mouse, the participants could switch between the sides of a card and move back and forth in the "pile of cards".

Fig. 3. The layout of the flashcards used in the study.

3.4 Statistical Analysis

The results of the study was the number of words each test participant could remember during the three test sessions. The two test groups were compared at group level by using the mean of correct answers for each group. The comparison has been made through statistical analyzes using the statistical program Minitab and a two tailed t-test, with a significance level of 0.05. Analyses of the variance has also been made using the standard deviation of the results.

4 Result

The results from the three test sessions were compiled separately for the two test groups, see Table 1. The summary is based on the average number of correct answers during the three posttests for each group. In order to compare the variance of the groups, the summary also includes the standard deviation. In addition to this, the overall mean were calculated and presented in the table. The remainder of this section contains analyzes of the outcomes of the study and results from the performed t-tests.

Test group	**Immediate posttest**	**1-Week posttest**	**2-Week posttest**
Gamification	12.75 (3.34)	11.13 (3.55)	10.50 (4.36)
Non gamification	16.38 (3.08)	14.13 (3.89)	14.75 (3.80)
Overall	14.56 (3.21)	12.63 (3.72)	12.63 (4.09)

Table 1. The mean results from the three posttests rounded to two decimals. The standard deviation is written in parentheses.

Immediate posttest

The results from the first posttest, performed immediately after the training sessions, showed that the average score was higher for the participants who had used the procedure not including gamification. When comparing the two groups through a t-test, it was shown that there was a significant difference between the results ($t(14) = -2.255$, $p = 0.019$). This means that the method not including any gaming elements helped the test participants remember more Hungarian words during the test. The results also show that the variance is slightly higher for the test group that used the method including gamification, but there was no significant difference between the variances.

1-Week posttest

The results from the second posttest revealed that the group that used the method including gamification still has a lower average score, compared to the

other group. The difference between the two groups, on the other hand, had decreased during this test. When comparing the two test groups through a t-test, it could be shown that there was no longer a significant difference between them ($t(14) = 1.623$, $p > 0.05$).

In addition to this it could also be seen that the variance for both groups had increased during the second test, meaning that the results varied more during this test. It could also be noticed that the test group not learning through gamification had the higher variance this time, but there was still no significant difference between the variances.

2-Week posttest

The results from the last posttest were similar to the ones gained from the second test. The noticed difference was that the group using gamification had performed a lower mean score than before, while the other group had a higher mean score.

When comparing the results with a t-test it could be seen that there was no significant difference between the results of two groups ($t(14) = -2.079$, $p > 0.05$). The variance of both groups had increased during this test, but there was still no significant difference between them.

5 Discussion

One of the conclusions that can be made based on the results from the conducted study, is that there was a significant difference between the two methods, when it comes to the short-term memory. The group that used the method not including gamification performed significantly better than the group using gamification, on the test carried out immediately after the training sessions.

However, this was not the case during the subsequent tests. During the two later tests, the t-tests did not show any difference between the performance of the two groups. This indicates that the used methods had an equal effect on the participants long-term memory and can therefore be classified as equally good learning methods for acquiring a new language. The same holds for the variances of the two groups. Even though the variance differed between the three posttests, it was not significant.

Although the study does not show a relation between gamification and an improved acquisition of a third language, it cannot be excluded that a relation exists. In contrary to the results of this study, both [17] and [18] indicates that gamification can be used as a tool to improve the performance of the user. In order to establish the effect gamification has on performance, additional research is required.

5.1 Limitations

One of the limitations of the conducted study is the limited number of test participants in the two groups. A sample size of 16 observations is not large enough

to eliminate the personal differences that may have affected the results. There is a possibility that the test group that used the method not including gamification had greater natural talent for acquiring a new language and therefore performed better during the tests. Another personal factor that may have affected the outcome is the study experience of a participant. A participant who is used to acquire new knowledge has more likely developed a personal study technique, that can help the person streamline the learning process. In order to reduce the effect of personal differences a more extensive study would have to be conducted that includes more observations.

