Measuring Behavior 2020-21

Volume 1

12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods

13-15 October, 2021, Kraków, Poland

Proceedings

Volume Editors

Andrew Spink

Noldus Information Technology; Andrew.Spink@noldus.nl

Jaroslaw Barski

Medical University of Silesia, Katowice; jbarski@sum.edu.pl

Anne-Marie Brouwer

Perceptual and Cognitive Systems, TNO; anne-marie.brouwer@tno.nl

Gernot Riedel

University of Aberdeen; g.riedel@abdn.ac.uk

Annesha Sil

University of Aberdeen; annesha.sil@abdn.ac.uk

ISBN: 978-90-74821-93-3

DOI: 10.6084/m9.figshare.13013717

Table of Contents

Preface to Volume 1 of Measuring Behavior 2020-21

Andrew Spink[1], Jarosław Barski[2], Anne-Marie Brouwer[3], Gernot Riedel[4], Annesha Sil[4]

1 Noldus Information Technology, Wageningen, The Netherlands, Andrew.Spink@noldus.nl

2 Medical University of Silesia, Katowice, Poland, Jarosław Barski, jbarski@sum.edu.pl

3 Perceptual and Cognitive Systems, TNO, Soesterberg, The Netherlands, anne-marie.brouwer@tno.nl,

4 University of Aberdeen, Aberdeen, UK, g.riedel@abdn.ac.uk & annesha.sil@abdn.ac.uk.

The current Measuring Behavior conference was originally scheduled for May 2020. As the pandemic swept across the world, we initially delayed it for six months and then again until October 2021. Nevertheless, a number of authors needed to have their contributions published in 2020, for instance so that they could complete their PhD and graduate. We have therefore brought out an initial first volume of the Proceedings specially for those authors. It is unusual to publish a conference Proceedings a whole year before the conference, but we are living in exceptional times. The papers will still all be presented at the conference in 2021. We will publish a second volume of Proceedings at that time, which will be a mixture of papers as they were submitted for 2020, updates on 2020 papers and completely new submissions. The special Research Topic of *Frontiers in Psychology* on 'Developments in Implicit Measurements' is going ahead as planned.

At the time of writing, the death toll associated with the COVID-19 coronavirus is one million and growing [1]. Understanding human behavior has never been more critical. The way that people interact with each other and the effect that government regulations have on that behavior is crucial in either controlling or increasing the spread of the virus. Cultural, psychological and sociological aspects have all contributede to give rise to different behavioral responses in different countries. Objective measurement of those behavioral responses is hard to come by. Several versions of a smartphone app have been developed to measure the extent to which people make contact, but it is still unclear if technologies like Bluetooth are really suitable for that sort of tracking, for example, Bluetooth signals go through walls which the virus does not.

The current volume doesn't address those issues directly. The Call for Papers leading to the contributions published here was published well before COVID-19 reared its head. Neverthless, the topics covered are all important, with new methods and techniques presented in a wide variety of disciplines including human factors, wildlife ecology, psychology, learning and memory, disease models, and behavioral tests. It is noticeable that there are quite a few papers focusing on integration of physiological and similar measures with other data streams obtained from the use of various sensors, which is in line with a trend we have seen in recent editions of Measuring Behavior.

The whole world is eagerly waiting for an effective vaccine against COVID-19. There is much uncertainty as to when one might be ready and in mass production but nevertheless we continue to be optimisticthat we will be able to hold the 12th *Measuring Behavior* conference jointly with the 6th *Seminar on Behavioral Methods* in the beautiful city of Kraków as planned on 13-15 October 2021. We hope this volume of the proceedings caters to many different interests.

References

1. John Hopkins Coronavirus Resource Centre, https://coronavirus.jhu.edu/, 29 September 2020.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Oral Presentations

Combining eye tracking and physiology for detection of emotion and workload

A.-M. Brouwer[1], I. Stuldreher[1], S. Huertas Penen[2,] K. Lingelbach[3] and M. Vukelić[4]

1 Perceptual and Cognitive Systems, TNO, Soesterberg, the Netherlands. anne-marie.brouwer@tno.nl. ivo.stuldreher@tno.nl

2 University of Twente, Enschede, the Netherlands. s.huertaspenen@student.utwente.nl

3 Institute of Human Factors and Technology Management, University of Stuttgart, Germany. katharina.lingelbach@iat.uni-stuttgart.de

4 Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, Germany. mathias.vukelic@iao.fraunhofer.de

Abstract

Peripheral physiological measures such as electrodermal activity (EDA), heart rate and pupil dilation, as well as neurophysiological measures such as electroencephalography (EEG), can inform us about individuals' cognitive and emotional state. We are interested in exploiting such measures in real life situations. A challenge of interpreting physiological measures as markers of mental state in real life is the lack of context information. We here approach this challenge by relating physiological measures to eye tracking. Participants scanned stimuli that induced different levels of workload (small sets of numbers that needed to be added or not) and different types of emotion (neutral, pleasant and unpleasant pictures). EDA, heart rate, pupil size and EEG were related to the first eye fixation on the stimulus. For peripheral measures, response traces across the following 10s were determined and signal amplitudes were compared between the different types of stimuli. EEG signals were compared for the different types of stimuli in the time interval from fixation onset to 1500 ms later using a cluster-based, non-parametric randomization approach. For the peripheral measures, high workload stimuli stood out from all other stimuli in all modalities, with patterns as expected from literature under more traditional experimental conditions: high values of EDA, heart rate, and pupil size for high compared to low workload stimuli. For emotional stimuli, peripheral physiological effects tended to be in the expected direction but were more modest in size. In the EEG signals, a significant late parieto-occipital cluster could be identified with higher amplitudes for high compared to low workload stimuli, as well as for emotional stimuli compared to the neutral stimuli. In future analyses we will combine fixation-locked signals from different modalities to detect mental states elicited by information that is being looked at. Our first results indicate that this may be especially helpful in situations related to cognitive workload, e.g. determining whether operators are not only looking at, but are also cognitively processing information that is presented on a screen.

Introduction

Physiological measures such as skin conductance or electrodermal activity (EDA), heart rate, pupil dilation and electroencephalography (EEG) reflect a range of physical and sensory processes. However, they also contain information about the cognitive and emotional state of individuals [1-3]. This has convincingly been shown in laboratory experiments where participants sit still and are presented by stimuli that are designed to elicit certain cognitive or emotional states. For instance, heart rate is higher and pupil size larger when performing a controlled task with a high memory load compared to performing a low load task [4], heart rate and skin conductance responses differ when pictures with different emotional content are shown [5], and a stronger P300 (a late component in the event-related potential) can be observed for stimuli that draw attention compared to ones that do not [6]. To reliably extract, study and use this information in real life environments, it is important to relate recorded physiological data to context, i.e. to what occurs in the outside world. One way in which this can be done is to use information from eye movements, such as recorded via eye tracking. Recorded fixation locations indicate what is being observed when combined with camera images, or when the visual environment is known in another way (e.g. because certain parts of an information display have fixed locations, or when information is known since it is presented on a monitor). Previous work studied EEG signals related to fixation.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

While this is challenging since eye movements strongly influence EEG, there is now a range of studies showing that higher order cognitive processing of stimulus information during reading and visual search are reflected in such fixation- or saccade related potentials (e.g. [7-9]). Little research has been dedicated to relating other physiological signals to fixations. An exception is [10], who not only related EEG to fixation onset, but also pupil dilation. They found that pupil dilation was higher after fixations on target objects relative to distractor objects, and that this information was helpful besides EEG to classify data as coming from fixations on targets versus distractors. Also, [11] examined both pupil dilation and EEG following fixations on targets and distractors, where participants were asked to remember the locations of the targets while performing an auditory math task. While EEG was especially informative to distinguish between fixations on target and distractor items, pupil size was informative as to whether the target location would be remembered correctly. Specifically, a large pupil size was associated with not remembering the target location, probably because of moments of high workload caused by the math task. We are not aware of work that related EDA and heart rate to fixation onset in paradigms where participants move the eyes around.

As suggested, combining modalities can be helpful to better identify mental state and therewith predict performance. Multimodal measurement techniques can be helpful for different reasons. Firstly, a mental state (e.g. that goes with finding a target, as in [10]) may be better identified when utilizing multiple measures of interest that reflect a similar mental state but are affected by different types of noise. Hence, a combination could result in a more robust identification. Secondly, different modalities can reflect different types of mental state (e.g. event-related potentials in the EEG reflect target detection and pupil size reflects workload as in [11]), enabling more fine-grained mental state estimation. EDA and pupil size are robustly correlated to states of bodily arousal, whereas the P300 component in the EEG reflects states of attention. Peripheral physiological measures may be relatively suitable for emotional engagement (arousal), and EEG (reflecting cortical activity) more for cognitive processes [12-16]. Note that in many cases, these types of processes are expected to coincide: an emotional stimulus is likely to draw attention and a difficult cognitive task likely elicits arousal.

In the current study, we recorded EDA, heart rate, pupil dilation and EEG and related these signals to fixation onset where stimuli are viewed that we expect to elicit different degrees of workload and emotion. Our ultimate aim is to identify affect and cognition in an environment where individuals freely look around so that humans and machines can interact more naturalistic and efficiently. As a first step in exploring and comparing different fixation-locked signals in response to different types of stimuli within the same participants and the same paradigm, we designed an experiment that induced quite predictable and limited amounts of large eye movements. This may be akin to situations of operators who subsequently view portions of a display with information that need to be taken in and induce different types of mental state.

Methods

Participants

A total of 20 healthy participants (5 men, 15 women) took part in this study. They were between 19 and 34 years old, with an average of 23 years. Participants were recruited through the participant pool of the research institute where the study took place (TNO) and received a monetary reward to compensate for time and travel costs. None of the participants wore glasses. All participants signed an informed consent form in accordance with the Helsinki Declaration [17], before participating in the study. This study was approved by the Human Research Protections Official (HRPO) and the TNO Institutional Review Board (TCPE). Eye recording failed in two of the participants, leaving us with fixation-locked data from 18 participants. For the first four participants that we recorded, EDA and heart rate data were lost. In an additional four participants, heart rate data was lost. Four participants were excluded in the EEG analysis due to poor signal quality. In sum, pupil size was obtained for 18 participants, EDA for 14 participants, heart rate for 10 participants, and EEG for 16 participants

Materials

For measuring eye gaze location and presenting stimuli, we used a Tobii Pro TX300 eye tracking system (Tobii Technology, Stockholm, Sweden). This system consists of a noninvasive standalone eye tracking recording unit fixed underneath a stimulus screen. Gaze location of both eyes was recorded at 60 Hz. The screen was a 23-inch flat-screen monitor, set at a resolution of 1920 * 1080 pixels. The monitor was about 40 cm from the participants' eyes.

EDA and ECG (electrocardiogram, to obtain heart rate) were recorded using a Biosemi ActiveTwo MkII system, with a sampling frequency of 512 Hz. EDA was measured by placing gelled electrodes on the fingertips of the index finger and the middle finger of the left hand. ECG electrodes were placed on the right clavicle and on the lowest floating left rib. Additionally, we measured neurophysiological activity using EEG. The scalp EEG potentials were recorded using an actiCap 32-channel system according to the extended international 10-05 system with a LiveAmp amplifier (Brain Products GmbH, Munich, Germany). The impedance of the electrodes was kept below 20 kΩ at the onset of each session. EEG data was digitized at 250 Hz, using the BrainVision Recorder Software (Brain Products GmbH, Munich, Germany). The unified collection of signals from the different recording systems and the stimulus presentation program were synchronized and stored for off-line data analysis using Lab Streaming Layer (LSL) [18].

Stimuli and design

Participants were presented with pictures that were expected to induce different levels of workload and types of emotion. There were five types of these pictures: 1) inducing workload: displaying three three-digit numbers arranged around the letter 'A' indicating that these numbers needed to be added (NumbersAdd), 2) inducing no workload: displaying three three-digit numbers arranged around the letter 'N' indicating that these numbers did not need to be added (NumbersNone), 3) inducing pleasant emotion: a picture from the International Affective Picture System (IAPS) [19] with high valence and high arousal (HVHA), 4) inducing unpleasant emotion: a picture from the IAPS with low valence and high arousal (LVHA), 5) inducing no emotion: a picture from the IAPS with neutral valence and low arousal (Neutral).

From the IAPS, the pictures were randomly drawn out of collections of 60 pictures with valence scores higher than 5.5 and arousal higher than 5.5 (pleasant), valence scores lower than 4.5 and arousal higher than 5.5 (unpleasant) and valence scores between 4.5 and 5.5 and arousal lower than 4.5 (no emotion or neutral).

All pictures were approximately 205 by 154 pixels in size and could appear at any of 9 locations on the screen, with a minimum distance of 191 pixels between two sequentially presented pictures. A picture was presented for 10 s. Nine seconds after picture onset, the next picture appeared. Workload inducing pictures could be followed by a screen prompting the participant for the result of the addition. This was intended to motivate participants to really perform the math during the workload inducing picture, and to allow them a short break. Blocks of stimuli separated by these questions consisted of 15 or 20 pictures, containing 3 or 5 pictures, respectively, of each type. Otherwise, the order of pictures was random. Participants finished up to 14 blocks (with a minimum of 12 blocks, and a median of 14 blocks).

Procedure

Participants received a short explanation about the study and were invited to ask any question they may have. They then signed the informed consent form. The Tobii eyetracker was calibrated using a nine-point calibration. Participants were fitted with the ECG, EDA and EEG electrodes. Participants were asked to not speak during the experiment unless absolutely necessary and to keep movements to a minimum.

Analysis

All data analysis was performed with custom written or adapted scripts in MATLAB® and Python^{TM} .

For the EDA, the phasic and tonic components were extracted using Continuous Decomposition Analysis [19] as implemented in the Ledalab toolbox for MATLAB®. The phasic component was z-score standardized following [21]. These data are further used in the analysis.

ECG measurements were processed to acquire the inter-beat interval (IBI, which is the inverse of heart rate). ECG was band-pass filtered between 5 and 15 Hz using a third order Butterworth filter. Peaks were detected following Pan and Tompkins [22]. The IBI semi-time series was transformed into a timeseries. This was done by interpolating consecutive IBIs and then resampling at 512 Hz. IBI was then transformed to heart rate and further used in the analysis.

To handle missing values in the raw eye tracking data (pupil size and gaze location), typically occurring due to blinks, the data were linearly interpolated in time windows of maximum 75 ms of consecutive missing data points [23]. Data from the left and right eye were averaged [23-24]. Additionally, gaze position was smoothed using a median filter with a sliding window of 20 ms [24]. Gaze position over time was used to determine fixation onset, where we are interested in the first fixation on the picture. In order to do this robustly without having to rely on more or less arbitrary temporal and spatial thresholds, we followed a previously adopted approach [11,25] and determined the time of the maximum velocity of the saccade of interest as a proxy of fixation onset, though note that the actual fixation starts in the order of 30 ms later. For convenience, we still refer to our data as 'fixation-locked' rather than 'saccade-locked'. Maximum saccade velocity was searched for in a 1.5 seconds window, starting at the onset of the new picture. Note that this method takes advantage of the design of our experiment by providing us knowledge of the approximate time of fixations of interest. It should be replaced by another method in situations with more unpredictable timing - e.g., in the case of a known display, fixations of interest are those associated with saccades that moved the gaze from outside into a certain spatial area of interest; or generic temporal and spatial thresholds should be used.

For each picture, EDA epochs were extracted, starting at time of fixation and ending 10 seconds later. For each participant, these epoched signals were aligned by subtracting the average value of the first 500 ms from the epoched signal, and averaged across pictures of each of the five picture types (HVHA, LVHA, Neutral, NumbersAdd and NumbersNone). The same procedure was followed for heart rate and pupil size. Next, for each participant and picture type, the response amplitude was determined by taking the maximum value in the epoch for EDA and heart rate, and by taking the average value in the epoch for pupil size. These data are used in statistical analyses.

To analyze the neurophysiological data, EEG signals were de-trended, zero-padded and re-referenced to mathematically linked mastoids [26]. We excluded two EEG channels (T8 and T7) from the analysis due to artefact contamination. Next, we band-pass filtered the EEG signals between 1 to 20 Hz to calculate fixation-locked event-related potentials (FERP). The filtering was done by using a first order zero-phase lag finite impulse response (FIR) filter.

For the analysis of FERPs, fixation-locked epochs ranging from 200 ms before and 1500 ms after the beginning of the fixation onset were created separately for the five picture types (HVHA, LVHA, Neutral, NumbersAdd and NumbersNone). We rejected epochs containing a a maximum deviation above 200 µV in any of the frontal EEG channels (AFp1, AFp2). Furthermore, for each remaining epoch we performed an independent component analysis (ICA) using the extended infomax ICA algorithm [27] as implemented in the MNE-Python toolbox [28]. The ICA was used to remove further cardiac-related artefacts, ocular movement and muscular artefacts. The selection of components indicating artefacts was done by careful visual inspection of the topography, times course and power spectral intensity of the components [29,30].

To study spatio-temporal changes of neurophysiological signals we baseline-corrected the artefact-free EEG epochs by subtracting the mean amplitude of the time interval between -200 ms and 0 ms before the fixation onset. FERPs were then calculated by averaging the EEG signal separately for each picture type (HVHA, LVHA, Neutral, NumbersAdd, NumbersNone) and each channel. For the statistical evaluation we performed a mass-univariate analysis. We chose a cluster-based, non-parametric randomization approach which included correction for multiple comparisons as described by [31] and implemented in the MNE-Python toolbox [28].We compared the baseline corrected data from all electrodes at all time points after the fixation onset to locate effects of emotional pictures (comparing Neutral vs HVHA and Neutral vs LVHA) and workload pictures (comparing NumbersAdd vs NumbersNone) in time and space. Clusters were identified as adjacent points in space (electrodes) and time (time point in the EEG segment) using a cluster-level threshold of p<.01 estimated via a t-test (uncorrected). The cluster-level statistics were defined as the sum of t-values within every cluster. The correction of multiple comparisons was realized by calculating the 95th percentile of the maximum values of summed t-values estimated from an empirical reference distribution. T-values exceeding this threshold were thus considered as significant at p<.05 (corrected). The reference distribution of maximum values was obtained by means of a permutation test (randomly permuting the data points across the compared condtions for 1000 times). Thereby, we perform the statistics separately for the emotional and workload pictures.

For each physiological modality, we specifically compare responses to pictures with numbers, and responses to neutral versus emotional pictures, since these types of pictures differ only with respect to the emotional or cognitive state that they are expected to induce, and are similar with respect to other, low-level stimulus characteristics.

Results

Figure 1 shows the traces, averaged across participants and picture type, for EDA (A), heart rate (B) and pupil size (C). Especially NumbersAdd stimuli elicit clear responses in all three modalities.

Figure 2 presents the average of the response amplitude for EDA, heart rate and pupil size. Because our data were not normally distributed, non-parametric tests were used for statistical comparison. Wilcoxon signed rank tests indicated that EDA amplitude is significantly higher for high workload pictures, with numbers to add (Mdn = 0.379) than for low workload pictures, with numbers not to add (Mdn = 0.172). The same result was found for HR amplitude (Mdn = 3.707 for NumbersAdd and Mdn = 3.125 for NumbersNone) and for pupil size amplitude (Mdn = 0.411 for NumbersAdd and Mdn = 0.138 for for NumbersNone). Regarding emotional pictures, EDA amplitude is higher for low valence pictures (Mdn = 0.066) than for neutral pictures (Mdn = 0.032). There is a small trend in the same direction for high valence versus neutral pictures. No statistically significant differences were found in heart rate and pupil size amplitudes when comparing high valence pictures to neutral pictures, and when comparing low valence to neutral pictures. An overview and details of the statistical results are given in Table 1.

For EEG, using the non-parametric cluster-based randomization test, we found one significant late parieto-occipital electrode cluster for the comparison between the neutral pictures against the pictures with high valence (Figure 3A). This cluster comprised 15 electrodes with a difference from 164 ms to 912 ms after fixation onset. Similarly comparing the neutral pictures versus the pictures with low valence, we observed one significant late parieto-occipital electrode cluster (Figure 3B). The cluster comprised 17 electrodes with a difference from 160 ms to 1000 ms after fixation onset. Comparing the NumbersAdd and NumbersNone pictures, we found one significant cluster over parieto-occipital electrode regions (Figure 4). The parieto-occipital comprised 12 electrodes with a difference from 252 ms to 672 ms after fixation onset.

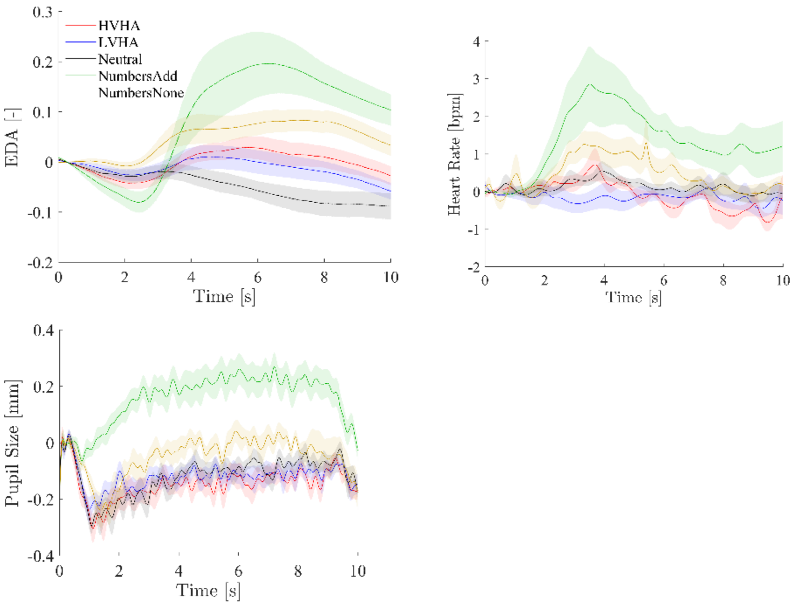


Figure 1. Response traces (average and standard error of the mean) time-locked to fixation onset for EDA (A), heart rate (B) and pupil size (C). Red traces represent HVHA pictures; blue LVHA; black Neutral; green NumbersAdd; yellow NumbersNone.

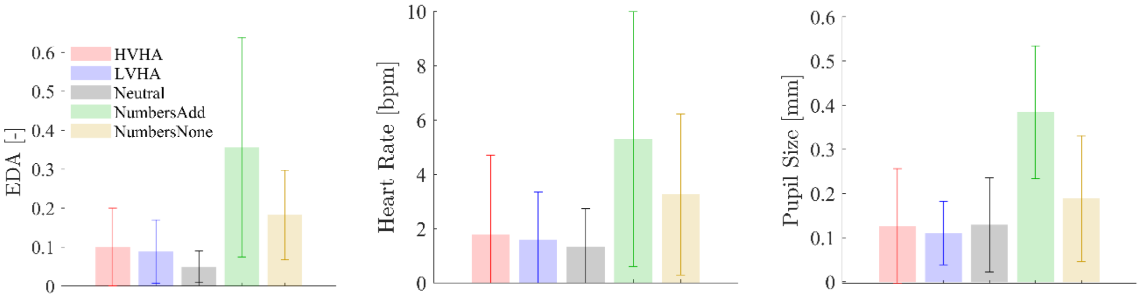


Figure 2. Average response amplitude in response traces time-locked to (from left to right) HVHA, LVHA, Neutral, NumbersAdd and NumberNone pictures, for EDA (A), heart rate (B) and pupil size (C). Error bars represent standard errors of the mean.

	EDA	Heart rate	Pupil size
NumbersAdd vs. NumbersNone	**W = 121, p =.006**	**W = 65, p =.043**	**W = 147, p =.007**
HVHA vs. Neutral	W = 100, p = .098	W = 44, p = .733	W = −82, p = .879
LVHA vs. Neutral	**W = 106, p =.049**	W = 50, p = .424	W = 66, p = .396

Table 1. Statistics for comparison between stimulus types, for EDA, heart rate and pupil size.

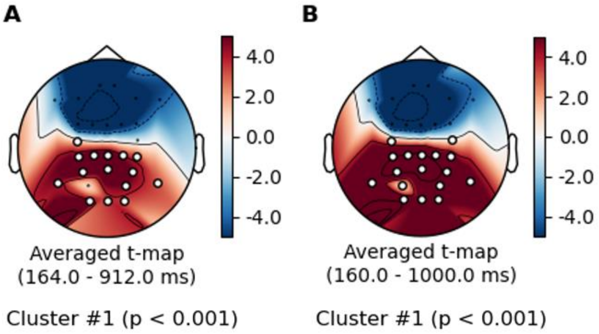


Figure 3. Spatio-temporal dynamics for the emotional pictures. The plots show the topographic maps of the t-values that represent the difference comparing the neutral with the positive (HVHA) pictures in (A) and the neutral with the negative (LVHA) pictures in (B). Electrode clusters showing significant differences in the non-parametric randomization test, are indicated by filled white circles. In both comparisons, an extended late parieto-occipital cluster was found. The amplitudes of the FERPs were larger for the positive (HVHA) and negative (LVHA) pictures than for the neutral pictures.

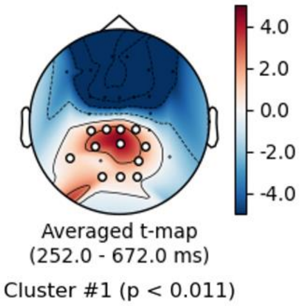


Figure 4. Spatio-temporal dynamics for the workload pictures. The plot show the topographic map of the t-values that represent the differences comparing the numbers not to add (NumbersNone) with the numbers to add (NumbersAdd). Electrode clusters showing significant differences in the non-parametric randomization test, are indicated by filled white circles. The non-parametric randomization test reveals an extended late parieto-occipital cluster. The amplitudes of the FERPs were larger for the NumbersAdd compared to the NumbersNone condition.

Discussion

We examined EDA, heart rate, pupil size and EEG related to fixations on stimuli that were expected to induce different levels of workload and types of emotion.

For all three peripheral physiological measures, we found the expected increase when comparing high workload stimuli (NumbersAdd) to stimuli with the same visual appearance, but without an associated mental workload task (NumbersNone). The average EDA stimulus traces for emotional and neutral pictures showed the expected pattern with larger values for emotional (high arousal) pictures compared to neutral pictures. The heart rate patterns are roughly consistent with those reported in [5], where heart rate was examined in response to IAPS pictures without having participants move their eyes towards the pictures. They also found an acceleration starting at around 2 seconds for pleasant pictures, followed by a strong deceleration at around 3.5 seconds, whereas unpleasant pictures show more of a deceleration, and heart rate responses to neutral pictures were closer to those to pleasant compared to unpleasant ones. However, our statistical analyses on the overall amplitude did not show significant effects for heart rate – only for EDA the comparison between neutral and high valence, high arousal pictures reached significance.

For the EEG fixation-locked dynamics, we found a late parieto-occipital cluster sensitive to pictures inducing high and low valence compared to neutral pictures. A similar effect was found for high versus low workload inducing pictures. These findings are consistent with earlier studies. Previous studies consistently find higher amplitudes in late components for high compared to low arousing affective pictures (where P300 and later slow wave potentials elicited with affective pictures are often denoted as late positive potential or LPP) [32-34]. Concerning the high and low workload pictures in our study, high workload pictures are expected to induce arousal, attentional and working memory processes, which are processes that are associated with higher amplitudes in late parieto-occipital components [6,35]. Note that in our case, the participant finds out during the first fixation on a high workload picture that a mental task has to be performed, i.e., that the picture is particularly relevant. This is a different case than most EEG workload studies, in which stimuli such as beeps are usually associated with low rather than high P300 amplitudes. In these studies, high workload as e.g. induced by a double task, prevents participants to allocate much attention to the presented stimuli [36,37], which is consistent with a low P300.

Performing a mental task (adding numbers) seems to induce immediate and stronger arousal compared to viewing pictures that may not have direct relevance to the particular participant, and are not related to any (upcoming) action. Arousal due to an upcoming, socially relevant emotional task can expected to be stronger than emotion induction through pictures [38]. Fixation-related physiology may especially be helpful in situations related to cognitive workload, e.g. determining whether operators are not only looking at, but are also cognitively processing information that is presented on a screen; or in situations involving strong, personally relevant emotions that are related to upcoming action.

To bring fixation-locked physiology closer to applications, several steps are required. In following analysis, we will examine whether on an individual level, for a single fixation, a classification algorithm can estimate which of the five picture types an observer is looking at – i.e., identify an individual's mental state using combined multimodal fixation locked measures. Combining modalities may aid to get a more strongly differentiating signal.

For this first EDA, heart rate and pupil size fixation locked study, we wanted to heighten the chance that participants dwelled on a certain stimulus for some time, which is why we presented stimuli on different locations but sequentially, with only a short time of simultaneous presence on the screen. Figure 1 suggests that minimum gaze times of 4 to 6 seconds would suffice to obtain an undisturbed maximum signal. A more ecological experiment would entail the presentation of multiple stimuli on the screen at once, e.g. in an operational setting where an observer has to monitor and interpret different parts of a display.

Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-19-1-7015. Any opinions, finding, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.

References

1. Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interact. Comput*. **21**: 133–145.

2. Picard, R. W. (1997). Affective Computing. Cambridge: MIT Press.

3. Parasuraman, R., and Rizzo, M. (2007). Neuroergonomics: The Brain at Work. Oxford; New York: Oxford University Press.

4. Brouwer, A.-M., Hogervorst, M. A., Holewijn, M., van Erp, J. B. F. (2014). Evidence for effects of task difficulty but not learning on neurophysiological variables associated with effort. *Int. J. Psychophysiol*. **93**: 242–252.

5. Bradley, M. M., Codispoti, M., Cuthbert, B. N., Lang, P. J. (2001). Emotion and Motivation I: Defensive and Appetitive Reactions in Picture Processing. *Emotion* **1(3)**: 276-298.

6. Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* **1**18: 2128–2148.

7. Wenzel, M. A., Golenia, J.-E., Blankertz, B. (2016). Classification of Eye Fixation Related Potentials for Variable Stimulus Saliency. *Front. Neurosci,* **10**: 23.

8. Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: Analyses and review. J. *Exp. Psychol. Gen.* **140(4)**: 552–572.

9. Brouwer, A-M., Reuderink, B., Vincent, J., van Gerven, M. A. J., van Erp, J. B. F. (2013). Distinguishing between target and nontarget fixations in a visual search task using fixation-related potentials. *Journal of Vision* **13(3)**:17, 1–10.

10. Jangraw, D. C., Wang, J., Lance, B. J., Chang, S.-F., Sajda, P. (2014). Neurally and ocularly informed graph-based models for searching 3D environments. *J. Neural Eng*. **11(4)**: 046003.

11. Brouwer, A.-M., Hogervorst, M. A., Oudejans, B., Ries, A. J., Touryan, J. (2017). EEG and Eye Tracking Signatures of Target Encoding during Structured Visual Search. *Front. Hum. Neurosci,* **11**: 264.

12. Cromwell, H.C., Panksepp, J. (2011). Rethinking the cognitive revolution from a neural perspective: How overuse/misuse of the term “cognition” and the neglect of affective controls in behavioral neuroscience could be delaying progress in understanding the Brain. *Mind. Neurosci. Biobehav. Rev*. **35**: 2026–2035.

13. Schupp, H.T., Flaisch, T., Stockburger, J., Junghöfer, M. (2006). Emotion and attention: event-related brain potential studies, in: *Progress in Brain Research*. Elsevier, pp. 31–51.

14. Scherer, K.R., Schorr, A., Johnstone, T. (Eds.) (2001). Appraisal processes in emotion: theory, methods, research, Series in affective science. Oxford University Press, Oxford, New York.

15. Posner, J., Russell, J.A., Peterson, B.S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol,* **17(3)**:715-34.

16. Hajcak, G., MacNamara, A., Foti, D., Ferri, J., Keil, A., (2013). The dynamic allocation of attention to emotion: Simultaneous and independent evidence from the late positive potential and steady state visual evoked potentials. *Biol. Psycho,*. **92**: 447–455.

17. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects (2014). *J. Korean Med. Assoc.* **81**: 14.

18. Kothe, C. (2014). Lab streaming layer (LSL). https://github. com/sccn/labstreaminglayer. Accessed on January 16, 2020.

19. Lang, P. J., Bradley, M. M., Cuthbert, B. N. (2008). International affective picture system (iaps): affective ratings of pictures and instruction manual. University of florida, Gainesville, Tech Rep A-8, Tech. Rep.

20. Benedek, M., Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods* **190(1)**: 80-91.

21. Ben-Shakhar, G. (1985). Standardization within individuals: A simple method to neutralize individual differences in skin conductance. *Psychophysiology***22(3)**: 292-299.

22. Pan, J., Tompkins, W. J. (1985). A real-time QRS detection algorithm. *IEEE Trans. Biomed. Eng* **32(3)**: 230-236.

23. Wass, S. V., Smith, T. J., Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults, *Behavior Research Methods* **45(1)**: 229–250.

24. Olsen, A. (2012). The tobii i-vt fixation filter, Tobii Technology.

25. Dias, J. C., Sajda, P., Dmochowski, J. P., Parra, L. C. (2013). EEG precursors of detected and missed targets during free-viewing search. *J. Vis.* **13**:13.

26. Nunez, P. L., Srinivasan, R. (2006). Electric fields of the brain: the neurophysics of EEG, 2nd ed. Oxford ; New York: Oxford University Press.

27. Lee, T.W., Girolami, M., Sejnowski, T.J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation* **11(2)**: 417-441

28. Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Parkkonen, L., Hämäläinen, M. (2014). MNE software for processing MEG and EEG data, *NeuroImage* **86**: 446-460.

29. Chaumon, M., Bishop, D.V.M., Busch, N.A. (2015). A practical guide to the selection of independent components of the electroencephalogram for artifact correction, *J. Neurosci. Methods* **250**: 47–63.

30. Hipp, J.P., Siegel, M. (2013). Dissociating neuronal gamma-band activity from cranial and ocular muscle activity in EEG. *Front Hum Neurosci* **7**: 338.

31. Maris E., Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J of Neurosci Methods* **164(1)**: 177–190.

32. Cuthbert, B.N., Schupp, H.T., Bradley, M.M., Birbaumer, N., Lang, P.J. (2000). Brain potentials in affective picture processing: Covariation with autonomic arousal and affective report. *Biol. Psychol.* **52(2)**: 95–111.

33. Olofsson, J. K., Nordin, S., Sequeira, H., Polich, J. (2008). Affective picture processing: an integrative review of ERP findings. *Biol Psychol* **77**: 247-265.

34. Schilling, T., Sipatchin, A., Chuang, L., Wahl, S. (2018). Tinted lenses affect our physiological responses to affective pictures: An EEG/ERP study. In 2nd International Neuroergonomics Conference: The brain at work and in everyday life. Frontiers Research Foundation.

35. Polich, J., Kok, A. (1995). Cognitive and biological determinants of P300: an integrative review. *Biol. Psychol.* **41**: 103–146.

36. Allison, B. Z., Polich, J. (2008). Workload assessment of computer gaming using a single-stimulus event-related potential paradigm. *Biol. Psychol.* **77**: 277–283.

37. Dehais, F., Duprès, A., Blum, S., Drougard, N., Scannella, S., Roy, R., Lotte, F. (2019). Monitoring Pilot's Mental Workload Using ERPs and Spectral Power with a Six-Dry-Electrode EEG System in Real Flight Conditions. *Sensors* 19.

38. Brouwer, A.-M, Hogervorst, M. A. (2014). A new paradigm to induce mental stress: The Sing-a-Song Stress Test (SSST). *Frontiers in Neuroscience* **8**: 224.

Physiological synchrony in EEG, electrodermal activity and heart rate for the assessment of shared attention

Ivo Stuldreher[1], Nattapong Thammasan[2], Jan van Erp[1,2] and Anne-Marie Brouwer[1]

1 Perceptual and Cognitive Systems, TNO, Soesterberg, The Netherlands. ivo.stuldreher@tno.nl

2 Research Group Human Media Interaction, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands.

Abstract

Physiological measures such as brain potentials as measured through electroencephalography (EEG), skin conductance or electrodermal activity (EDA) and heart rate can be informative of individuals' attentional engagement. We are interested in exploiting such measures in real life situations. A challenge of interpreting physiological signals as markers of attention in real life, is the lack of context information. When studying groups of individuals, a suitable approach may be to determine physiological synchrony (PS) – the degree to which physiological measures of multiple people uniformly change. We explored to potential of PS as measure of attentional engagement. We here determined PS in EEG, EDA and heart rate of participants that where selectively attentinding to the narrative of an audiobook, or to short auditory stimuli that were interspersed. We found that PS increased upon presentation of emotionally or cognitively relevant short stimuli, so that it may be used for detection of relevant events in cases where timing is unknown. Using the synchrony of a participant's EEG, EDA or heart rate with the two groups attending to the narrative or short-stimuli as a predictor of attentional instruction allowed for the correct identification in 96%, 73% and 73% of the cases, respectively. PS may thus be of value when interested in monitoring attentional engagement. As PS does not dependent on event markers, this approach is well-suited for real-life settings.

Introduction

Continuous and implicit measures of attentional or emotional engagement may be extracted from physiological signals, such as brain potentials as measured though electroencephalography (EEG), skin conductance or electrodermal activity (EDA) and heart rate. The modulation of physiological responses by attentional engagement has long been studied in reductionistic laboratory paradigms with repeated presentations of the same stimuli. Event-locked metrics are then extracted by relating physiological signals to the time that stimuli of interest occur. However, research physiological measures as measure of mental state is shifting to more uncontrolled real-life environments. Here it is often difficult to relate physiological signals to known stimuli; it is often unclear what the stimulus of interest is and practical issues prevent the implementation of a-priori known event markers. A novel way to circumvent these problems is to determine the synchrony of physiological responses across individuals rather than looking at specific events and individual observers. High physiological synchrony (PS) – i.e., the degree to which physiological measures of multiple people uniformly change – would indicate shared attention to an apparently generally relevant event. Indeed, EEG signals were found to synchronize more across participants with shared attentional engagement to narrative video or audio-clips than to scrambled clips [1]. PS in EEG further predicted general expressions of interest and attention and correlated with memory retention [2,3]. There is also an (older) body of literature on PS in measures of autonomic nervous system activity, such as heart rate and EDA, reviewed by [4]. Rather than indicators of shared attention, these have been interpreted as more general indicators of some form of connectedness between people. Up to date, these bodies of literature have remained separate.

In the current study, we compare PS in neural and autonomic physiological measures to determine selective attentional focus of individuals who are all presented with the same auditory stimulus and are all attending to it, but are attending to different stimulus aspects. As a proof-of-concept of the potential value of PS as informative index of attentional or emotional engagement in real-life settings, we aim to answer three research questions.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

First, does PS in EEG, EDA and heart rate across shared attending individuals increase during (in this case, known) emotionally or cognitively relevant events? If so, PS may be used to detect relevant events in cases where timing of relevant events is unknown. Second, can we classify the selective attentional focus of individuals using PS (in this case, does PS tell us which general attentional instruction the participant received)? If we can detect relevant events using PS (research question 1), this would allow us to zoom in on those events, which subsequently may enhance classification of selective attentional focus (research question 2). Research question three is whether using knowledge of the timing of relevant events could enhance classification performance.