5.2 Future research

For further research it would be suitable to develop an application that better utilizes the benefits that gamification offers. Due to the limited resources, not all desired parts could be included in the implemented application used in the test. One suggestion is to create a mixed learning style by involving more senses in the learning process, which is preferable according to [22]. This could be done by including, for example, audio in terms of pronouns or figures to facilitate recollection.

6 Acknowledgement

I would especially like to thank the people who participated in the survey and made the study possible. I would also like to thank supervisors and the peer review group for all valuable feedback and support.

References

1. Patricia A. Alexander, Diane L. Schalert, and Ralph E. Reynolds. What Is Learning Anyway? A Topographical Perspective Considered. *Educational Psychologist*, 44(3):176–192, 2009.
2. Lawrence W. Barsalou. Language comprehension: Archival memory or preparation for situated action? *Discourse Processes*, 28(1):61–80, 1999.
3. H. Douglas Brown. *Principles of Language Learning and Teaching*. Course Technology Crisp, 1993.
4. Luma Da Rocha Seixas, Alex Sandro Gomes, Filho de Melo, and José Ivanildo. Effectiveness of gamification in the engagement of students. *Computers in Human Behavior*, 58:48–63, 2016.
5. Charlotte Danielson. *Enhancing Professional Practice: A Framework for Teaching*. Association for Supervision & Curriculum Development, 2007.
6. Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From Game Design Elements to Gamefulness: Defining "Gamification". In *Proceedings of The 15th International Academic MindTrek Conference*, pages 9–15, 2011.
7. Adrián Domínguez, Joseba Saenz-de Navarrete, Luis de Marcos, Luis Fernández-Sanz, Carmen Pagés, and José-Javier Martńez-Herráiz. Gamifying learning experiences: Practical implications and outcomes. *Computers & Education*, 63:380–392, 2013.

8. Zoltán Dörnyei and Richard Schmidt. *Motivation and Second Language Acquisition.* Natl Foreign Lg Resource Ctr, 2001.
9. Hilary Ellis, Stephen Heppell, Aleks Krotoski, John Kirriemuir, and Angela McFarlane. Unlimited learning. computer and video games in the learning landscape. Technical report, London: ELSPA (Entertainment and Leisure Software Publishers Association), 2006.
10. Figueroa Flores and Jorge Francisco. Using Gamification to Enhance Second Language Learning. *Digital Education Review*, 27:32–54, 2015.
11. James Paul Gee. *What video games have to teach us about learning and literacy.* New York: Palgrave Macmillan, 2003.
12. Juho Hamari, David J. Shernoff, Elizabeth Rowe, Brianno Coller, Jodi Asbell-Clarke, and Teon Edwards. Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior*, 54:170–179, 2016.
13. Jiying Han and Hongbiao Yin. Teacher motivation: Definition, research development and implications for teachers. *Cogent Education*, 3(1), 2016.
14. Karl M. Kapp. *The Gamification of Learning and Instruction: Game – based Methods and Strategies for Training and Education.* John Wiley & Sons, 2012.
15. Andrew S. Latham. Learning through Feedback. *Educational Leadership*, 54(8):86–87, 1997.
16. Uelaine Lengefeld. *Study Skills Strategies: Your Guide to Critical Thinking.* Course Technology Crisp, 1994.
17. Elisa D. Mekler, Florian Brühlmann, Alexandre N. Tuch, and Klaus Opwis. Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*, 71:525–534, 2017.
18. Manuel Ninaus, Goncalo Pereira, René Stefitz, Rui Prada, Christa Neuper, and Guilherme Wood. Game elements improve performance in a working memory training task. *International Journal of Serious Games*, 2(1):3–16, 2015.
19. Harold Pashler, Mark Mcdaniel, Doug Rohrer, and Robert Bjork. Learning Styles: Concepts and Evidence. *Psychological Science in the Public Interest*, 9(3):105–119, 2008.
20. Peter Skehan. *Individual differences in second-language learning.* London : Edward Arnold, 1989.
21. Scott Thiebes, Sebastian Lins, and Dirk Basten. Gamifying Information Systems - A Synthesis of Gamification Mechanics and Dynamics. In *Proceedings of the European Conference on Information Systems (ECIS) 2014*, pages 1–18, 2014.
22. Daniel G. Tight. Perceptual Learning Style Matching and L2 Vocabulary Acquisition. *Language Learning*, 60(4):792–833, 2010.
23. Maarten Vansteenkiste, Mingming Zhou, Willy Lens, and Bart Soenens. Experiences of Autonomy and Control Among Chinese Learners: Vitalizing or Immobilizing? *Journal of Educational Psychology*, 97:468–483, 2005.
24. Snežana Šćepanović, Nada Žarić, and Tripo Matijević. Gamification in higher education learning - State of the art, challenges and opportunities. In *Proceedings of the Sixth International Conference on e-Learning, Belgrade*, pages 128–134, 2015.