Methods and Materials

Participants

Twenty-seven participants (17 female), aged between 18 and 48 (M = 31.6, SD = 9.8) years, were recruited through the institute's participant pool. The study was approved by the TNO Institutional Review Board (TCPE). Prior to the study, all participants signed informed consent, in accordance with the Declaration of Helsinki. Participants received a small reimbursement for their time and travel costs. Data of one participant were discarded due to failed physiological records.

Materials

EEG, EDA and electrocardiogram (ECG) were recorded at 1024 Hz using an ActiveTwo Mk II system (BioSemi, Amsterdam, Netherlands). EEG was recorded with 32 active Ag-AgCl electrodes, placed on the scalp according to the 10-20 system, together with a common mode sense active electrode and driven right leg passive electrode for referencing. The electrode impedance threshold was set at 20 kOhm. For EDA, two passive gelled Nihon Kohden electrodes were placed on the ventral side of the distal phalanges of the middle and index finger. For ECG, two active gelled Ag-AgCl electrodes were placed at the right clavicle and lowest floating left rib.

Stimuli and Design

Participants performed the experiment one by one. Each participant was presented with the same audio file, composed of a 66 min audiobook (a Dutch thriller 'Zure Koekjes', written by Corine Hartman) interspersed with short, auditory stimuli. We instructed half of the participants to focus on the narrative of the short story and ignore all the other stimuli or instructions (narrative attending – NA) and we instructed the other half of the participants to focus their attention on the interspersed short stimuli and ignore the narrative (short-stimuli attending – SSA). The auditory stimuli were emotional sounds from the international affective digitized sounds (IADS) [5] and beeps that SSA participants needed to keep track of [6]. A more elaborate description of the used stimuli and order of presentation can be found in [7].

Pre-processing

Data processing was done using Matlab R2019a software (Mathworks, Natick, MA, USA). For EEG processing, we also used EEGLAB v14.1.2. for Matlab [8]. We performed logistic infomax independent component analysis to remove ocular or muscle-related artifacts [9]. Then, the multiple artifact rejection algorithm (MARA) [10] was used to identify artifactual independent components. These components were removed from the data after re-referencing to the average channel value. Samples whose squared amplitude magnitude exceeded the mean-squared amplitude of that channel by more than four standard deviations were marked as missing data in an iterative way with four repititions.

EDA was downsampled to 32 Hz. Using continuous decomposition analysis as implemented in the Ledalab toolbox for Matlab [11], the tonic (slowly varying) and phasic (fast peaks) components were separated. In further analyses we use the phasic components as they are mainly related to responses to external stimuli.

ECG measurements were processed to acquire the inter-beat interval (IBI), the interval between consecutive R-peaks. ECG was downsampled to 256 Hz and high-pass filtered at 0.5 Hz. The Pan-Tompkins algorithm was used to detect R-peaks [12]. The obtained IBI semi-time series was then transformed into a time series by interpolating consecutive intervals and resampling at 32 Hz.

Analysis

We assessed PS in the EEG by measuring inter-subject correlations (ISC) between participants. We evaluated the ISC in the correlated components of EEG, following [1,3]. The goal of correlated component analysis is to find underlying sources of neural activation using linear combinations of electrodes that are maximally correlated between participants. The technique is similar to principal component analysis, but rather than maximizing variance in a dataset, projections of correlated component analysis capture maximal correlations between datasets.

In the phasic component of EDA and IBI we also assessed PS based on ISC. We followed the moving window approach of [13]. We computed Pearson correlations over running 15 second windows at one second increments. The overall correlation between two responses was computed as the natural logarithm of the sum of all positive correlations divided by the sum of the absolute values of all negative correlations.

To answer our first research question (does PS in EEG, EDA and heart rate across shared attending individuals increase during emotionally or cognitively relevant events), we investigated whether the degree of PS was different during intervals in the narrative audiobook with interspersed stimuli compared to parts of the narrative audiobook without such interspersed stimuli. We computed PS separately for participants in the NA and SSA conditions. This was done for experiment parts where only the narrative audiobook was presented, parts that also contained sequences of beeps, parts that also contained affective sounds and parts that contained either beeps or affective sounds. Correlated components for EEG were always extracted considering the entire experiment duration. We conducted paired-sample t-tests to test whether PS during sequences of beeps, affective sound or either of these two were higher than during narrative parts without interspersed short stimuli.

To answer our second research question, we investigated whether we could identify the selective attentional focus of a participant (either attending to the narrative or to the short stimuli) based on PS. For EEG, EDA and IBI we classified the attentional focus of each participant based on the group she or he had higher PS with. To avoid biases in the component extraction step for EEG, data from the to-be tested participant were excluded in this step. Chance level classification performance was determined using surrogate data with 100 renditions of randomly shuffled attentional condition label. In each shuffle, we followed the same procedure as above. Two sample one-tailed t-tests were conducted to test whether classification performance was higher than chance level classification performance.

To answer our third research question on whether using knowledge of the timing of relevant events could enhance classification performance, we did the same as above, except rather than using data as recorded during the whole stimulus we focused on parts of the data containing beeps, affective sounds or narrative only.

Results

Inter-subject correlations during interspersed stimuli compared to narrative only parts

Figure 1. shows inter-subject correlations as measure of PS for EEG, EDA and IBI during experiment parts where only the narrative audiobook was presented compared to experiment parts where short stimuli interspersed the audiobook (beeps, affective sounds, or either one of them), for SSA participants (A) and NA participants (B). Test statistics of the paired-sample t-tests (comparing narrative only to one of the interspersed stimulus parts) are shown in the figure.

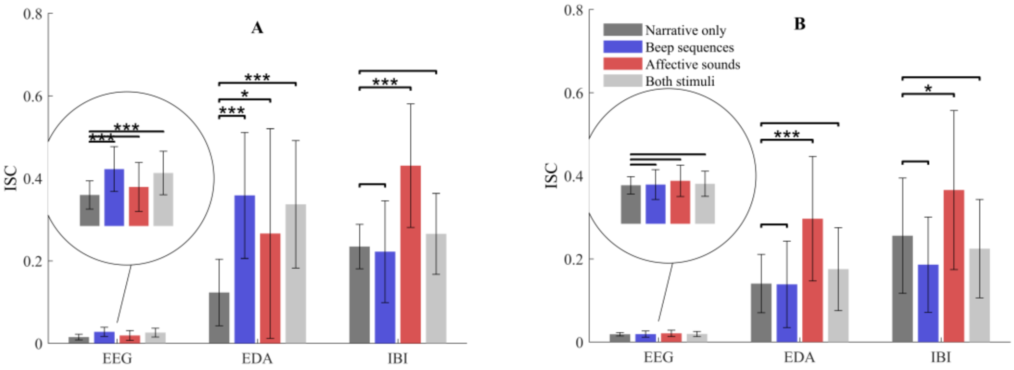


Figure 1. Inter-subject correlations (ISC) for EEG, electrodermal activity (EDA) and inter-beat interval (IBI) during parts of the narrative audiobook without interspersed stimuli compared to experiment parts with interspersed beeps, interspersed affective sounds or both stimuli combined averaged over NA participants (A) or SSA participants (B). Paired-sample t-test were used to test ISC during one of the interspersed stimuli vs. ISC in the narrative only condition are shown. (* p < .05, ** p < .01, *** p < .001).

Results are in line with our hypotheses. PS of SSA participants, who were instructed to focus on the interspersed stimuli, is generally higher during interspersed stimuli than during parts of the audiobook book where only the narrative audiobook was presented. For EEG this is the case for interspersed beeps ($p < .001$) or when considering any type of stimulus ($p < .001$). In EDA the effect is found for all types of stimuli, i.e. beeps ($p < .001$), affective sounds ($p = .026$) or when considering any thype of stimulus ($p < .001$). In IBI PS is only higher during interspersed affective sounds ($p < .001$). PS of NA participants, who were instructed to ignore the interspersed stimuli is not higher during beeps than during narrative only, but is higher during affective sounds for EDA ($p < .001$) and IBI ($p = .017$).

Identification of selective attention based on inter-subject correlations

Table 1. presents classification performance as to which attentional instruction group a participant belongs (NA or SSA) based on PS. Our assumption is that the attentional condition of participants can be identified based on the group with whom she or he shows the highest average synchrony. Classification performance was substantially higher than chance for all modalities (EEG, EDA and IBI), when using data from the whole stimulus without any knowledge about timing of relevant stimuli. EEG scores best (96%), for EDA and IBI performance is 73% (first column in Table 1). When we consider the subsets of data centered around relevant events (beeps or affective sounds), or narrative only, classification performance does not increase for any of the measures (second to fourth column in Table 1). Interestingly, each metric seems to perform especially well for one of the three subsets. EEG performs especially well when considering sequences of beeps (88%), EDA during intervals interspersed with affective sounds (73%) and IBI during parts of the narrative audiobook that are not interspersed (73%).

Discussion

We computed PS through inter-subject correlations in EEG, EDA and heart rate to investigate its potential for determining attentional or emotional engagement in real-world environments.

Our first research question was whether PS in EEG, EDA and heart rate across shared attending individuals increased during emotionally or cognitively relevant events so that PS may be used for event detection. We found that PS indeed tended to increase during relevant events. Participants that were specifically instructed to focus their attention on the interspersed stimuli had higher generally higher PS during interspersed stimuli as compared to parts of the stimulus where only the narrative audiobook was presented. PS across participants instructed to focus on the narrative of the audiobook, on the other hand, was only higher during interspersed

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

affective sounds as compared to narrative only parts. Although these participants were ‘top-down’ instructed to ignore any interspersed stimulus, the affective sounds can be expected to draw attention through a bottom-up fashion based on mechanisms of emotional relevance [14], resulting in increased PS during these moments. Considering this, we can conclude that PS increases during emotionally or cognitively relevant events. PS may thus be used to detect relevant events in cases where timing of relevant events is unknown.

Table 1. The percentage of participants of which the attentional condition is correctly identified using inter-subject correlations in EEG, electrodermal activity (EDA) and inter-beat interval (IBI) considering all four time intervals. In brackets the chance level classification performance is shown as means and standard deviations. Grey cells depict classification accuracies significantly higher than this chance level distribution. p-values are shown in the table.

	Whole stimulus	**Beep sequences**	**Affective sounds**	**Narrative only**
EEG	96 (49 ± 11) $p < .001$	88 (52 ± 13) $p = .003$	73 (50 ± 13) $p = .037$	73 (50 ± 10) $p = .010$
EDA	73 (50 ± 10) $p = .009$	69 (50 ± 10) $p = .032$	73 (49 ± 10) $p = .009$	62 (49 ± 11) $p = .115$

Our second research question was whether we could classify the selective attentional focus of individuals using PS. Classification accuracies considering the whole stimulus were high and well above chance level. Using subsets of data based on known interspersed stimuli (research question three) did not result in higher classification accuracies. Findings presented in Figure 1A indicated that PS in EEG appeared to mainly reflect attention to the beeps, whereas PS in IBI mainly reflected attention to affective sounds. Similar variation between the physiological metrics was found in the classification results. Although classification accuracies of selective attention identification did not increase when using subsets of data based on event information, we did observe that each physiological measure performed relatively well for one of the three stimulus types. PS in EEG performed relatively well considering the sequences of beeps, PS in EDA performed relatively well considering the affective sounds and PS in IBI performed relatively well considering the narrative audiobook parts where no stimuli were interspersed. These findings support our vision that multimodal PS, i.e., combining multiple physiological measures into a single index of PS, may be advantageous for mental state monitoring [15]. Combining metrics may result in better classification performance and better event detection. We want to further investigate this in future work.

References

1. Dmochowski, J. P., Sajda, P., Dias, J., Parra, L. C. (2012). Correlated components of ongoing EEG point to emotionally laden attention–a possible marker of engagement? *Frontiers in human neuroscience* **6**: 112.

2. Dmochowski, J. P., Bezdek, M. A., Abelson, B. P., Johnson, J. S., Schumacher, E. H., Parra, L. C. (2014). Audience preferences are predicted by temporal reliability of neural processing. *Nature Communications* **5**: 4567.

3. Cohen, S. S., Madsen, J., Touchan, G., Robles, D., Lima, S. F., Henin, S., Parra, L. C. (2018). Neural engagement with online educational videos predicts learning performance for individual students. *Neurobiology of learning and memory* **155**: 60-64.

4. Palumbo, R. V., Marraccini, M. E., Weyandt, L. L., Wilder-Smith, O., McGee, H. A., Liu, S., Goodwin, M. S. (2017). Interpersonal autonomic physiology: A systematic review of the literature. *Personality and Social Psychology Review* **21(2)**: 99-141.

5. Bradley, M. M., Lang, P. J. (2007). The International Affective Digitized Sounds (IADS-2): Affective ratings of sounds and instruction manual. University of Florida, Gainesville, FL, Tech. Rep. B-3.

6. De Dieuleveult, A. L., Brouwer, A. M., Siemonsma, P. C., Van Erp, J. B., Brenner, E. (2018). Aging and sensitivity to illusory target motion with or without secondary tasks. *Multisensory research* **31(3-4)**: 227-249.

7. Brouwer, A. M., Stuldreher, I. V., Thammasan, N. (2019). Shared attention reflected in EEG, electrodermal activity and heart rate. *Proceedings of the 1st workshop on Socio-Affective Technologies: an interdisciplinary approach* (Bari, 7 October 2019)

8. Delorme, A., Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods* **134(1)**: 9-21.

9. Bell, A. J., Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation* **7(6)**: 1129-1159.

10. Winkler, I., Haufe, S., Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions* **7(1)**: 30.

11. Benedek, M., Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods* **190(1)**: 80-91.

12. Pan, J., Tompkins, W. J. (1985). A real-time QRS detection algorithm. *IEEE Trans. Biomed. Eng* **32(3)**:230-236.

13. Marci, C. D., Ham, J., Moran, E., Orr, S. P. (2007). Physiologic correlates of perceived therapist empathy and social-emotional process during psychotherapy. *The Journal of nervous and mental disease* **195(2)**:103-111.

14. Öhman, A., Flykt, A., Esteves, F. (2001). Emotion drives attention: detecting the snake in the grass. *Journal of experimental psychology: general* **130(3)**: 466.

15. Stuldreher, I. V., de Winter, J.C.F., Thammasan, N., & Brouwer, A. M. (2019). Analytic approaches for the combination of autonomic and neural activity in the assessment of physiological synchrony. *Proceedings of 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* 4143-4148.

Validation of wearables for electrodermal activity (EdaMove) and heart rate (Wahoo Tickr)

Ana Borovac [1], Ivo Stuldreher [2], Nattapong Thammasan [3] and Anne-Marie Brouwer [2]

1 Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Reykjavík, Iceland. anb48@hi.is.

2 Perceptual and Cognitive Systems, TNO, Soesterberg, The Netherlands. ivo.stuldreher@tno.nl. anne-marie.brouwer@tno.nl

3 Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands. n.thammasan@utwente.nl

We are interested in monitoring physiology out of the lab because of its potential informative value about mental state, such as stress. For real life studies or applications utilizing physiology, we need to equip individuals with wearables that are designed to be compatible with individuals performing their daily activities, which can come at the cost of a reduction in signal quality. In the current study, we compare wearable sensors of electrodermal activity (EdaMove) and heart rate (Wahoo Tickr) to their high-end, laboratory counterparts (ActiveTwo). Signals were compared at a general level as well as in relation to their response to emotional sounds. EdaMove showed more general drift and responses returned slower to baseline than ActiveTwo. Responses to emotional sounds were about equally clear as ActiveTwo. Apart from a delay of around 6.7 seconds, Wahoo Ticker accurately recorded general heart rate levels. However, it did not capture fast changes which also resulted in less clear responses to emotional sounds. Both wearables are potentially suitable to record physiology in real life, but for synchrony recordings, EdaMove is expected to be suitable whereas results based on Wahoo Tickr are expected to be less clear.

Introduction

Monitoring physiological measures such as heart rate and skin conductance (or electrodermal activity - EDA) in daily life can be of interest for measuring human behavior. Monitoring a physiological variable can be helpful for tracking certain medical condition (such as cardiac activity for people with a cardiac disease) and also for gaining insights about a person's mental state, such as stress, which has a major influence on behavior and health. While relating physiology to mental state is not without challenges [1], in a recent systematic review, De Looff et al. [2] report that high levels of job stress are associated with an increased heart rate, and decreased heart rate variability measures. For EDA, they reported that such evidence was not available - no EDA studies were found that met the inclusion criteria of the review.

Monitoring physiology outside the lab requires wearable devices. For heart rate, many types of (affordable) wearable devices are available. For EDA this is not the case, which may partly explain the scarcity of real-life studies involving EDA. However, given that EDA is uniquely innervated by the sympathetic *fight or flight* part of the Autonomic Nervous System, it may offer valuable information related to stress. Indeed, recent studies show the first evidence that EDA as recorded during daily life can be related to individuals' burnout symptoms over time [3] and EDA increases preceding aggressive behavior [4]. Sano et al. [5] showed EDA features to be informative about students' level of stress and mental health. These three studies all used a particular EDA wearable that records EDA with sensors embedded in a wristband. While this is convenient for the user, EDA responses to arousing events as recorded using sensors at the wrist have been found to be much less sensitive compared to sensors at the palm of the hand or fingertips, the latter being the usual location to record EDA in the laboratory [6,7]. Specifically, Van Lier et al. [8] show in a thorough study on validating wearables that a wristband could pick up effects caused by a strong sustained stressor, but, in contrast to the *gold standard* sensors placed at the finger, not to short, more mild ones. These findings probably reflect a lower density of sweat glands at the wrist compared to the fingers or palm [9].

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

We are interested in recording physiological responses to events that may be arousing for an individual in real life. In the current study, we tested heart rate and EDA wearables that we estimated to be relatively sensitive to pick up such responses. Therefore, we did not use wearables from the side of the spectrum that may have favored user comfort at the cost of signal quality (i.e., we did not use wristbands for recording EDA and heart rate) but wearables that are less comfortable to use on a daily basis, but rely on higher quality signals. These wearables are Wahoo Tickr, a wearable that extracts heart rate from electrical signals as measured at the chest, and EdaMove, a wearable that measures skin conductance at the hand palm. While wristbands are easier to attach, and in the case of EDA, less visible and interfering than a chestband and sensors at the palm, we found both of these wearables to be working well from a user perspective in recordings in which 90 young teenagers were recorded during a school day.

We collected EDA and HR in an experiment aimed at examining the effect of attentional instruction in multiple physiological data streams. It involved auditory emotional stimuli, and EDA and HR were recorded using both high-end laboratory equipment as well as the EdaMove and Wahoo Tickr. We utilize the data of this experiment here to compare signals coming from both types of equipment at a general level as well as at a response level, i.e. in relation to an emotional sound.

Methods

Participants

We recorded from 27 participants (aged between 18 and 48) with no self-reported problems in hearing or attention. Prior to the experiment all participants signed informed consent and after the experiment they received a small monetary award for their time and traveling costs. Data of three participants were discarded due to failed physiological recordings. The study was approved by TNO Institutional Review Board (TCPE) and TU Delft Human Research Ethics Committee.

Materials

Standard EDA, ECG (electrocardiogram) and EEG were recorded using an ActiveTwo system (BioSemi, Amsterdam, Netherlands) at 1024 Hz. EEG data is not further discussed here. Additionally, a wearable device, EdaMove 4 (Movisens GmbH, Karlsruhe, Germany), was also used to recorded EDA at different positions at 32 Hz. Figure 1 shows a picture of the simultaneous recording of EDA using both lab-grade and consumer-grade equipment with recording locations at the fingers and the palm respectively. HR was also recorded using Wahoo Tickr (Wahoo Fitness, Atlanta, GA, USA).

For EDA recorded with ActiveTwo, two passive gelled Nihon Kohden electrodes were placed on the ventral side of the distal phalanges of the middle and index finger. For EDA recorded with EdaMove, two self-adhesive electrodes were placed on the palm. Electrodes of both devices were placed on the non-dominant hand.

For ECG recorded with ActiveTwo, two active gelled Ag-AgCl electrodes were placed at the right clavicle and lowest floating left rib. For recordings with Wahoo Tickr, a band was fitted around the participant’s chest after applying gel on its sensors.

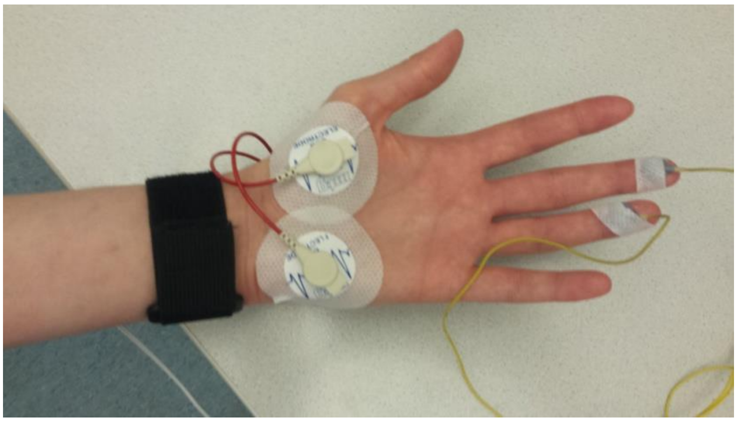


Figure 1. Electrodes on two fingers record EDA with ActiveTwo and electrodes on palm record EDA with EdaMove.

Stimuli and design

Each participant listened to an audio file, composed of a 66 min audiobook interspersed with other auditory stimuli including emotional sounds. Emotional sounds were taken from the IADS (International Affective Digitized Sounds-[10]: 12 neutral, 12 pleasant and 12 unpleasant sounds. Half of the participants were asked to attend to the audiobook and half to the other stimuli, where we expect the emotional stimuli to draw the attention of all participants.

Analysis

Data processing was done using MATLAB 2018b software (The Mathworks, Natick, MA, USA). ActiveTwo EDA was first downsampled to 32 Hz. The phasic (fast) and tonic (slow) components were extracted for both ActiveTwo and EdaMove EDA signals using Continuous Decomposition Analysis [11] as implemented in the Ledalab toolbox for MATLAB (available on www.ledalab.de).

The output from Wahoo Tickr is HR at 1 Hz. ActiveTwo ECG was downsampled to 256 Hz and high-pass filtered at 0.5 Hz. R-peaks were detected from a squared version of the reconstructed frequency-localized version of the ECG waveform using wavelets, following The Mathworks (2015). The R-to-R interval, or inter-beat interval (IBI) was extracted. The IBI semi-time series was transformed into a time-series HR. This was done by interpolating consecutive IBIs and then resampling at 2 Hz.

For both EDA and HR, plots of the raw wearable and ActiveTwo signals across the whole experiment were inspected for all participants.

To examine their sensitivity to emotional stimuli, EDA and HR signals were also aligned to the onset of the emotional sounds. For EDA, the phasic component was used since here, we are interested in the fast changes. In addition, the number of phasic EDA peaks larger than 0.1 μS was determined for both signals from the 1st to the 65th minute of the experiment.

Results

EDA

Figure 2 shows a typical sample of raw EDA traces from both devices. We observed that the return to baseline after a peak is slower in EDA recorded with EdaMove than with ActiveTwo. Figure 3 shows this in averaged data.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

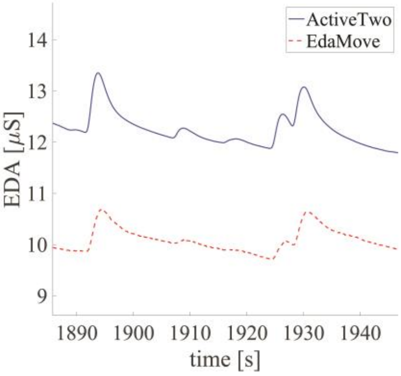


Figure 2. Example of raw EDA traces.

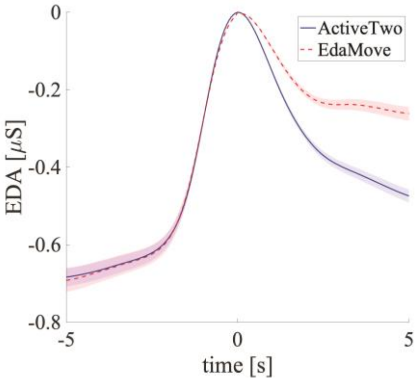


Figure 3. Averages of peaks with the standard error of the mean.

In this figure, epochs of raw signals were aligned to the time and the height of the maxima of the peaks. In order to include the same peaks present in both EdaMove and ActiveTwo signals, only peaks present in phasic components of both devices were used. A peak was considered to be present in both phasic components if the time difference of its maximum value was not larger than 1 s. Total number of peaks varied strongly between participants (a range of 0-570 for EdaMove and 1-358 for ActiveTwo). A paired *t*-test indicated no significant difference (t_{23}=1.85, *p*=0.08).

Besides the difference in the response shape, visual inspection also revealed more slow drift in the signal of the EdaMove than in ActiveTwo. An impression of this is given by the averages of the tonic components of all participants in Figure 4.

Figure 5 shows phasic EDA baselined on the onset of the emotional stimulus that elicited the strongest EDA response in our experiment (EroticFem2).

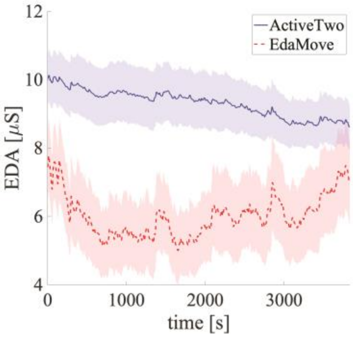


Figure 4. Averages of the tonic component of EDA signal with the standard error of the mean.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

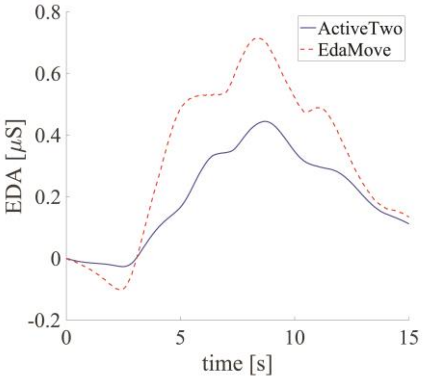


Figure 5. Averages of the phasic component of EDA signal after the IADS EroticFem2 stimulus.

Inspecting these plots for all sounds shows that averaged EDA phasic responses to emotional sounds are similar for both devices. If anything, EdaMove averaged phasic responses are somewhat stronger than those recorded by ActiveTwo. This may be caused by the slower drop of the signal as described above, causing a strong average response even when response latencies between individuals differ.

HR

Figure 6 shows a typical sample of HR traces from both devices. ActiveTwo reflects the typical, fluctuating pattern of HR that is associated with breathing [12]. HR from Wahoo Tickr does not show this. Since HR as obtained from Wahoo Tickr was delayed with respect to that obtained from ActiveTwo, Figure 6 also shows a shifted version of the Wahoo Tickr data. Shifting data was done using the determined delay of that participant. For this we filtered Wahoo Tickr data with a moving average (window size of 20) and filtered ActiveTwo HR with a 4th order low pass digital Butterworth filter with a normalized cutoff frequency of 0.03. After filtering, minima and maxima with a prominence larger than 1 were detected. We then calculated the time difference between peaks in Wahoo Tickr and ActiveTwo signals. Extrema are counted as the same if the delay of the Wahoo Ticker is not larger than 10 s. Finally, the median of the found delays of an individual participant was defined as the final delay of the Wahoo Tickr. The mean delay across participants was 6.7 s (sd: 1.3).

The mean absolute HR difference was 3.4 bpm (median: 2.7, sd: 2.2, range: 1.9-11.1). Mean Wahoo Tickr HR was 69.5 bpm and ActiveTwo HR was 70.1 bpm. This was not significantly different (paired t-test: t_{23}=1.89, p=0.07).

Figure 7 shows HR baselined on the onset of EroticFem2. A decrease in HR in response to the sound can be observed for both devices (consistent with Bradley & Lang [13]). Consistent with our finding that Wahoo Tickr does not capture fast fluctuating changes in HR, this pattern is more pronounced for ActiveTwo.

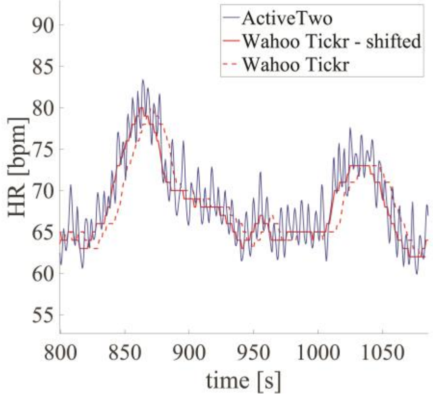


Figure 6. Example of raw HR traces. The dashed line indicates the Wahoo Tickr HR before shifting for the calculated delay.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

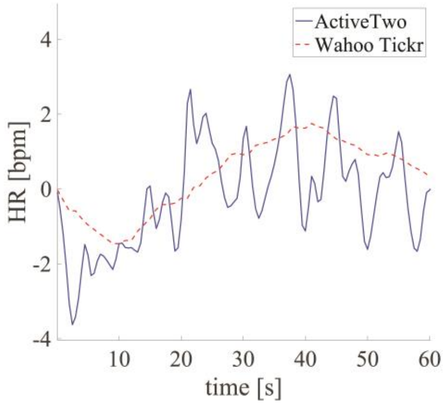


Figure 7. HR with the standard error of the mean after the IADS EroticFem2 stimulus.

Discussion

We observed some systematic differences between the EDA and HR signals from the wearable and lab-grade equipment.

Return to baseline seemed to be slower for EDA signals recorded with EdaMove than ActiveTwo. The underlying reason is yet to be investigated but we speculate that the size of EdaMove's self-adhesive electrodes affect the return part of a peak by slowing down evaporation of sweat. This might also induce more drift in the signal. The number of peaks found in phasic component of EDA did not significantly differ, though EdaMove tends to show more peaks. We observed for some parts of the data that EdaMove showed peaks that were completely absent for ActiveTwo, suggesting that this has something to do with the distribution of sweat glands at the lower end of the palm versus the fingers rather than with sensitivity of the sensors.

HR from Wahoo Tickr was delayed with respect to ActiveTwo and did not capture fast changes. This is likely caused by filtering and other processing by the Wahoo Tickr equipment. The difference in absolute HR between the systems are caused by the fact that only ActiveTwo captured fast changes in HR such as those caused by breathing, but also by ActiveTwo capturing ectopic beats (‘skipping’ a beat, or an ‘extra’ beat) that are usually excluded from the data for a more robust estimate of HR. This happened relatively often in three participants.

We found both wearables to perform quite well. Slow drifts (EdaMove) or a delay (Wahoo Tickr) are not necessarily big problems for analyses of emotional responses as long as they are known. General HR as captured by Wahoo Tickr corresponds to the ActiveTwo gold standard. EdaMove captured quick changes as those caused by emotional sounds equally well as ActiveTwo, while this was more difficult Wahoo Tickr. When the aim is to capture physiological responses to potentially emotional events, we therefore expect EdaMove to perform about as well as ActiveTwo and Wahoo Tickr as less well. In addition, the lack of systematic fast variation in HR (Figure 6) indicates that Wahoo Tickr is not suited to capture heart rate variability.

Note that in the present study, we recorded from participants who were sitting still. In real life circumstances involving movement, recordings of EDA and HR will be affected by noise and inherent associations of EDA and HR with body movement. This is a topic of ongoing research. See e.g. [14] for a study of EdaMove under conditions of movement, and a way to identify movement artefacts, and [15] for potential ways ot cope with the effect of movement on HR.

References

1. Brouwer A.-M., Zander T.O., van Erp J.B.F., Korteling J.E., Bronkhorst A.W. (2015). Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. *Frontiers in Neuroscience* **9**:136.

2. de Looff P.C., Cornet L.J.M., Embregts P.J.C.M., Nijman H.L.I., Didden H.C.M. (2018). Associations of sympathetic and parasympathetic activity in job stress and burnout: A systematic review. *PLOS ONE* **13**(**10**):e0205741.

3. de Looff P., Didden R., Embregts P., Nijman H. (2019). Burnout symptoms in forensic mental health nurses: Results from a longitudinal study. *International Journal of Mental Health Nursing* **28(1)**:306-317.

4. de Looff P., Noordzij M.L., Moerbeek M., Nijman H., Didden R., Embregts P. (2019). Changes in heart rate and skin conductance in the 30 min preceding aggressive behavior. *Psychophysiology* **56(10)**: e13420.

5. Sano A., Taylor S., McHill A.W., Phillips A.J., Barger L.K., Klerman E., Picard R. (2018). Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: Observational study. *Journal of Medical Internet Research* **20(6)**: e210.

6. Payne A.F.H., Schell A.M., Dawson M.E. (20182016 Lapses in skin conductance responding across anatomical sites: Comparison of fingers, feet, forehead, and wrist. *Psychophysiology* **53(7)**:1084-1092.

7. van Dooren M., de Vries J.J.G., Janssen J.H. (2012). Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology and Behavior* **106(2)**: 298-304.

8. van Lier H.G., Pieterse M.E., Garde A., Postel M.G., de Haan H.A., Vollenbroek-Hutten M.M.R., Schraagen J.M., Noordzij M.L. (2019). A standardized validity assessment protocol for physiological signals from wearable technology: Methodological underpinnings and an application to the e4 biosensor. *Behavior Research Methods*, 1-23.

9. Boucsein W. (2012). *Electrodermal activity*. New York: Springer Science+Business Media, LLC.

10. Bradley M.M., Lang P.J. (2007). The international affective digitized sounds (IADS-2): Affective ratings of sounds and instruction manual. *Technical Report B-3*, University of Florida , Gainesville, FL.

11. Benedek M., Kaernbach C. (2010). A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods* **190(1)**: 80-91.

12. Aasman J., Mulder G., Mulder L.J. (1987). Operator effort and the measurement of heart-rate variability. *Human factors* **29(2)**: 161-170.

13. Bradley M.M., Lang P.J. (2000). Affective reactions to acoustic stimuli. *Psychophysiology* **37(2):** 204-215.

14. Thammasan N., Stuldreher I.V., Wismeijer D., Poel M., van Erp J.B.F., Brouwer A.-M. (2019). A novel, simple and objective method to detect movement artefacts in electrodermal activity. *Proceedings of 8th International Conference on Affective Computing and Intelligent Interaction* (Cambridge, 3-6 September 2019), 1-7.

15. Brouwer A.-M., van Dam E., van Erp J.B.F., Spangler D.P., Brooks J.R. (2018). Improving real-life estimates of emotion based on heart rate: A perspective on taking metabolic heart rate into account. *Frontiers in Human Neuroscience* **12**, 284.

Data Synchronisation and Processing in Multimodal Research

T.C. Dolmans[1], M. Poel[2], J.W.J.R van 't Klooster[3], and B.P. Veldkamp[4]

1 Data Management & Biometrics, University of Twente, Enschede, The Netherlands. t.c.dolmans@utwente.nl
2 Data Management & Biometrics, University of Twente, Enschede, The Netherlands. m.poel@utwente.nl
3 Behavioural, Management and Social Sciences, University of Twente, Enschede, The Netherlands. j.vantklooster@utwente.nl
4 Onderzoeksmethodologie, Meetmethoden en Data-analyse, University of Twente, Enschede, The Netherlands. b.p.veldkamp@utwente.nl

Abstract

Gathering physiological information with the help of various, sometimes wearable, devices can help gain insight into mental states and constructs, such as mental workload. These devices measure, for example, galvanic skin response, heart rate, or neural activity. However, challenges arise during the collection and synchronisation when combining multiple devices to evaluate the mental state of the participant. Devices use different timestamps, communicate only with certain hardware or software, and do not always provide reliable avenues of data synchronisation after recording. Furthermore, data processing of multiple modalities requires new ways of thinking about the data in order to maximise yield. In this paper, we discuss different types of approaches to a workflow and pipeline that can help overcome these problems when identifying mental workload. Modularity and generalisability are identified as key criteria for both data collection and processing. They are embodied through the usage of LabStreamingLayer for the collection and synchronisation of data, and lightweight convolutional neural networks that interlink using mixture-of-experts modules for processing of data.

Introduction

Mental workload (MWL) is a topic that has gained a lot of attention in a variety of fields, such as neuroscience [1, 2], human factors and ergonomics [3], and human factors in computing systems [4]. What these papers have in common is that they seek to gain insight into individual perceived workload. They work towards understanding how workload explicitly manifests in the body based on implicitly measured bio-signals. In this context, implicit measurement refers to the collection of data without explicitly asking participants to provide information. However, since participants are still aware of sensors that are physically attached to them, the measurement is not unobtrusive. The University of Twente (the Netherlands) is currently engaged in a collaborative project with several smaller and bigger companies in the East of the Netherlands. The objective is real-time classification of MWL through implicit measurement. Several bio-signals are being explored to do this. Some of the signals we focus on are brain activity, through electroencephalography (EEG) [5] and functional near-infrared spectrometry (fNIRS) [6], galvanic skin response (GSR) [7], heart rate (HR) [3], and pupillary activity (PA) [4]. All of these modalities have individually proven to shed new light on the collective understanding of MWL and its manifestations in our measurements. With the help of novel techniques, such as centralised recording and machine and deep learning, research that uses all the above-mentioned modalities becomes possible. Such research may lead to novel insights into the physical manifestations of MWL. However, this requires an understanding of a multitude of physiological principles, software communication, and data processing. Hence, the collection, synchronization, and processing of bio-signals from multiple devices simultaneously is a non-trivial task. Our objective is to design a workflow and pipeline that can help overcome this task. This paper discusses different possible approaches towards reaching our objective. Furthermore, it defines key principles and suggests verified ways of constructing a pipeline and workflow for multimodal research in naturalistic environments. First, we discuss relevant literature around MWL. Thereafter, the data synchronisation problem is broken down into two main streams of solutions. Lastly, data processing is discussed, again, with a focus on the two main streams of options. The discussed meta-principles generalise well across modalities, processing methods, and research objectives and should thus be applicable to any multimodal research.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Methods and Results

Mental Workload

Mental workload (MWL) is defined as the subjective measure of required (mental) capacity for a given tasks. The construct of MWL depends on two variables: the cognitive resources that one has, and the cognitive resources required for a task. A state of "flow", as described by Csíkszentmihályi (1975) [8], wherein the required versus available cognitive resource (RCR/ACR, or α) of the task are aligned, is considered a proper balance. In flow, one experiences full emersion with the task at hand. This is desirable in, for example, sports, music, or education. Determining when one resides in such a state requires explicit information. The ACR and prior knowledge of a participant can be determined, by, for example, taking a placement test before experimentation begins. It is important to collect this information because it allows for calibration of the perceived difficulty level of your task; the RCR differs per participant and hence determine task difficulty. During an experiment, we would like to gain insight into the MWL of the participant in order to assess α. In turn, we can then determine various things, e.g. the growth of a skill of a participant. Formula 1 summarizes such growth, based in the discussed factors.

$$\beta = P + \alpha t$$

Formula 1: Growth of skill in a given domain, signified by β, modelled by a function of prior knowledge (P), plus ratio of the required cognitive resources for the task over the available cognitive resources $\frac{RCR}{ACR}$, or α, into time spent performing said task, t. For a state of flow, $\frac{RCR}{ACR}$ should stay between 0.8 and 1.2 [9].

Traditionally, the assessment of growth of a skill is also performed with explicit information provided by the participant. In other words, participants are actively involved in (self)assessing their degree of flow and their progress in a skill. However, the mere act of performing a measurement of a phenomenon through reflection can interfere with the phenomenon [10]. Hence, requiring subjects to reflect on their experiences may reduce objectivity. We want to prevent this reflective blur, as well as the interruption of flow. Implicit physiological measurements can satisfy this requirement and provide new insight into flow, since they allow us to asses MWL without interfering with the participant's activities. However, we're continuing to strive for unobtrusive measurements to further reduce the interruption of participants' flow. Understanding the development of MWL will yield valuable insight into the correlation between key moments of learning during the experimental task and the respective perceived workload at those times. In a mature setting, users would receive feedback from an unobtrusive system in real time, nudging them in the direction of a desirable α. This can, for example, be done by adapting the environment's intensity to trigger a state of flow. However, the participant can also be adapted to the environment by modulating the participant's arousal. This can, for example, be done with the help of visual, audial, or even olfactory stimulation [11, 12]. Now that we have outlined our objective, namely the identification of a state of "flow" in participants, we will discuss several considerations and obstacles that need to be overcome during data synchronisation and processing.

Data Synchronisation

Two key aspects to any solution are modularity and generalisability (MG). Concretely, new devices should be easy to add to the setup and their data interpretation should fit within the pipeline with minimal structural implications; modularity. Then, the additional data that becomes available from added modalities should contribute to our classification accuracy in MWL related tasks. Furthermore, the pipeline should be applicable to other topics besides MWL, regardless of research objective or design; generalisability. If we wish to use data from a variety of devices to determine MWL, it is important to synchronise all incoming data properly. In other words, to line up the data such that it overlaps for the duration of the experiment. Despite the widespread usage of the UNIX timestamp on many devices, the exact formatting and calibration can vary, in our experience. Synchronisation of data recorded across devices can become very challenging and time consuming because of this. Hence, we require different methods for data synchronisation. Several approaches to tackling this problem present themselves. Time synchronisation in our case refers to the synchronisation of time-series data received from multiple devices. The synchronisation is often a hurdle when working with distributed networks. Differences in sampling rate, internal time-keeping and package-size all contribute to desynchronization. The most common ways to prevent asynchrony among devices, rely on network time protocols, de-jittering, and/or smoothing of time differences [13, 14].

The first consideration to make is whether data recording should be centralised or decentralised. This means that all data are recorded using one application, or several, respectively. Figure 1 presents an overview of the main branches of options for data synchronisation. In a decentralised setup, several options become available: pre-and-post recording synchronisation. Pre-recording synchronisation requires all devices to be set up in such a way that they communicate about their perception of time the same way. Furthermore, The degree of (a)synchrony should always be verified after the recording. This is requires relatively a lot of attention from researchers and lab technicians, thus increasing the chance of human-error. Post-recording synchronisation does not require the manual setting of timestamps beforehand but does require the addition of synchronisation markers to each device. The challenges with this are generating such markers on each recording device or software and merging the data from different file formats. Practically, most studies that use multimodal data, use decentralised recording, meaning that data synchronisation is done at least partially manually [15, 16].

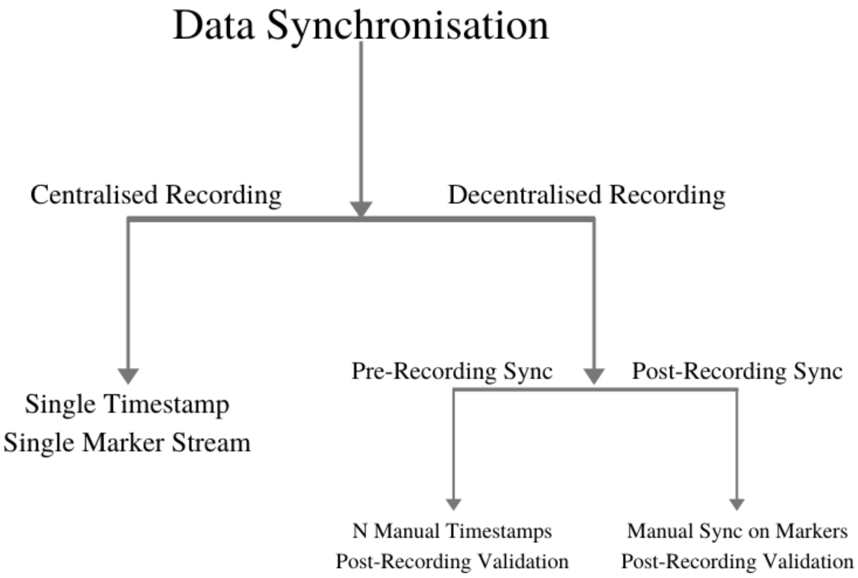


Figure 1: Data synchronisation options flowchart for multimodal data. Under "Pre-Recording Sync", N refers to the number of devices that are being used for the research.

In a centralised setup, all devices are recorded on the same computer or server. Creating the ensemble of data in a single data format will allow for easier and more generalisable data processing. However, the biggest challenge with centralised recording is accessing the data from all devices on a single location. Often, hardware manufacturers keep their data streams behind "closed doors". This means that setting up a centralised recording system requires detailed knowledge in various domains, raising the threshold for doing multimodal research. In our research, we use open source software, such as LabStreamingLayer (LSL) to construct a centralised recording method. LSL is a toolbox that can stream and access data using small "apps" for each device [13]. These apps are either developed and available on the internet, or new ones can be made. Given a device, these LSL apps can stream recorded data to a local network in several abstractions, such as samples, chunks, and metadata. The streams of all apps on a network can then be detected and recorded by any device connected to the same network. The resulting recording is a single extensible data format file that contains all available network streams [17]. This data format contains a stream for each device and useful information is described in the meta-data of the streams, such as stream type, name, and sampling rate [13]. However, not every device has a (working) app, meaning that some apps may have to be developed and compiled into executables, which can require a lot of time. By now, it is evident that all knowledge needed for the creation of such a setup requires multi-disciplinary teams. The expertise of physiological research as well as data gathering and interpretation calls for a level of specialisation that is seldom found in one person or specialty group. Part of this can be attributed to the complexity of the problem, another part to the structure under which many devices still operate: single modality, supported soft and hardware only. Researches that have successfully set up a centralised recording system using LSL still report slight issues with synchronisation of up to several hundreds of milliseconds due to, for example, the internal buffer of the recording app [18]. This difference is too large in EEG research, which is the case in [14]. However, other studies do not report any issues at all [19]. Newer versions of LSL's recording software "LabRecorder" makes use of built-in synchronisation tools based on Network Time Protocol, such as offset correction and de-jittering [14], which can help overcome the majority of desynchronization across devices and even correct for time drift. In summary, using LSL allows for the addition of new devices, without needing to radically change the existing infrastructure for the collection of data. However, a barrier to overcome is the creation of apps for devices that do not yet have support. In terms of synchronisation, there appears to be no consensus on whether small data synchronisation issues are common, hence it is advised to append a marker channel to several data streams as a reference. In our case, when we want to assess MWL using multimodal implicit data, the centralised recording method is preferred. This method satisfies the modularity requirement since it allows for the addition or removal of recording devices. Furthermore, less manual work, and thus less chance of human error, is associated with the compilation and synchronisation of data. Lastly, since we are dealing with relatively small segments of data of 3 to 5 seconds, we are less concerned by time drift when using markers to synchronise.

Data Processing

Once all data have been collected and synchronised, whether it be in a centralised or decentralised way, data analysis commences. Assuming synchronisation was successful, we now wish to classify MWL during different stages of the experiment. Once we have succeeded in this, we can evaluate whether our task was suitable for our goal of inducing flow in the participant. Similarly to the collection of data, several key considerations have to be made for processing and interpretation of it. Some data streams, for example those that contain GSR and HR, can be interpreted with methods that make use of relatively straightforward computations. For instance, when interested in MWL, GSR data can be analysed with a decision tree and learnt thresholds, making it manageable and computationally lightweight. Nonetheless, the results provide valuable insights into perceived MWL, as illustrated by [20]. However, there are some modalities that benefit from less straightforward processing methods, such as EEG [21-23]. State of the art techniques, such as filter-bank common spatial pattern (FBSCP) or deep neural networks (DNN), use learnt features to classify EEG tasks to achieve superior results to static solutions. Furthermore, the combination of EEG and fNIRS has proven to lead to higher classification accuracies than the modalities separately [6, 24]. Therefore, it can be expected that the addition of multiple other modalities will further increase the classification accuracy when using more such processing methods. However, determining to what extent the addition of a modality influences classification accuracy is difficult, since there are many ways to involve these newly gained data during processing. For example, in an EEG and fNIRS context, we run into issues like incongruence in the sampling rate and number of channels. Other issues could be the amplitude or dimensionality of the signal. Pre-processing and fusion of data can help overcome such issues. In general, there are two main categories in data fusion: early fusion, meaning data streams are transformed and fused before most of the feature learning takes place, and late fusion, meaning that the results of the analysis on individual modalities are fused. Shin and colleagues (2018) present a way of designing a hybrid late fusion system by pre-processing EEG and NIRS data separately, and merging them after performing analysis on the individual streams [6]. The results of FBCSP for EEG and shrinkage linear discriminant analysis (sLDA) for NIRS data are fed through meta sLDA separately. These results are then combined to create three feature vectors that are classified with a majority vote, as can be seen in Figure 2. Effectively, the data is processed and classified separately. This means that accuracy is improved by checking the classifications against each other and choosing the most common one.

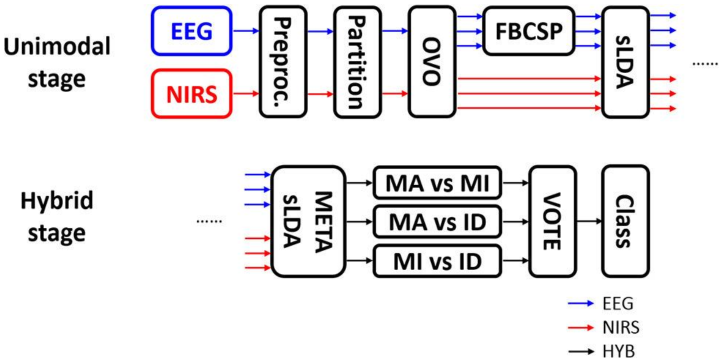


Figure 2: “Data processing flow. EEG and NIRS data were separately processed at the unimodal stage and were combined at the hybrid stage. OVO, FBCSP, and sLDA indicate “one-versus-one” strategy, filter-bank common spatial pattern, and shrinkage linear discriminant analysis, respectively” Quote and image retrieved from Shin, Kwon, & Chang-Hwan, 2018 [6], page 4. MA, MI, and ID refer to mental arithmetic, motor imagery, and idle state, respectively.

The second major category in data fusion is early fusion. Early fusion is still relatively new in this context, meaning that very little research has been done to explore the additional value over late fusion. In deep learning, multi-task learning describes a similar problem in that it attempts to optimise several loss functions by leveraging domain-specific information contained in training data of peripheral tasks [25]. However, in a multimodal setting, instead of attempting to classify several different things form one data stream, we are attempting to classify several things from multiple streams. Kaiser and colleagues (2017) describe the MultiModel architecture, which is able to perform well in different domains [26]. They use several smaller convolutional, attention, and mixture-of-expert (MoE) blocks to formulate four modality networks (MNets) that classify written language, images, audio, and categorical data. The high-level architecture of the MultiModel can be seen in figure 3. MNets are lightweight, usually convolutional, neural networks that are built to perform heavy feature extraction on raw data. They output a new data stream that can be interpreted by an input encoder. The end result is essentially one DNN that is able to perform well on tasks in different domains, as well as improve results within domains when sharing features learnt in peripheral domains.

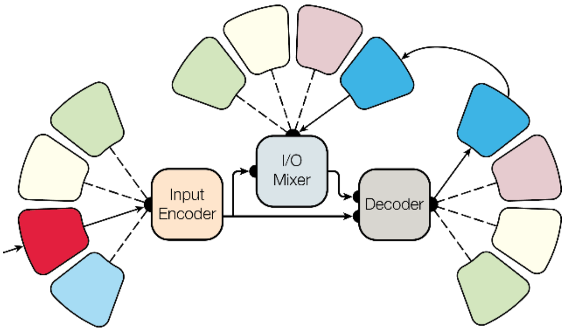


Figure 3: The MultiModel, with MNets, the input encoder, and an autoregressive decoder. Image retrieved from Kaiser et al., 2017, [26], page 3.

To summarise, in early fusion, using MNets over larger, more specialised DNN has several advantages. MNets are better able to generalise across tasks, since they make use of task-peripheral data to learn features. Furthermore, they learn which features to share when used in conjunction with MoE blocks, which reduces computational cost [26]. From the viewpoint of the earlier postulated MG (modularity & generalisability) criterium, early fusion is only attractive when the early convolutional layers feed data into the same input encoder; only then can we leverage the added benefit of the attention and MoE blocks that are common to all modalities. In late fusion, more developed methods, such as the aforementioned FBCSP and sLDA can be used to generate multiple binary classification problems. The combination of these problems then results in 2^N states, where N represents the number of modalities. Both early and late fusion fit our MG criterium to some extent and determining which one is “best” depends highly on the context. Long-term studies are required to evaluate the classification accuracy, compute cost, and MG of both paths.

A common hurdle with training DNN is the lack of annotated data. Properly annotated data is hard to come by, since such datasets are usually domain specific and handcrafted. Recent advancements in deep learning have led to deeper models with more trainable parameters. However, the creation of datasets required for these bigger models has not always followed suit in these advancements [27]. Therefore, it is essential to carefully think about an experimental paradigm that can provide accurately annotated data as it runs. This holds for any research paradigm, but becomes especially important when considering to use neural networks for data processing, regardless of how data collection and fusion was done. Hence, we recommend only using stimulus presentation software that is able to work in conjunction with your collection method. In our case this collection is done with LSL, since it can handle both the creation of annotated data and prevent the bulk of synchronisation issues [13, 17]. Furthermore, we recommend building in redundancy by streaming not only event markers, but also other metrics that can be formulated for a specific research paradigm. For example, performance of participants can be tracked to assess the difficulty of the task. Task-specific performance indicators can for example be the ratio of correct/incorrect answers, or the response time to stimuli. These performance indicators can then be used during the classification of MWL.

Conclusions

First, we postulated our interest in implicitly measuring MWL through multimodal data. The MG criterium was put forward for the synchronisation and processing of multi-modal data. Modularity implies the easy addition of modalities with minimal structural implications. Generalisability implies poly-applicability of the designed pipeline, as well as increased accuracy with an increased number of modalities. For our purposes, a centralised collection method is preferred for recording and synchronisation. However, this brings its own set of problems, such as apps that still have to be developed for several devices before they can interface with LSL. In terms of data processing, both early and late fusion show potential and can satisfy the MG criterium. Hence, we conclude that the eventual choice of fusion should be context dependant. Furthermore, we believe that modular and generalisable DNN architectures, such as the MultiModel will play an increasingly large role over the coming years, in a multitude of fields.

References

16. Lim, W., O. Sourina, and L. Wang, STEW (2018). Simultaneous Task EEG Workload Data Set. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **26(11)**: 2106-2114.

1. Toppi, J., et al., (2016). Investigating cooperative behavior in ecological settings: an EEG hyperscanning study. *PloS one,*. **11(4)**:. e0154236.
2. Schmalfuß, F., et al., (2018). Potential of wearable devices for mental workload detection in different physiological activity conditions. *Proceedings of the Human Factors and Ergonomics Society Europe*: 179-191.
3. Duchowski, A.T., et al. (2018). The index of pupillary activity: measuring cognitive load vis-à-vis task difficulty with pupil oscillation. in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM.
4. Gerjets, P., et al., (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in neuroscience*. **8**:. 385.

5. Shin, J., J. Kwon, and C.-H (2018). Im A ternary hybrid EEG-NIRS brain-computer interface for the classification of brain activation patterns during mental arithmetic, motor imagery, and idle state. *Frontiers in neuroinformatics* **12**: 5.

6. Nourbakhsh, N., et al., (2017). Detecting users' cognitive load by galvanic skin response with affective interference. *ACM Transactions on Interactive Intelligent Systems* **7**(**3**): 1-20.

7. Csikszentmihalyi, M. (1975). Beyond Boredom and Anxiety Jossey-Bass Inc. San Francisco, USA.

8. Csikszentmihalyi, M. (1997). Finding flow: The psychology of engagement with everyday life. Basic Books.

9. Mahtani, K., et al., (2018). Catalogue of bias: observer bias. *BMJ evidence-based medicine* **23(**1): 23.

10. Weinbach, N., et al., (2015). Can arousal modulate response inhibition? *Journal of Experimental Psychology: Learning, Memory, and Cognition* **41(6)**: 1873.

11. Hughes, M. (2004). Olfaction, Emotion & the Amygdala: arousal-dependent modulation of long-term autobiographical memory and its association with olfaction: beginning to unravel the Proust phenomenon? *Impulse: The Premier Journal for Undergraduate Publications in the Neurosciences* **1(1)**: 1-58.

12. Kothe, C. (2015). Lab streaming layer (LSL*).* https://github.com/sccn/labstreaminglayer. Accessed on October, 2014.

13. Boulay, C. (2019). Time Synchronisation. [cited 2020 March 13 2020]; Available from: https://github.com/sccn/labstreaminglayer/blob/master/docs/info/time_synchronization.rst.

14. Spitzley, L. et al (2018). Using multimodal data to infer group dynamics in an adversarial group game. in *13th Annual Symposium on Information Assurance (ASIA '18)*.

15. Born, J., et al. (2019). Multimodal Study of the Effects of Varying Task Load Utilizing EEG, GSR and Eye-Tracking. *bioRxiv*: 798496.

16. Ojeda, A. and C. Kothe (2015). Extensible Data Format (XDF). [cited 2020; Available from: https://github.com/sccn/xdf.

17. Bleichner, M.G., B. Mirkovic, and S. Debener (2016). Identifying auditory attention with ear-EEG: cEEGrid versus high-density cap-EEG comparison. *Journal of neural engineering* **13(6)**: 066004

18. Mendonca, P. and M.A. Abreu. (2019). A Hybrid System for Assessing Mental Workload. in 2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG). *IEEE.*

19. Martinez, R. et al. (2019). A Self-Paced Relaxation Response Detection System Based on Galvanic Skin Response Analysis. *IEEE Access* **7**: 43730-43741.

20. Lotte, F. et al., (2018). A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of neural engineering* **15(3)**: 031005.

21. Craik, A., Y. He, and J.L. Contreras-Vidal (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of neural engineering* **16(3**): 031001.

22. Yang, B., et al. (2018). Motor Imagery EEG Recognition Based on FBCSP and PCA. in *International Conference on Brain Inspired Cognitive Systems*, Springer.

23. Fazli, S. et al. (2012). Enhanced performance by a hybrid NIRS–EEG brain computer interface. *Neuroimage* **59(1)**: 519-529.

24. Caruana, R. (1998). Multitask Learning. Autonomous Agents and Multi-Agent Systems

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

25. Kaiser, L., et al. (2017). One model to learn them all. *arXiv preprint* arXiv:1706.05137

26. Yu, F., et al (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint* arXiv:1506.03365.

Baboons on the Move: Enhancing Understanding of Collective Decision Making through Automated Motion Detection from Aerial Drone Footage

Christopher L. Crutchfield, Jake Sutton, Anh Ngo, Emmanuel Zadorian, Gabrielle Hourany, Dylan Nelson, Alvin Wang, Fiona McHenry-Crutchfield, Deborah Forster, Shirley C. Strum, Ryan Kastner, Curt Schurgers

Engineers for Exploration, University of California, San Diego
ccrutchf@ucsd.edu

Introduction

Collective and distributed decision-making has long been a topic of interest in animal research since it is a complex process in many nonhuman animal species. Long-lived social mammals that interact within societies have much in common with humans. Within these particular societies, individuals and their connections within their social network have a critical impact on group-level behavior. This is particularly true of nonhuman primates. In this paper, we examine tracking of baboon troop movements using a combination of human observers and computer vision techniques to aid in the study of the group-level behaviors that impact troop movement and collective decision-making.

To understand group-decision and the context thereof, one would ideally be able to continuously (1) monitor identified individuals and their activities, (2) track the relational dynamics and social networks, (3) monitor group-level behavior and (4) monitor the environment.

In order to find and analyze the moment when a decision is made—for example, what direction a troop will head—it is necessary to track the entire group of baboons individually. The decision is a complex negotiation that can originate from both local dynamics and relational history between individuals. These interactions between individuals may have more impact than other independent variables. Since it is difficult to know all of the variables involved in a decision, it is necessary to understand the moment in which a decision is made.

Currently, researchers in the field take notes about observations. Unfortunately, these notes do not provide the full context as a single researcher can only see a portion of the entire sleeping site. GPS radio collars have been used to augment the notes taken by field researchers, but they only allow for observing partial group membership and thus require attempting to fill in gaps left by having only partial results. The use of drones would allow for monitoring of individuals and small groups, but also the troop as a whole in a way that is not achievable with GPS collars or field observations alone.

Aerial drone footage can help to fill in some of these gaps, as it can provide a complete view of the entire site. As there may be hours' worth of drone footage to review, it is impractical to do it by hand. Instead, computer vision techniques can be used to reduce the amount of video that must be reviewed. By annotating the beginning of a video, an individual identification can be maintained through the majority of the video through automated means. This ensures that the context of a decision is not lost.

Automated computer vision techniques, however, are challenged by the movement and distortion of drone-mounted cameras. As these cameras have six degrees of freedom (DoF)—the freedom of movement a non-fixed camera has in three-dimensional space—it becomes necessary to compensate for the potential movement of said camera. Additionally, individual identification currently requires a significant effort in the field as experienced field researchers must annotate individual baboon IDs in as close to real-time as possible. This motivates the minimization of the time it takes to process the drone footage.

In this paper, we report on methodological and computational developments that show promise towards solving these problems, with the primary goal of being able to use the processed footage to identify the moment of decision. Drone footage is less invasive than other methods (e.g. radio collars) and allows researchers to more

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

easily view the entire group. This footage also provides additional context that collars cannot, as a collective decision may originate from an individual's agenda.

Background

Birds and honeybees have previously been used to study collective decision-making, but nonhuman primates are of particular interest to study because of their cognitive and behavioral similarities to humans [1]. With the development and popularization of the global positioning system (GPS), researchers were better able to investigate leadership, decision-making, and model troop movement in groups of animals [2], leading to findings of group movement governed by a majority rule. When GPS data was overlaid on a detailed environment map, topography was also found to influence group movement [2].

The Uaso Ngiro Baboon Project (UNBP) encompasses the most complete baboon socio-ecological field research for close to half a century, with a more focused study on troop movement since 1994. While this project has many of the above elements, it does not currently capture the moment the collective decision is made. The troop's first major, collective decision of the day—movement from the sleeping site—has the potential to provide this missing insight. As baboons' preferred sleeping sites are rocky outcrops (see Figure 1), the initial movement is more constrained compared to decisions later in the day, therefore allowing the use of aerial imagery to be more practical.

In order to be able to better understand all of these complex factors, and building on the comprehensive history of baboons in the field within "Darwin's monkey: Why baboons can't become human" by Dr. Shirley Strum [1], we established an interdisciplinary effort that includes engineering students within the Computer Science and Engineering, the Electrical and Computer Engineering, Mathematics, and Data Science departments at University of California, San Diego, and field primatologists. While field methods in tracking have benefited over the years from technology augmented tools, there are still gaps in the capabilities of these tools. Here we discuss the progress of an on-going effort to bridge one of these gaps by augmenting field observations from the ground with aerial footage from drones processed using computer vision.

Related Works

This paper has similar goals to the paper by Haalck et al [3]; we want to be able to detect moving subjects even when given an unstable camera and low pixel density of those individual subjects. While their goal was to study, follow, and map the paths of individuals, our objective, rather than follow an object already in motion, was to study the group dynamic of our subjects and how they collectively decide before their movement is smooth.

This makes the method of reducing noise employed by Haalck et al difficult for us to use. While Haalck et al are able to eliminate almost all noise, they do so by employing statistical models that require the animals to have smooth motion [3]. Where we differ is in the number of animals and kinds of movements that we are able to pick up. As our baboons will typically stop or change direction on a whim, our data does not fit this smooth motion requirement. This necessitates implementing a different means of following the decision-making progress.

This period of decision-making is relatively hard to study with current methods of manual field research, not only because of the sheer quantity of baboons involved but also because of the poor viewing angles of people taking notes on the ground. Any single researcher on the ground can see only a partial view of the big picture. With drone footage, not only can it be easier to skim through hours of irrelevant footage, as a computer can assist, but it can also be digitized and pieced together so that the group can be understood as completely as possible. The moment we are trying to define has multiple stages, including instances where smaller portions of the group or even individuals advocate for different directions, until they collectively decide on one final direction. Field researchers currently are uncertain of how they collectively make that decision, so by identifying when this is happening, making it easier to visualize, and giving insight into things we can't visualize (such as baboon direction and future trajectory prediction) we hope to make understanding their process much more feasible.

Methods and Results

In our testing, we used a DJI Mavic 2 Pro drone equipped with an L1D-20C RGB, 4K camera. We found that flying at an altitude of 50 meters allows for baboons to have a frame size of about 25 by 25 pixels, which provides for sufficient detail for the animals to be resolved by humans and our algorithm. In the initial test footage we acquired, it appeared that the baboons habituated to the drone within a day or two.

We encountered some drawbacks from this method of data collection. First of all, even though the drone is very stable, it still moves slightly, having rotated $2.64°$ and 0.5 meters during a minute-long video. This is can be accounted for with our computer algorithm.

Another drawback that we have yet to fully solve is the relatively short flight time of the most commercial drones. As this flight time is often limited to under 30 minutes, it will be necessary to fly a second drone to fill in the gaps.

The algorithm that produces this result happens in multiple distinct steps, as can be seen in Figure 1.

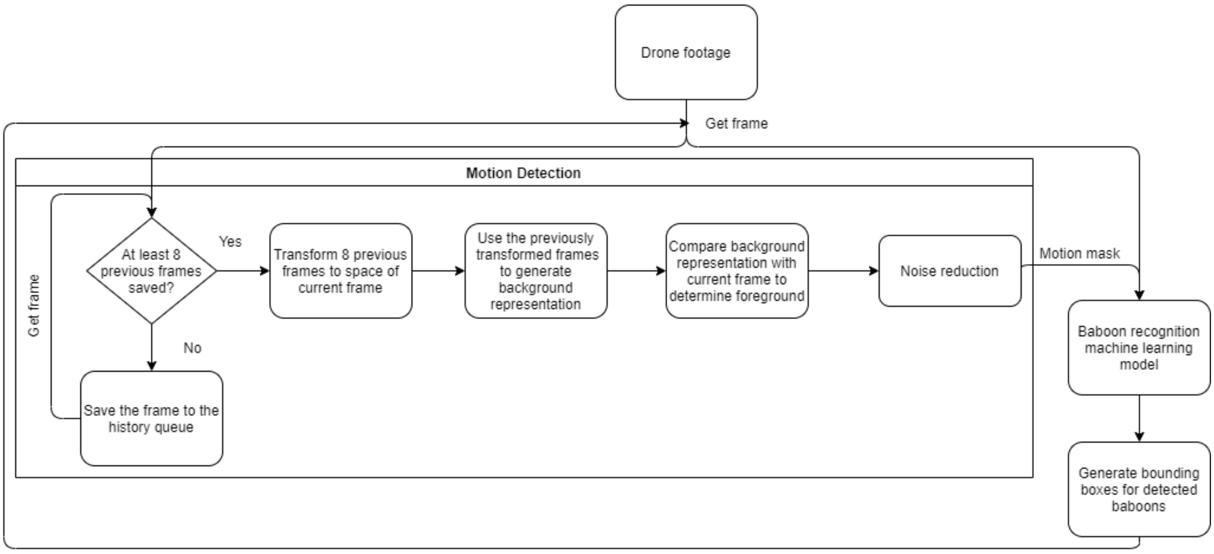

Drone footage
Get frame
Motion Detection
Get frame
At least 8 previous frames saved?
Yes
No
Transform 8 previous frames to space of current frame
Use the previously transformed frames to generate background representation
Compare background representation with current frame to determine foreground
Noise reduction
Motion mask
Save the frame to the history queue
Baboon recognition machine learning model
Generate bounding boxes for detected baboons

Figure 1. Proposed pipeline of the baboon detection algorithm

At this point in time, much of Figure 1 is currently implemented. The algorithm can successfully segment motion given a couple of constraints. (1) The background of the video must be sufficiently full of features. This means that the background must have regions of sufficiently different contrast levels to be able to select unique, common features between frames. (2) The drone camera must be relatively stable so that when the frames are wrapped, image artifacts are not a significant issue.

Motion Detection

In order to detect motion, it is first necessary to generate a representation of the background. We do so by implementing the following steps.

Frame transformation - In order to align the previous eight frames to the space of the current frame, it is necessary to generate a transformation matrix that can be used to warp the previous frame's space to that of the current frame. If we define the previous frame, f_{t-i} and the current frame, f_t, an ORB feature detector [4] can be used to find like features between these two frames. The n most similar features are then chosen to estimate a

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

transformation matrix from f_{t-i} to f_t, through the use of the RANSAC algorithm. This transformation matrix is then applied to wrap the previous frames into the space of the current frame.

Intersection - Once the previous eight frames are warped to match the current frame, the previous eight frames are then intersected as described in “An Efficient Approach for Object Detection and Tracking of Objects in a Video with Variable Background,” equation (8) (Ray & Chakraborty, 2017). Instead of using the formula listed for quantizing (see equation 7) the frames, we use equation (1) listed below instead.

$$H_i^Q = \{h_l^Q : h_l^Q = q_j * 40, \text{for } q_{j-1} < h_l^N \le q_j\} \tag{1}$$

Quantization here compresses the image so that all pixel values are between 1 and 40. This effectively thresholds what pixels are considered to be the same. Originally, Ray & Chakraborty (2017) had compressed the values to be between 1 and 10. Our change here allows for increased sensitivity to contrast changes. H_i is defined as the i^{th} historical transformed frame produced by applying the transformation matrix to f_{t-i}. h_l^Q represents the quantized value of the pixel at the l^{th} location. Finally, h_l^N is defined as $h_l^N \in \frac{H_i}{2^8-1}$.

The goal of the intersections is to remove parts of the image that are changing between the frames. The intersection operation leaves pixels that have changed between the two frames being compared with a value of 0 intensity.

Union - Once we have intersected each pair of eight frames, the seven intersected frames are then combined by a union operation as listed in “An Efficient Approach for Object Detection and Tracking of Objects in a Video with Variable Background, equation (9) [5].

The goal of the union operation is to fill the gaps created by the intersection operation with the true background. By filling in the pixel values with 0 intensity with values from other frames, we estimate the actual background.

Foreground Selection - Foreground selection is also implemented as defined in the same paper by Ray & Chakraborty. For this operation, section III, part C until equation (14) [5]. We deviate after this equation in an attempt to reduce noise.

The goal of foreground selection is to choose candidate pixels for the moving foreground. Current pixels selected include potential noise and thus we must compensate for this.

Noise reduction - In order to reduce noise, we first perform a morphological opening operation as defined below in equation (2)

$$A_{opened} = A_{foreground} \circ B_{opening} = (A_{foreground} \ominus B_{opening}) \oplus B_{opening} \tag{2}$$

where $\ominus$ and $\bigoplus$ represent erosion and dilation respectively. $A_{foreground}$ refers to the mask generated by the foreground selection step and $B_{opening}$ is an ellipse kernel of size 6x6 pixels. This is a very lossy operation, as such, it is necessary to dilate this mask to make the elements of the remaining motion mask significantly larger. This is done by the following operation.

$$A_{dilated} = A_{opened \oplus B_{dilation}} \qquad ()$$

where $B_{dilation}$ refers to another ellipse kernel with a size of 30x30 pixels. It is next necessary to combine the $A_{foreground}$ mask with $A_{dilated}$ mask to a less noisy mask. We do this by performing a boolean and on the two masks.

$$A_{reduced} = A_{dilated} \wedge A_{foreground} \qquad ()$$

Since the motion represented in $A_{dilated}$ has a much larger radius, it is expected that the regions covered by $A_{dilated}$ will encompass those of true motion in $A_{foreground}$.

Connecting the blobs - The remaining mask may not have connected blobs. We address this by performing the following operation (Ray & Chakraborty, 2017).

$$A_{motion} = (A_{reduced} \oplus B_{kernel}) \ominus B_{kernel}$$

where B_{kernel} is an ellipse kernel of size 12x12 pixels. A_{motion} represents the motion mask output by the motion detection part of our pipeline.

The left image of Figure 2 displays a frame from one of the videos that were used to produce test results, while the right image is the same rock formation from the ground to provide a sense of scale. This is one of the main sleeping sites of the baboons observed by UNBP.

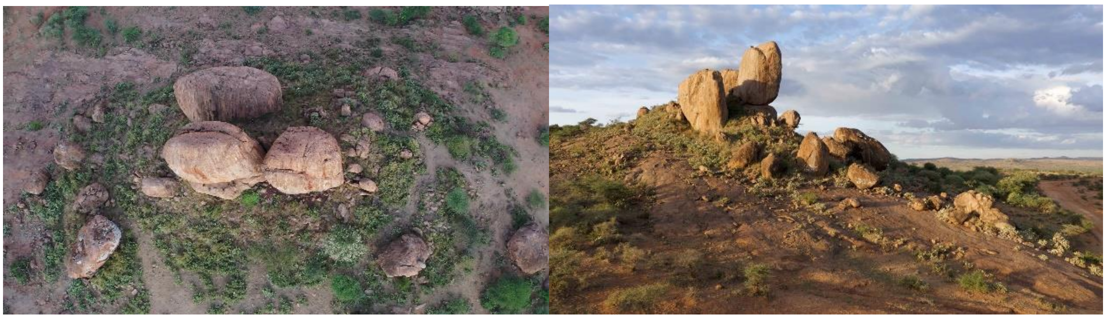


Figure 2. White Rock Sleeping Site. Right-Top View. Left – ground view

Figure 3 provides an example of flagged movement pixels as generated by the above algorithm. This image provides a small cop of a video taken from the same height as Figure 2. As can be seen, these moving baboons are correctly flagged as areas of interest.

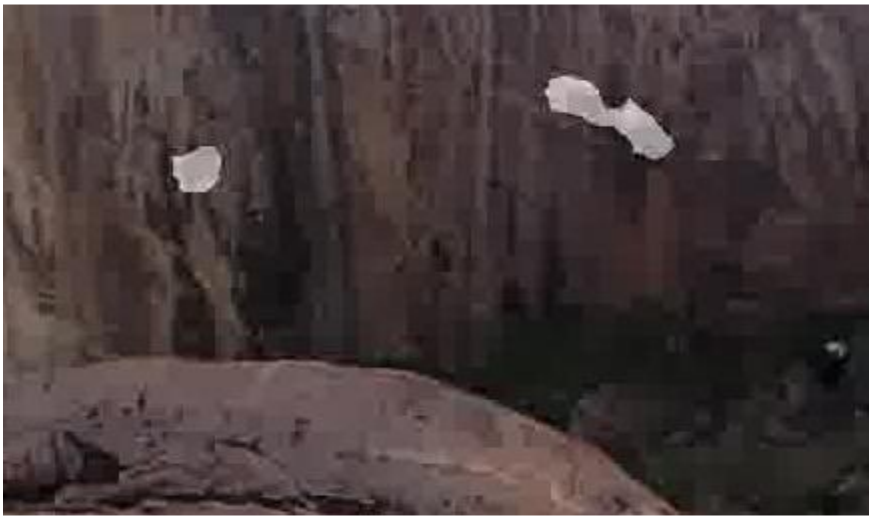


Figure 3. Baboon movement detected and flagged

Baboon Recognition Machine Learning Model

We currently do not have this part of the pipeline implemented. This stage of the pipeline will allow us to detect non-moving baboons and filter out additional noise.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Generate Bounding Boxes for Detected Baboons

After we implement the machine learning model, we will use its output to produce bounding boxes that define where the baboons are.

Discussion

The five main constraints for acquiring the footage are (1) the video must capture the entire baboon sleeping site, approximately 100 meters in length, in order to observe troop-level decision making, (2) the resolution of the video must be large enough so that each baboon is sufficiently large such that it can be resolved so that its individual movement can be registered and tracked, (3) the footage needs to be relatively stable, since it uses motion detection as a primary means of tracking, (4) the footage must be obtained in a way that does not disrupt the baboons' normal behavior, and (5) the footage needs to be sufficiently long to ensure that all necessary information is collected to understand the group-level decision.

Since the DJI MAVIC 2 PRO has about a 25 to 30 minute flight time, it is necessary for us to compensate for this flight time as it may not be sufficient to collect the necessary data. As a result, we will investigate the idea of flying a second drone near the end of the flight time of the first. We will design a hand-off concept to allow the information from the first video to flow into the second video.

In order to be able to fully understand the collective-decision made by the baboon troop, it is necessary to be able to identify individuals within the group. Our current plan for doing so is to pre-label the first frame of the video with individual identification. It will then be possible to propagate these individual identifiers through the video using the pipeline discussed in methods. If the tracking of the individuals begins to drift, it will be necessary to have the researchers relabel at the point where drift occurred, allowing a course-correct to happen mid-processing.

Conclusion

As we evolve the method proposed here, we hope to be able to better understand how baboons make collective decisions. Understanding how nonhuman primates make group-level decisions has the opportunity to help inform our knowledge of how other primates, including humans, make decisions. This knowledge allows us to improve the group-level decisions that we make, as well as the ones we instruct our technology to make.

As we continue to create more adaptive technology that integrates deeper into everyday life, mixed groups of humans and other intelligent agents—such as a fleet of autonomous vehicles mixed with human-operated vehicles in traffic—are expected to perform collaborative tasks. The understanding of emergent, group-level behavior is invaluable. Understanding how different entities within a group vary in skill, knowledge, and social influence decisions will be important to our future as a species.

Ethical Statement

The baboons are wild and except for an intervention to translocate them in 1984, we do not interact, touch or interfere with them. The Uaso Ngiro Baboon Project (previously the Gilgil Baboon Project) has been studying these baboons since 1972, long before formal ethical statements were required. We continue to act in accordance with ethical best practices.