Physical Approach with Varying Degrees of Intention Expression with the Pepper Robot

Georgios Tsiatsios

Department of Computing Science
Umeå University, Sweden
gets0002@student.umu.se

Abstract. With the evolution of robots and their entrance to people's personal lives, the need for effective and natural human-robot interaction has become essential. This paper examines how people's comfort level is affected when the Pepper robot is approaching them from the back while expressing varying degrees of intention. At first, a summary of related existing literature studies is presented followed by three research hypotheses that are explored with experiments in a controlled environment. N=15 people volunteered to participate in this 1 (approach from the back) x 3 (no intention expression vs intention expression vs excessive intention expression) user study. The data collected from the participants were analyzed in various combinations. The result of the data analysis showed that the degree of intention expression of the robot affects the comfort level of the person interacting with it. Additionally, when comparing people who stated that had previous interaction with a robot in the past and people that had not, we observed a similar attitude towards the robot in varying degrees of intention expression. It was also observed that people, regardless of their gender, have mixed feelings towards excessive intention expression while in general, the normal intention expression was the preferred robot approach. The results obtained show that human-robot interaction is affected by the degree of intention expression of a robot, thus this parameter must be carefully evaluated in the robot design process.

1 Introduction

The area of personal robotics has been studied and developed intensively in the last years. Instead of only being used for industrial purposes, robots are becoming more personal for collaborative and everyday task purposes. As robots co-exist with humans in the society, effective and meaningful communication and human-robot interaction between them is essential.

Considerable research has focused on robots understanding of human intentions. For example, in [9], the researchers propose a method for a robot to predict human intention by analyzing gaze information. In [3], an omni-directional cane robot designed to support for elderly walking uses a hybrid model to analyze the motion of the human to support walking in a comfortable and natural way.

S. Bensch, T. Hellström (Eds.): Umeå's 23rd Student Conference in Computing Science USCCS 2019, pp. 63–72, January 2019.

Several methods have been developed for robots to express their intentions. In [4], Eun-Sook Jee et al. created different sound samples by using musical assistance and sound effects to set the timbre, so that the robot expresses different intentions such as affirmation, denial, encouragement, introduction and question. They found that pitch, intonation and timbre are features to be respected in robot language sound design for intention expression and emotion. In [1], the researchers propose an on-board intention projection on the floor where automated guided vehicles (AGVs) co-exist with human workers, in order to increase AGV's predictability and reduce human stress and potential collision.

Additionally, for coordination purposes, people count heavily on speech [11]. By designing robots that are able to produce human-like communication, the coordination between them becomes more effective since, people do not need any sort of training in order to interact with them, resulting in easier adaptation and acceptance to co-existing environments.