Acknowledgments

The authors would like to thank: Neil Thomas for providing test data, Eric Lo for providing additional guidance on the algorithm, Allison Busa for initial experiments on machine learning processes, Charmaine Beluso for reviewing annotation tools used in testing, Charles Kinzel for his work improving performance, Davin Tjong for his work testing other avenues not discussed in this paper. The authors would also like to thank Corina Most for her useful feedback on the project and the other field staff for their support.

This work was funded in part by the REU Site Engineers for Exploration, supported by the NSF under Grant No. 1852403. Additional support was provided by the UCSD ECE and CSE departments and the UCSD Qualcomm Institute.

References

1. Strum, S. (2012). Darwin’s monkey: Why baboons can’t become human. *Yearbook of Physical Anthropology* **55**: 3-23.

2. Strandburg-Peshkin, A., Farine, D., Couzin, I., & Crofoot, M. (2015). Shared decision-making drives collective movement in wild baboons. *Science* **348:**1359-1361.

3. Hallck, L., Mangan, M., Webb, B., & Risse, B. (2020). Towards image-based animal tracking in natural environments using a freely moving camera. *Journal of Neuroscience Methods* **330.**

4. Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. *International Conference on Computer Vision*, (pp. 2564-2571). Barcelona, Spain.

5. Ray, K. S., & Chakraborty, S. (2017). An Efficient Approach for Object Detection and Tracking of Objects in a Video with Variable Background. arXiv:1706.02672

Assessing disinhibition behaviour in behavioural variant frontotemporal dementia patients using ecological, cognitive and anatomical tasks

Delphine Tanguy*[1,4], Valérie Godefroy*[1], Idil Sezer[1], Arabella Bouzigues[1], Carole Azuar[3], David Bendetowicz[1,2,3], Guilhem Carle[3], Armelle Rametti-Lacroux[1], Pierre Jannin[4], Xavier Morandi[4], Richard Levy[1,2,3], Bénédicte Batrancourt[1] and Raffaella 'Lara' Migliaccio[1,2]

1 Inserm U 1127, CNRS UMR 7225, Sorbonne Universités, UPMC Univ Paris 06, Institut du Cerveau et de la Moelle épinière, FRONTlab Paris, France. delphine.tanguy@icm-institute.org

2 AP-HP, Hôpital de la Salpêtrière, Institut de la Mémoire et de la Maladie d'Alzheimer, Paris, France.

3 AP-HP, Hôpital de la Salpêtrière, Behavioural Neuropsychiatry Unit, Paris, France.

4 Univ Rennes, CHU Rennes, Inserm, LTSI – UMR 1099, F-35000 Rennes, France

**Equal contribution*

Introduction

Behavioural variant frontotemporal dementia (bvFTD) is an early-onset neurodegenerative disease (the second most common form of dementia after Alzheimer's disease) associated with behavioural disturbances. These troubles result from a frontotemporal lobar neurodegeneration [1]. Patients suffering from the behavioural variant of FTD show a progressive deterioration of personality, social conduct and cognition, characterised by difficulties to modulate their behaviours. These changes include "positive" (hyperactivity) symptoms such as disinhibition, perseveration or compulsivity, and "negative" (hypoactivity) symptoms, such as apathy or inertia [2]. Most of these symptoms are difficult to recognise and poorly studied because of the lack of adapted clinical tools, misdiagnoses are thus frequent and FTD is often confused with other psychiatric pathologies [2]. Without definitive biomarkers, bvFTD diagnosis only relies on clinical diagnostic criteria, and in average, six years pass between the first symptom and a correct diagnosis.

In our study, we focused on the inhibition deficits. We can distinguish two types of inhibition:

- **T**he cognitive inhibition is defined as the ability to resist to an exogenous or endogenous interference, inhibit cognitive contents or processes previously activated and to suppress inappropriate or irrelevant responses [3].
- **T**he behavioural inhibition refers to the control of emotional and social behaviours in a social context, and the control of overt behaviours. It is the ability to adapt actions to environmental changes, to suppress impulsions to violate norms, to delay gratification and to control impulsive actions [4].

Both cognitive and behavioural disinhibition are reported in bvFTD patients and represent the core of current clinical criteria [2, 6]. Therefore, disinhibition has been identified as one of the major causes for caregiver distress with apathy [5]. However, its assessment, characterisation and neural correlates are still poorly known [7].

Tools currently used to measure behavioural disinhibition are often incomplete and inappropriate for bvFTD patients. It consists in standardised caregiver questionnaires containing subscales related to disinhibition. Caregiver are often asked to provide insight into behavioural changes of patients, but this presents a non-negligible bias. Besides, these tools are mostly old paper-pencil tests, which lack ecological validity. Thus, our main objective was to design a new tool for the measurement of this critical symptom of bvFTD, using a methodology allowing the recording and the quantification of behaviours in a more objective manner.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

The aim of our study is two-folded: first, to identify a signature of inhibition troubles on a behavioural and cognitive plan, characterising different bvFTD subtypes, in order to generate new clinical tools to assess these symptoms, and second, to study the neuronal networks associated with the different subtypes of disinhibition in bvFTD patients.

Method

Project

We studied behavioural and cognitive inhibition troubles in bvFTD using a population of 17 patients (Mini Mental State Evaluation >20 for a correct cognition) and 16 control subjects, following three approaches:

- A **behavioural approach** with an investigation reproducing a close-to-real-life situation (ecological task)

From 2015, the laboratory has developed a research program called ECOCAPTURE (http://clinicaltrials.gov/NCT02496312) which aims at defining and assessing more precisely the behavioural markers of apathy and disinhibition, in patients with neuropsychiatric conditions [8]. ECOCAPTURE is characterised by its original methodological approach using behavioural sensing (video, accelerometer sensor, eye-tracking glasses) under ecological conditions. ECOCAPTURE is used here to assess in a quantitative and objective way the disinhibition behaviours, in a “real-life” situation (waiting room).

The subject is in a fully furnished waiting room (chairs, sofa, dresser…) during 45 minutes, instructions are to get comfortable and to enjoy the room (see Figure 1). A lot of objects are available to pass the time (games, magazines, refreshments, food…). After 20 minutes, a questionnaire has to be filled, making the subject interacting with the environment. This semi-ecological situation is relevant to study social disinhibition and compulsivity as it can favour the generation of inappropriate behaviours, especially when participants forget that they are observed. They may for instance inappropriately interact with the experimenter or the room and its objects, show inappropriate reactions when they are asked to complete a task or reveal strange reactions to some unexpected events programmed in the scenario (such as music broadcasting). For details on the scenario, see Batrancourt et al., 2019 [8].

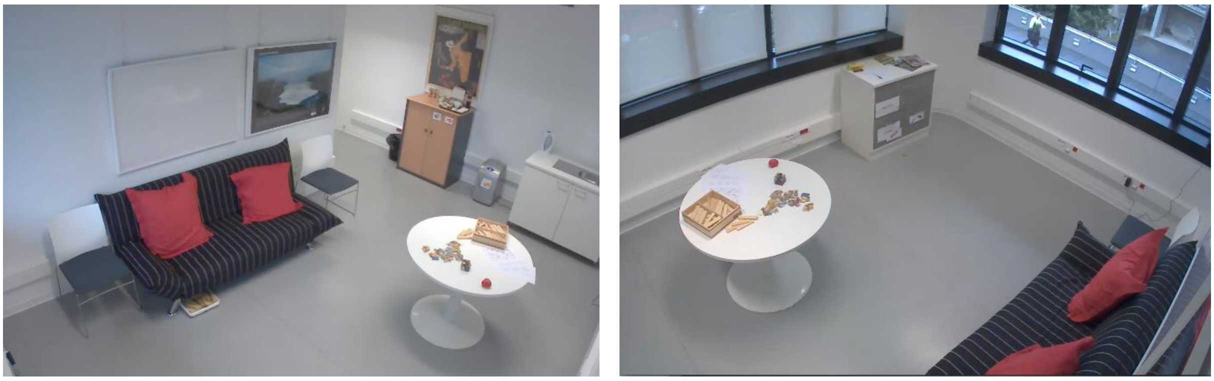


Figure 1. Ecological setting (PRISME platform, Brain and Spine Institute, Paris): view of the waiting-room (The Observer XT®, Noldus).

In the room, a video recording system allows us to see the subject in live and to encode and analyse some disinhibition behaviours.

Defining disinhibited behaviours is not an easy task, disinhibition, impulsivity and compulsivity are all concepts related to inhibition troubles. For some authors, impulsivity is a subcomponent of the disinhibition syndrome [2]. For others, disinhibition belongs to the concept of impulsive behaviours [9]. Finally, while compulsivity was mostly associated with harm-avoidance and impulsivity with risk-seeking, these two concepts share neuropsychological mechanisms involving dysfunctional inhibition of thoughts and behaviours. We decided here to study these three components of inhibition troubles.

We thus selected disinhibited behaviours potentially observable in the context of the ECOCAPTURE scenario according to the definitions of symptoms by Rascovsky et al. [2]. In particular, disinhibition includes socially inappropriate behaviours (such as verbal or physical aggression), loss of manners or decorum (such as inappropriate laughter) and impulsive, rash or careless actions (such as stealing). On the other hand, perseverative/compulsive behaviours are repetitive movements, compulsive or ritualistic behaviours (e.g. tapping, rocking, compulsive eating…) [2].

In this way, we retained and defined 20 behaviours related to social disinhibition (e.g. familiar or rude behaviour towards investigator, lack of manners or decorum) and to impulsivity/compulsivity (e.g. singing, utilisation behaviour, repetitive movements) to complete the ECOCAPTURE ethogram. See Table 1 for the complete list.

The video-based behavioural data are generated by a manual video annotation tool (The Observer XT®, Noldus), using the ethogram listing behavioural categories.
With all these data, the aim wass to retain some metrics leading to a differentiation in disinhibition between patients and control subjects and to differentiate disinhibition subtypes within patients.

Table 1 - Ethogram listing the 20 quantified behaviours and their definition

Behaviour label	Definition	Example
SOCIAL DISINHIBITION		
Rude behaviour toward investigator	Showing anger, hostility or aggressiveness towards the investigator	Yelling "Enter" with anger when the investigator is knocking several times at the door
Familiar behaviour towards investigator	Showing inappropriate familiarity towards the investigator	Speaking in colloquial language
Nudity[1]	Exposing inappropriate part of one's body	Removing one's pants
Inappropriate comfort-enhancing	Showing strange behaviours aiming at enhancing comfort while waiting	Taking off one's shoes and walking around shoeless in the room
Lack of decorum	Failing to respect norms of politeness of local culture	Yawning, sneezing or couching without putting hand on the mouth
Disregards for investigator[2]	Ignoring investigator in the room	Not answering questions
Inappropriate gesture or posture[2]	Showing strange gestures towards oneself or strange postures	Picking one's nose/teeth
Harsh handling of objects	Handling an object of the room without care, implying potential damage	Trying to break the locker box instead of searching for the key
IMPULSIVITY/ COMPULSIVITY		
Self-talking[2]	Speaking for oneself aloud when alone	Commenting on the environment

Behaviour label	Definition	Example
	in the room	when entering the room
Laughing[3]	Laughing alone in the room	Laughing at the sight of the locked box
Dancing	Dancing alone in the room	Doing a few dance steps
Singing[3]	Singing alone in the room	Singing "O Christmas Tree" without any reason
Use of profanity or swearing	Swearing or using obscene language alone in the room	Saying "Oh! Fuck!" when getting bored
Inappropriate use of objects[3]	Using an object of the room in an unconventional way, without taking account of the proper value of the object	Discarding the content of a beverage in the sink
Utilisation behaviour	Grabbing objects of the environment and starting the "appropriate" behaviour associated with it at an "inappropriate" time	Opening and closing the window without any real purpose
Inability to shift behaviour	Persevering in a useless action and showing inability to shift attention to change action	Keeping trying to open the tap unsuccessfully (no running water in the room)
Compulsive eating	Abnormal eating behaviour (eating large amounts of food and/or strange foods in the absence of hunger)	Eating sardines just after breakfast
Repetitive movements[2]	Repeating ritualistic or stereotyped behaviours	Rubbing hands
Compulsive room exit	Showing compulsive desire to exit the room (instead of waiting in it)	Persistently trying to exit the room
Inability to focus on guidelines	Inability to concentrate on and follow the instructions given by the investigator	Not keeping eye-tracking glasses

[1]Behaviour never observed
[2]Behaviours suppressed after reliability analysis (similar loadings on several components)
[3]Behaviour initially categorized as compulsion but loading on social disinhibition component

Two different examiners coded the video to quantify the behaviours and the intercoder reliability was assessed through the calculation of the Intraclass Correlation Coefficient (ICC). All the calculated ICC were between 0.80 and 1, indicating a very good reliability.

- A **cognitive approach** via a neuropsychological evaluation (cognitive task)

Cognitive inhibition is assessed using a francophone version of the Hayling Sentence Completion Test (HSCT), a classical tool evaluating cognitive inhibition with completion of sentences [10]. As rapidly as possible,

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

participants are asked to complete 15 sentences using the appropriate word (automatic condition, part A), and 15 sentences using a completely unconnected word (inhibition condition, part B). Three scores will be available: completion time in part A (time A), completion time in part B (time B) and errors score in part B (HSCT score). An augmentation of completion time in part B and an augmentation of errors score are expected with bvFTD patients, indicative of inhibition troubles.

The mini-Social cognition & Emotional Assessment (mini-SEA) orbitofrontal battery assesses affective and emotional functions depending on the limbic system. The whole battery is composed from two subsets: a reduced version of the Faux-Pas test, assessing theory of mind deficits, and a facial emotions recognition test in which participants must identified which emotion is being expressed [11]. A lower score is expected in patients than in control subjects.

In 2018, 13 bvFTD and 12 controls were explored combining behavioural and cognitive approaches, and results were already discussed in a congress. A multivariate power analysis showed that nine subjects per group were necessary and sufficient to get valid significant results.

- An **anatomical approach** using Magnetic Resonance Imaging (MRI) analyses (anatomical task)

Brain MRI protocol includes a 3D T1 and resting state fMRI allowing the study of structural and functional abnormalities respectively. We used SPM12 (Statistical Parametric Mapping) running under ®MATLAB R2015b for Voxel-Based Morphometry (VBM) analyses. A larger group of healthy controls was used for the VBM. The following set of contrasts was thus performed implementing a two-sample *t*-test: bvFTD-G1 vs HC, bvFTD-G2 vs. HC, bvFTD-G1 vs bvFTD-G2. A significance threshold of $p \leq 0.05$ corrected for multiple comparisons (family-wise error) was accepted when comparing the patients vs. controls and of $p \leq 0.001$ uncorrected when comparing the patients each other's.

Comparison between patients and control subjects related to behavioural and cognitive outcomes should lead to neural correlates of disinhibition.

These three approaches are complementary and provide a cognitive and behavioural signature (composite score) of different inhibition troubles, associated with different frontotemporal atrophy patterns, leading to a new and objective diagnostic tool. This triple assessment will allow us to define the whole profile of disinhibition in bvFTD patients. Furthermore, a validation study will be necessary for a clinical utilisation.

Statistical analyses

Regarding the behavioural task, we extracted principal components through a principal component analysis (PCA) on the behavioural metrics. We compared mean scores on the extracted behavioural dimensions between bvFTD patients and control subjects through Wilcoxon tests. A clustering approach based on the components discriminating patients and control subjects isolated two subgroups of patients. These two subgroups showed distinct patterns of neurocognitive functioning (Hayling test and mini-SEA), and distinct frontotemporal atrophy (analyses performed with *SPM12 Matlab*). Cognitive scores were compared between HC and the subgroups of bvFTD patients through a Kruskal-Wallis test followed by post-hoc comparisons with Wilcoxon tests.

Results

PCA and behavioral comparisons between patients and HC

Considering results of the PCA (eigenvalues and percentages of total and common variance explained), we extracted three principal components. After performing a reliability analysis, we suppressed four behaviors because they showed similar loadings on several components, so that all extracted components presented high internal consistency (Cronbach's alphas >.70). One behavioral item (nudity) was never observed and therefore removed. As table 2 shows, we thus obtained 15 behavior variables loading on three behavioral components

(PC1, PC2 and PC3) with eigenvalues greater than 1, accounting for 79% of the total variability. PC1 included behaviors initially categorized as impulsivity/compulsivity whereas PC2 and PC3 mostly contained behaviors related to social disinhibition (except for inappropriate use of objects, singing and laughing) (see table 1). Comparisons of mean scores for PC1, PC2 and PC3 revealed that bvFTD patients present significantly higher scores than HC but only on PC1 (W= 40; p=1.62e-04) and PC2 (W=82; p=0.014). To make reading easier, PC1 and PC2 are hereafter labelled according to their conceptual content as "Compulsivity" and "Social disinhibition" respectively.

Table 2: Results of the PCA[1]

Behaviour variables	Factor loadings		
	PC1	PC2	PC3
Utilisation behaviour	**0.89**	0.06	0.05
Inability to shift behaviour	**0.91**	0.00	0.00
Compulsive eating	**0.89**	0.06	0.05
Dancing	**0.81**	0.51	0.05
Inability to focus on guidelines	**0.66**	-0.10	-0.09
Compulsive room exit	**0.70**	0.01	0.00
Inappropriate use of object	0.53	**0.84**	0.01
Harsh handling of object	0.36	**0.91**	-0.02
Singing	-0.07	**0.95**	-0.07
Rude behaviour towards investigator	-0.21	**0.67**	0.32
Familiar behaviour towards investigator	-0.05	**0.96**	0.05
Laughing	0.37	-0.02	**0.87**
Use of profanity or swearing	-0.06	-0.05	**0.94**
Lack of decorum	-0.17	0.42	**0.57**
Inappropriate comfort-enhancing	-0.06	-0.03	**0.97**

[1]Values are the standardized factor loadings of the PCA for the 15 behaviours (selected after reliability analysis); Coefficients in bold denote the behaviours corresponding to each of the 3 principal components (PC1, PC2 and PC3); N=33 with 16 HC and 17 bvFTD.

Clustering and multi-approach comparisons between subgroups of patients and HC

For the clustering approach, we used only the two discriminatory dimensions of so-called "Compulsivity" and "Social disinhibition". Based on the results of the hierarchical cluster analysis, we identified two subgroups of bvFTD patients: *bvFTD-G1* (n=6) and *bvFTD-G2* (n=9). Two patients were "unclassifiable" because of strong behavioural dissimilarities and were excluded from the study. K-means analysis allowed to confirm the structure of the two subgroups previously identified. After checking that *bvFTD-G1* and *bvFTD-G2* were equivalent in terms of age, education level and reported duration of disease, we compared the two subgroups and HC on behavioural, cognitive and neuroimaging data.

Behavioural results

We observed a significant effect of group on the two behavioural dimensions of Compulsivity (p< .001) and Social disinhibition (p< .001). Compulsivity was significantly higher in bvFTD-G2 compared with bvFTD-G1 (p= .0032) and HC group (p< .001), but it is similar in bvFTD-G1 and HC. Social disinhibition was significantly higher in bvFTD-G1 compared with bvFTD-G2 (p< .001) and HC group (p< .001), but it was similar in bvFTD-G2 and HC. BvFTD-G1 patients are therefore characterized by more socially disinhibited behaviours whereas bvFTD-G2 patients are more prone to impulsive and compulsive behaviours.

Cognitive results

By looking at the two groups of patients, the score of errors on the Hayling test was significantly higher in bvFTD-G2 compared with bvFTD-G1 (p= .0032) and HC (p = .0025) and in bvFTD-G1 compared with HC (p= .0055) (Figure 2.a). Compared with HC, the mini-SEA score was lower for bvFTD-G2 only (p< .001). There was no significant difference between bvFTD-G1 and bvFTD-G2 for mini-SEA. The two subgroups are thus different in terms of cognitive inhibition but not for social cognition capacity.

Neuroimaging results

Figure 2.b shows the results of VBM analysis. Two main patterns of grey matter atrophy were identified for *bvFTD-G1* and *bvFTD-G2*. *BvFTD-G1* compared with HC, showed atrophy in the left amygdala, lingual gyrus, postcentral gyrus and thalamus. *BvFTD-G2* revealed a huge and extensive pattern of atrophy in bilateral frontal and temporal lobes classically described in literature. No difference was found in the contrast of *bvFTD-G1* and *bvFTD-G2*.

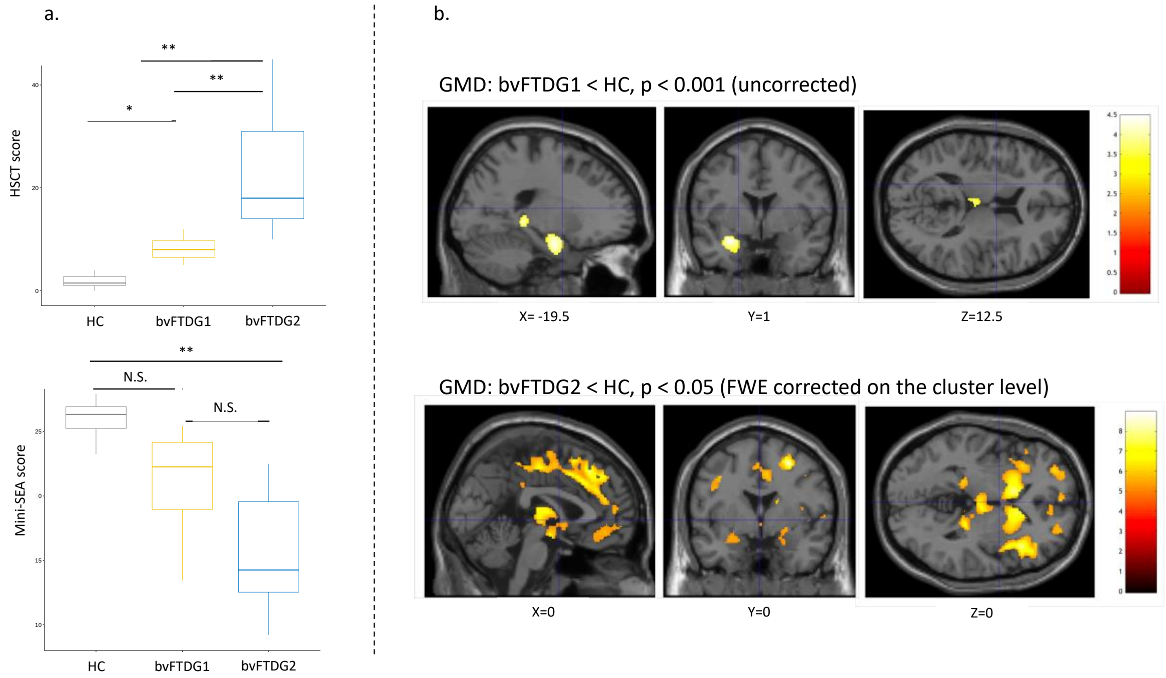


Figure 2: Neurocognitive and neuroanatomical comparisons between bvFTD subgroups and controls.
a. Group comparison for the HSCT score measuring cognitive inhibition (upper part) and for the mini-SEA score assessing social cognition capacity (lower part). **p<0.01; *p<0.5; N.S.=non-significant. *bvFTD-G1*: N=4; *bvFTD-G2*: N=5; *HC*: N=10.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

b. Group comparisons for gray matter density (GMD) in *bvFTD-G1* vs *HC* without correction (upper part) and in *bvFTD-G2* vs *HC* with family wise error (FWE) correction (lower part). T-score scale is shown on the right. Coordinates in MNI space. Left side of the brain is shown on the left. *bvFTD-G1*: N=4; *bvFTD-G2*: N=5; *HC*: N=40.

Discussion

In this study, we investigated the assessment of two symptoms related to inhibition troubles in bvFTD (social disinhibition and impulsivity/compulsivity) in order to make it more efficient and useful for bvFTD diagnosis, prognosis and treatment.

We extracted three principal components from the 20 initial behavioural variables retained in the study-ethogram: one was conceptually related to impulsivity/compulsivity and the two others to social disinhibition. Three behaviours (i.e. laughing, singing, inappropriate use of objects) theoretically classified as impulsive or compulsive behaviours loaded on social disinhibition dimensions, suggesting that these behaviours may also be interpreted as inappropriate social behaviours.

The clustering allowed isolation of two subgroups of bvFTD patients (*bvFTD-G1* and *bvFTD-G2*) with distinguished behavioural, neurocognitive and neuroanatomical patterns. *BvFTD-G1* patients are behaviourally characterized by high social disinhibition. Compared to controls, they show little cognitive impairment and a brain atrophy localized in few regions including the medial anterior and posterior temporal lobe (e.g., amygdala and thalamus) that are compatible with their behavioural profile. Indeed, social disinhibition seems to result from an alteration of social cognitive processes as the Theory of Mind, which is associated with a network involving the amygdala [7]. *BvFTD-G2* patients are behaviourally characterized by compulsivity. They also present higher cognitive inhibition troubles than *bvFTD-G1* and a larger atrophy of the frontotemporal regions as such the orbitofrontal cortex, medial frontal, and temporal areas. A previous study in bvFTD also showed correlations between orbitofrontal hypometabolism and stereotypic responses with indifference to rules [6].

BvFTD-G1 and *bvFTD-G2* therefore seem to correspond to more or less severe forms of structural and functional brain damage and compulsive behaviours may be considered as an indicator of higher severity of bvFTD. However, comparisons of individual factors (duration of illness, demographic and genetic data) between the two subgroups could not account for their distinct levels of impairment, as already noticed by Zamboni et al. who found no correlation between severity of disinhibition and disease duration [12]. Future research should dig into the etiological bases and in particular environmental factors associated to these behavioural subtypes of bvFTD.

Conclusion

For the first time, behavioural measurement under close-to-real-life situation has been used to complete the clinical description of bvFTD symptoms. Based on the association of classical paper-and-pencil test and this behavioural measurement, we isolated two subgroups of bvFTD patients with distinct patterns of neurocognitive functioning and frontotemporal atrophy. Our results therefore suggest the possibility to detect two forms of behavioural expression in bvFTD even just through simple observation under ecological conditions.

For a clinical investigation, this type of ecological behavioural observation could be used to reach a more reliable and accurate evaluation of bvFTD symptoms, allowing a better diagnostic and treatment for patients.

Ethical statement

This study is part of clinical trial C16-87 sponsored by INSERM. It was granted approval by the local Ethics Committee, or “Comité de Protection des Personnes”, on 17/05/2017 and registered in a public clinical trial registry (clinicaltrials.gov: NCT03272230, NCT02496312). All study participants gave their written informed consent to participate, in line with French ethical guidelines.

Funding

Study funded by the program “Investissements d’avenir” ANR-10-IAIHU-06; the foundation “Fondation pour la recherche médicale”, FRM DEQ20150331725, frm.org; and the ENEDIS company, enedis.fr

References

1. 1. Massimo et al. (2009). Neuroanatomy of apathy and disinhibition in frontotemporal lobar degeneration. *Dement Geriatr Cogn Disord*, 27:96–104.

2. 2. Rascovsky et al. (2011). Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain, 134:2456–2477.

3. 3. Wilson, S.P., and Kipp, K. (1998). The Development of Efficient Inhibition: Evidence from Directed-Forgetting Tasks. Developmental Review 18, 86–123.

4. 4. Harnishfeger, K.K. (1995). Theories, Definitions, and Research Evidence. 30.

5. 5. O'Connor et al. (2017). Behavioral-variant frontotemporal dementia: Distinct phenotypes with unique functional profiles. Neurology, 89(6), 570-577.

6. 6. Peters et al. (2006). Orbitofrontal dysfunction related to both apathy and disinhibition in frontotemporal dementia. Dement Geriatr Cogn Disord, 21:373–379.

7. 7. Paholpak et al. (2016). Person-Based Versus Generalized Impulsivity Disinhibition in Frontotemporal Dementia and Alzheimer Disease. Journal of geriatric psychiatry and neurology, 29(6), 344-351.

8. 8. Batrancourt B, Lecouturier K, Ferrand-Verdejo J, Guillemot V, Azuar C, Bendetowicz D, Migliaccio R, Rametti-Lacroux A, Dubois and Levy R (2019) Exploration Deficits Under Ecological Conditions as a Marker of Apathy in Frontotemporal Dementia. Frontiers in Neurology, 10:941.

9. 9. Rochat, L., Billieux, J., Juillerat Van der Linden, A.-C., Annoni, J.-M., Zekry, D., Gold, G., and Van der Linden, M. (2013). A multidimensional approach to impulsivity changes in mild Alzheimer’s disease and control participants: cognitive correlates. Cortex 49, 90–100.

10. 10. Burgess, P.W., and Shallice, T. (1996). Response suppression, initiation and strategy use following frontal lobe lesions. Neuropsychologia 34, 263–272.

11. 11. Funkiewiez, A., Bertoux, M., de Souza, L.C., Lévy, R., and Dubois, B. (2012). The SEA (Social cognition and Emotional Assessment): a clinical neuropsychological tool for early diagnosis of frontal variant of frontotemporal lobar degeneration. Neuropsychology 26, 81–90.

12. 12. Zamboni, G., Huey, E.D., Krueger, F., Nichelli, P.F., and Grafman, J. (2008). Apathy and disinhibition in frontotemporal dementia. Neurology 71, 736–742.

Applying Entropy to Understand Drivers' Uncertainty during Car-following

Wei Lyu[1,2], Rafael C. Gonçalves[2], Fu Guo[1], Guilhermina A. Torrão[2], Vishnu Radhakrishnan[2], Pablo Puente Guillen[3], Tyron L. Louw[2] and Natasha Merat[2]

1 School of Business Administration, Northeastern University, Shenyang, China. lvweineu@gmail.com, fguo@mail.neu.edu.cn

2 Institute for Transport Studies, University of Leeds, Leeds, UK. tswl@leeds.ac.uk, tsrg@leeds.ac.uk, gtorrao@gmail.com, mn16vr@leeds.ac.uk, t.l.louw@leeds.ac.uk, n.merat@its.leeds.ac.uk

3 Toyota Motor Europe NV/SA, Hoge wei 33, 1930 Zaventem, BE. pablo.puenteg@gmail.com

Abstract

As one of the main processes in most microscopic simulation models and modern traffic flow theory, car-following has drawn huge academic attention from the engineering and physiological domains. However, given the inherently uncertain and unpredictable nature of human behaviour, car-following models have always faced challenges in capturing drivers' behaviour accurately and objectively. Therefore, to better capture drivers' uncertainty in car-following, this paper contrasts four different entropy algorithms (Shannon Entropy, Steering Wheel Entropy, Approximate Entropy and Sample Entropy) as a novel measure, based on time headway data during car following. Results showed that not all the entropy measures tested are suitable for the context of car-following, especially when it comes to measuring uncertainty in time headway data. Approximate and Sample entropy algorithms in a moving time window seem to be the most appropriate, as they consider drivers' prior time headway data as a factor in the perceived uncertainty. This paper contributes to the fields of microsimulation and human factors, as it demonstrates how entropy can be a precise and replicable measure of changes in behaviour, as well as anomalies in patterns of time headway data in car-following situations.

Introduction

The concept of car-following was first introduced by Pipes [1], and can be defined as '*the decision of the driver to follow the preceding vehicle efficiently and safely*'. Here, both efficiency and safety can be described in terms of time and distance between the preceding and following vehicle. Time headway (THW) is one way to characterise the safety margin in car-following, which is the extent to which the following vehicle is susceptible to unpredictable decelerations of the preceding vehicle [2]. Car-following models aim to explain the interplay between phenomena at the microscopic level of individual driver behaviour and the macroscopic level of traffic flow. Over the past few decades, there have been numerous attempts to apply Newtonian-based models from the engineering domain, to approximate and interpret car-following (for a full review see [3]). However, one of the issues with car-following models is that human behaviour is inherently random and unpredictable [4], which is challenging to capture in classical mathematical models. In an attempt to further improve the accuracy of these models, there was a trend to incorporate psychological factors, such as motivation and attitude [3]. However, models which included psychological variables tended to be no better at explaining drivers' car-following behaviour, partly due to the substantial intra-driver variability, or 'uncertainty', in terms of changes in driving behaviour or strategies in different driving stages or scenarios, and inter-driver differences in terms of risk-taking behaviour and demographics [5]. Among these factors, the uncertainty of drivers' behaviour, as a ubiquitously inherent nature of drivers, is claimed to be taken into account in characterising car-following behaviour [6]. Therefore, there are clearly challenges with incorporating psychological metrics into these models. However, there is still a need to capture the 'uncertainty' element in drivers' behaviour. An alternative approach is to analyse the patterns of vehicle-based measures, such as time headway, to see how drivers' uncertainty has changed during the car-following process. Entropy theory is a possible candidate to illustrate 'uncertainty' in drivers' car-following behaviour.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Information entropy, as proposed by Shannon, provides a mathematical expression of the amount of uncertainty associated with a variable *X*, where the 'uncertainty' is the summation of all possible outcomes, where the outcomes are unknown. In the context of car-following, uncertainty can refer to the variability in the patterns of fluctuation in a driver's relative position to a lead vehicle. Mathematically, the measure of information entropy associated with each possible data value is the negative logarithm of the probability mass function for the value. According to Shannon's (1948) definition, events with high or low probability with p of approximately 0 or 1, will not contribute substantially to the final entropy value. By contrast, a uniform event where *p* equals approximately 0.5, will result in a much higher final entropy value. It is worth mentioning that the entropy *H* (expressed below) is a function of the probability distribution $\{p_1, p_2, \cdots\}$ rather than a function of values or statistical indicators of the original series $\{x_1, x_2, \ldots\}$. In this way, it can be concluded that the less regular the original series, the higher the uncertainty and the higher the entropy.

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

Based on the entropy concept by Shannon, researchers have proposed various modifications of entropy and applied them to time series data in different fields. Nakayama et al. [7], for example, introduced Steering Wheel Entropy based on prediction error, to quantify drivers' steering behaviour while performing non-driving-related-tasks (NDRT). Steering Wheel Entropy yielded significance where classical statistical methods failed [8]. In addition, many researchers have adopted entropy-based methods to detect abnormal changes of drivers' physiological signals (such as the ones provided by electroencephalogram and electrocardiogram tests) in different driving experiments, especially drowsy driving [9], fatigue driving [10] and distracted driving [11]. These studies demonstrate the usefulness of using entropy in driving-related research to quantitatively measure randomness and uncertainty in time series data. Therefore, it may be helpful here in the analysis of vehicle metrics to understand underlying changes in drivers' car-following behaviour.

The aim of the paper is to investigate the suitability of different entropy algorithms to understand drivers' uncertainty in their car-following behaviour. To the best of the authors' knowledge, this is the first attempt to use entropy to understand drivers' uncertainty in a car-following context. This approach may provide a new perspective to measure and understand drivers' car-following behaviour.

Methods

In this section, four entropy algorithms will be introduced, including Shannon Entropy, Steering Wheel Entropy, Approximate Entropy and Sample Entropy. As the theoretical foundation of many other entropy algorithms, the concept proposed by Shannon is easy to understand and implement. Based on the framework of Shannon Entropy, Nakayama (1999) first introduced Steering Wheel Entropy in driving context to measure drivers' smoothness under different non-driving-related-tasks, and this entropy algorithm proved effective in detecting drivers' behavioural changes from the steering wheel angle. In addition, numerous authors have shown the usefulness of Approximate Entropy and Sample Entropy to show uncertainty and irregularity in time series data [12, 13]. The aim of this paper is to explore the use of entropy to capture drivers' behavioural uncertainty from time headway data series in the context of car-following, so we have chosen these four entropy algorithms as they were each designed to capture different situational-contexts.

Shannon Entropy. Shannon (1948) defined entropy as the negative logarithm of the probability mass function for a particular value. Once we obtain a time series (for example, time headway data), we can plot the histogram of the original data and compute the frequency of data points in each bin of the histogram. The frequency can serve as the probability in the entropy formula. Thus, we can calculate the entropy value of the original data series without having prior knowledge or other statistical characteristics of the raw data.

Steering Wheel Entropy. Steering Wheel Entropy was first proposed by Nakayama et al. [7]. It was based on the assumption that in free driving, a driver tends to control the steering wheel smoothly and predictably because of the anticipatory nature of preview control (i.e., a driver's continuous steering wheel corrections before entering a

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

curve). In consideration of the validity of this entropy in quantitatively measuring drivers' lateral control of the vehicle, this paper uses it as a potential approach to assessing drivers' longitudinal control in car-following. Mathematically, this algorithm calculates the entropy value based on the prediction errors between prediction value and real value of the steering wheel angle time series. The second-order Taylor expansion was used to obtain the predicted steering angle. The prediction error $e(n)$ is defined as the difference between $\theta(n)$ and $\theta_p(n)$ (see Figure 1a). The 90th percentile value α is calculated from the frequency distribution of the computed prediction errors, which is then used to divide the distribution of $e(n)$ is into nine bins, as shown in Figure 1b. The proportion of prediction errors falling into each bin is computed and the steering entropy value *Hp* is calculated using the Shannon Entropy formula, while a log base is changed to 9 to assure the final entropy value falls between 0 and 1. As can be seen from Figure 1b, driving with NDRT results in a broader frequency distribution and a consequently higher *Hp* value.

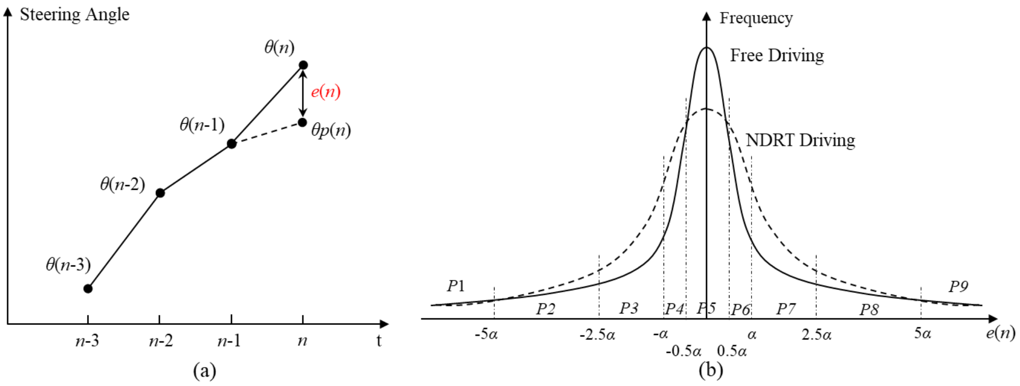


Figure 1 Steering Wheel Entropy (Nakayama et al., 1999), (a) diagram of the prediction error $e(n)$, (b) frequency distribution of the prediction error in driving with/without NDRT.

Approximate entropy (ApEn). ApEn measures the logarithmic probability that nearby pattern runs remain close in the next incremental comparison [14]. If there is a greater likelihood of the data remaining close and similar to the next incremental comparison, which would indicate regularity, it will yield lower ApEn values and vice versa. In the ApEn algorithm, a pair of parameters (m, r) are set as input. More specifically, m is the length of the template (length of the window of the different vector comparisons), and r is a noise filter (superposition of noise much smaller in magnitude than r barely affects the calculation). ApEn (m, r, N) measures the logarithmic frequency with which blocks of longitude m that are close together stay together for the next position, if possible, with the same number of observations N due to the bias mentioned below. Delgado-Bonal & Marshak provide a comprehensive tutorial on how ApEn is defined, calculated and applied.

Sample Entropy (SampEn). In the calculation of ApEn, there are two important limitations due to its bias. The first is that the ApEn result with different r may be different, and the second is that ApEn is influenced by the length of the data series. To eliminate the potential bias of ApEn, Richman and Moorman (2000) proposed SampEn as an updated version which does not involve self-counting. In other words, in the ApEn algorithm, the comparison vector $x(i)$ counts itself to avoid the presence of log0, which will present when there are no similar patterns. SampEn (m,r,N) is the negative value of the logarithm of the conditional probability that two similar sequences of m points remain similar at the next point m+1, counting each vector over all the other vectors except on itself. The calculation of the SampEn is highly similar to ApEn, and the main differences lie in whether the sum of all template vectors is inside or outside other logarithms, as can be seen from the below formula. For the detailed calculation and illustration, Delgado-Bonal & Marshak (2019) provided a comprehensive tutorial for both ApEn and SampEn.

$$ApEn(m,r,N) \square -\frac{1}{N-m}\sum_{i=1}^{N-m}\log\frac{\sum_{j=1}^{N-m}\left[number\ of\ times\ that\ d\left[\left|x_{m+1}(j)-x_{m+1}(i)\right|\right]<r\right]}{\sum_{j=1}^{N-m}\left[number\ of\ times\ that\ d\left[\left|x_{m}(j)-x_{m}(i)\right|\right]<r\right]}$$

$$SampEn(m,r,N) = \text{-log}\frac{\sum_{i=1}^{N-m}\sum_{j=1,j\neq i}^{N-m}\left[number\ of\ times\ that\ d\left[\left|x_{m+1}(j)-x_{m+1}(i)\right|\right]<r\right]}{\sum_{i=1}^{N-m}\sum_{j=1,j\neq i}^{N-m}\left[number\ of\ times\ that\ d\left[\left|x_{m}(j)-x_{m}(i)\right|\right]<r\right]}$$

Application & Discussion

In this section, the four entropy algorithms will be applied to the time headway data in a car-following driving scenario. Data used here is from an experiment which aimed to assess changes in driver behaviour after exposure to automation in car-following situations. In this experiment, the participant was instructed to resume control of the vehicle from an automated driving system, and then follow the preceding vehicle until the end of the trail. The lead vehicle maintained a constant speed of 40 mph (64.4 km/h). In this paper, the time headway data of one participant is selected to serve as an example for the application of the four algorithms discussed in the previous section. The data used here include 200 s of time headway data, starting at the point the participant resumes manual control from the automated driving systems. The sample rate of driving data collection is 60 frames per second, which resulted in 12 000 time headway data points.

Entropy in a fixed time window

If one wants to measure how the variation of time headway over time affects the progression of the level of entropy in a given stream of data, it is feasible to assume that using a fixed accumulated time window might be a possible option. The advantages of this approach is that by comparing the difference of the accumulated entropy values from the beginning of the series to a certain point in the time series (e.g. from 0 to 10s) and its previous iterations (from 0 to 8; from 0 to 6, etc.) it is possible to see how the entropy in time headway progresses over time. As shown in Figure 2, all the four entropy algorithms are sensitive to the changes in time headway data at the beginning when the computed data series is relatively short. However, after 90 seconds after take-over, values from all the entropy algorithms tend to be flat, ignoring the change of time headway data. These entropy algorithms are somewhat dependent on the length of the original data. It appears that when the computed time headway data series is long enough (for example, more than 90 seconds in this case), a small incremental time window (2 seconds in this case) will bring little additional information and influence to the existing data. This may explain the absence of fluctuations in all four entropy measures. One possible explanation for this phenomenon is that all entropy algorithms are based on previous observations in the same dataset. That being said, the larger is the dataset, the less likely a new incremental piece of data would affect the overall distribution.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

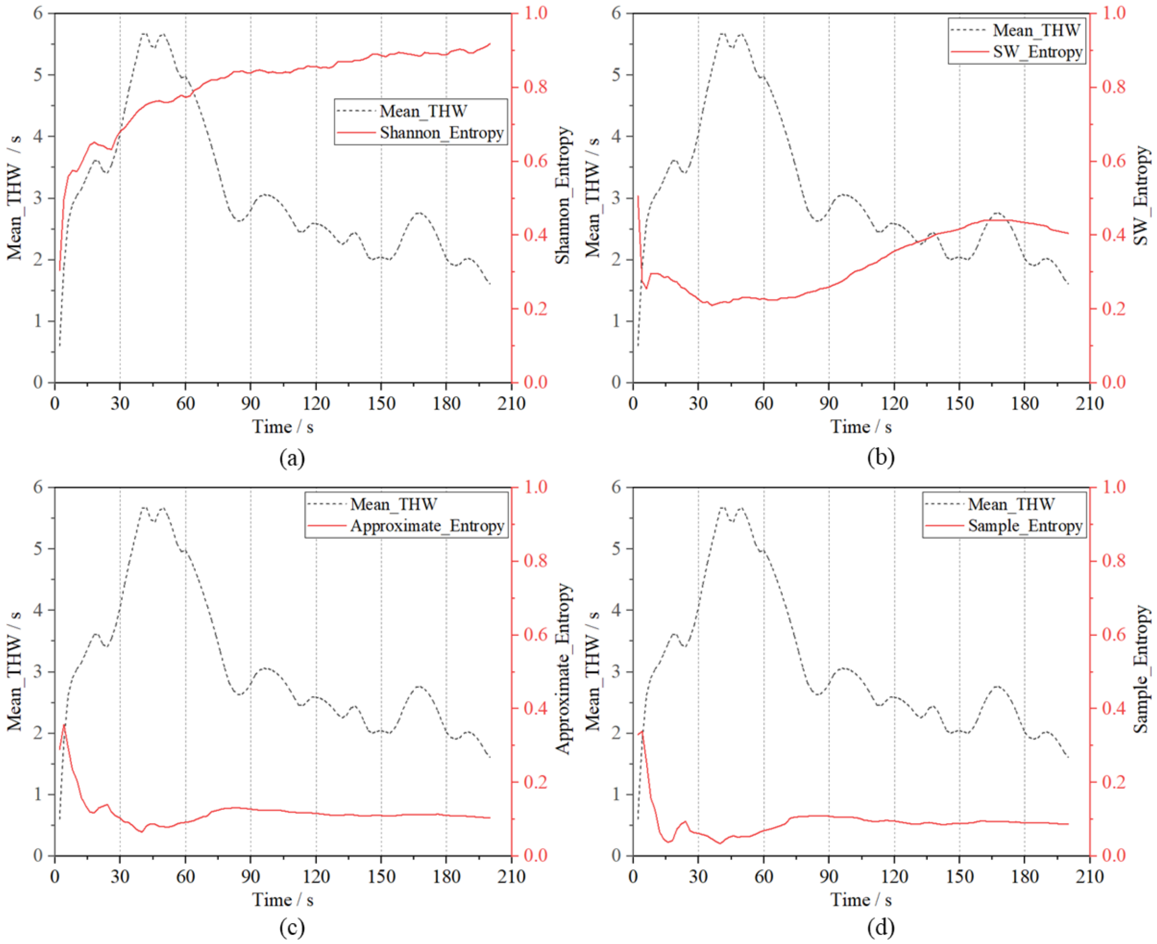


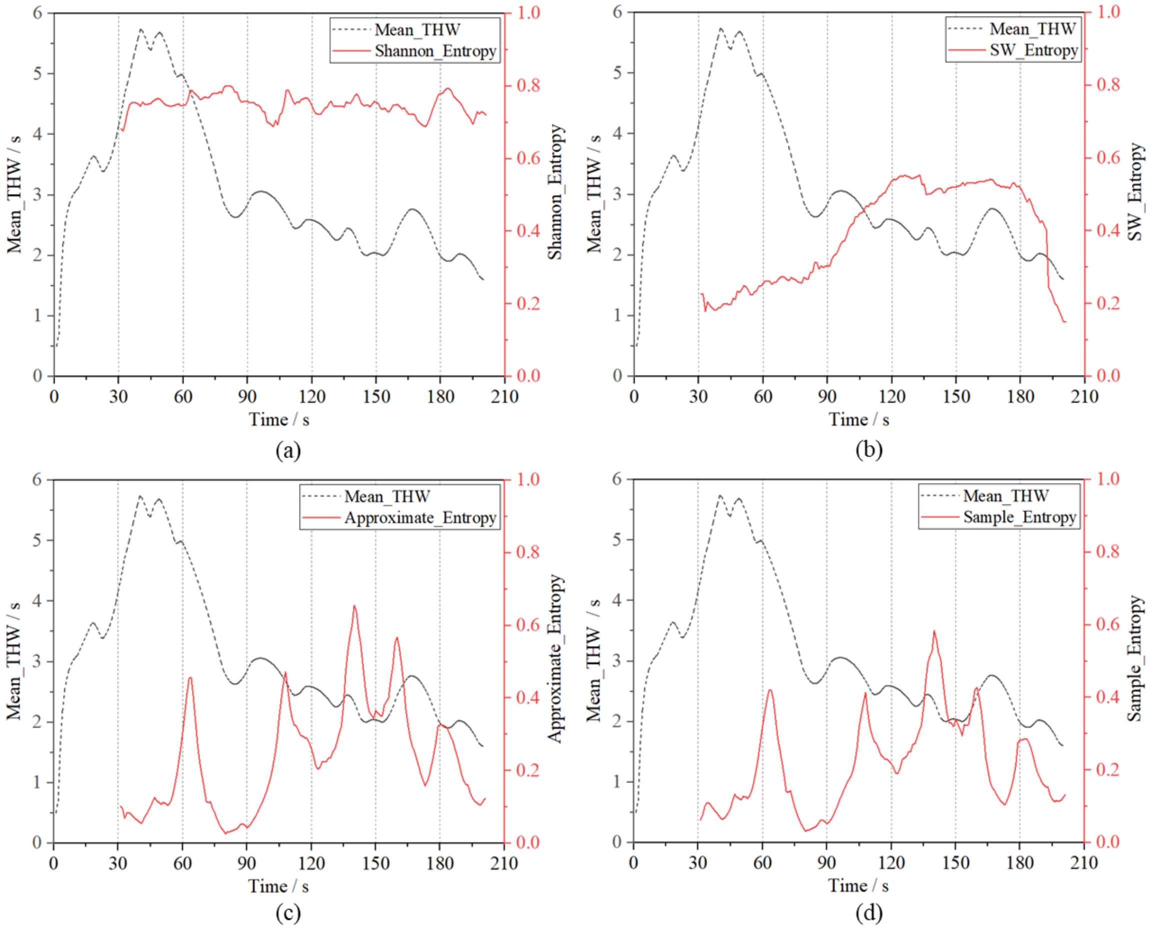
Figure 2 Mean time headway and corresponding entropy of different algorithms in accumulated fixed time window, (a) Shannon Entropy, (b) Steering Wheel Entropy, (c) Approximate Entropy, (d) Sample Entropy. Note: To improve readability, the values of ApEn and SampEn are multiplied by 40.

Entropy in a moving time window

Instead of a fixed time window, a moving window strategy can also be used to detect drivers' behaviour in a comparable length of each segmentation. As this paper focuses on the measurement of drivers' uncertainty in car-following, the fixed time window strategy would assume that the uncertainty or changes in driver behaviour are based on their previous actions during the whole period of car-following, as the calculation involves all the driver's prior time headway data. This approach is unlikely to explain drivers' real behaviour, as control of a vehicle on a 'manoeuvring level' is mainly based on constantly-updated information in a short period (seconds), and not based on long-term information [15]. That being said, shorter and moving time windows would theoretically be more suitable, as it accounts just for the relevant information to the driver when it comes to changes in their behaviour. Figure 3 presents the mean time headway and its corresponding entropy of different algorithms with a moving time window of 30 s and a step size of 1 second. Based on the results shown in Figure 3, overall, the moving window strategy seems more appropriate to distinguishing the entropy of the time headway data, as compared to the fixed window strategy.

As a direct implementation of the information entropy, Shannon's Entropy is easy to understand and implement, and it can reveal the randomness to a certain degree. However, according to its definition, Shannon entropy is highly dependent on the distribution of the time series and ignores the order of the values within the series, which makes it barely possible to detect patterns of variation. This can be seen in Figure 3a, where the above-mentioned differences in time headway before and after stabilisation are not reflected in the entropy values. As a comparison, Steering Wheel Entropy (SW entropy) inverts the logic, since it does not handle the raw data directly but tries to calculate the predicted values from the raw data. From Figure 3b, it can be seen that immediately after regaining control, the time headway increases and then decreases linearly, making the pattern less uncertain and easier to predict, which, therefore, results in a low entropy value. Conversely, if the driver’s time headway fluctuates more during car-following, the entropy value increases. However, the SW entropy measure does not seem to be sensitive enough to detect small changes in the time headway data, which is not entirely consistent with the result from Nakayama [7]. This entropy algorithm was initially devised for assessing drivers’ lateral control of the vehicle, more specifically, the high-frequency steering corrections [7]. As there are more adjustments and fluctuations in steering behaviour, the algorithm will yield higher entropy value, compared with the time headway data in our car-following experiment, which describes the longitudinal accelerating or decelerating behaviour with less uncertainty.

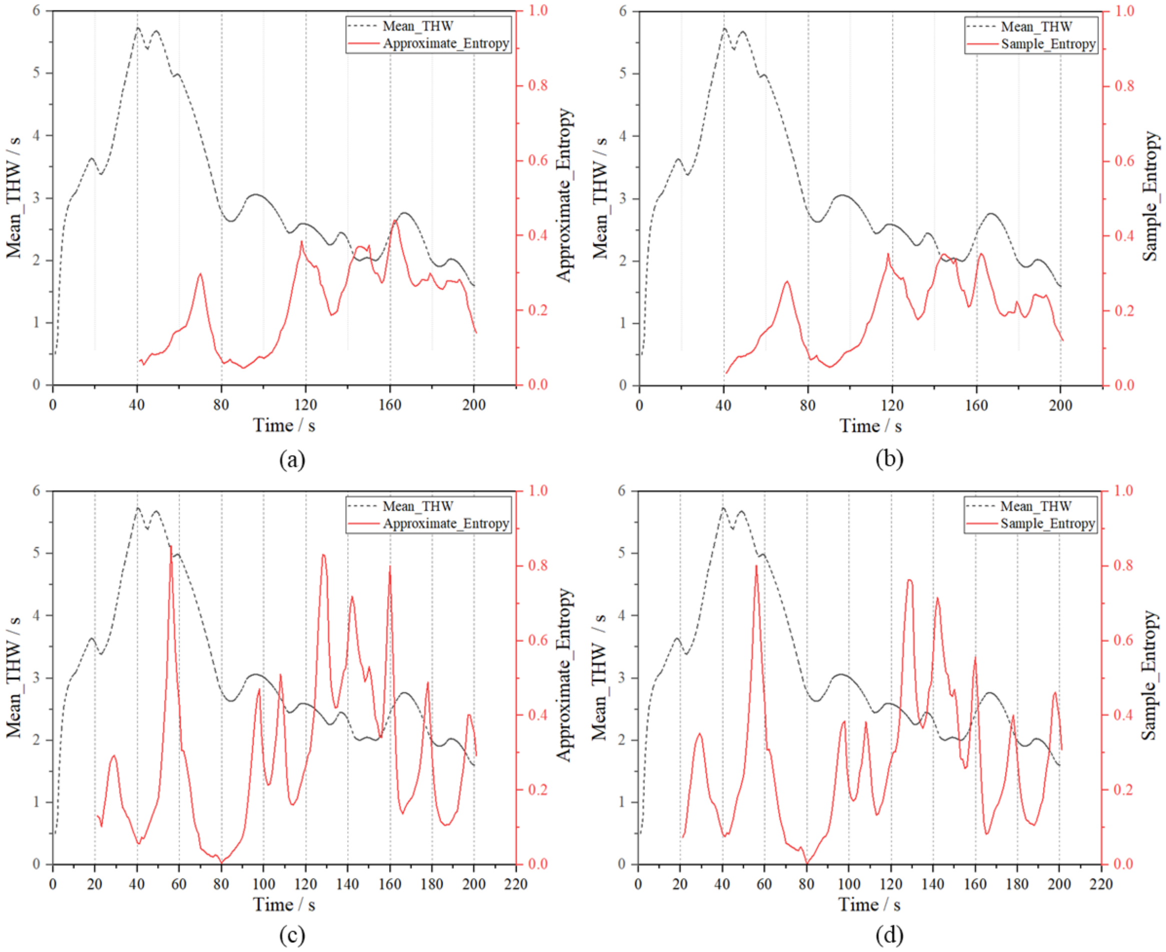
Figure 3c and 3d show the time headway and Approximate Entropy and its updated version, Sample Entropy. Due to their construction, these two entropy algorithms are more generic and independent of the nature of the dataset, ignoring the distribution of the original data and focusing more on the patterns of the series. Additionally, the value of ApEn and SampEn is non-negative and finite for deterministic processes with noise, as the parameter r serves as the noise filter as well. It can be seen from the plot that, using the sliding window strategy, the entropy value changes correspondingly. More specifically, the entropy value increases when there are more new patterns in the computed window, while the entropy value decreases if the period is more predictable and regular. They are sensitive enough to tell the changes in series and can be used to detect variance when drivers change their behaviour. Considering the theoretical definition of uncertainty from Boer [2], as unpredictable or unaccounted changes on human behaviour when it comes to their adopted time headway to a lead vehicle, both Approximate and Sample entropy can highlight the moments where variations in the fluctuation pattern happen, represented by sudden spikes in the entropy value. In other words, it is possible to assume that those two algorithms are a good surrogate metric for the degree of uncertainty in drivers’ behaviour, with the advantage of being directly quantifiable and replicable for comparison of different experimental datasets, adding scientific value for a previously subjective concept.



(a) (b) (c) (d)

Figure 3 Mean time headway and corresponding entropy of different algorithms in moving time window (window=30s), (a) Shannon Entropy, (b) Steering Wheel Entropy, (c) Approximate Entropy, (d) Sample Entropy. Note: To improve readability, the values of ApEn and SampEn are multiplied by 40.

Figure 4 shows the Approximate Entropy and Sample Entropy in different moving time windows (40 s or 20 s). The motivation for the experimentation with different time windows is that entropy is affected by the patterns seen in the dataset. As car-following datasets are characterised by a constant fluctuation on time headway values, including a larger amount of data might allow the algorithms to identify patterns which could not be seen in smaller samples, as a tradeoff for sensitivity (i.e. the larger the dataset, less likely it is for small changes to make a difference to the entropy values). Based on this observation, it is worthy to note that the size of the time window used for the calculation of entropy must be in line with the experimental design. In datasets with slower fluctuations, a longer time window should be required in order to reveal a pattern in driver's car-following. Conversely, in datasets with more constant fluctuations, a larger time window would affect the variable's sensitivity to highlight uncertainty on the observed pattern.



(a)

(b)

(c)

(d)

Figure 4 Mean time headway and corresponding entropy of different algorithms in moving time window, (a) Approximate Entropy (window=40s), (b) Sample Entropy (window=40s); (c) Approximate Entropy (window=20s), (d) Sample Entropy (window=20s). Note: To improve readability, the values of ApEn and SampEn are multiplied by 40.

Conclusion

The aim of this paper was to contrast different entropy algorithms as a measure of drivers' uncertainty in car-following. Four different entropy algorithms were applied to a time headway data series, and their results and suitability were contrasted and discussed. Our analysis showed that not all the entropy measures tested were suitable in the context of car-following, especially when it comes to measuring uncertainty in time headway data. Sample and Approximate entropy seem to be the most appropriate, especially when it comes to the moving time windows, as they consider the drivers' prior time headway data as a factor in the perceived uncertainty, and avoid a flattening of the entropy progression as the sample size increases. This paper contributes to the field of human factors and automation, as it demonstrates how entropy can be a precise and replicable measure of changes in behaviour, as well as anomalies in patterns of time headway data. The entropy algorithms used here can be used in future data analysis of time headway datasets, as a proxy to directly access drivers' level of uncertainty during the car-following. However, it is important to note that entropy measures are affected not only by the size of the entire data set but also the size of the time window sample used in the calculation of entropy.

References

13. Pipes, L. A. (1953). An operational analysis of traffic dynamics. *Journal of Applied Physics*, **24(3)**: 274–281.

14. Boer, E. R. (1999). Car following from the driver's perspective. *Transportation Research Part F: Traffic Psychology and Behaviour* **2(4)**: 201–206.

15. Brackstone, M., & McDonald, M. (1999). Car-following: A historical review. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 2, pp. 181–196. https://doi.org/10.1016/S1369-8478(00)00005-X

16. Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation* **55(3)**: 249–270.

17. Saifuzzaman, M., & Zheng, Z. (2014). Incorporating human-factors in car-following models: A review of recent developments and research needs. *Transportation Research Part C: Emerging Technologies* **48**: 379–403

18. Sheu, J. B., & Wu, H. J. (2015). Driver perception uncertainty in perceived relative speed and reaction time in car following - A quantum optical flow perspective. *Transportation Research Part B: Methodological* **80**: 257–274.

19. Nakayama, O., Futami, T., Nakamura, T., & Boer, E. R. (1999). Development of a steering entropy method for evaluating driver workload. *SAE transactions* 1686-1695.

20. Nemoto, H., Yanagishima, T., Taguchi, M., & Wood, J. M. (2002). Driving workload comparison between older and younger drivers using the steering entropy method. *SAE Transactions* 2040–2047.

21. Huang, C. S., Pal, N. R., Chuang, C. H., & Lin, C. T. (2015). Identifying changes in EEG information transfer during drowsy driving by transfer entropy. *Frontiers in Human Neuroscience* 9(OCTOBER), 1–12.

22. Wang, F., Wang, H., & Fu, R. (2018). Real-Time ECG-based detection of fatigue driving using sample entropy. *Entropy* **20(3)**: 196.

23. Yu, L., Sun, X., & Zhang, K. (2011, July). Driving distraction analysis by ECG signals: an entropy analysis. In *International Conference on Internationalization, Design and Global Development* (pp. 258-264). Springer, Berlin, Heidelberg.

24. Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy maturity in premature infants Physiological time-series analysis using approximate entropy and sample entropy. *Americal Journal of Physiology Heart and Circulatory Physiology* **278**: H2039–H2049.

25. Xie, H. B., He, W. X., & Liu, H. (2008). Measuring time series regularity using nonlinear similarity-based sample entropy. *Physics Letters, Section A: General, Atomic and Solid State Physics* **372(48)**: 7140–7146.

26. Delgado-Bonal, A., & Marshak, A. (2019). Approximate entropy and sample entropy: A comprehensive tutorial. *Entropy* **21(6)**: 541.

27. Michon, J. A. (1985). A critical view of driver behavior models: what do we know, what should we do?. In *Human behavior and traffic safety* (pp. 485-524). Springer, Boston, MA.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Feasibility study of magnetoencephalographic inter-subject synchrony during music listening

Nattapong Thammasan [1], Ayaka Uesaka [2], Tsukasa Kimura [3], Ken-ichi Fukui [3] and Masayuki Numao [3]

1 Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands. n.thammasan@utwente.nl.

2 Institute for Datability Science, Osaka University, Osaka, Japan. auesaka@ids.osaka-u.ac.jp.

3 Department of Architecture for Intelligence, Osaka University, Osaka, Japan. {kimura,fukui,numao}@ai.sanken.osaka-u.ac.jp

Introduction

Common behaviors of human can be driven by having similar cognitive or emotional responses. Shared emotions elicited by the same stimuli have been found showing synchronization in some brain areas of multiple subjects by using functional magnetic resonance imaging (fMRI) [1,2]. Understanding inter-brain synchrony may pave a way toward insights on emotional dynamics when exposing emotional stimuli. The relatively low temporal resolution of fMRI that limits the study of brain dynamics led to the emergence of estimating inter-subject correlation (ISC) from signals captured by higher temporal resolution tools including electroencephalogram (EEG) and magnetoencephalogram (MEG). Particularly, MEG that provides higher spatial and temporal resolution is proven to be tolerant to noise and artifacts and hence gain higher attention in recent years; many works have demonstrated that MEG ISC is viable approach to discover the dynamics of speech [3], movie perception [4], social interaction [5], but very few attentions have been paid to emotional dynamics.

While emotion can be induced by a variety of stimuli, music is considered an extraordinary material to elicit strong emotions and evoke a wide variety of emotions [6]. Besides, studying neural and physiological correlates with music could enable the application of music in therapy and uplifting emotional state. Music is thus used as stimuli in this research. The only prior work on MEG ISC in music-emotion [7] focused on the concurrent changes in power spectral density of MEG signals across participants. Unfortunately, the frequency domain analysis may sacrifice the advantage of the excellent temporal resolution of MEG.

We, therefore, conduct another study of MEG-based music emotion with the application of correlated component analysis (CCA) approach [8] that allows the analysis of inter-subject synchrony in time-domain to retain temporal resolution. This approach is under our ultimate goal to employ synchrony metrics to identify interesting events during music listening paving the way to understand the interplay between music and emotion. MRI images were also recorded to enable the analysis at the source level at higher anatomical precision. Under this research direction, this paper presents a preliminary result that suggests the feasibility of employing MEG ISC to examine shared emotions occurred when experiencing the same naturalistic stimuli.

Methods

Data acquisition

Thirty-six healthy adults (11 females, 25 males) participated in the study, which had been approved by the ethics committee of Center for Information and Neural Networks, Suita, Japan. All subjects gave written informed consent to participate. In the MEG experiment, a subject participated in six sessions of music listening task; one session encompassed the listening of four 45-second musical excerpts, each of which was preceded by 5-second white-noise sound listening. Four excerpts were derived from literature in music-emotion research and were expected to elicit four different types of emotions: high-arousal-positive-valence, low-arousal-positive-valence, low-arousal-negative-valence, and high-arousal-negative-valence. At the end of each song, the subject had 25 seconds to annotate emotional valence and arousal by pressing left and right mouse buttons respectively at the

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

number of times corresponding to the felt valence and arousal scores (discrete scale from 1 to 9). The music was delivered by using Presentation software (Neurobehavioral Systems Inc., Albany CA), and a picture of arousal-valence plane was also projected to the screen located 60 cm in the front of the subject to aid annotation.

The MEG signals were recorded in sitting position with 360-channel neuromagnetometers by Elekta Neuromag system (Helsinki, Finland) comprising of 103 magnetometers, 206 planar gradiometers, and 51 vertical magnetic sensors, at the sampling frequency of 1000 Hz, and all MEG channels were band-pass filtered to 0.03-330 Hz. In addition, vertical and horizontal electrooculogram, electrocardiogram, stimulus triggers, digital timing signals for synchronization, microphone sound, and audio signals were recorded simultaneously at the same sampling frequency as MEG. Six head-position-indicator (HPI) coil were also attached to mastoids and forehead, together with the registration of 3D anatomic landmarks digitizer before the measurement, to help determine position of the subject's head with respect to the sensor helmet before each stimulus session.

Additionally, structural MRI of 35 out of 36 subjects was also performed with Siemens 3-Tesla MRI Scanner (Siemens Healthcare, Erlangen, Germany). Each structural MRI consisted of 275 slices with a slice thickness of 1mm. Then, the MRI and MEG were co-registered with the function of Neuromag software; the head-shape points reconstructed from anatomical MRI were aligned with the digitized head shapes and the HPI coils used in the MEG sessions. This alignment would help precisely correct the position with regard to MEG sensor helmet, compensate the movement, and allow source localization in the future analysis of source localization.

External magnetic artifacts on MEG signals were suppressed offline using the spatiotemporal signal space separation (tSSS) method [9] implemented within the Elekta Neuromag Maxfilter system. The time delay for audio presentation was approximately 158 ms determined by an analysis of sound recorded from the microphone in the recording room with respect to the stimuli triggering signal, to ensure the time alignment.

Inter-subject synchrony

To assess the extent of synchrony across subjects, we applied the CCA approach proposed by Dmochowski et al. [8], which is the generalization of canonical correlation analysis to multiple subjects. The aim of the approach is to uncover the projection matrices that maximize inter-subject correlation (ISC) between subject-aggregated data, yielding the underlying neural sources that show maximal correlation with other subjects. In this study, ISC is computed at the song level from the aggregated data from all subjects listening to the same song. The correlation coefficients of the three most correlated components are averaged to determine the overall MEG synchrony of the group for a particular song.

Results

To demonstrate the feasibility of analyzing MEG synchrony when listening to music, the computed ISCs were associated with the average scores of arousal and valence annotation across subjects. The results from Spearman's rank correlation coefficient, depicted in Figure 1, analysis reveals the statistical significant correlation of the average ISCs with valence score (r = 0.52, p = 0.01), while similar trend is also found in arousal annotation but is not significantly correlated (r = 0.29, p = 0.18). The result implies that when listening to emotionally-more positive songs, the synchrony of brain activity among listening population is generally higher inferring similar brain responses to these positive songs. In contrast, the songs that subjects reported having higher aroused feelings seem to drive the synchronous MEG signals compared to lower-aroused songs, but this heightened synchrony is not prominent.

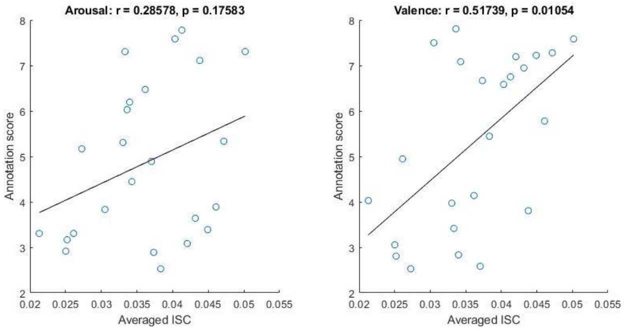


Figure 1. Correlation between ISCs averaged from top-three components and arousal/valence score.

Discussion

The results should be, however, interpreted with sufficient precaution, as numerous points await elucidation. In particular, the underlying mechanism of this homogenous change in neural activities is yet to be further investigated. Although music is a highly complex audio signal, the increased synchrony might be due to either low-level auditory processing or high-level musical feature perception or both. For instance, musical mode and tempo were found to contribute to the change in EEG activity in music cognition [10], and the association between valence score and average ISC in our results might be owing to the fact that musical excerpts with similar valence annotation tend to have similar tempo, intensive, or mode; this hypothesis necessitates the analysis on musical features in our future work. In the bottom-up view, emotion process can also play a key role in enhancing the inter-brain synchrony by eliciting momentary shared emotion. Further investigation on the scalp projection on underlying sources that drive synchrony or source localization by incorporating MR images are highly encouraged and included in our future work in order to reveal which regions of the brain involve with the inter-brain synchrony. Besides, to exploit the excellence in temporal resolution of MEG, the analysis of synchrony over the course of time is also worthy for future study to determine the exact temporal position that the synchrony occurs, rather than analyzing at song-level granularity. Moment-to-moment ISC would enable possibility to explore the links between instantaneous musical structure and momentary emotion, the mental process that was demonstrated to be changing over time when listening to music [11].

References

1. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R. (2004). Intersubject Synchronization of Cortical Activity During Natural Vision. *Science* **303(5664)**: 1634-1640.

2. Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I. P., Hari, R., Sams, M. (2012). Emotions promote social interaction by synchronizing brain activity across individuals. *Proceedings of the National Academy of Sciences of the United States of America* **109(24):**9599–604.

3. Thiede, A., Glerean, E., Kujala, T., Parkkonen, L. (2019). Atypical brain-to-brain synchronization during listening to continuous natural speech in dyslexia. *bioRxiv* 677674.

4. Lankinen, K., Saari, J., Hari, R., Koskinen, M. (2014). Intersubject consistency of cortical MEG signals during movie viewing. *NeuroImage* **92**: 217-224.

5. Hasegawa, C., Ikeda, T., Yoshimura, Y., Hiraishi, H., Takahashi, T., Furutani, N., Hayashi, N., Minabe, Y., Hirata, M., Asada, M., Kikuchi, M. (2016). Mu rhythm suppression reflects mother-child face-to-face interactions: a pilot study with simultaneous MEG recording. *Scientific Reports* **6**: 34977.

6. Koelsch, S. (2012). Brain and Music. Wiley.

7. Thiede, A. (2014). Magnetoencephalographic (MEG) Inter-subject Correlation using Continuous Music Stimuli. Master thesis. Aalto University, Finland.

8. Dmochowski, J. P., Sajda, P., Dias, J., Parra, L. C. (2012). Correlated components of ongoing EEG point to emotionally laden attention - a possible marker of engagement?. *Frontiers in human neuroscience* **6**(112).

9. Taulu S., Simola J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine & Biology* **51**:1759–1768.

10. Sebastian, S. (2017). Toward Studying Music Cognition with Information Retrieval Techniques: Lessons Learned from the OpenMIIR Initiative. *Frontiers in Psychology* **8**(1255).

11. Popescu, M., Otsuka, A., Ioannides, A. A. (2004). Dynamics of brain activity in motor and frontal cortical areas during music listening: a magnetoencephalographic study. *NeuroImage* **21(4)**:1622-1638.

Development of an algorithm to identify stabilisation time for car-following after transitions of control from vehicle automation

Rafael C. Gonçalves[1], Wei Lyu[1,2], Guilhermina A. Torrão[1], Pablo Puente Guillen[3] , Tyron L. Louw[1] and Natasha Merat[1]

1 Institute for Transport Studies, University of Leeds, Leeds, UK. tsrg@leeds.ac.uk, t.l.louw@leeds.ac.uk, gtorrao@gmail.com, n.merat@its.leeds.ac.uk

2 School of Business Administration, Northeastern University, Shenyang, China. lvweineu@gmail.com

3 Toyota Motor Europe NV/SA, Hoge wei 33, 1930 Zaventem, BE. pablo.puenteg@gmail.com

Abstract

The goal of this paper was to describe the development and validation of an algorithm able to detect the beginning of a car-following task engagement inside a time headway (THW) dataset. The motivation for this paper comes from the sensitivity of car-following models to noise inside the datasets, leading to unreliable results. Another aggravating factor for the noise inside such models is that nowadays, more studies are being developed considering the context of vehicle automation and transitions of control, where it is expected that drivers will have a certain delay until the time they recover motor coordination of the driving task and become able to follow a lead vehicle. The algorithm uses the concept of "stability" in car-following, which is defined as a constant but small fluctuation in vehicle's THW as the criteria to identify the beginning of the task. In the end, a nested loop approach was applied, and the tool reached a performance of 89.2% reliability when tested against an experimental dataset, providing statistically significant improvement in the data quality by reducing unnecessary noise.

Introduction

Car-following is an area of research in the field of human factors and road safety that has been developed for over 60 years [1], including the development of models that replicate how vehicles follow each other in a constant flow of traffic (see [2] for a complete literature review). Those models are largely applied in driver safety research (e.g. [3]), as they outline how close or far drivers are willing to maintain their distance from a lead vehicle, in order to avoid possible collisions, while maintaining a steady flow of travel. The goal of this paper is to describe the development and validation of a filtering algorithm that may improve such models by reduction of noise within dataset relevant to a resumption of control from automated driving.

According to Gipps [4], car-following is a particular driving-related task, where drivers are continually meditating and adjusting their distance from a lead vehicle, as a tradeoff between safety and efficiency. It is important to note that this constant balance is what distinguishes car-following from other activities, such as collision avoidance manoeuvres - where there is no mediation of distance, but rather an eminent need to brake/avoid an impending obstacle; or a free drive – where there is no lead vehicle close enough to be followed, and drivers do not need to adjust their speed or relative position, but just keep a reasonable, preferred, cruising speed. In Gipps's model, as in many others in the field (c.f. [5,6,7]), the primary input data used for the parameter estimation of the model are 1) the vehicle's speed, 2) the relative speed of the lead vehicle, 3) drivers' expected "reaction time" [SIC], and 4) the driver's maximum and minimum desired acceleration profiles. This data is then used to estimate how closely a driver is willing to follow a lead vehicle, which is translated into time headway (THW), as a function of relative acceleration and time $s_f(t+\tau)$, as expressed in the following formula:

$$s_f(t+\tau) = \min\{s_{f\ acc}(t+\tau), s_{f\ dec}(t+\tau)\}$$

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Gipps's (1981) formula for estimation of safe car-following distance, where t is time, τ is their expected reaction time to an impending obstacle/situation, and s_f is the driver's willingness to increase or decrease their relative speed, compared to a lead vehicle.

Boer (1999) [8] further complemented this general perspective of car-following models, arguing that drivers are not able to accurately maintain a specific distance from a lead vehicle in constant moving traffic. The author suggests that drivers are more likely to establish 'safe boundaries' of minimum and maximum THW; which they are willing to maintain, and fluctuate their vehicle position inside these boundaries, with constant, but quick, adjustments. With that in mind, generally, THW datasets from a car-following task/experiment show a characteristic stream, with sine wave-shaped pattern, representing these small adjustments (see Fig. 1). This sort of data is then used by modellers to estimate the 'safe boundaries' adopted by drivers during car-following conditions.

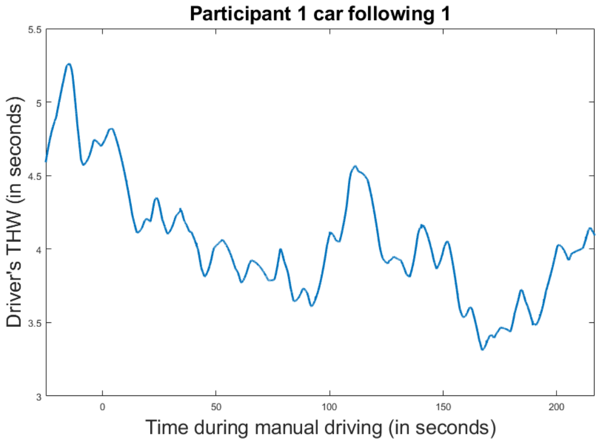


Figure 1: Example of a driver car-following dataset generated from the Leeds Driving Simulator Study in partnership with Toyota Motor Europe, under the L3Pilot Project.

According to Ciuffo et al. [9], these estimated 'safe boundaries' are generally used as data for the development and implementation of safety features in vehicles, or even control modules of automated vehicles (see [10], as an example). By understanding driver preferences in this context, new models can be developed to better adjust to the drivers' safety boundaries, leading to a more comfortable, and pleasant, automated driving experience. However, the use of data from real drivers for the implementation of such models is prone to noise, caused by the unpredictable nature of human behaviour. One factor that contributes to this noise in the model is the inclusion of non-car-following data, which may lead to inaccurate outputs. For example, car-following is strictly defined as an activity that involves the mediation of drivers' gap to a lead vehicle. However, this includes acceleration data from free driving, or braking data for collision avoidance, which are natural activities commonly accompanying car-following, but might skew the overall mean and standard deviation of the THW dataset, leading to model parameters which may not match actual human behaviour.