Currently, robots are evolving and becoming more autonomous. Therefore, there is a rising need to research the contrary situation where, the robot expresses its intentions while approaching a human, to make its presence known.

This paper aims to answer the following question: "How do different degrees of intention expression of an approaching robot (from the back) using verbal communication affect the comfort level of people in human-robot interaction?". The approach from the back is chosen so that people do not have prior knowledge of the robot's existence. Since the comfort level will reach a maximum peak (we assume that happens with normal intention expression), there is a need to examine whether the comfort level is decreased, increased or remains the same after it reaches that peak. That can be accomplished by using excessive intention expression to examine if there is a change in the comfort level outcome. A visual representation of the possible hypothesized outcomes of the comfort level with the increase of intention expression of the robot is presented in Figure 1. The red enumerated dots represent the places in the x-axis where we performed the experiments.

Based on Figure 1, we posed the following three hypotheses before conducting the experiments:

1. The participants will not find appropriate the approach of the robot without intention expression (low comfort level).
2. It is considered that the robot, by expressing normal intention will produce a satisfactory human-robot interaction.
3. It is believed that with excessive intention expression (intention and additional information), the comfort level of the participants will be decreased compared to Hypothesis 2, and it might create a repulsive motion from the robot.

Finally, we compared the outcome of the comfort level based on the gender of the participants and based on previous interaction with a robot or not.

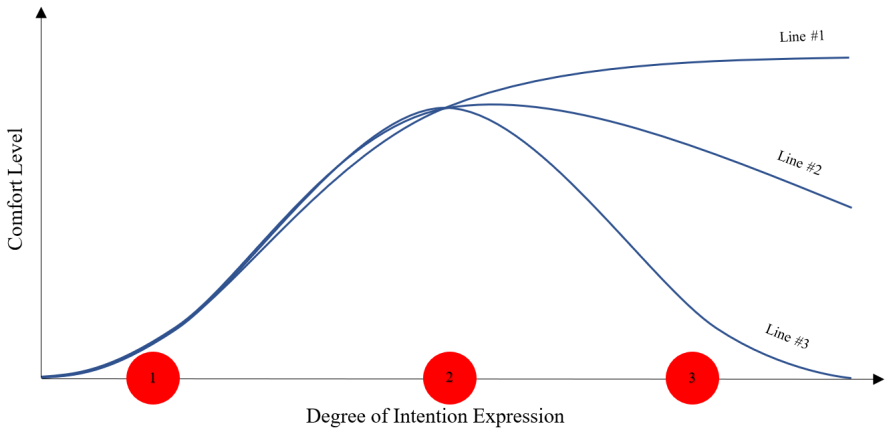

Fig. 1. Hypothesized comfort level outcomes.

2 Related earlier Work

Plenty of studies and research are focused on ways of how a robot should approach a person. In 1966, Edward T. Hall introduced the concept of proxemics [2]. This term refers to the personal space that people maintain around themselves.

Leila Takayama et al. [12] made a controlled experiment to find about human and robot factors that influence proxemics in different scenarios such as people approaching a robot, people being approached by an autonomous robot and people being approached by a tele-operated robot. They found that people with at least one year of experience with robots were more comfortable being closer to the robot that people with less than one year of experience. They also found that people who held more negative attitudes toward robots felt less safe when interacting with the robot.

In [5], Michiel P. Joosse et al. created an online study to investigate cross-cultural HRI proxemics preferences and proved that different cultures indeed have a different behavior toward an HRI. In [14], S. N. Woods et al. conducted a study to investigate how a robot should approach people in order to handle them an object. They found that people were against frontal approach by a robot in a seating context and seemed to prefer a front left or front aright approach.