Defining the beginning of "true car-following" for a specific experimental dataset may be particularly challenging in experiments that consider manual car-following tasks, in the period immediately after a transition of control from vehicle automation. To date, many studies have demonstrated that, compared to manual driving, there are delays in drivers' response to hazards after resumption of control from automation, and that this change in control from system to human results in lateral and longitudinal jerks of the vehicle, which must be adjusted and brought under control by the human, in order to achieve a steady driving state (see [11,12,13] for examples). Therefore, when attempting to study car-following behaviour in this context, the inclusion of data in the model immediately after the transition of control may produce errors for parameter estimation. To our knowledge, this issue has not yet been addressed in the literature and is, therefore, the focus of this paper. We describe the development, implementation and testing of an algorithm, designed to systematically identify the exact time when drivers initiate the car-following task, following the transition of control from automation.

Methods

Data reduction techniques for car-following analysis

To date, many experimental studies investigating driver behaviour in car-following have used a range of approaches to remove undesirable/noisy data from the dataset, before embarking on data analysis. This section of the paper outlines a number of these techniques, highlighting their advantages and limitations, which then motivated the development of the algorithm developed and described by us.

The first, and most common technique to identify whether or not drivers are engaged in car-following, is the observation of drivers' average THW. For instance, using a large dataset from a naturalistic driving study (168053 samples) Loulizi et al. [14], found that drivers' THW during car-following ranges between 1.2 and 2.5 s (seconds). Based on this data, it was proposed that drivers are not engaged in car-following whenever their THW is above or below this given threshold. Therefore, data present values not in this range were treated as noise and removed from the analysis. However, it can be argued that this technique is based on an arbitrary threshold, which changes based on different scenarios. For example, THW in car-following may vary depending on traffic density and driving style of the individual driver [15]. Another issue of this approach is that it is very sensitive to individual differences and outliers. Time headway is also sensitive to individual driver characteristics, such as risk-taking propensity, with safe drivers wishing to maintain a longer headway than risk-taking drivers, which this technique dismisses.

A second solution is to remove part of the dataset, from the beginning and end of the data collection period, so that the data used for parameter estimation does not contain periods which involve the driver approaching the lead vehicle or intentionally keeping a longer headway. Again, this is an arbitrary procedure, which is likely to diminish the statistical power of the model, due to a reduced volume of data. Another problem with this approach is that it might be susceptible to type two statistical errors, where, by chance, the removed data represents a valuable difference between the experimental groups and masks the outcome of the treatment under investigation.

The third, and last, approach is to select an experimental design for the data collection which forces a car-following situation, by placing a vehicle in front and behind the subject's vehicle (see [3] as an example), which creates a less naturalistic behaviour for drivers. Another issue with this approach is that it does not reduce the noise caused by transitions of control from vehicle automation [12], as drivers would still need to take over control, and, therefore, have a certain delay in their ability to drive.

Considering the above limitations, four main criteria were established for the design of a new technique to tackle this issue. That it should: 1) The technique should be structured enough to be replicable in different experiments, and be consistent in the results across datasets, 2) use the overall distribution of every single data point and account for any individual differences among drivers, 3) be flexible enough to reduce the data loss to a minimum, whenever trimming the datasets is required; and 4) consider the issues inherent to the process of transitions of control from vehicle automation.

Theoretical foundation for the algorithm design

In one of the core studies for car-following analysis, Herman et al. [16] introduced the term "stability" in car-following, considering a stable time where the Δ (delta), or variability in drivers' position remains relatively constant, with small fluctuations, not affecting the overall microstructure of the surrounding traffic. According to these authors, a moment of stable fluctuation of the position behind the lead vehicle is one of the core characteristics of the car-following process. We believe that this concept is key to identifying when drivers engage in car-following, after taking over from vehicle automation, due to the theoretical similarity between the process of motor coordination reacquisition and the process of stabilisation.

According to Mole et al. [12], whenever drivers are removed from the motor control loop of the driving task (see [11] for a complete definition of the term), there is a "transition lag" for the recovery of motor control coordination, which means that drivers might take some time between the moment of automation disengagement, until the point they are entirely in control of driving again. As empirical evidence for such scenarios, [13] and [11] have reported that drivers display an erratic steering and acceleration profile, in the first few seconds after the transition of control (which may take up to 30 s). Therefore, when considering car-following, drivers are expected to have a large Δ in their acceleration/deceleration profile, right after the transition from automation. Over time, as they approach their comfortable, and desired, following distance, this fluctuation would be minimised to a stable point. Based on the above assumption, we based our algorithm on the concept of stabilisation, assuming that drivers are car-following whenever this initial fluctuation is stabilised.

Detection of the stabilisation point

To identify when drivers' THW fluctuation starts to become stable, it was first necessary to understand the nature of the experimental data collected for this analysis. Fig. 2 shows a histogram of the variance in THW for one participant, used for the development of our algorithm, which comprehends the period of car-following of one drive and the transition of control referent to that particular task.

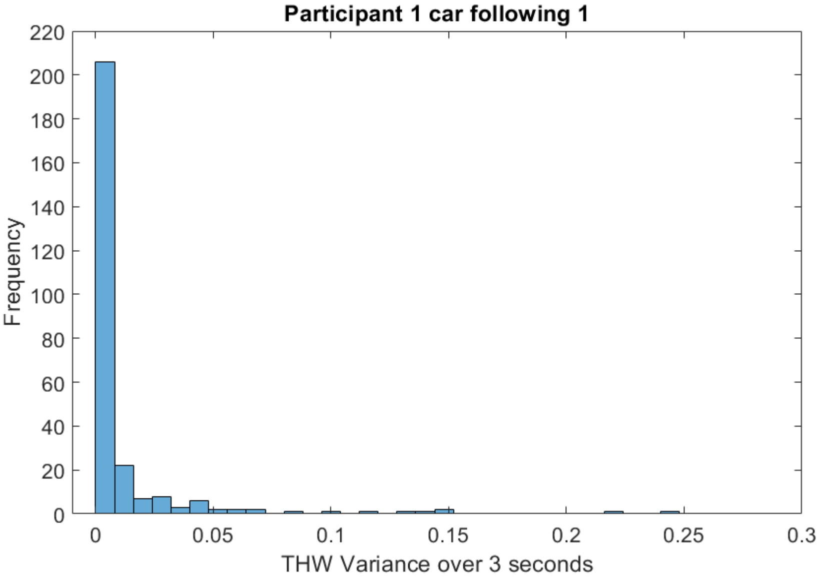


Figure 2 Histogram of one driver's THW over a ~300 s period. The variance was calculated within a time window of 3 seconds.

As shown in Fig. 2, the histogram of THW variance data during car following follows somewhat logarithm distribution, where there is a high frequency of data points with low values, and a low frequency the further the values get from 0. In other words, there is only a small proportion of time when there is a large variance of drivers' THW. If the stable period of car-following is defined by a low fluctuation in relation to the overall variance distribution inside one individual data sample (the same driver), we defined as stable the moment in which drivers were able to maintain the variance of their THW under the inflexion point of the overall log distribution (which is calculated as the mean plus standard deviation of the distribution). This method would satisfy the four conditions defined at the end of the previous section, as 1) it is consistent and replicable across different datasets; 2) it accounts for individual drivers' THW variability; 3) minimises the data lost by filtering the dataset, and; 4) considers the issues related to the transition of control from vehicle automation.

Algorithm description and application

To implement the solution described above, a nested loop approach was applied, where, for each point in time (for a frequency of 60 Hz), the algorithm calculated the variance of drivers' time headway, for a period of S seconds, and checked if this variance was below the inflexion point for the overall variance of the distribution of the drivers' car-following (which is characterised by mean + standard deviation of the distribution). In case the observation passed this criterion, the algorithm would check the next *S* seconds until the same condition was satisfied *I* times. In case any of the *I*-second time windows did not satisfy the criteria, the loop was interrupted, and restarted again, moving forward by one frame. Fig. 3 shows a flowchart with a graphical representation of how the algorithm works.

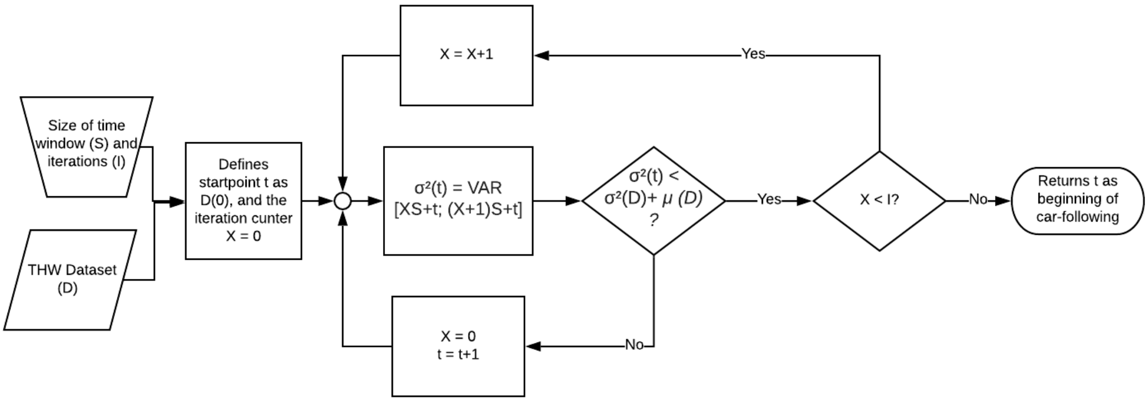

Size of time window (S) and iterations (I)
THW Dataset (D)
Defines startpoint t as D(0), and the iteration cunter X = 0
X = X+1
σ²(t) = VAR [XS+t; (X+1)S+t]
X = 0
t = t+1
σ²(t) < σ²(D)+ μ (D) ?
Yes
No
X < I?
Yes
No
Returns t as beginning of car-following

Figure 3 Flowchart for the stabilisation algorithm.

Algorithm use and access

The program was written in Matlab programming language, using the Matlab R2018 version (Mathworks, 2020). The program is divided into two ".m" functions, which are available to download using the following link: https://github.com/rafaelCirinoGoncalves/RCGoncalves.git. In order to use these functions, both must be placed in the same file location as the dataset, which contains the input variables. As a reminder, the dataset must be set to 60hz frequency; otherwise, resampling methods might be necessary to adjust the dataset to the code.

The first function, called "find threshold.m" will be used to calculate the inflexion point inside the given dataset, which shall be manually entered as an entry value for the second function. It has as entry values the THW dataset of a given participant, and the size in frames of the time window used to calculate the variance. The second function, called "findStabilizationTime.m" is the final function, which returns the index for the beginning of the car-following in the given dataset. It receives as input values the threshold calculated by the previous function, the number of iterations desired as a requirement for the beginning of the car-following, the size of the time window for variance calculation and the given THW dataset of an experiment participant.

Algorithm parameter estimation and validation

In order to ensure the quality of the algorithm, different parameters were tested for the values S (size of the time window used to calculate the variance on drivers THW), and I (number of iterations required to satisfy the criteria of a stable car-following) as described on Fig. 3. The outputs of the algorithm were then manually examined to check the accuracy of the method. We considered that the algorithm failed to detect the precise beginning of a car-following task of a given dataset whenever it returned its start point during periods of sharp increments or decrements on drivers' THW, or whenever the algorithm removed more than 3 s of small fluctuation inside the dataset, which is already considered to be a car-following. Examples of both failure cases can be seen in Fig. 4.

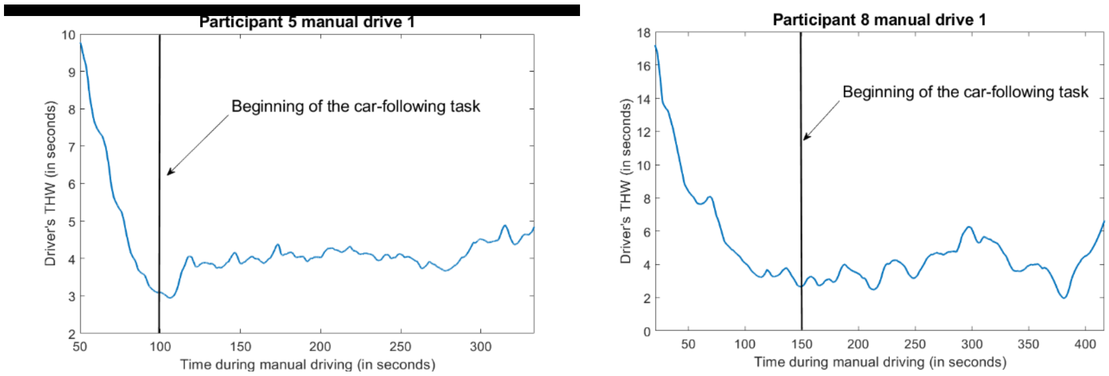


Figure 4. Examples of data points in which the algorithm failed to identify the precise beginning of the task.

To test the possible improvements for the dataset, this study compared the results of mean and SD of THW with the full length of drivers' car following task for the experiment described below, and then again, after the application of the algorithm.

Experimental design for the dataset

The experiment used data from 28 participants (9 F), with age varying from 21 to 86 years-old (M = 39, SD =14), who drove one section of road, which included four manual car-following tasks, interspersed with the same task performed by an automated vehicle. An urban environment was used for the driving scenario, where drivers were asked to follow a lead vehicle, and turn the automated system on/off whenever requested. The experiment began with a manual drive, which included 5-minutes of car following, after which they were required to turn on the automation. This resulted in an automated car following period, with no need for human intervention. After about five minutes of automated car following, the drivers were required to take back control from the vehicle automation, due to fading road markings, and maintain the car following in manual mode for another 5 minutes. Failure to resume control did not result in a collision. This cycle of the last two steps was repeated four times, resulting in eight takeovers, overall. For half of the takeovers (which were presented in a counterbalanced order), there was no lead vehicle immediately after resumption of control, and drivers approached a new vehicle, which joined from a side road, right after the driver resumed control. For the other half, there was a lead vehicle during the transition of control, and drivers had to respond accordingly.

Results

After extensive testing of parameters, the combination which yielded the best accuracy was a 3 s time window for the calculation of the THW variance and repetition of a 10-time loop. In order words, in order for the driver to be considered engaged in the car following task, the variance of their THW to the lead vehicle had to be lower the mean + SD of the overall THW distribution for this single participant during 10 cycles of 3 s (totalling a time of 30 s).

As outlined above, this combination of parameters was manually validated for the entire data sample: 4 runs of 28 participants, resulting in 112 samples. In total, the algorithm failed to identify the precise start point of the car following task in 9 samples. For 2 of these samples, the algorithm defined the start point of the car-following data too early, whenever drivers' were still stabilising motor control of the vehicle (including non-car following data), and for 7 it considered the stabilisation point too late (removing too much car-following data), leading to an 89.2% accuracy. For the cases where the algorithm failed to find the beginning of the car following scenario, the average error margin was of 8.2 s when including acceleration data, and 9.1 s when part of the car following task was removed. Therefore, even considering those failure cases, the algorithm itself was able to remove a considerable amount of noise from the data, by cutting most of the non-car following task from the dataset.

For the comparison between the datasets, before and after the application of the filtering algorithm, a 2 (dataset with and without algorithm) by 2 (transition of control with or without the lead vehicle) ANOVA test was applied, using the mean and SD of THW of the dataset as the dependent variables. Results showed a significant main effect of the application of the filtering algorithm on drivers' average THW [F (1, 27) = 57.42, p<.001, $\mathbf{p\eta^2}$ =.68], where *posthoc Bonferroni* (see Fig. 5) tests showed that the overall THW mean was lower after the application of the algorithm. The ANOVA results also identified a significant interaction effect between the two independent variables [F (1, 27) = 8.119, p<.001, $\mathbf{p\eta^2}$ =.907], indicating that the observed THW for drivers was lower after the application of the algorithm in cases where the lead car was not present during the take-over (3.87 s without the algorithm and 3.18 s with the algorithm), and higher whenever the lead car was present (3.46 s without the algorithm and 3.54 s with the algorithm). When drivers were required to take over in the presence of the lead, they reduced their speed, to increase their gap. On the other hand, whenever there was no car in front during the take-over, the drivers needed to accelerate, and reduce the gap between them and the vehicle in front, to engage in car-following. Therefore, the fact that the filtering algorithm was able to increase the mean THW during the situations when drivers increased their gap with the lead vehicle, and decreased the mean THW during the times when drivers decreased their gap with the lead vehilce, indicates that the noise inherent to the process of transition of control was successfully removed from the sample, increasing the overall data quality.

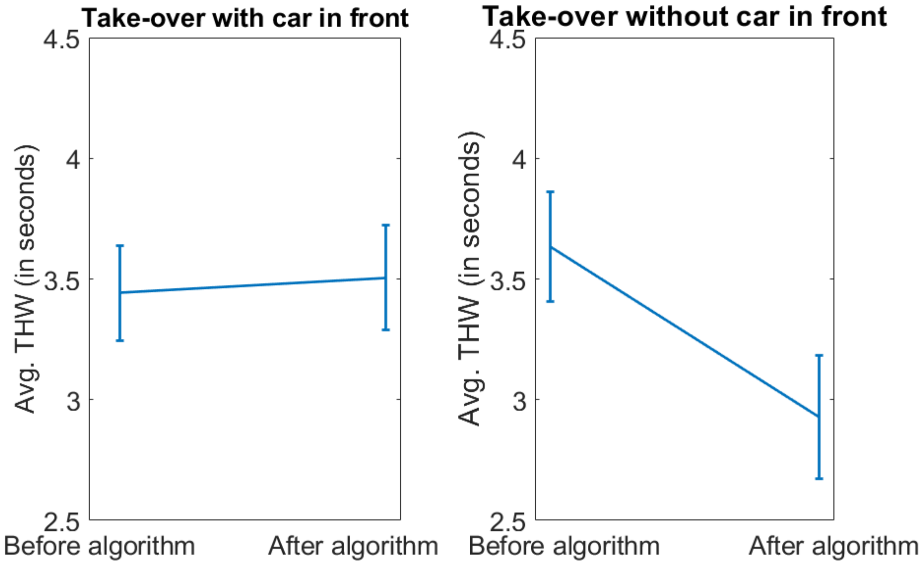


Figure 5 Differences in THW between the experimental conditions and algorithm application.

The analysis also showed a significant main effect of the application of the filtering algorithm on the SD of THW [F (1, 27) = 517.222, p<.001, $\mathbf{p\eta^2}$ =.950], where the SD of divers' THW was significantly lower for the dataset with the application of the algorithm (.81 s), compared to the one without (1.73 s) – see Fig. 6. This suggests that the algorithm was not only able to remove the moments of higher Δ for vehicle position (which indicates a sharp acceleration/deceleration and not a car following position adjustment) but also reduced the internal variability of the data sample, which improves the power of the statistical tests that may be applied for further analysis.

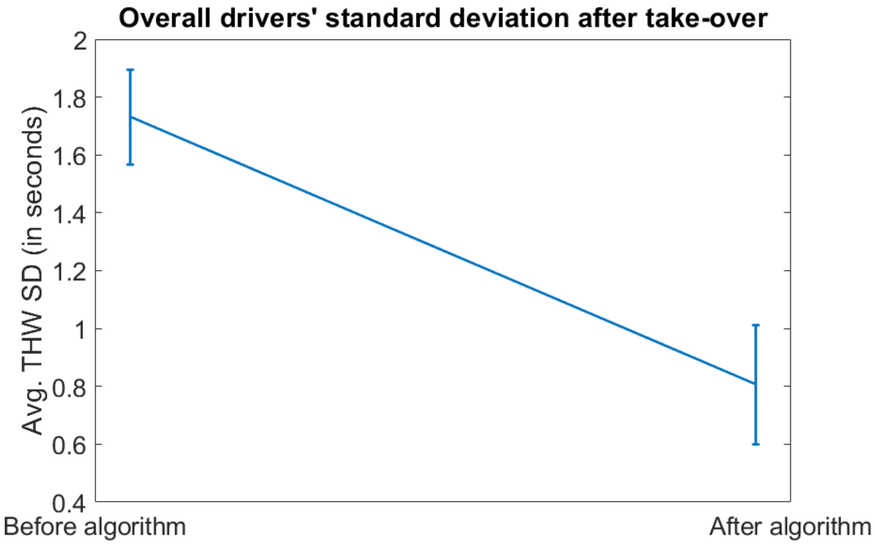


Figure 6 Differences of the standard deviation of THW between the experimental conditions and algorithm application.

Conclusion

The objective of this paper was to describe the development and validation of a filtering algorithm for car-following THW-based experimental datasets, particularly relevant to the resumption of control from automated vehicles. The algorithm was capable of accounting for the process of transitions of control from vehicle automation when defining the beginning of a car-following task. It used the concept of "stability" in car-following as a theoretical basis, to define when drivers started to engage in car-following.

For testing and validation of its applicability, the algorithm was applied to an existing dataset, involving a transition of control from vehicle automation followed by car-following. The results were compared in terms of accuracy of the method for identifying the start of car-following, and how much noise was removed from the dataset. Results showed that the success of the algorithm was high (89.2% reliability), and could successfully reduce noise in the dataset, based on our proposed theoretical foundation.

The solution proposed in this paper is an improvement over other common approaches, as it considers the whole distribution of every driver's THW (accounting for individual differences), reduces the amount of data loss in the process of trimming, and is specifically tailored to deal with data related to transitions of control from vehicle automation. The algorithm is easy to implement, and generic, which can be applied to different datasets, regardless of the experimental conditions. Due to its generic approach, and proven reliability, we believe that this tool can be applied in future research to generate more accurate results in car-following models, as more and more research is trying to understand how humans behave after experiencing transitions from vehicle automation.

References

1. Pipes, L. A. (1953). Kinetic theory of vehicular traffic. *Journal of applied physics.* volume *24,* 54. https://doi.org/10.1063/1.1721265

2. Brackstone, M., & McDonald, M. (1999). Car-following: A historical review. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 2, pp. 181–196. https://doi.org/10.1016/S1369-8478(00)00005-X

3. Zhang, J., Wang, Y., & Lu, G. (2019). Impact of heterogeneity of car-following behavior on rear-end crash risk. *Accident Analysis and Prevention, 125,* 275–289. https://doi.org/10.1016/j.aap.2019.02.018

4. Gipps, P. G. (1981). A behavioural car-following model for computer simulation. *Transportation Research Part B* ***15*****(2)**: 105–111. https://doi.org/10.1016/0191-2615(81)90037-0

5. Reiter, U. (1994). Traffic Performance Models for Unsignalised Intersections and Fixed- Time Signals. *Proceedings of the Second International Symposium on Highway Capacity* Volume **1**: 21–50

6. Ranjitkar, P., Nakatsuji, T., & Kawamua, A. (2005). Car-Following Models: An Experiment Based Benchmarking. *Journal of the Eastern Asia Society for Transportation Studies* **6**: 1582–1596. https://doi.org/10.11175/easts.6.1582

7. Treiterer, J., & Myers, J. (1974). The hysteresis phenomenon in traffic flow. Transportation and traffic theory, **6**: 13-38

8. Boer, E. R. (1999). Car following from the driver's perspective. *Transportation Research Part F: Traffic Psychology and Behaviour* ***2*****(4)**: 201–206. https://doi.org/10.1016/S1369-8478(00)00007-3

9. Ciuffo, B., Punzo, V., & Montanino, M. (2012). Thirty years of Gipps' car-following model. *Transportation Research Record* **2315**: 89–99. https://doi.org/10.3141/2315-10

10. Wei, C., Romano, R., Merat, N., Wang, Y., Hu, C., Taghavifar, H., … Boer, E. R. (2019). Risk-based autonomous vehicle motion control with considering human driver's behaviour. *Transportation Research Part C: Emerging Technologies* **107**:1–14. https://doi.org/10.1016/j.trc.2019.08.003

11. Merat, N., Jamson, A. H., Lai, F. C., Daly, M., & & Carsten, O. M. (2014). Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation Research Part F: Traffic Psychology and Behaviour* **27**: 274–282. https://doi.org/10.1016/J.TRF.2014.09.005

12. Mole, C. D., Lappi, O., Giles, O., Markkula, G., Mars, F., & Wilkie, R. M. (2019). Getting Back Into the Loop: The Perceptual-Motor Determinants of Successful Transitions out of Automated Driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, (January), 001872081982959. https://doi.org/10.1177/0018720819829594

13. Russell, H. E., Harbott, L. K., Nisky, I., Pan, S., Okamura, A. M., & Gerdes, J. C. (2016). Motor learning affects car-to-driver handover in automated vehicles. trials **6(6)**: 6.

14. Loulizi, A., Bichiou, Y., & Rakha, H. (2019). Steady-State Car-Following Time Gaps: An Empirical Study Using Naturalistic Driving Data. *Journal of Advanced Transportation* https://doi.org/10.1155/2019/7659496

15. Edie, L. C. (1961). Car-Following and Steady-State Theory for Noncongested Traffic. *Operations Research***9(1)**: 66–76. https://doi.org/10.1287/opre.9.1.66

16. Herman, R., Montroll, E. W., Potts, R. B., & Rothery, R. W. (1959). Traffic Dynamics: Analysis of Stability in Car Following. *Operations Research* **7(1)**: 86–106. https://doi.org/10.1287/opre.7.1.86

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Deep learning systems for automated rodent behavior recognition suffer from observer bias: Time to raise the bars

E.A. van Dam[1,3], L.P.J.J. Noldus[1,2] and M.A.J. van Gerven[3]

1 Noldus Information Technology BV, Wageningen, The Netherlands. e.vandam@noldus.nl

2 Department of Biophysics, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands.

3 Department of Artificial Intelligence, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands.

Abstract

In order to be useful in behavioral research, automated systems that can recognize high-level behavioral activities must be able to recognize them independent of animal genetic background, drug treatment or laboratory setup. However, just as human observers, deep learning systems suffer from observer bias.

The current recipe for training a recognition system is to record a dataset, annotate the dataset and use supervised learning to train a classifier to recognize the behaviors. The classifier iteratively finds the best optimization path to get as close to the ground truth as it can, using all the cues it can find. Hence, the quality and robustness of the resulting classifier is always dependent on the variance and representational value of the data trained on. In order to be robust to using cues that are only coincidentally or concurrently related to the behavior classes, data augmentation is applied to the input: typically image transformations like flipping, scaling and rotation. Deep learning models are very good at finding informative cues, but this also means they are sensitive to using biased cues that only apply within the training dataset.

Bad recognition performance on a test dataset reveals this bias. However, in almost all studies that describe behavior recognition systems the test set is recorded in the same setup, with animals from the same strain and treatment as those in the training set, hence revealing only some of the over-fitting. It was recently shown that although deep models can reach better performance than conventional methods, the performance is less transferable to different settings [1]. We argue that good performance on a within-dataset setup is important but not enough for useful deployable systems.

Three approaches are known to deal with dataset diversity. First is to standardize laboratory setups. This limits the variance but leaves the animal and treatment related variation. Second is to aim for quick adaptation of the recognition system towards a new setup with minimal annotation effort, i.e. fine-tuning or retraining. But retraining supervised systems on a new setup requires new ground truth data and brings back the manual annotation task for a significant number of video segments. Moreover and more importantly, researchers who need to compare the behavior between treatment groups simply cannot use differently trained observation models. The third is to explicitly strive for generic recognition with robust methods, which is in principle possible as is proven by humans.

We propose to raise the bar for future automated behavior recognition systems to be useful. Behavior recognition beyond body point tracking and pose estimation needs to deal with the following:

- Appearance differences: For behavior it is irrelevant whether the animal is white or black, thick or slim, long or short-haired. The same holds for the appearance of the environment that enables or limits behavior, such as the walls, floor, feeder, drink spout, enrichment objects.
- Behavior style differences: The same behavior can be performed in many different ways. It varies per behavior bout, for example in duration, pace and sub-behavioral pattern. It varies per physical or

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

emotional state, but can also vary per animal, depending on strain, gender, age, history and medication. It also can vary due to different layout of the environment, such as the height of the drinking spout.

- Behavior sequence differences: Finally, the order of the behavior bouts is subject to change and in particular dependent on animal treatment. For example, grooming bouts are mostly in between resting events. Rearing events are normally followed by a direction change followed by a walk. Behavior recognition systems that use history or recurrence (HMMs, LSTMs, 3D-CNNs) train on temporal context and hence on behavioral context, and will have difficulty to recognize the behavior bouts when applied in a different context.

In order to meet these demands we need to find new ways of behavior augmentation or set new demands on the network output, so networks can learn to differentiate between relevant and irrelevant cues. More and more diverse data always helps but will not be sufficient to deal with abnormal behavior. Understanding the composition of behavior by identifying the syllables and grammar that it comprises of [2] can be a useful intermediate step.

Conclusion

Currently available automated systems for the recognition of semantic behavior activities suffer from observer bias with respect to animal treatment and environment setup. Increasingly sophisticated deep learning models provide a promising future [3], but the biggest hurdle still is generalization. In this abstract, we set the bars for future systems more explicitly: In order to be useful in behavioral research, they must 1) recognize control and treated group without fine-tuning and 2) be able to deal with differences in appearance, behavioral style and behavior sequences.

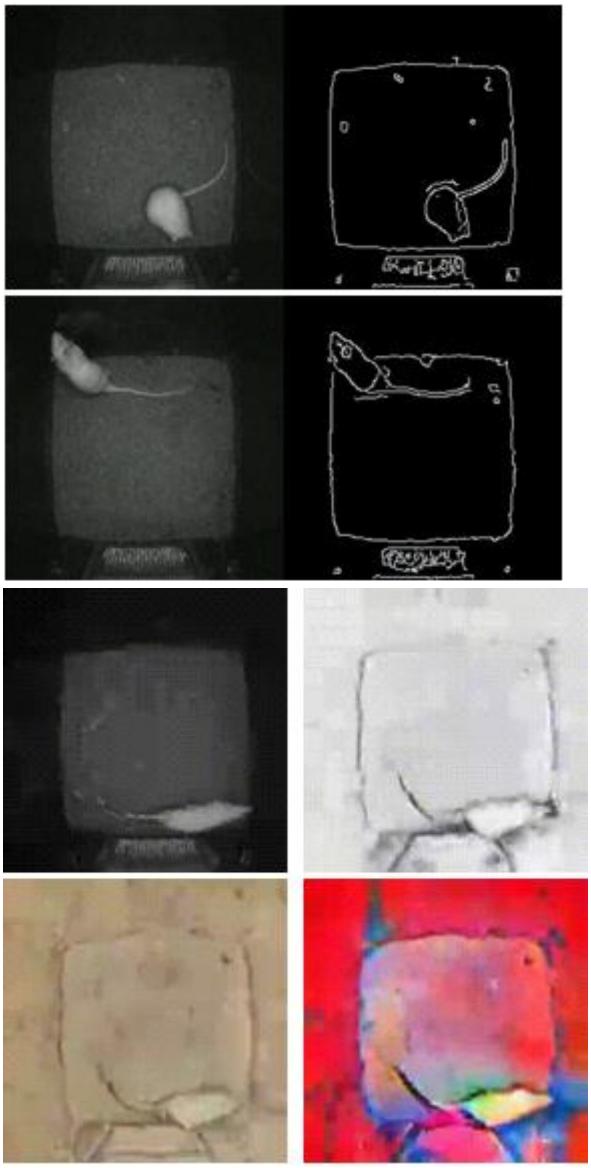


Fig 1. Diversifying the appearance of the video does not influence the ability of humans to recognize the behavioral activities. The same should hold for automated recognition systems.

References

1. van Dam, E. A., Noldus, L. P. J. J., & van Gerven, M. A. J. (2020). Deep learning improves automated rodent behavior recognition within a specific experimental setup. *Journal of Neuroscience Methods* **332**: 108536.

2. Datta, S. R. (2019). Q&A: Understanding the composition of behavior. *BMC biology* **17(1)**: 44.

3. Mathis, M. W., & Mathis, A. (2020). Deep learning tools for the measurement of animal behavior in neuroscience. *Current Opinion in Neurobiology* **60**:1-11.

Using Smartphone and Wearable Sensors to Track and Monitor Smoking Episodes Quantitatively in Daily Life

Donghui Zhai[1,2], Erika Lutin[1,2], Giuseppina.Schiavone[2,3], Walter De Raedt[2], and Chris Van Hoof[1,2,3]

1 Department of Electrical Engineering (ESAT), KU Leuven, Kasteelpark Arenberg 10, Leuven, Belgium. donghui.zhai@kuleuven.be.

2 imec, Kapeldreef 75, Leuven, Belgium.

3 Holst Centre/imec, High Tech Campus 31, Eindhoven, The Netherlands. giuseppina.Schiavone@imec-nl.nl.

Introduction

Smoking has been the leading cause of morbidity and mortality over the past decades, yet the prevalence of smoking remains high. Globally, there are currently 1.1 billion smokers [1]. To smooth the way towards health promotion and disease prevention of these large at-risk smoking populations at an affordable expense, more and more researchers advise to exploit solutions enabled by mHealth systems, which can deliver tailored counselling and behavioral change treatment in a scalable way. So far, most mHealth applications for smoking cessation are primarily founded upon evidences from experiments and theoretical works conducted decades ago, either in constrained settings or based on outdated research methods such as interviews and paper diaries [2]. In the absence of objective methods to monitor smoking, it is often difficult to design appropriate and efficacious smoking-cessation services, because of its temporal nature and contextual dependence. Recent advances in wearable sensors and digital technologies are beginning to change the priority in smoking habits monitoring towards ambulatory study, which refers to the collection and analysis of social, physical and physiological data relevant with this behavior in situ. Many investigators advocate quantitative data collected from these wearable sensors and smartphones, which are ubiquitous and pervasive in our daily life, can contribute to the better understanding of smoking habits [3], [4].

Methods

To conduct studies in an ambulant mode, we have developed a research platform which consists of several wearable sensors, a smartphone and a mobile phone app called ASSIST, as shown Fig. 1. This current study is designed to run in the real-life setting as an observational trial and does not intend to alter and intervene participants' normal routines. This aim is fulfilled by the passive collection of smoking episodes, concurrent contextual and physiological data. By exploiting and integrating the latest wearable and mobile technologies into one application, this system goes beyond the boundary of laboratory settings and can collect enriched information about smoking habits as they naturally occur in real life environment 24/7. The post analysis and fusion of this authentic information will pave the way for the development of intelligent agents to address smoking cessation in a large population. The various sensors and methods employed to collect these multi-modal data are explained in the following sections.

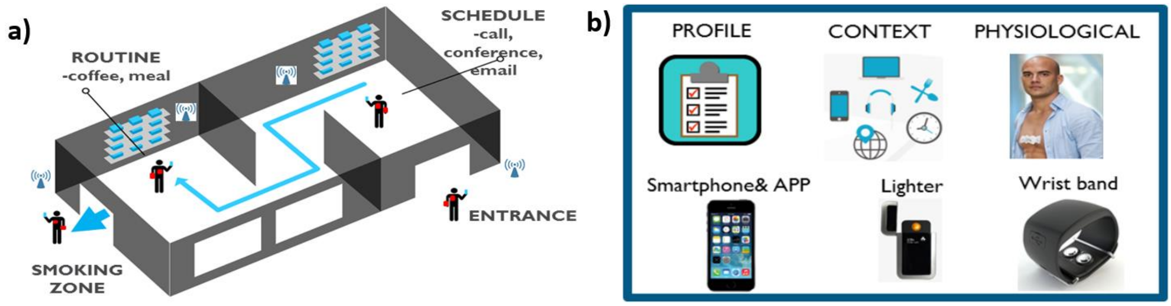
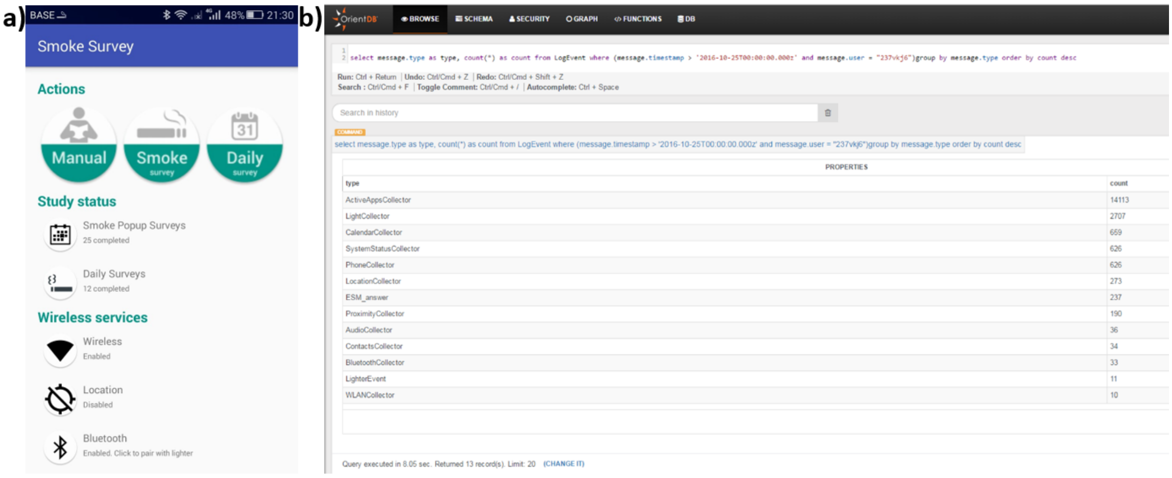


Figure 1. a) the schematics of the study design, and b) sensors and multimodalities of information collected in the study.

App and database

The Android smartphone app that constitutes the front-end of the configurable platform collects both subjective inputs such as Ecological Momentary Assessment (EMA) surveys and daily diaries, and objective data such as GPS, phone status (screen on and off, active apps, etc.) and accelerometer data in their raw form. The ASSIST back end, which is based on an OrientDB cloud database, collects, stores, and processes the collected data. The client-server configuration and modular design, making it both scalable and extensible, are flexible enough to incorporate new features and new studies in the future.

The advantage of tracking with smartphones is high geographical precision (due to GPS), exact temporal resolution, and the possibility to collect additional information on movement, phone logs and Internet use records. Most importantly, however, it allows sending inquiries to the respondent via EMA and making location-specific inquiries.



Smoke Popup Surveys
25 completed

Daily Surveys
12 completed

Wireless
Enabled

Location
Disabled

Bluetooth
Enabled. Click to pair with lighter

```
select message.type as type, count(*) as count from LogEvent where (message.timestamp > '2016-10-25T00:00:00.000z' and message.user = "237vkj6")group by message.type order by count desc
```

Run: Ctrl + Return | Undo: Ctrl/Cmd + Z | Redo: Ctrl/Cmd + Shift + Z
Search : Ctrl/Cmd + F | Toggle Comment: Ctrl/Cmd + / | Autocomplete: Ctrl + Space

select message.type as type, count(*) as count from LogEvent where (message.timestamp > '2016-10-25T00:00:00.000z' and message.user = "237vkj6")group by message.type order by count desc

PROPERTIES

type	count
ActiveAppsCollector	14113
LightCollector	2707
CalendarCollector	659
SystemStatusCollector	626
PhoneCollector	626
LocationCollector	273
ESM_answer	237
ProximityCollector	190
AudioCollector	36
ContactsCollector	34
BluetoothCollector	33
LighterEvent	11
WLANCollector	10

Query executed in 8.05 sec. Returned 13 record(s). Limit: 20 (CHANGE IT)

Figure 2. a) the ASSIST study app, and b) the back-end database dashboard showing the overview of the collected records.

Wearable sensors and lighters

In the course of the study, to automate smoking events tracking, participants are required to use a battery-powered electric lighter when they are going to smoke. This device can register smoking events when users press the side button to light up cigarettes. It is also paired with the ASSSIST app installed on smokers' smartphone through Bluetooth, so that the timestamps of smoking events are synchronized with concurrent contexts sampled by the phone. In the meantime, to capture possible biomarkers of smoking, we employed two wearable sensors to collect physiological signals. One is worn on the wrist to measure skin conductance and skin temperature, and the other is attached to the chest collecting electrocardiograms. These signals are mediated by sympathetic nervous system and often reported to be affected by smoking.

EMA surveys

EMA survey is often used to overcome cofounders and the lack of context reported in daily diaries, which are also susceptible to recall errors. This is an event-centered sampling method that focuses on observing and interviewing people while they are smoking in the context. Due to its contingency with the smoking event, such inquiries are highly reliable and result in more detailed information compared to daily diaries.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

In the current design of this study, participants are prompted to make annotations about their emotions, social context, activity, smoking urges and so on. These prompts are primarily triggered by smoking events captured by the electric lighter. However, when the Bluetooth connection is down, the triggering can fall back on predefined randomization mechanism. In such cases, users will receive at most five randomized surveys a day.

Data Privacy and Security Measures

As there is participant-centered information collected in this study, several measures are taken to protect users' privacy and data security. First of all, to authenticate users' identification, each participant is assigned a unique 6-length activation code when accessing the application for the first time. Second, privacy sensitive EMA answers and contextual data collected from users' smartphone are encoded as numbers instead of real texts. In addition, to avoid data leakage, the application only stores a portion of history records, and it is set to upload all data onto an encrypted Orient DB cloud database. Data stored on the cloud are anonymised and ported into the local database once data collection is finished. Access to both cloud and local databases are password- protected so entry is permitted to authorised users. Only researchers directly involved in the study have access to the database, other external researchers can only apply for access to the preprocessed data where all identifiable and privacy-sensitive information is hashed or removed.

Results and Discussion

This research platform was tested and evaluated in a pilot study during a development phase, and it proved to be especially reliable and convenient in capturing the precise time of smoking in our ambulatory studies. Then it was rolled out for a real-life study which has completed the trial of 53 smoker volunteers. We have observed a difference between the initial self-reported cigarette intake and the actually measured numbers [5]. When confronting difference like this, we can get a set of new insights, which we otherwise would not have discovered by only relying on the retrospective interviews. The collected data, when coupled with appropriate data analysis methods, will enable the study of behavioral patterns, social interactions, physical mobility and among other interesting aspects.

Conclusion

This platform has been realized using a modular design and validated to be able to adapt to various needs of studies. With its multiple usage options and scalability, it can facilitate data collection and behavior monitoring in ambulatory study. The small form factor and omnipresence of such sensors ensure researchers to conduct various alike studies in real-life setting, such as stress [6], eating disorders [7], autism and depression, with high fidelity of data and minimal impact on user comfort. On the other hand, as data gathered in ambulatory settings fill in gaps between researchers and subjects regarding what happens in a real-world setting, they have the potential to not only foster a deeper understanding of smoking behaviour and patterns, but also accelerate the development of more efficacious cessation applications [8].

Ethical statement

All the experimental procedures and the set-up were performed in accordance with the protocol detailed in [9] . We have obtained formal approval of this protocol from the ethical committee of UZ Leuven, and the approval number is S60078. All the smoker volunteers who participated in our study signed an informed consent form in which they allowed the researchers to collect and use their personal data for research purposes.

References

1. World Health Organization. (2019). WHO report on the global tobacco epidemic 2019: Offer help to quit tobacco use.

2. Tyas, S. L., & Pederson, L. L. (1998). Psychosocial factors related to adolescent smoking: a critical review of the literature. *Tobacco control* **7(4)**: 409-420.

3. Onnela, J. P., & Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* **41(7)**: 1691.

4. Smets, A., Lievens, B., & D’Hauwers, R. (2018). Context-Aware Experience Sampling Method to Understand Human Behavior in a Smart City: a Case Study. *Measuring Behavior 2018*.

5. Zhai, D., Schiavone, G., De Raedt, W., & Van Hoof, C. (2019). Investigation of Heart Rate Changes before and during/after Smoking Events in Free Living Conditions. *2019 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)* (pp. 137-140).

6. Smets, E., Velazquez, E. R., Schiavone, G., Chakroun, I., D’Hondt, E., De Raedt, W., ... & Van Diest, I. (2018). Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ digital medicine* **1(1)**: 1-10.

7. Simões-Capela, N., Schiavone, G., De Raedt, W., Vrieze, E., & Van Hoof, C. (2019). Toward Quantifying the Psychopathology of Eating Disorders From the Autonomic Nervous System Perspective: A Methodological Approach. *Frontiers in Neuroscience* **13**.

8. Vinci, C., Haslam, A., Lam, C. Y., Kumar, S., & Wetter, D. W. (2018). The use of ambulatory assessment in smoking cessation. *Addictive behaviors* **83**: 18-24.

9. Zhai, D., Schiavone, G., Van Diest, I., Vrieze, E., DeRaedt, W., & Van Hoof, C. (2019). Ambulatory Smoking Habits Investigation based on Physiology and Context (ASSIST) using wearable sensors and mobile phones: protocol for an observational study. *BMJ open* **9(9)**: e028284.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Start Making Sense: Predicting confidence in virtual human interactions using biometric signals.

S. Dalzel-Job[1], R.L. Hill[1], R. Petrick[2]

1 Department of Informatics, University of Edinburgh, Edinburgh, Scotland. sdalzel@ed.ac.uk, r.l.hill@ed.ac.uk

2 School of Mathematical & Computer Sciences, Heriot-Watt University, Edinburgh, Scotland. R.Petrick@hw.ac.uk

Introduction and Aims

This project investigates the use of biometric data to predict confidence levels during task-focused interaction between humans and virtual humans. The project comprises of two main studies, the first of which examines the relationship between biometric signals – galvanic skin response (GSR), heart rate, facial expression and eye movements – and self-report levels of confidence during a task-oriented interaction between a human and a virtual human. Through the manipulation of the feedback and task demands, participants were exposed to unexpected situations and varying levels of ambiguity, resulting in a measurable range of perceived confidence as well as more implicit biometric and behavioural indicators of confidence and success. The second study utilises the paradigm and results from Experiment 1 to train an AI instruction giver to identify instances where behavioural and biometric feedback from a human signal low confidence, enabling it to modify or supplement its instructions accordingly. To ensure that the AI is acting in a useful way, and that the experimental manipulation and behavioural demonstrations of confidence are valid, the participant judges the perceived success of the interaction, as well as their trust in the AI under varying levels of feedback. This paradigm can then be adapted for use across a wide range of situations and scenarios; from interactions with virtual human avatars or agents via AR, VR, desktop or mobile devices, to fully embodied conversational agents or robots, this paradigm will enable a successful, smooth interactions between humans and Ais.

Background

Virtual humans – whether computer-controlled agents or human-controlled avatars – are widely used during online interactions. Not only are they utilised during social interaction (e.g. gaming), but also in important joint-action or task-oriented communication. Historically, virtual humans have been used in support and health [1-5], as well as in areas such as teaching and training [6-10]. To date, there has been a wealth of research into how virtual humans should behave during interactions with users in order to maximise success [11-16]. Our previous research has discovered that the optimum behaviour of a virtual human, specifically its eye movements, varies depending on the purpose of the interaction [15, 16]. This study aims to expand and extend these findings by investigating which combination of behaviours maximise positive perceptions of a virtual agent, as well as maximising any given task performance, with the aim of developing trustworthy, likeable and useful virtual humans. Furthermore, it aims to develop a system that can utilise real-time non-verbal feedback from a user to indicate confusion or occasions of uncertainty. This will enable the system to supplement or alter instructions to maximise the possibility of a smooth, successful interaction.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Methods

Experiment 1

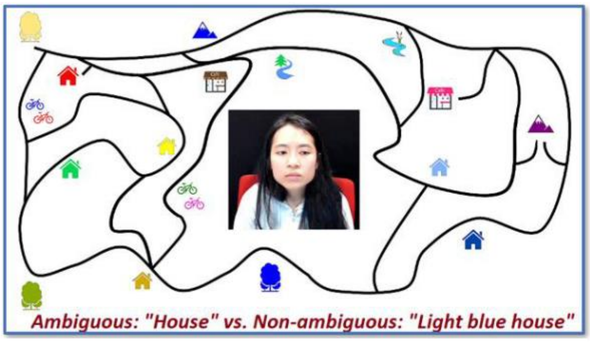


Figure 1:The human follower looks at the Landmark when she has located it.

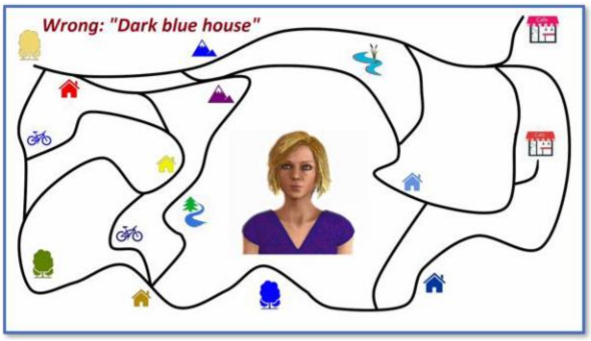


Figure 2: The speaker guides both a human and a virtual human.

A human speaker guides a listener through the map to locate a landmark. Each speaker guides both a human listener and a virtual human listener (Figure 1 and Figure 2, respectively). The listener may indicate that the target has been successfully located by looking at it (correct condition) or that an incorrect target was chosen (wrong condition). It can be seen in Figure 2 that the listener has not found the correct target, the dark blue house, but is instead looking at the other side of the map. Furthermore, the speaker may have insufficient information to uniquely identify a target, resulting in them having to choose between multiple possible choices to guide the listener towards. In Figure 1, for example, the target landmark given to the speaker may be ‘House’, but there is more than one house on the map, leading to an ambiguous situation where the speaker has to decide which house to guide the listener to. These manipulations deliberately generate situations of uncertainty or ambiguity. The speaker believes that the virtual human listener is either controlled by a human (avatar condition) or by a computer “AI” (agent condition). In all conditions, the listener is actually a video, and is non-interactive, although results from the study suggest that this was not identified by the participants, and that they treated the listener like an interactive virtual human.

The amount of time the listener looks at the user is also systematically varied. It has been found that the optimum amount of looking by a virtual human at a human can vary, depending on the purpose of the interaction [16]. The previous research examined the impact of looking at the user 0%, 25%, 75% and 100% of the interaction; this was adapted slightly in the current study, with the listener looking at the user during either 0%, 30% or 70% of the interaction. This was intended to identify more about the effects on the user of the listener’s looking behaviour; some research has suggested that there may be a threshold amount of looking, at around 70%, at which point the social impact on the user is at its highest [16].

Measures

The users respond to questions relating to their confidence in their instructions, and they also report if they were confident that the listener found the target. They are also asked questions relating to their social perceptions of the listener.

The eye movements of the users are recorded using an Eyetribe remote eye tracker [17]. This particular device was chosen for its non-invasive nature, and its portability. Galvanic skin response (GSR) is measured using two Shimmer sensors attached to the tips of two fingers, and another sensor attached to a third finger to detect changes in heart rate [18]. Changes in GSR and heart rate can indicate a change in arousal; these changes could be positive or negative in valence (it could indicate joy or anger, happiness or frustration, for example, but in isolation the GSR data does not allow you to identify which). The facial expressions of the users are also detected during the interaction [20], allowing for a fuller understanding of the nature of the arousal. An indication of an angry facial expression in conjunction with a large GSR peak, for example, is more informative about the effect of any stimuli on a user than the GSR alone.

iMotions is software that allows the presentation of the stimuli and collation, time-stamping and processing of all the behavioural, biometric and survey data in preparation for analysis [19]. Examining these behaviours and biometrics together rather than independently allows the identification of the behaviour, or combination of behaviours associated with varying levels of confidence, as indicated by their relationship with the responses to the survey. This combination of objective and subjective measures enables us to begin to build a model of how users respond and adjust to different types of feedback, and to use this information to design behaviourally appropriate virtual humans, responding in real-time to non-verbal feedback that may indicate anything other than a smooth interaction.

Experiment 2

The non-verbal behaviours identified in Experiment 1, which are associated with confidence – in self and in the interlocutor – can be used by a planning system to identify instances of confusion, or where the user may require extra information. The facial expressions and eye movements are fed into the planning system in real-time, and upon breaching a pre-specified threshold, the system responds accordingly, signalling the system to provide extra information where low confidence or confusion is indicated, and continuing without additional clarification when biometric responses suggest that the interaction is going well.

Outcomes and Applications

This research can be applied to several different situations: wherever it is desirable for a system to respond in real-time to a user’s emotional state, the system can be trained to identify signals of confusion or uncertainty and respond immediately to remedy the situation. With the advent of more mobile eye trackers and the increasing popularity and affordability of smart watches, as well as other devices that already detect heart rate, which could potentially be developed to identify changes in GSR, this paradigm presents the possibility of interactive systems responding in real-time to behavioural and biometric cues provided via our everyday devices. Interactive and ubiquitous, virtual companions, advisors, teachers, coaches or even mediators could soon be available to provide customisable, interactive, responsive and truly trustworthy, effective virtual humans.

Ethical Statement

Ethics approval for this study was granted by the Informatics Ethics Committee, University of Edinburgh (rt #3690),

References

1. Robillard, G., et al., (2010). Using virtual humans to alleviate social anxiety: preliminary report from a comparative outcome study. *Stud Health Technol Inform* **154**: 57-60.

2. Kang, S.-H., et al. (2008). Does the contingency of agents' nonverbal feedback affect users' social anxiety? in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1. International Foundation for Autonomous Agents and Multiagent Systems*.

3. Lok, B., (2006). Teaching communication skills with virtual humans. *IEEE Computer Graphics and Applications* **26(3)**: 10-13.

4. Yuen, E.K., et al., (2013). Treatment of social anxiety disorder using online virtual environments in second life. *Behavior therapy* 44(1): p. 51-61.

5. Kenny, P., et al. (2007). Virtual patients for clinical therapist skills training. in *International Workshop on Intelligent Virtual Agents*. Springer.

6. Kim, Y., J. Thayne, and Q. Wei (2017) An embodied agent helps anxious students in mathematics learning. *Educational Technology Research and Development* **65**(1): 219-235.

7. Johnson, W.L. and J. Rickel, (1997) Steve: An Animated Pedagogical Agent for Procedural Training in Virtual Environments. *Sigart Bulletin* **8(1-4)**: p. 12-16.

8. Johnson, W.L., J.W. Rickel, and J.C. Lester (2000). Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. *International Journal of Artificial Intelligence in Education* **11**: p. 47-78.

9. Johnson, W.L., et al. (2003). Evolution of User Interaction: The Case of Agent Adele in *8th International Conference on Intelligent user interfaces* (Miami, Florida, USA).

10. Rickel, J. and W.L. Johnson (1999). Virtual humans for team training in virtual reality in *Proceedings of the ninth international conference on artificial intelligence in education*.

11. Cassell, J., J. Sullivan, and S.e. Prevost (1999). Embodied Conversational Agents, ed. M. Cambridge.: MIT Press.

12. Cassell, J. and K.R. Thorisson, (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence* **13(4-5)**: 519-538.

13. Dalzel-Job, S., C. Nicol, and J. Oberlander (2008). Comparing behavioural and self-report measures of engagement with an embodied conversational agent: A first report on eye tracking in Second Life. in *The 2008 Symposium on Eye Tracking Research & Applications* (Savannah, Georgia)

14. Dalzel-Job, S., J. Oberlander, and T.J. Smith (2011) Contested staring: issues and the use of mutual gaze as an on-line measure of social presence.

15. Dalzel-Job, S., J. Oberlander, and T.J. Smith (2011). Don't Look Now: The relationship between mutual gaze, task performance and staring in Second Life. in *The 33rd Annual Conference of the Cognitive Science* (Boston, Massachusetts, USA).

16. Dalzel-Job, S., (2015). Social interaction in virtual environments: the relationship between mutual gaze, task performance and social presence.

17. Eyetribe. *The Eyetribe*. [cited 2020 30.01.2020]; Available from: https://theeyetribe.com/theeyetribe.com/about/index.html.

18. *Shimmer*. [cited 2020 07/02/2020]; Available from: https://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor.

19. *iMotions*. [cited 2020 07/02/2020]; Available from: https://imotions.com/.

20. *Affectiva*. [cited 2020 07/02/2020]; Available from: https://www.affectiva.com/.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Measuring Behavior in Counseling Clinic Waiting Areas

Lauralee Wikkerink, Shireen Kanakri[1]

1 Department of Construction Management and Interior Design, Ball State University, Muncie, Indiana. smkanakri@bsu.edu

Abstract

The built environment can have an impact of patient mental health and overall well-being, especially when it comes to healthcare environments. Often overlooked, this research investigates the behaviors associated with counseling clinic waiting areas and how these behaviors relate to the occupants' perceptions of their own behaviors. Knowledge from this study will help designers and clinicians develop waiting areas that are conducive to individuals' entire well-being: mental, physical, spiritual, and emotional. The study consists of two forms of data collection: observations and surveys. Observations are used to study client behavior and activity in the waiting room and the surveys are to gather client perceptions about the built environment. Findings from this study reveal the highest frequency of behaviors of counseling clients (unable to relax/fidgety) and activity (use of phone/tablet) to reveal helpful changes to the design of counseling clinic waiting rooms. Based on the knowledge gained from this study, the researcher recommends several design changes and inclusions such as limiting cell phone use, providing beverage stations, and creating an environment conducive to private meditation and self-care.

Introduction

According to the National Institute of Mental Health, an estimated 16.2 million adults in the United States has had at least one major depressive episode and an estimated 44 percent received care from a health professional. The need for counseling is tremendous and counseling practices require research to determine design choices that are most beneficial for their clients. The counseling clinic waiting room is unique: unlike many waiting rooms, counseling client visits are often recurring and anxiety levels—already present in most healthcare waiting rooms—may be heightened by depression, stress, nervousness, frustration, fear, etc. as well as what is about to be experienced and discussed in the therapy session.

Natural elements including live plants have been shown to be beneficial to individuals in the built environment including the hospital, home, and workplace. The work of Beukeboom et al. [1] is widely cited for the stress-reducing effects of plants, both natural and artificial. When combining this idea with the counseling clinic waiting room, knowledge from this study will help designers and clinicians develop waiting areas that are conducive to individuals' entire well-being: mental, physical, spiritual, and emotional.

Methods

Research indicates that the built environment can affect behaviors [2]. It is hypothesized that clients' observed behaviors in the counseling clinic waiting room will correspond to their survey responses about the environment, and their behaviors and comfort. The population under study consisted of people who spend time in a couseling clinic environment. Specific pathologies of this group were not considered as the focus of the study was the clinic interior itself. In order to fuly analyze the behaviors of individuals inside a couseling clinic waiting area, in-person observations where conducted and personal client surveys were distributed with regards to ethical considerations put in place by the Institutional Review Board. Surveys where gathered to determine how the occupant perceived their own behaviors. The survey sought to determine the most influential elements of the waiting room environment including the significance of plants or natural elements. In this way, both quantitative and qualitative data was collected.

Surveys gathered this information:

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

- Age
- Gender
- Activities while waiting (Reading books/magazines; Using smart phone/tablet; Talking with others (strangers); Talking with receptionist; Playing with child; Walk; Use restroom; Other (please include)
- Behaviors/emotions: Excited; Anxious; Relaxed; Stressed; Unable to relax; Nervous; Sad; Angry; Worried; Thankful; Bored; Optimistic; Other (please write current emotions here)
- Preferences regarding the environment with five-point Likert-scale responses (Agree; Slightly agree; Neutral; Slightly disagree; Disagree):
    - The amount of privacy I have in the waiting room is adequate.
    - The waiting room is neat and clean.
    - The artwork and décor are suitable to me.
    - It is important to me to have plants in the waiting room.
    - The waiting room environment is soothing.
    - The amenities in the waiting room (Wi-Fi, magazines, beverages, etc.) are convenient for me.
    - I feel comfortable in this waiting room.
    - I receive quality care at this clinic.
    - The waiting room is attractive.
    - I prefer live plants to artificial plants in the waiting room.
    - I prefer a waiting room with natural light.
- Fill in the blank: I wish this waiting room ____.
- Open space for comments/suggestions regarding the waiting room environment. Survey data will be gathered and coded. SPSS will be used to analyze descriptive statistics. The two environmental conditions will be compared to one another.

In-person observations of waiting room occupants allow for real-time analysis of how they act within the waiting space including use of their time, perceived emotions, and use of amenities offered in the environment. The researcher completed observations in 5 hour increments for a total of 40 hours and used The Observer XT Noldus software to record observation data. The following activity frequencies were collected:

- Read
- Using phone/tablet
- Talk with friends/family
- Talk with others
- Talk with staff
- Play with child

- Walk
- Use restroom
- Drink/get beverage
- Sit
- Other activity

The following behavior frequency was also collected for analysis during observations:

- Enter
- Exit
- Talkative
- Unable to relax/fidgety
- Anxious/seeks social support
- Happy/pleasant
- Other behavior

Results

The researcher observed ten activities during the observation period which were: Read, Use phone/tablet, Talk with friends/family, Talk with others, Talk with staff, Play with child, Walk, Use restroom, Drink/get beverage, Sit, Other activity. Other activities observed were: fill out paperwork, complete survey, read about survey, stand, put phone away, take off jacket, gargle water, blow nose, write/journal, file/examine nails, write check, stretch, throw out trash, look through provided reading materials/put magazine back, tie shoes, spill beverage, move to different seat, watch tv, work, water plants. Activities were distinguished from behaviors as being a movement or action separate from an emotion with the exception of enter/exit which was included as a behavior for ease of use in the observational software.

All the survey respondents (100%) agreed that the waiting room is neat and clean. Eleven (91.67%) of the respondents agreed that the artwork and décor is suitable; the client feels comfortable in the waiting room; and they receive quality care. Ten respondents (83.33%) agreed that the waiting room is attractive. Ten respondents (83.33%) agreed that the waiting rooms amenities were convenient for them (one respondent did not complete the back portion of the survey). Eight respondents (61.54%) agreed that waiting room's privacy is adequate with 1 respondent (7.69%) disagreeing. Eleven respondents (91.67%) agreed that they receive quality care at the clinic with one respondent (8.33%) slightly agreeing.

Regarding the waiting room's natural environment, four respondents (30.77%) agreed that it is important to have plants in the waiting room. Two respondents (15.38%) slightly agreed, and 7 (53.85%) were neutral. Regarding the preference between live plants to artificial plants in the waiting room, 3 (25.0%) agreed, 5 (41.67%) slightly agreed, and 4 (33.3%) were neutral. Regarding preferring a waiting room with natural light, 4 (33.33%) agreed, 4 (33.33%) slightly agreed, and 4 (33.33%) were neutral.

Discussion

This research sought to determine how client perceptions of the waiting room environment differ from their behaviors; if clients perceive the inclusion of plants as beneficial to the waiting area; and what amenities in the waiting area are important to waiting individuals (versus what amenities are used). Client perceptions, as indicated through survey responses, indicated a positive perception of the waiting room and the quality of care they receive. The inclusion of plants could not be deemed beneficial, as there was no test site to compare the results. Lastly, observations and client survey responses indicated a high use of mobile devices while observations also concluded a high use of the provided amenity/beverage station.

The suvey responses in this study indicated a lower level of importance of plants/nature in the waiting area than in other studies. This difference could be indicative of the short time individuals are in the counseling clinic waiting room and/or the lush nature surrounding the entrance to the site location.

The study’s findings of high frequency of cell phone use in the waiting area and the knowledge that overuse of devices can be detrimental to mental health, prompts the design of a waiting room that limits or prohibits cell phone use. Although the use of mobile devices may be a positive distraction for some individuals, it is detrimental to others. If the use of mobile devices is contributing to the inability to relax, it may benefit both these clients and their counselors to change the way the individual spends his or her time prior to the counseling appointment. Further comparative studies are needed to see the direct impact of specific environmental factors and how these relate to a given behavior.

References

1. Beukeboom, C. J., Langeveld, D., & Tanja-Dijkstra, K. (2012). Stress reducing effects of real and artificial nature in a hospital waiting room. *Journal of Alternative and Complementary Medicine* **18(4)**: 329-333.

2. Berke, E. M., & Vernez-Moudon, A. (2014). Built environment change: A framework to support health-enhancing behaviour through environmental policy and health research. *Journal of Epidemiology and Community Health* **68(6)**: 586-590.

Measuring Behavior of Low-Vision Populations Using Virtual Reality

Lauren Ashley Hughes, Shireen Kanakri[1]

1 Department of Construction Management and Interior Design, Ball State University, Muncie, Indiana. smkanakri@bsu.edu

Abstract

While the low-vision population in American continues to increase, few empirical studies have been completed investigating how environmental factors affect a low-vision person's ability to accurately perceive the interior environment. This study uses quantitative research methods to understand the critical relationship between contrast levels within the built environment and the ability of the low-vision and normal-sighted population to accurately perceive that environment, while also investigating the environmental factor of contrast and how varying levels of contrast within interior spaces affect the behavior of low-vision participants within the interior environment. This study consists of two phases. The first phase was a survey designed to evaluate the preferences and potential behavioral impacts caused by perceived contrast levels present within the interior environment on people with all visual abilities. The second stage of the study involves observing and videotaping low-vision and normal-sighted participants within a laboratory setting and documenting preferred contrast levels and changes in behavior under different contrast scenarios. The findings of this study demonstrate the preference among normal-sighted and low-vision participants for high-contrast environments. High-contrast between the wall and floor surfaces are helpful cues to the low-vision population's ability to perceive the environment accurately. This researcher also found that low-vision participants exhibited fewer behaviors in high contrast environments than in medium and low contrast environments. The researcher recommends that high-contrast environments are an appropriate inclusive design measure that interior designers should consider when designing the built interior, and that they decrease the risk of behavioral reactions exhibited within an interior space.

Introduction

Inclusive Design has been in the forefront as designers seek to create spaces that everyone can use. While this can mean designing for children, adults, individuals with physical impairments, or learning disabilities, the focus for this research is individuals with low-vision. Low-vision is defined as a visual impairment that makes performing everyday activities difficult and that cannot be corrected with glasses or surgery [1]. In order to gain a better understanding of this population, it is crucial to investigate the relationship between contrast levels of the built environment and the ability of low-vision and normal-sighted populations to navigate and perceive that environment. The question becomes, how do researchers gain a better understanding of the relationship between contrast levels in an interior space and human behavior while inside that environment? The number of people with visual limitations is expected to increase by 6.31 million by the year 2020 due to the aging population of baby boomers [2] plus the addition of those who are genetically predisposed to have low-vision. This drastic increase elevates the need for consideration in the design of interior spaces to allow this combined population of people to safely and independently navigate and participate with the built environment.

Methods

In order to compare results from low-vision and normal-vision people, the researcher's goal was to collect at least 100 participants identifying with each group for the first phase of the research. This large number would allow for more data to analyze before prepping for the next research phase where a physical presence would be required. To accomplish this, emails were composed detailing the objectives of the research, what would qualify someone to participate, and what benefits or risks might be associated. Per ethical considerations, informed consent was also discussed and a link to complete the survey was provided.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Low-vision individuals were recruited via email lists from the National Research and Training Center for Blindess and Low Vision at Mississippi State University as well as the National Federation for the Blind. By agreeing to distribute the survey request through their network, the researcher was able to access the population directly impacted by the research. As for the normal-vision population, social media outlets like Facebook and Twitter allowed for distribution of the survey in which people could send the link to others who might be willing to participate.

Investigating the impact of the environment on human behavior was divided into two key parts, evaluation of preferences for research subjects and then direct observation of the subjects within an environment designed around those preferences. Prior research has been lacking in this field surrounding the interior environment and its occupants, but it has indicated that the environment can impact an individuals willingness to participate in that environment. For many, safety is a concern in research surrounding a low-vision population. This research effort was tasked with highlighting specific risk factors for this population so that future design of interior spaces might be more inclusive of individuals with low-vision, but it was completed in a way that minimized risk to participants.

The initial phase of this research study involved distribution of an online survey to a mixed population of people with normal vision and people with low-vision. This method of an online survey was decided upon for safety and convenience factors. People would be able to complete the survey in their own home or workplace without the risks associated with going to a new location to complete the survey in physical form. In addition, ease of access would make people more likely to participate in the survey as there would be less barriers to entry.

At first glance, observing low-vision and normal vision in different environments poses several research design challenges. The researcher is posed with the challenge of testing subjects in completely different interior spaces while also being aware of potential risks associated swith a low-vision population. A way to get around these challenges was to utilize virtual reality technology in combination with one physical space. A single room was modeled virtually while finish materials of the room were changed out. In this way, the researcher was able to document preferred contrast levels between floor, base, and wall materials as well as the behaviors exhibited by participants while wearing the VR equipment.

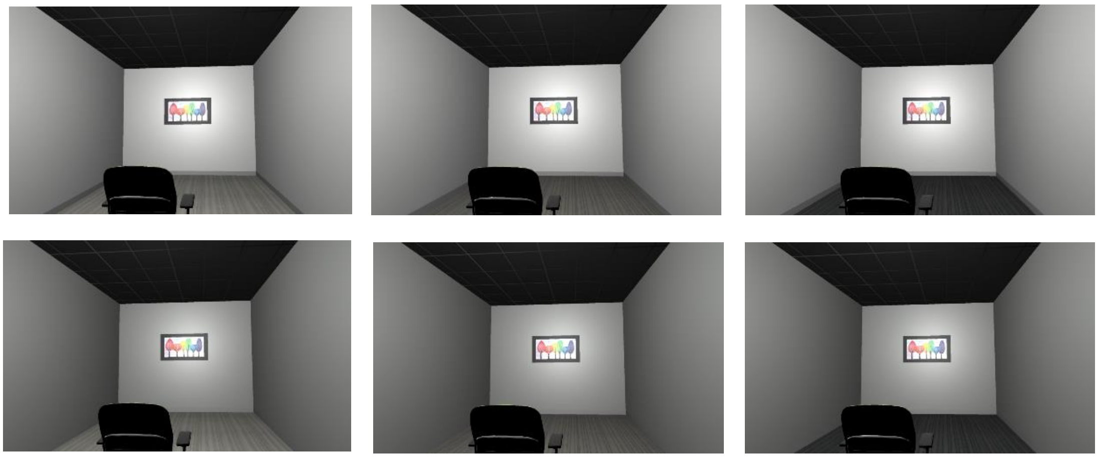


Figure 1. Illustrates the study environments with various contrast changes as they were used in the VR equipment.

Since this study aimed to pinpoint the preferred level of contrast desired by people with low-vision in wall to wall and floor to floor transitions, nine different study environments were created to obtain feedback from research participants. Each study environment was classified as a high, medium or low-contrast environment. These classifications were developed by collecting the light reflectance values of the wall and floor materials and using subtraction to determine the difference between the light reflectance values [3]. The following number ranges were used to determine the classification of low, medium and high contrast environments: 1) High (30 –

45); 2) Medium (15 – 29.9); and 3) Low (0 – 14.9). Scenes that feature a light wall and floor material were categorized as a low-contrast environment while scenes featuring a light wall with a dark floor were categorized as a high-contrast environment. Scenes with a combination of light wall materials and medium floor materials were categorized as a medium contrast environment.

Equipment

Video recording. Per ethical guidelines, permission to video record the study was requested prior to conducting the observation. If the participant was not comfortable with being recorded, then the researcher took notes diligently throughout the study making sure to include detailed descriptions of what was observed and discussed. If permission to video-record the observation was granted, then the observation was recorded in its entirety, and the video was transcribed within a week from the time the study was completed. Videos were then deleted from the computer.

The researcher recorded the observations with a HP Wide Vision HD Camera with integrated dual array digital microphone that is integrated into a laptop. The observation was recorded in its entirety so the researcher could watch the videos at a later time to verify the behavior frequencies recorded were accurate. This contributes to the validation of the study.

Virtual reality equipment. The researcher chose to test the hypothesis for this study by utilizing a virtual reality environment so that the research participant could experience different materials within the same room. A virtual reality environment allowed for a fast and efficient method of interchanging finish materials and still allowing the participant to experience the materials in a more realistic and holistic setting. The use of the virtual reality backpack computer allows for the participant to move around the space freely without fear of tripping over any chords.

HP Z VR backpack computer. For this observational study, the researcher utilized an HP Z Virtual Reality computer. This computer can be mounted in a stationary dock and connected to the VR goggle headset by a twelve-foot-long connection cord, or it can be attached to a backpack for a free roam experience [2]. Portable batteries are stored in a holster in the backpack to provide power to the computer in the backpack without any needed power cords (Hewlett Packard, 2019). The researcher chose to use the backpack computer option because it would reduce the risk of participants tripping on a long cord. A research assistant held the backpack and followed the research participant so that they would have a more authentic experience while wearing the headset and traveling through the virtual reality environment without the weight and restriction of motion they may have exhibited if wearing the backpack. This also allowed for the research participant to provide assistance to the participant if needed.

The virtual reality environment is powered through the use of SteamVR [2]. Once the application file is selected and opened, the researcher can toggle through each of the nine study environments by using the right and left arrows on the keyboard. A monitor was connected to the computer so that the researcher could see exactly what the research participant was seeing inside of the virtual reality headset.

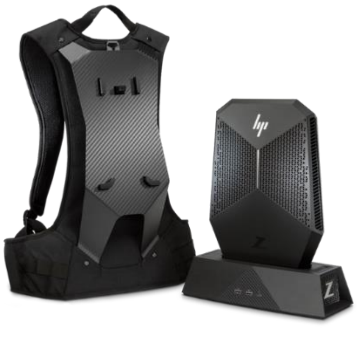


Figure 2. Illustrates HP Z VR backpack computer and docking station.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

HTC Vive virtual reality headset system. The HTC Vive virtual reality headset system includes a headset, two room boundary sensors, and hand controls [4]. The headset connects to three ports at the top of the computer system. It is powered by the portable batteries located within the backpack. Due to the simplistic nature of the study environment used for this research, the hand controls were not needed and therefore were not utilized.

Two room sensors were placed at approximately eight feet above the finished floor in two opposite corners of the room, and the room dimensions were synchronized with the SteamVR program. This allowed participants to move around the room naturally without the danger of running into any walls. If a participant reaches the sensor's established boundaries, the room scene goes away and a blue grid appears warning the participant to stop.

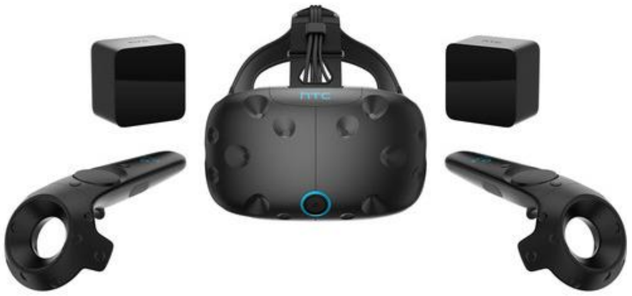


Figure 3. Illustrates HTC Vive's virtual reality headset system used in this study.

Results/Discussion

This study investigated the preferred level of contrast between floor and wall finishes within a space for both normal-sighted and low-vision participants. The findings suggest that a high contrast environment is preferred by people of all visual abilities. While a high contrast environment can be helpful for a person with low-vision's ability to perceive the environment accurately, designers should be careful when specifying dark floor colors, as they may cause people with low-vision to have anxiety toward walking on those floors. This validates a study conducted by Hughes, Carroll & Miller [5].

This study also investigated the effect of contrast levels within the interior environment on a person's behavior within that environment. Participants with severe/profound vision loss exhibited a higher number of behavioral reactions to contrast levels within the interior environment than normal-sighted participants who exhibited no behavioral reactions. Low-vision participants exhibited the highest number of behavioral reactions in medium contrast environments and the lowest number of behavioral reactions in high contrast environments. Therefore, high contrast environments proved to be the easiest to navigate while triggering the least number of behavioral reactions to contrast levels within the environment.

References

1. National Eye Institute. (2019, July). Low-Vision. Retrieved October 16, 2019, from https://nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/low-vision.

2. Akpek, E. K., Smith, R. A. (2013). Overview of Age Related Ocular Conditions. The American Journal of Managed Care. 19. 5. 67-75.

3. Schambureck, E. & Parkinson, S. (2018) Design for Sight: A Typology System for Low-Vision Design Factors. Journal of Interior Design, 43(2), 33–54.

4. HTC Corporation. (2019). Vive VR. Retrieved October 21, 2019 from https://www.vive.com/us/product/vive-virtual-reality-system/.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

5. Hughes, A., Carroll, R., & Miller, B. (2019) One Step Closer to Inclusive Design: Designing for Clients with Low Vision. Interior Design Educators International Conference, March 6-9, Charlotte, NC.

Identifying Canine Posture from a Wearable Sensor: Application of Cross-Disciplinary Methods to Companion Animal Monitoring

Jack O’Sullivan, Cameron Smith, Lucy Asher

School of Natural and Environmental Sciences, Newcastle University, Newcastle, United Kingdom.
J.O'Sullivan4@newcastle.ac.uk

Abstract

Companion animals present unique challenges and opportunities in the application of remote monitoring and biologging technologies for health, welfare and behavioural assessment. Developing effective methods of monitoring, while accounting for the unique constraints imposed, necessitates compromise. Accelerometers in particular have been utilised extensively within human and veterinary medicine, and behavioural ecology. However, often these devices require proprietary software, present measures without transparent validation, or suffer from insufficient validation samples. As posture constrains the behavioural repertoire of an individual it can be used within behaviour recognition algorithms. The assessment of postural state is common in human and animal applications for this reason, but differences in attachment method and in species’ body conformation makes it difficult to transfer existing methods to new species or contexts. Two strategies for postural recognition were tested in dogs; the recognition of transition events, and the direct recognition of postural states. The use of tri-axial accelerometer data for posture recognition was explored using a sample of 20, kennel housed, Labradors with ~10 hours of annotated accelerometer and video recordings. Relevant features, sourced from multi-disciplinary literature, were calculated from the accelerometers. A series of binary classification tasks were formulated to explore the ease with which each posture or postural transition could be identified from accelerometer features. Linear Discriminant Analysis was used to explore the classification of postures using different feature permutations. Direct recognition of postural states performed better than classifying based on transitions between postures. The most promising approach involved identification of the “Standing” posture, followed by binary separation of “Sitting” and “Lying” postures. “Stationary-Standing” and “Locomoting-Standing” were also classified. We were able to achieve acceptable levels of posture classification in dogs using methods which would be fast to compute, however this may be further refined with more sophisticated classification methodologies or additional features.