In [10], there is a proposed model for a robot approaching people to initiate conversation. The experiment took place in a real field (shopping mall) and the robot's task was to advertise shops. The authors also mention that they faced some failures regarding the robot not being able to reach the target person, to people not seeing or listening to the robot, communication lag and error in responses of the robot and to people unwillingness to have a conversation with the robot. In a study by J. Mumm et al. [7], the likeability and gaze behavior of the robot were manipulated in order to explore the physical and psychological distancing. The results showed that when the participants disliked the robot,

they kept a greater distance from it but when they liked it, regardless of the gaze condition, the distancing did not differ.

Furthermore, a paper from Mohammad Obaid et al. [8] showed that the posture of the robot has a significant impact in the attitude of the people towards it. The experiment had a person approaching a robot and the opposite, while the human and robot were in either a standing or seated position. The results showed that the distances were impacted by the robot's posture and not by the human's. A sitting posture of the robot seems to be more approving and less threatening. Finally, K. L. Koay et al. [6] made an analysis of the comfort level of people interacting with a robot in a variety of behaviors, where the subjects had a comfort level device to state their discomfort at any moment during of the experiment. The results showed that subjects expressed discomfort in a three-meter proximity when the robot was blocking the path and when the participants thought that it may collide with them.

3 Study design

3.1 Methodology

In order to explore the research question and test the hypotheses made, a 1 (approach from the back) x 3 (no intention expression vs intention expression vs excessive intention expression) user study was conducted.

N=15 people participated in this user study, where all of them were students at Umeå University at that time. The gender distribution was almost equal since 8 of the participants were males and 7 were females. Their ages ranged from 20 to 31 with Mean = 24.46. Finally, 40% of the participants stated that they had a previous interaction with a robot in the past.

The experiments took place in a controlled environment (classroom) with one participant, the robot and the experimenter present. The room had enough space for the robot to move, eliminating possible collision.

3.2 Equipment

The social robot used in this user study was the Pepper from SoftBank Robotics. Pepper is an autonomous humanoid robot designed by Aldebaran Robotics and released in 2015 by SoftBank Robotics. It was designed with the ability to read human emotions by analyzing expressions and voice tones. It has 4 microphones in the head and 2 High-Definition (HD) cameras, one in the forehead and one in the mouth. Furthermore, a 3-Dimensional (3-D) depth sensor is located behind its eyes, a gyroscope in the torso and touch sensors in the hands and head. Finally, 2 sonars, 6 lasers, 3 bumper sensors and another gyroscope are included in the mobile base.

The Pepper robot is available for educational and research purposes in schools and universities for teaching programming and conducting research experiments and user studies in the human-robot interaction field.

3.3 Procedure:

Each participant was welcomed to the study and was informed about the experiment procedure. They were given a consent form to fill in and the experiment started only after they agreed and signed it.

The participant was seated in a marked, pre-defined position. The Pepper was placed 4 meters away from the user and approached him/her from the back while moving with its default speed in three different ways:

1. The Pepper approached the participants without verbal communication (no intention expression).
2. The Pepper approached the participants using normal verbal communication by saying "Hello human, I am approaching you!" (expected degree of intention expression).
3. The Pepper approached the participants using excessive verbal communication by saying "Hello human, I am a robot and I see you sitting there, so I am approaching you to ask you how you are doing today because I am programmed to do so!" (increased degree of intention expression).

The marked positions were evaluated so that the robot would stop by keeping 1.2 meters (with a small offset due to imprecise robot movement) distance from the user. When the Pepper reached its final position, it waited to make eye contact with the user and then, it proceeded into asking two questions, "How are you feeling today?" and "Do you like talking to a robot?" with relevant answers for each reply. Finally, at the end of the conversation, the Pepper said goodbye to the participant and returned to its initial position.

3.4 Data Analysis:

The data acquisition was accomplished by a questionnaire that the participants filled in after the end of the experiment. In the beginning, there were asked about their gender, their age and any sort of previous interaction with a robot. Furthermore, a question about their overall experience with the Pepper robot on a scale from 1 to 10 helped the researcher identifying the participant's overall satisfaction with the conducted experiment. Finally, they were asked to rate the three approaches based on their comfort level on a scale from 1 to 10. The data are presented in Table 1 along with a Box and Whiskers chart representation for better visualization which is presented in Figure 2.