Introduction

The remote monitoring of health and welfare has seen a continued boom in popularity over recent years, a trend that appears set to continue both in human and non-human animal contexts [1, 2, 3]. However, often the fields of behavioural ecology, human medicine, and veterinary medicine develop techniques relevant to the remote monitoring of their subjects independently from one another, likely due to a perceived difference in contexts within which the monitoring techniques are applied.

Companion animals present unique challenges in the application of remote monitoring and biologging technologies, such as accounting for owner perception and habits, or the large variability in breed activity and conformation. However, the unique situation of these animals also presents a valuable potential sample for the implementation of cross-disciplinary ideas while simultaneously addressing the previously mentioned domestic constraints, and constraints present within other contexts. This has been addressed in the past with veterinary applications of remote monitoring technologies which frequently draw on techniques originating within the human medicine literature. We have posed the aims and associated issues of the successful remote monitoring of animal health and welfare with regard to the unique behavioural repertoire of domestic dogs. This forms the core of a “behavioural informatics” approach to remote animal monitoring that could present measures of highly descriptive behavioural information beyond that of activity level measures currently extrapolated from to form conclusions.

Volume 1 of the Proceedings of the joint 12th International Conference on Methods and Techniques in Behavioral Research and 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18, 2021. www.measuringbehavior.org.
Editors: Andrew Spink, Jarosław Barski, Anne-Marie Brouwer, Gernot Riedel, & Annesha Sil (2020).
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

We aimed to perform an initial exploration of an approach to address remote animal monitoring, using a single tri-axial accelerometer sensor, with consideration of past literature across the spectrum of related disciplines. The predominant focus of the methodology is on extracting valuable behavioural information that could potentially provide further details on the current health and welfare of the focal dogs. To achieve this computationally intensive and complex classification solutions were avoided to allow optimal processing speed through multiple permutations and to ensure process transparency. The features used were based on the assessment and implementation of similar methods as they appear in both human and animal literature with an aim to identify a satisfactory path forward to be developed and built upon further.

Methods

Acceleration data were collected over a single day (~10 hours) from a single, collar-attached, tri-axial accelerometer (AX3, Axivity, York, UK: 23.0 x 32.5 x 7.6mm), initially aligned ventrally. The sample consisted of 20 (9 male, 11 female), labradors, all aged between 3 and 6 years of age. The accelerometer sampling frequency was 100Hz and data were resampled prior to processing to ensure this was consistent. Videos were taken of the dogs during their daily walk and periods where dogs were able to access a communal paddock area.

Drawing upon both the behavioural ecology, and human medicine literature we proposed the identification of postural state as an initial stage of behavioural classification. This allows the reduction of potentially meaningful behaviours, possible for the focal individual to perform, within a time period and simplifies the construction of classification tasks. Following from this, the further identification of locomotion-related behaviours could further help to isolate meaningful behavioural states applicable to health and welfare.

To achieve the initial postural identification two distinct approaches were used; the identification of the transitional periods between postural states, and the identification of postural states themselves. Three postures were identified from the video data and annotated using ELAN Version 5.7 software [4]. Accelerometer data were then synchronised with the video data and each second of acceleration was labelled for both postural state and postural transition, if appropriate. Tri-axial acceleration data were separated into the static and dynamic components of acceleration using a 4th order, zero phase, low pass Butterworth filter [5]. The static components were retained and were converted into a suite of unique features, informed by the literature, across each of the 3 axes and the combined measure of the signal vector magnitude (VM3) of the three axes. This resulted in 51 total features for the classification of postural transitions. For the classification of postural states a suite of 110 features were calculated.

In the detection of transitional events the shape of the acceleration as it progresses between postures has been previously used successfully in human and dog-related literature [6, 7]. As such a 100 coefficient Estimated Cumulative Difference Function (ECDF) representation of each transitional window was calculated and included as additional features for the relevant classification tasks. The number of ECDF coefficients used was optimised for performance during cross validation.

To identify locomotion the same methods were used as in the classification of postural state, but applied to periods of locomotion, considered a singular category as differing gaits were not labelled due to the increased training requirements to ensure observer accuracy. Periods not labelled as standing were excluded to alleviate the degree of data imbalance including non-standing stationary periods would introduce. Additional features relating to the rhythmicity of locomotion were calculated using a Fast Fourier Transformation (FFT) of each window of dynamic component data.

The identification of posture, locomotion and postural transition events were each proposed as binary classification tasks. Postural transition event identification was constructed in a series of one vs all tasks (i.e. Stand to Sit transitions identified from a generic class of all other classes) and included the detection of any transitional event from among an equal number of randomly selected non-transitional data windows. Postural state recognition was also arranged in this way, however, also included were the three possible binary classification tasks between each of the three possible postural states (i.e. Standing vs Sitting). Locomotion detection was constructed as a binary decision between Standing-Locomotion and Standing-Stationary.

A linear discriminant analysis (LDA) classifier was selected over more complex methods due to the computational simplicity and rapid implementation of the technique across the multitude of classification tasks presented. Features were assessed for collinearity per task by stepwise elimination of the highest Variance Influence Factor (VIF), until a threshold of VIF= 5 was reached. Remaining features were further selected using 4 simplistic methods of feature selection; Manually selected, Forward Stepwise, Backward Stepwise, and Correlation-based Feature Selection (CFS) [8]. Feature selection occurred within a 5-fold cross-validation.

The performance of models was assessed using both the Area Under the ROC Curve (AUC), and the F-Score. This provided an overall view of performance for rapid assessment. The frequency of selection for each feature during cross validation was used as a simplistic measure of importance. Inclusion in greater than 50% of the cross validation models, for the best performing feature selection method, resulted in the features inclusion in a final validation assessment performed on a portion of, previously excluded, data to which the models processed so far were naïve. The same performance measures were used for the assessment of the validation model performance.

Results

Measures of performance are given for validation models with feature lists and other model variables selected within the prior cross-validation. Validation data was withheld prior to cross-validation to ensure model naivety.

For postural transition event identification, from among randomly selected non-transitional data, an LDA model including 28 coefficient ECDF representations of both the Z axis and VM3 data was selected. The AUC of the resultant LDA model was 0.560, and the F score was 0.504.

Attempting to classify specific transitional events from among other transitional events, excluding non-transitional data, performed approximately as well, or worse, than chance across the various configurations of the classification process. The Lie to Stand transition model, with an additional 9 VM3 ECDF coefficients, gave an AUC of 0.463 and an F score of 0.256. The Sit to Stand transition model, with an additional 53 VM3 ECDF coefficients, gave an AUC of 0.551 and an F score of 0.439. The Stand to Lie transition model, with an additional 11 Z axis ECDF coefficients, gave an AUC of 0.660 and an F score of 0.265. The Stand to Sit transition model, with an additional 13 Y axis ECDF coefficients, gave an AUC of 0.571 and an F score of 0.554.

The classification of specific postural states from among the other two possible postures displayed consistently higher performance than the classification of transitional events. The identification of periods of standing gave an AUC of 0.787 and an F score of 0.784. The Sitting posture classification achieved an AUC of 0.742 and an F score of 0.425. The classification of the Lying posture gave an AUC of 0.794 and an F score of 0.506. Reconfiguring the classification tasks to instead be a binary choices between two postural states gave AUC results of; 0.773, 0.820, and 0.784, and F Score results of 0.804, 0.822, and 0.743 (for Stand vs Sit, Stand vs Lie, and Lie vs Sit binary classifications respectively).

The detection of locomotion while the dog was observed as being in a standing posture, from among stationary standing periods, gave an AUC of 0.782 and an F Score of 0.544.

Discussion

The exploration of postural classification from tri-axial accelerometer data using the computationally light LDA classifier produced promising results in terms of performance. It is clear here that transitional events are difficult to identify from among randomly selected non-transitional data as well as from each other. This is likely due to the transient nature of these behaviours, despite this being a common method when attempting to remotely monitor such transitions within human medicine [9]. This could also be explained by the differences in postural transitions between quadrupeds and bipeds. The pronograde nature of the dog results in a much less distinct transitional acceleration signal, particularly from a lie to a stand. This transition would occur predominantly within the anterior/posterior and dorsal-ventral axes, up and forward as the dog rises, without consistent or dramatic shifts in the orientation of the device. However, humans performing the equivalent transitions would produce a consistent and distinct shift in device orientation, which is clearly observed in a waist mounted sensor [9]. Non-collar sensor attachment may improve performance of such a method with domestic dogs. However, the attachment methods are constrained by what is deemed acceptable for long-term wear by the owner of the dogs.

The identification of postural states directly was more successful with classification results performing consistently above chance. This methodology draws more directly from the animal-focused methods of behavioural ecology [e.g. 10] but adapts aspects of the sensors to more directly suit the companion animal context, such as by attaching the sensor to the collar rather than to the core of the body. Further refinement and validation of the methodology would likely present an effective postural detection mechanism that could be highly informative in terms of canine health and welfare. The use of more sophisticated and robust feature selection protocols would allow a more detailed examination of the degree of information provided by each of the generated features. Longitudinal sensor deployment within wild animal populations often features devices consisting of multiple complementary sensors, with little additional power or weight-burden [11]. Such an approach could assist further here, with the inclusion of sensors such as magnetometers adding further informative data with which to derive additional features or to supplement those calculated here. Additionally the implantation of more intensive classification or machine learning algorithms would likely further improve performance with sufficient optimization however this should be tailored to specific research problems as such methods can become a “black-box” with useful information in terms of how a classification is made often being lost. The results presented here suggest a hierarchical method of classification with the best performing being implemented first [as used in 12, 13].

To conclude. we were able to achieve acceptable levels of posture classification in dogs using methods which would be fast to compute, based on the detection of postures, rather than on transitions between postures. These methods will need further validation in different breeds and contexts. Posture detection classification from accelerometer data may be further refined from the approach presented here with more sophisticated classification methodologies or the use of additional relevant features from accelerometers or additional complementary sensors.

Ethical Statement

The proposal was approved by the Newcastle University Animal Welfare and Ethical Review Body, project ID number 557.

References

1. Whitham, J., Miller, L. (2016). Using technology to monitor and improve zoo animal welfare,. Anim. Welf. 25, 395–409.

2. Jukan, A., Masip-Bruin, X., Amla, N. (2017). Smart Computing and Sensing Technologies for Animal Welfare: A Systematic Review. *ACM Computing Surveys* 50.

3. Nweke, H.F., Teh, Y.W., Mujtaba, G., Al-garadi, M.A. (2019). Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Inf. Fusion* **46**: 147–170.

4. Lane, D.M., Hill, S.A., Huntingford,J.L., Lafuente, P., Wall, R., Jones, K.A. (2015). Effectiveness of slow motion video compared to real time video in improving the accuracy and consistency of subjective gait analysis in dogs. *Open Vet. J.* 5, 158–165.

5. Ladha, C., Belshaw, Z., O'Sullivan, J., Asher, L. (2018). A step in the right direction: an open-design pedometer algorithm for dogs. *BMC Vet. Res.* **14**: 107.

6. Hammerla, N.Y., Kirkham, R., Andras, P., PÏotz, T. (2013). On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. *ISWC 2013 - Proceedings of the 2013 ACM International Symposium on Wearable Computers*, 65–68.

7. Ladha, C., Hammerla, N.Y., Hughes, E., Olivier, P., PÏotz, T. (2013). Dog's life: Wearable Activity Recognition for Dogs. *Proc. 2013 ACM Int. Jt. Conf. Pervasive ubiquitous Comput. - UbiComp '13*, 415**.**

8. Hall, M.A. (1999). Correlation-based Feature Selection for Machine Learning.

9. Rodríguez-Martín, D., Samà, A., Pérez-López, C., Cabestany, J., Català, A., Rodríguez-Molinero, A. (2015). Posture transition identification on PD patients through a SVM-based technique and a single waist-worn accelerometer. *Neurocomputing* **164**: 144–153.

10. Shepard, E.L.C., Wilson, R.P., Quitana, F., Gómez Laich, A., Liebsch, N., Albareda, D.A., Halsey, L.G., Gleiss, A., Morgan, D.T, Myers, A.E., Newman, C., Macdonald, D.W. (2010). Identification of animal movement patterns using tri-axial accelerometry. *Endanger. Species Res.* **10**: 47–60.

11. Williams, H.J., Holton, M.D., Shepard, E.L.C., Largey, N., Norman, B., Ryan, P.G., Duriez, O., Scantlebury, M., Quintana, F., Magowan, E.A., Marks, N.J., Alagaili, A.N., Bennett, N.C., Wilson, R.P. (2017). Identification of animal movement patterns using tri-axial magnetometry. *Mov. Ecol.* 5.

12. Chakravarty, P., Cozzi, G., Ozgul, A., Aminian, K. (2019). A novel biomechanical approach for animal behaviour recognition using accelerometers. *Methods Ecol. Evol.* **10**: 802–814.

13. Zhang, S., McCullagh, P., Nugent, C., Zheng, H. (2010). Activity monitoring using a smart phone's accelerometer with hierarchical classification. *Proceedings - 2010 6th International Conference on Intelligent Environments, IE 2010*, 158–163.

Posters

Behaviour- and cognition-based methods to assess feeding motivation in dairy cows around dry-off

Guilherme Amorim Franchi, Mette S. Herskin and Margit Bak Jensen

Aarhus University, Department of Animal Science, Blichers Allé 20, 8830, Tjele, Denmark

Hunger can be defined as a negative emotional state caused by undernourishment (adapted from [1]). In intensive dairy production systems worldwide, one way to reduce milk synthesis before the day of last milking as part of a dry-off process (i.e. the artificial milking cessation generally 60 d before the expected day of calving [2]) is through restrictive feeding. As part of a larger study investigating the effects of various dry-off management on productivity, metabolism and welfare of cows, we developed a series of novel methods to assess the degree of feeding motivation (i.e. hunger) in those cows. The listed methods all aimed to stimulate meaningful behavioural responses specifically related to the cows' potential degree of hunger and the alleviation hereof. The first method was an operant push-gate test [3], which relied on a trade-off between a desired feed resource and the performance of an operant response to gain access to the resource in question. The second method was a feed-related attention bias test [4], which was based on the assessment of cows' vigilance for feed cues and engagement with such cues. Furthermore, together with the previous test, we carried out a visual lateralisation test [4]. The respective test relied on the concept of brain asymmetry and resultant lateralised sensory function to assess the cows' perception of feed. Finally, we developed two feeding frustration tests [5], which consisted of assessing cows' behaviours directed towards inaccessible feed. All procedures involving cows were approved by the Danish Animal Experiments Inspectorate in accordance with the Danish Ministry of Justice Act No. 1306 (November 23, 2007). Overall, the tests are relatively quick, account for cow's motivational and emotional states, and require little or no training prior to testing, except the operant push-gate test that might be more challenging for some cows. Conversely, they generally require a controlled test environment and, if conducted successively, might lead to habituation and consequently reduced behaviour responses. The concepts behind the different methods, the measures we used in each test, as well as their respective advantages and disadvantages will be outlined and discussed.

References

1. D'Eath, R. B., Tolkamp, B. J., Kyriazakis, I., & Lawrence, A. B. (2009). 'Freedom from hunger' and preventing obesity: the animal welfare implications of reducing food quantity or quality. *Anim. Behav* **77(2**): 275-288.

2. Collier, R. J., Annen-Dawson, E. L., & Pezeshki, A. (2012). Effects of continuous lactation and short dry periods on mammary function and animal health. *Animal* **6(3)**: 403-414.

3. Franchi, G. A., Herskin, M. S., & Jensen, M. B. (2019). Dairy cows fed a low energy diet before dry-off show signs of hunger despite ad libitum access. *Sci. Rep*. **9(1)**: 1-9.

4. Franchi, G. A., Herskin, M. S., & Jensen, M. B. (2020). Do dietary and milking frequency changes during a gradual dry-off affect feed-related attention bias and visual lateralisation in dairy cows? *Appl. Anim. Behav. Sci.* **223**: 104923.

5. Franchi, G. A., Herskin, M. S., Tucker, C. B., Larsen, M., & Jensen, M. B. (2020). Assessment of feeding motivation in dairy cows during gradual dry-off using two feeding frustration tests. Manuscript in preparation.

Volume 1 of the Proceedings of the joint meeting of the 12th International Conference on Measuring Behavior and the 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18 2021. www.measuringbehavior.org.
Editors: Spink, Andrew, Barski Jarosław, Brouwer, Anne-Marie, Riedel Gernot, Sil Annesha (2020). Pp 100.
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Validation of the Arabic Version of the Depression Anxiety Stress Scales (DASS-42) among Undergraduates in Kuwait

B.M. Alansari

Department of Psychology,Faculty of Social Sciences, Kuwait University, baderansari@gmail.com

Introduction

Psychological distress is largely defined as a state of emotional suffering characterized by symptoms of depression (e.g., lost interest; sadness; hopelessness) and anxiety (e.g.,. restlessness; feeling tense) [8]. These symptoms may be tied in with somatic symptoms (e.g., insomnia; headaches; lack of energy) that are likely to vary across cultures [5,6]. The relationship between distress and depression - and to a lesser extent, anxiety - raises the issue of whether psychological distress lays in the pathway to depression if left untreated [4]. Unfortunately, the course of psychological distress is largely unknown. Psychological distress is based on tripartite framework that includes anxiety, depression and stress. Psychological distress is assessed with standardized scales that are either self-administered or administered by a research interviewer or a clinician. The Depression Anxiety Stress Scales DASS-42 [7], specifically designed to provide relatively pure measures of, the three related constructs including anxiety, depression and stress**.**

Objectives

To examine the reliability, validity and factor structure of the Arabic adaptation DASS-42 in an undergraduate sample.

Methods

Sample

The sample consisted of (1108) Kuwait University students (487) males and (621) females, with a mean age of (21.28 ±1.22).

Design

Descriptive study based upon self-reported The Depression Anxiety Stress Scales DASS-42, Beck Depression Inventory-II, Beck Anxiety Inventory –BAI, and BFI 2 Anxiety & Depression facet scales.

Measures

The Depression Anxiety Stress Scales DASS-42 [7] consisting of 42 items (4-point Likert scale), The Beck Depression Inventory-II [2] consisting of 21items (4-point Likert scale), the Beck Anxiety Inventory –BAI [1] consisting of 21 items (4-point Likert scale), and the Second Big Five Inventory – BFI 2 [9] a 4-item (5-point Likert scale) facet scales for Anxiety & Depression facet scales, were administered to assess psychological distress.

Procedures

In order to participate, students were required to return the signed consent form to their teacher by the date of the study. Dates for participation in the study were considered to ensure that the mind and final exams did not interfere with the emotional state of the participants. Participants completed the Arabic versions of DASS-42, BDI-II, BAI & BFI-2) Anxiety & Depression facet scales including demographics.

Volume 1 of the Proceedings of the joint meeting of the 12th International Conference on Measuring Behavior and the 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18 2021. www.measuringbehavior.org.
Editors: Spink, Andrew, Barski Jarosław, Brouwer, Anne-Marie, Riedel Gernot, Sil Annesha (2020). Pp 100.
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

Statistical analysis

One-way ANOVA to assess gender differences, Cronbach's Alpha were computed for DASS-42 to assess the internal consistency, Pearson’s correlations between DASS-42 and BDI-II, BAI & BFI-2) Anxiety & Depression facet scales were computed to assess concurrent validity, exploratory factor analysis were used to examine the structure DASS-42, BDI-II, BAI & BFI-2 (Anxiety & Depression) scales, and Kaiser-Meyer-Olkin (KMO) test of sampling adequacy were calculated in this study using IBM SPSS Statistics.

Results

Table 1. One-way ANOVA of the scores of two gender groups in PD variable scores

Measures	Sum of squares	df	Mean square	*f-test*	*P-Level*
Stress (DASS)	5974.75	1	5974.75	132.58	.000
Depression (DASS)	3587.96	1	3587.96	70.56	.000
Anxiety (DASS)	2347.08	1	2347.08	51.28	.000
Depression(BDI-II)	2442.83	1	2442.83	25.72	.000
Anxiety (BAI)	1287.52	1	1287.52	7.87	.005
Anxiety (BFI-2)	146.27	1	146.27	19.52	.000
Depression (BFI-2)	323.31	1	323.31	35.27	.000

Table 1 shows significant gender differences between males and females in DASS Stress, Depression (DASS), Anxiety (DASS), Depression(BDI-II), Anxiety (BAI), Anxiety (BFI-2), and BFI-2 depression in which females have the highest means, therefore we run the following exploratory and confirmatory factor analysis using the separate samples (see table 2). Kaiser-Meyer-Olkin (KMO) Test value for the sample and was average (0.84) which indicate the sampling is adequate and thus may suggest the suitability of the data for factor analysis [3].

Table 2: Alpha Reliability, correlations, the explanatory (PCA) factor analysis of DASS-42, BDI-II, BAI and BFI-2 Anxiety & Depression facet scales extracts one factor solutions

Scales	Alpha r	Alpha r	Stress DASS	Depression	Anxiety DASS	BDI-II	BAI	Anxiety BFI2	Depression	EFA Factor (male)	EFA Factor (female)
Stress (DASS)	.91	.81		.77	.69	.61	.39	.41	.49	.88	.85
Depression (DASS)	.90	,89	.88		.74	.58	.39	.38	.51	.89	.86
Anxiety (DASS)	.86	.79	.87	.84		.52	.45	.44	.41	.88	.83
Depression (BDI-II)	.90	.89	.51	.59	.51		.39	.34	.58	.69	.78
Anxiety (BAI)	.94	.92	.42	.51	.54	.41		.39	.36	.67	.47
Anxiety (BFI-2)	.69	.76	.39	.40	.38	.37	.42		.49	.51	.52
Depression (BFI-2)	.79	.70	.44	.43	.44	.45	.39	.61		.61	.72
Eigen – Value										3.93	3.75
% Variance										56.14 %	53.54 %

	***All Correlation is significant at the 0.01 level (2-tailed).** **Matrix above diameter for females.** **Matrix below diameter for males.**

The positive correlation between the scales ranging from (r=0.88 to 0.37) for males and from (r=0.77 to 0.34) for females which suggests that the DASS-42 shows good evidence of convergent validity of the DASS-42. The explanatory factor analysis (EFA) using principal components method extracted one unipolar factor with the following loading for males: (.89) for Depression (DASS) , (.88) for Stress (DASS), (.88) for Anxiety (DASS), (.69) for Depression (BDI-II), (.67) for Anxiety (BAI) (.61) for Depression (BFI-2), and (.51) for Anxiety (BFI-2), which explains 56.14% of the total variance. The explanatory factor analysis (EFA) extracted one unipolar best factor solution for females, with the following loading: (.86), for Depression (DASS), (.85) for Stress (DASS) (.83) for Anxiety (DASS), (.78) for Depression (BDI-II), (.72) for Depression (BFI-2), and (.52) for Anxiety (BFI-2), and (.47) for Anxiety (BAI), which explains 53.54% of the total variance. Cronbach's alpha reliability was satisfactory for DASS-42 scales ranging between (0.91 to 0.86) for males and between (0.89 to 0.87) for females. Overall, the internal consistency of each scale in this study was satisfactory.

Conclusions

The construct level analysis suggested that Psychological distress (PD) based on tripartite framework psychological distress, is a single construct of unipolar factor. We conclude that assessment of mental health in general populations should use concomitant measures of psychological distress. The factor structure of the Arabic DASS-42, BDI-II, BAI & BFI-2 Depression & Anxiety scales was tested with explanatory factor analysis, which indicated that the seven scales provided a better fit to the data in a one-factor solution. Moreover, the highest explanatory factor loading for males: Depression (DASS), Stress (DASS), Anxiety (DASS), Depression (BDI-II), Anxiety (BAI), Depression (BFI-2), and Anxiety (BFI-2.). While the highest explanatory factor loading for females: Stress (DASS), Depression (DASS), Anxiety (DASS), Depression (BDI-II), Depression (BFI-2), Anxiety (BFI-2), and Anxiety (BAI). Thus, supporting the hypothesis that the PD based on tripartite model is the most appropriate Structure of Psychological Distress. The study provides evidence for the universality of the psychological distress syndromes measured by DASS-42, scales, and supports the development of culturally sensitive translations and adaptations of existing measurement tools in cross-cultural research. Findings confirm that the DASS-42, provides satisfactory validation, and thus it can be recommended as a measure of Psychological distress among nonclinical Arab samples.

References

1. Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology* **56(6)**: 893-897.

2. Beck, A. T., Steer, R. A., & Brown, G. K. (1996). BDI–II, Beck Depression Inventory: Manual (2nd ed.). Boston: Harcourt Brace.

3. Cerny, C.A., & Kaiser, H.F. (1977). A study of a measure of sampling adequacy for factor-analytic correlation matrices. *Multivariate Behavioral Research* **12:1**: 43-47.

4. Horwitz, A.V. (2007). "Distinguishing distress from disorder as psychological outcomes of stressful social arrangements." *Health* **11**:273-289.

5. Kirmayer, L.J. (1989). "Cultural variations in the response to psychiatric disorders and
1. Psychological distress." *Social Science and Medicine* **29**:327-339.

6. Klein man, A. (1991). Rethinking Psychiatry. From Cultural Category to Personal Experience. New York: The Free Press.

7. Lovibond, S. H., & Lovibond, P. F. (1995*)*. Manual for the Depression Anxiety Stress Scales. Sydney, Australia: Psychology Foundation.

8. Mirowsky, J., and C.E. Ross. (2002). Selecting outcomes for the sociology of mental health: Issues of measurement and dimensionality. *Journal of Health and Social Behavior* **43**:152-170.

9. Soto, C.J. & John, O.P. (2017). The Next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology* **113**: 117-143.

Measuring cohabitating partner’s attitude during social interaction

Julia M. Panfilova

Institute of Psychology, Russian Academy of Science, Moscow, Russia

Introduction

Cohabitation is a widely spread alternative union to formal marriage. We define cohabitation as partners’ who have intimate personal relations and also living together and share joint housekeeping duties without formalizing their relations in state registry services or signing a partnership contract.

Cohabiting couples who have an agreed trajectory toward marriage appear to do as well, or better, than other types of couples. It was also found that couples with ambiguity regarding marriage were at risk for negative relationship outcomes compared to other couples types [8]. Among young people in Europe by 2015, 15% lived in consensual union [1]. In Russia by 2015, 12,5 % of all spouse couples were not officially registered [2].

In the Family Code of Russian Federation only marriage formalized in registry services is officially recognized [3]. Due to absence of legal framework for such unions in Russia, in most cases cohabitation does not have an official status. Thus, if or when cohabitation ends, no guarantees or provisions can be offered to partners with children or with joint property. The wide spread nature of cohabitation as well as the risks outline has pushed Russian society and science to ask questions about the prospects of cohabitation and its role in maintaining family values [4].

Despite persistent social and scientific attention towards cohabitation, there is a lack of tools and approaches for analyzing and measuring cohabitation by measuring interactions within the cohabitating couple.

Methods

We analyzed attitudes during interaction with cohabitating partner, as well as factors, determining attitudes during interaction with cohabitating partner. The questionnaire is based on research in done on personal relationships within the framework of economic behaviour psychology [5, 6].

In our study of attitude during interaction with cohabitating partner, we rely upon provisions of resource and value approach.

Resource and value approach towards analysis of social interaction was proposed by V.P. Poznyakov and T.S. Vavakina within economic psychology behavioural research. According to that approach, significant partner’s interaction characteristics are formed by their values and norms that add certain and definite sense to the intra-partner interaction. [4,6,7].

Authors emphasize on the constructiveness of a value-resource approach for analysis of psychological relations and social interaction in different spheres of joint life activities. We suggest that reflection of a partner’s motives, theirshared values and norms are revealed through their acts during social interaction and these could be termed as attitude towards social interaction. We highlight economic-resource, social-resource, egocentric-recourse and social-value attitudes in interaction. Economic-resource attitudes are closely connected to value of finances in relations, where attitudes towards financial gains are measured during interaction with the partner. Social-resource attitudes express a partner’s aspiration to rise his or her social status, to join relevant circles of society through relations with the partner. Egocentric-resource attitudes are suggested to characterize a partner’s aims in interaction, connected with satisfying emotional demands in cohabitation: sexual satisfaction, home amenities and so on. In other words, egocentric-recourse attitudes assumes that partner focuses on himself or herself and that is tied to obtaining emotional satisfaction from cohabitation. Subject-value orientation presumes that partner is perceived as a person with own interests, aspirations, demands, while interaction with the partner is not aimed

Volume 1 of the Proceedings of the joint meeting of the 12th International Conference on Measuring Behavior and the 6th Seminar on Behavioral Methods to be held in Krakow, Poland, October 15-18 2021. www.measuringbehavior.org.
Editors: Spink, Andrew, Barski Jarosław, Brouwer, Anne-Marie, Riedel Gernot, Sil Annesha (2020). Pp 100.
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717

at any kind of financial profit or social status changing. The key features of subject-value orientation are sensitivity towards partner's demands, benevolence and commitment to universal values.

Description of questionnaire "Attitude during interaction with cohabitation partner"

We introduce the questionnaire "Attitude during interaction with cohabitation partner" that determines leading attitude. The questionnaire is aimed at measuring each expression strength of each attitude, as well as determining dominating attitude. The questionnaire consists of 12 questions, presented as scenarios with 4 behavioral choices in each, corresponding to 4 attitudes during interaction with the partner.

The questionnaire is based on partner's perception of the suggested scenario and the choices made by them. To increase the reliability of questionnaire, especially in an emotionally charged context such as cohabitation, the questionnaire was formulated in a neutral context and suggested scenarios are not connected to each other.

The questionnaire covers main aspects of cohabitation: joint life situations, partner's family and friends, financial area, partner's job, parties and anniversaries, sexual life, gifts and wealth.

Respondents need to choose 1 answer in each of 12 questions, each answer matches 1 score. After answering all the questions, the scores for each of 4 attitudes are summed up and expression strength of each attitude is defined as well as dominating attitude/s.

Determination of dominating attitude

One dominating attitude is defined when the score difference is 2 points or more between attitudes. If difference among the attitudes equals 0, expression strength of all the attitudes are equal.

Two attitudes are defined dominating if difference between their expression strengths equals 1 point or less, while differences between the other attitudes are by 2 points or more.

The validity of the questionnaire was checked by an expert panel that involved 1 doctorate of science, 2 science graduates and 5 specialist psychologists.

Statistical analysis was conducted with SPSS 22.0 by the items internal consistency method α-Cronbach. Reliability of the questionnaire equals to 0.76 that demonstrates acceptable internal consistency.

Text of the questionnaire

Text of the questionnaire and keys are outlined in Table 1 and Table 2 respectively.

Table 1: Questionnaire "Attitudes during interaction with cohabitating partner".

<table>
<tr><td colspan="2">You are offered 12 hypothetical situations on the topics of vacation, money, partner's family, common friends, sex, holidays, celebrations and 4 options are provide for each situation. Think and decide how would you behave in each situation. Please mind that there is no «true» or «false» answers, any answer that you choose is</td></tr>
<tr><td>#</td><td>Question and behaviour options</td></tr>
<tr><td>1.</td><td>Your partner invites you to celebrate his parent's anniversary, they live in the neighbour city, but you have already planned and prepaid lessons for your favourite hobby. If you miss lessons, costs will not be refunded. How would you behave?</td></tr>
<tr><td>A.</td><td>I will not go to my partner's parents. If I miss the lesson, I will lose money.</td></tr>
<tr><td>B.</td><td>I'll cancel the lesson despite the costs not being refunded and will go to his/her parents. On such an important day, I will choose to be with my partner.</td></tr>
<tr><td>C.</td><td>I'll go to partner's parents only if there are people whose contacts will be useful for my future life.</td></tr>
</table>

D.	I'll go to my partner's parents only if I have fun there.
2.	Your partner tells you, that he/she is offered to head a department, that would entail additional workload as well as salary increase. What would you ask first?
A.	Would we be able to spend as much time together as usual?
B.	How much money will we spend if his salary is increased?
C.	What privileges will my partner get?
D.	Will my partner's workload be extreme?
3.	How important is sex in your life?
A.	Sex makes us closer to each other.
B.	For me it's a way of getting pleasure.
C.	After sex my partner gets me expensive gifts.
D.	It makes me feel that I'm a good lover.
4.	Your partner's birthday is in a couple of weeks, and he/she asks your advice about the amount of guests and budget volume for the party. What would you advise?
A.	I'll suggest to the partner to invite to the colleagues and business partners that could be useful to him/her in the future.
B.	I'll suggest not to spend much money and to celebrate their birthday with a few people.
C.	I'll support whatever my partner wants to do irrespective of the expenses or number of people
D.	I'll suggest we go somewhere where we could have fun, as for expenses and guests - I'll leave it to the partner to decide.
5.	Your partner wants to introduce you to his childhood friend, who has a career in science, but he lives far away from you. How would you behave?
A.	I'll go to meet my partner's friend only if that person is interesting and will be a useful contact for me.
B.	I won't go since I'm sure that it'll be a dull visit.
C.	I'll refuse to go. For me it's a waste of time and money meeting an unknown person.
D.	I'll go, since I understand that my presence is very important for my partner.
6.	You presented an expensive mobile phone to your partner as a birthday gift, but he/she unwittingly broke its screen next day. How would you react?
A.	I'll expect that my partner will buy himself/herself an equivalent mobile phone, since I want my partner to have an expensive mobile phone.
B.	I wish I didn't give him/her that expensive mobile phone.
C.	I'll get the phone fixed.
D.	I'll be upset that my partner doesn't appreciate my presents.
7.	Your partner was fired some time ago, and he/she can't find new job for quite a long time. That was his/her only income source. How would you react?
A.	I'll wait for a bit longer and if my partner does not find a job I'll break up, since I don't want to cohabitate with a person with no income.
B.	We will live at my expense and I'll help my partner find a new job.
C.	I'll help my partner to find a worthy place and I will say to everyone that my partner is a freelancer to hide the fact that he is with no job.
D.	I'll keep them busy with other things and keep them feeling lighthearted while he/she is in a search of
8.	You are planning a joint vacation with your partner. What is the key thing for you?
A.	The main thing is that the dates of my vacation suites my schedule. If my partner can't make it for them, I'll go alone.
B.	Details of the vacation are not important for me if my partner will pay for everything.
C.	We will find dates and place that will suite us both and in the end I'll agree with my partner's opinion.

D.	I'd prefer a place where people of a certain status go and where there is a chance to get a useful acquaintance.
9.	You've learned that your partner broke relations with their close relatives, who are founders of the company where he works. How would you react?
A.	I'll recommend him/her to maintain good relations with his/her relatives so that's his/her status in the company stayes at the same level.
B.	I'll support partner's decision.
C.	I'll recommend him/her to maintain good relations, since if his position is changed or if he is fired, that would reflect on his/her income.
D.	I'll get upset, since I will miss parties that were organised by my partner's relatives.
10.	You live in apartments rented by your partner. Next month he/she would not be able to pay for it and asks you to pay the rent. How would you react?
A.	I'll lend money to my partner with the condition that he/she will return it to me the month after.
B.	I'll pay for the rent without any conditions.
C.	That would make me angry, since I didn't plan to pay for the rent.
D.	I'll suggest sharing the rent, but to move to a more prestigious district.
11.	Your partner and you want to start to do fitness, how will you choose fitness centre?
A.	I'll choose fitness programme by myself and will make my partner join me.
B.	We will do fitness at home so as not to pay for anything.
C.	We will choose fitness programme and schedule that will suit us both.
D.	We will pick a prestigious fitness-centre such that we improve our status level and make useful acquaintances.
12.	Your partner invites you to his/her favourite music band show, but you are not a fan of it. How would you behave?
A.	I'll go to the show only if my partner will buy another ticket for me.
B.	Sure I'll go and will buy tickets for us both, since it'll make my partner happy.
C.	I'll go to the show only if it'll be in a famous club or other prestigious place.
D.	I'll refuse, since I'm not ready to go to a show that I won't like.

Table 2: Keys to the questionnaire

Keys:			
Economic-resource orientation	Social-resource orientation	Egocentric-resource orientation	Subject-value orientation
1A, 2B, 3C, 4B, 5C, 6B, 7A, 8B, 9C, 10A, 11B, 12A	1C, 2C, 3D, 4A, 5A, 6A, 7C, 8D, 9A, 10D, 11D, 12C	1D, 2D, 3B, 4D, 5B, 6D, 7D, 8A, 9D, 10C, 11A, 12D	1B, 2A, 3A, 4C, 5D, 6C, 7B, 8C, 9B, 10B, 11C, 12B
Total:	Total:	Total:	Total:

Test subjects and results

Subjects involved 72 cohabitating men and 72 cohabitating women at the age of 18-35, living in Moscow and Moscow area. Respondents didn't have children, were Orthodox christians or atheists. Results showed that the value attitude dominated in the overall results of both men and women which could be explained by the affectionate and emotional nature of cohabitation.

Within cohabitation analysis, we consider such a notion as being a successful cohabitation – with a high probability of partners entering into family relations. A successful indicator is indicated by a high level of subject-value orientation, as well as satisfactory relations, readiness to have children with the current partner, as well as intention to formalize union with the current partner.

Acknowledgments

The research is funded by Russian Foundation for Basic Research (scientific project # 19-013-00735).

References

1. Corselli-Nordblad L., Gereoffy A. (2015) Marriage and birth statistics - new ways of living together in the EU. Eurostat Source: Statistics Explained Available at: http://ec.europa.eu/eurostat/statisticsexplained/ (Accessed 20.01.2020).
2. Federal State Statistic Service (2015). Population micro census. Available at: https://gks.ru/free_doc/new_site/population/demo/micro-perepis/finish/micro-perepis.html (Accessed 14.10.2019).
3. Family code of Russian Federation (1995). Federal Law: Accepted by the State Duma 08.12.1995 Collection of legislation RF 1, article 1, item 2, 6.
4. Panfilova Y.M. (2019). Orientation in interaction with the partner in non-registered marriage as indicator of partner's family values. *Human factor: Social psychologist* **1 (37):** 328-334.
5. Pozniakov V.P., Titova O.I. (2005). Russian enterpreneurs' relations of competition and partnership: regional and sex features]. *Issues of Economical Psychology* **2**: 181–204.
6. Poznyakov V.P. Vavakina T.S. (2016). Psychology of business partnership: theory and empirical researches. Institute of Psychology RAS publishing house, 320.
7. Vavakina T.S. (2011). Types of psychological position of Russian entrepreneurs towards business partnership. *Abstract of diss. ... Candidate of psychology* (Moscow) 187.
8. Willoughby B.J., Carroll J.S., Busby D.M. (2012) The different effects of "living together". Determining and comparing types of cohabiting couples. *Journal of Social and Personal Relationships*, **29**: 397–419.

Editors: Spink, Andrew, Barski Jarosław, Brouwer, Anne-Marie, Riedel Gernot, Sil Annesha (2020). Pp 100.
ISBN 978-90-74821-93-3. DOI: 10.6084/m9.figshare.13013717