The data analysis was performed using IBM SPSS Statistics Software[1]. The goal of the analysis was to examine if there are statistically significant differences between the three experiments, between males and females and finally between people who had a previous interaction with a robot or not based on their comfort level in all cases. In order to do that, there are some pre-analysis steps that need to be performed such as identifying if the samples are normally distributed and

[1] `https://www.ibm.com/analytics/spss-statistics-software`, IBM SPSS Statistics homepage

Participants	Sex	Age	Previous Interaction	Overall experience	1st Experiment	2nd Experiment	3rd Experiment
					No Intention	Expected Intention	Excessive Intention
1	F	23	YES	8	3	8	9
2	M	24	YES	10	8	10	9
3	F	25	NO	10	5	10	9
4	F	23	NO	9	5	7	10
5	F	31	YES	8	6	8	7
6	M	31	YES	8	2	9	6
7	F	21	NO	7	8	9	8
8	M	21	NO	7	5	9	4
9	F	20	NO	8	4	8	6
10	F	21	NO	10	6	9	7
11	M	27	YES	7	1	7	3
12	M	26	NO	9	4	9	6
13	M	25	NO	8	3	8	6
14	M	25	YES	8	4	6	7
15	M	24	NO	9	5	8	7
AVERAGE		24.47	40.00%	8.40	4.60	8.33	6.93
SD				1.06	1.96	1.11	1.91

Table 1. Raw experimental results.

depended or independent. The normality and dependency results allow us to choose the most appropriate test for examining the samples.

The samples were found to be normally distributed in all cases. Furthermore, the samples were treated as dependent when checking for significant differences between the three experiments and independent when checking for significant differences between males-females and previous interaction-not previous interaction. As a result, two-tailed paired t-tests were performed when checking for significant differences between the experiments (pairs: 1st-2nd, 1st-3rd, 2nd-3rd) and independent t-tests when checking for differences between males-females and previous interaction-not previous interaction. For all tests, the significance level (the probability for a given statistical model that, when the null hypothesis is true, the statistical summary would be greater or equal to the actual observed results) [13] was chosen to be 0.05.

4 Results and Discussion

4.1 Results

The two-tailed paired t-tests resulted in significant differences between all paired examined experiments with p-values 0.000 between 1st and 2nd experiment, 0.018 between 2nd and 3rd experiment and 0.000 between 1st and 3rd experiment. The results are reasonable and expected since their means are significantly different. These tests indicate that the degree of intention expression of the robot while approaching indeed affects the comfort level of the person interacting with the robot.

The two-tailed independent t-tests comparing males and females resulted in a significant difference (for both males and females) only in the experiment where the robot expressed excessive intention, as presented in Table 2, where "F" corresponds to females and "M" to males.

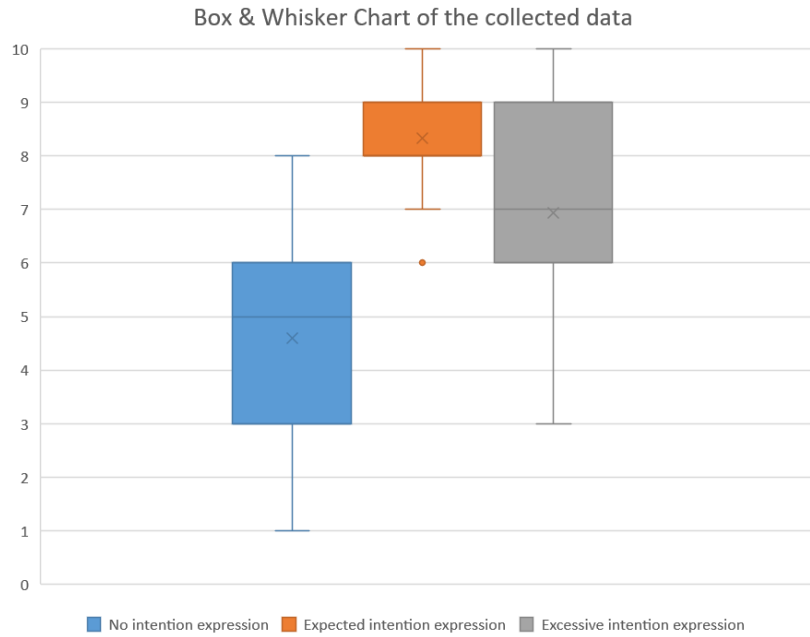

Fig. 2. Visual representation of the acquired data.

	1st Experiment	2nd Experiment	3rd Experiment
F	0.216	0.769	0.037
M	0.208	0.765	0.034

Table 2. P-values when comparing males and females.

It can be stated that people, regardless of their gender, have mixed feelings towards excessive intention expression, since some of them seem to have preferred it compared to normal intention expression, while others did not. Overall, it was observed that the participants were more satisfied with the expected intention expression of the robot.

Finally, the final test results when comparing people who had a previous interaction with a robot or not were somewhat surprising since there was not a significant difference between the two groups in neither of the experiments, as shown in Table 3, where "YES" corresponds to previous interaction and "NO" to not previous interaction. Apparently, humans regardless of previous interaction with a robot or not, seem to have the same attitude towards it in varying degrees of intention expression.

4.2 Discussion

From the experimental results, we found full support for two out of three hypotheses mentioned in Section 1. The data support that the participants did not

	1st Experiment	2nd Experiment	3rd Experiment
YES	0.351	0.363	0.875
NO	0.419	0.417	0.882

Table 3. P-values when comparing people with a previous and not previous interaction.

find the robot approach without intention expression appropriate (1st hypothesis). Although the Mean of the comfort level ratings was 4.60 and SD = 1.96, it is believed that the Mean could be lower because the room was quiet and some participants could hear the motors of the robot in the movement process.

Regarding the 2nd hypothesis, the data support that expected intention expression is the preferred way of the robot approach, having the highest Mean = 8.33 of all three experiments and the lowest SD = 1.11.

Furthermore, the 3rd hypothesis made is not fully supported because, although the average comfort level from the excessive intention expression of the robot was decreased compared to the comfort level of the normal intention expression (Mean = 6.93, SD = 1.91), we did not detect a repulsive motion from the robot.

Finally, the comparison based on previous interaction or not did not give proof that affects the comfort level whereas, based on gender, the excessive intention expression (3rd experiment) performed differently for both males and females.

5 Conclusions

The conducted user study presented in this paper allowed us to acquire data and analyze them in order to examine how varying degrees of intention expression of the approaching Pepper robot affect people's comfort level. Based on our results and the possible hypothesized outcomes presented in Figure 1, it can be seen that with excessive intention expression of the robot, the 2nd curved line better represents our findings. Also, it was shown that varying degrees of intention expression affect HRI. Additionally, the comfort level of males and females was the same way for no intention and normal intention expression and different for excessive intention expression. Finally, previous interaction with a robot did not affect the comfort level of the person taking the experiment. Parameters like these must be evaluated when designing autonomous social robots for satisfactory HRI in dynamic environments.

6 Limitations

As with most user studies, there are numerous limitations that need to be acknowledged. Firstly, the experiment was conducted in a controlled environment which may have affected the results that would have been obtained if it took place in the real world e.g. a shopping mall. Furthermore, the findings are limited to a specific age range (Mean = 24.46) and a small number of participants.

With a greater amount of participants, more data would have been collected and the results may have been more dominant. Additionally, the order of the experiments was the same for every participant (from 1st to 3rd), leading to feedback that may have been different if the order was not fixed for everyone. Finally, due to the fact that the experiment was conducted during an exam period, the available time of the participants was limited and the experimental steps were performed consecutively. Thus, a possible familiarity with the robot could affect the end results.

7 Future work

This user study can be extended to a higher degree such as including more participants, improving the software for a more robust robot movement in space while evaluating more complex scenarios (obstacle avoidance, target detection) and perform the experiments in the real field instead of a controlled environment.

8 Acknowledgements

I gratefully acknowledge the help, guidance and feedback of professors Thomas Hellström and Suna Bensch and students Chaitanya Ganesh Kudaka, Senthilkumar Konnaiyandi Sivakumaran and Christoffer Edlund. Finally, I would like to thank the people who volunteered to participate in the conducted experiment.

References

1. T. Fukuda, J. Taguri, F. Arai, M. Nakashima, D. Tachibana, and Y. Hasegawa. Facial expression of robot face for human-robot mutual communication. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, volume 1, pages 46–51 vol.1, May 2002.
2. Edward T. Hall. *The hidden dimension / Edward T. Hall.* Doubleday Garden City, N.Y, 1966.
3. Jian Huang, Pei Di, Toshio Fukuda, and Takayuki Matsuno. Motion control of omni-directional type cane robot based on human intention. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pages 273–278, 2008.
4. Eun-Sook Jee, Yong-Jeon Jeong, Chong Hui Kim, and Hisato Kobayashi. Sound design for emotion and intention expression of socially interactive robots. *Intelligent Service Robotics*, 3(3):199–206, Jul 2010.
5. Michiel P. Joosse, Ronald W. Poppe, Manja Lohse, and Vanessa Evers. Cultural differences in how an engagement-seeking robot should approach a group of people. In *Proceedings of the 5th ACM International Conference on Collaboration Across Boundaries: Culture, Distance & Technology*, CABS '14, pages 121–130, New York, NY, USA, 2014. ACM.
6. K. L. Koay, K. Dautenhahn, S. N. Woods, and M. L. Walters. Empirical results from using a comfort level device in human-robot interaction studies. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, HRI '06, pages 194–201, New York, NY, USA, 2006. ACM.

7. J. Mumm and B. Mutlu. Human-robot proxemics: Physical and psychological distancing in human-robot interaction. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 331–338, March 2011.
8. Mohammad Obaid, Eduardo B. Sandoval, Jakub Zlotowski, Elena Moltchanova, and C. Bartneck. Stop! that is close enough. how body postures influence human-robot proximity. In *25th IEEE International Symposium on Robot and Human Interactive Communication*, pages 354–361. IEEE, 2017.
9. K. Sakita, K. Ogawara, S. Murakami, K. Kawamura, and K. Ikeuchi. Flexible cooperation between human and robot by interpreting human intention from gaze information. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, volume 1, pages 846–851 vol.1, Sept 2004.
10. Satoru Satake, Takayuki Kanda, Dylan F. Glas, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. How to approach humans?: Strategies for social robots to initiate interaction. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, HRI '09, pages 109–116. ACM, 2009.
11. Aaron St. Clair and Maja Mataric. How robot verbal feedback can improve team performance in human-robot task collaborations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 213–220, New York, NY, USA, 2015. ACM.
12. L. Takayama and C. Pantofaru. Influences on proxemic behaviors in human-robot interaction. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5495–5502, Oct 2009.
13. Ronald L. Wasserstein and Nicole A. Lazar. The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.
14. S. N. Woods, M. L. Walters, K. L. Koay, and K. Dautenhahn. Methodological issues in hri: A comparison of live and video-based methods in robot to human approach direction trials. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pages 51–58, Sept 2006.

Author Index

Ganesh, Chaitanya, 1

Haggag, Omar, 11

Konradsson, Hanna, 23

Lyxzen, Ivan, 41

Nystedt, Anna, 51

Tsiatsios, Georgios, 63