Monitoring Mental State During Real Life Office Work

Anne-Marie Brouwer(✉), Loïs van de Water, Maarten Hogervorst, Wessel Kraaij, Jan Maarten Schraagen, and Koen Hogenelst

TNO, The Hague, The Netherlands
anne-marie.brouwer@tno.nl

Abstract. Monitoring an individual's mental state using unobtrusively measured signals is regarded as an essential element in symbiotic human-machine systems. However, it is not straightforward to model the relation between mental state and such signals in real life, without resorting to (unnatural) emotion induction. We recorded heart rate, facial expression and computer activity of nineteen participants while working at the computer for ten days. In order to obtain 'ground truth' emotional state, participants indicated their current emotion using a valence-arousal affect grid every 15 min. We found associations between valence/arousal and the unobtrusively measured variables. There was some first success to predict subjective valence/arousal using personal classification models. Thus, real-life office emotions appear to vary enough, and can be reported well enough, to uncover relations with unobtrusively measured variables. This is required to be able to monitor individuals in real life more fine-grained than the frequency with which emotion is probed.

Keywords: Emotion · Affective computing · Heart rate · Facial expression
Ecological Momentary Assessment · Experience sampling · Privacy

1 Introduction

Human-machine interaction could benefit from continuous information about an individual user's affective or cognitive state that is extracted from the user in an implicit, unobtrusive way, i.e. without requiring him or her to repeatedly provide self-reports. Especially automated, 'smart' systems that aim to support the user could use this information to optimize support in real time (adaptive automation – [1] or symbiotic systems). In an offline fashion, such information could be used to evaluate the interaction between (certain types of) users with (certain types of) systems. An example is examining mental effort in novices and experts working with information presented through different types of display designs. Other potential applications of continuous mental state monitoring are in the health domain: continuous information about an employee's state may lead to timely detection of increased workload or stress levels, allowing for adaptive interventions to promote a healthier working environment. Traditional ways to probe individuals' mental state such as self-report questionnaires, (neuro)psychological tests, or neuroimaging tools, often provide only snapshots of information, are burdensome, time-consuming, and/or obtrusive. Technological

J. Ham et al. (Eds.): Symbiotic 2017, LNCS 10727, pp. 18–29, 2018.
https://doi.org/10.1007/978-3-319-91593-7_3

advances now allow us to record, store and process a multitude of physiological and behavioral data from individuals in their daily environment [2–4]. Examples of such data are movements as recorded by a wristband, activity using a GPS system, heart rate and features related to facial expression from camera images. All of these techniques may allow us to monitor individuals' physiological and physical state continuously, implicitly and unobtrusively and, on the basis of those data, allow us to make inferences on individuals' mental state.

We here focus on mental state in an office situation. Previous studies in the laboratory (including simulated office environments) showed that a range of variables can be used to distinguish between different levels of induced stress or workload (e.g. [5–7]). However, it is not known whether naturally occurring stress, in a real life office environment, can also be detected implicitly and unobtrusively [8]. We do not know of any study in a real-life office situation in which an individual's emotional state is estimated and validated continuously using continuous, implicit, and unobtrusive sources of information.

There are several challenges in the topic of monitoring an individual's emotion continuously and implicitly in real life. Sensors can be affected by noise from the environment at unknown times and in unknown ways. There is no known mapping of variables to emotion that holds across changing contexts. In real life, context changes all the time - people perform different activities which go together with different types of noise and different types of valuable information. In addition, relations between implicitly measured variables and mental state differ between individuals [5, 9]. Arguably, strong, invariant associations are not expected since physiology and behavior are not there to inform researchers about emotion, but reflect processes that help the individual interact with and survive in the world. A final and important challenge we want to stress is that especially when one wants to refrain from experimentally inducing emotions, it is difficult to obtain ground truth ('true') emotion as experienced in real life while this information is essential to train and validate models that infer emotions from variables.

In order to meet these challenges to the largest extent possible, we designed our real-life office study so as to potentially enable personalized models by obtaining a relatively large number of ground truth emotion labels. A large number of labels would also help to capture potentially small variations in (real life) emotions. An office study forms a good case for mental state monitoring research in general since the amount and severity of sensor noise and changes in context in an office situation is expected to be low compared to many other real life situations. Our approach is to follow a number of individuals doing their work at a computer in office rooms at our institute over 10 days, recording unobtrusive variables that may convey information about emotional state (heart rate, facial features and keyboard and mouse activity). In addition, the participants are asked to rate their current emotional state every 15 min. These ratings (approximately 290 per person) are treated as ground truth emotion. Repeatedly asking individuals to rate their current feelings is referred to as Ecological Momentary Assessment or experience sampling [10–12]. Experience sampling minimizes retrospective biases that may typically occur in longitudinal studies using questionnaires on few occasions [12, 13].

Our main research question is whether we can estimate emotion using unobtrusively measured variables in an office setting. In this case, this would be operationalized to predicting an individual's subjective rating using data that was collected implicitly in the minutes prior to this subjective rating. We present first results on this, but we also hope that the present paper can serve as an anchor point for designing other real-life (office) monitoring studies, given the scarcity of these studies in the literature which makes it difficult to estimate what can be asked of participants in terms of experience sampling and being monitored in general. While for research and monitoring purposes, large amounts of data are preferred, it is of the utmost importance that participants adhere to the instructions and do not drop out due to experienced obtrusiveness (having to answer questions repeatedly) or perceived violations of their privacy. Drop-out is a severe problem in real-life monitoring studies. We therefore report participants' study experiences, their thoughts regarding privacy, and their ideas towards real-life implementation. In addition, we describe how rated emotion varies during regular office work, where it is of interest to know whether it seems to vary enough to expect that modelling is ever possible. Finally, we present associations that we find between rated emotion and the different variables that were recorded, to give an impression of whether and how they are related and can be expected to be useful in a model for estimating emotion.

Implicit variables that may reflect emotional state during office work include a range of physiological variables, facial expression, body posture and computer interaction (e.g. [5, 8, 14, 15]). Based on considerations concerning ease of use, obtrusiveness and budget, we chose to include heart rate as recorded at the wrist, facial expression and computer interaction.

2 Methods

2.1 Participants

Nineteen participants were recruited by advertising within the participant pool of the Netherlands Organization for Applied Scientific Research (TNO), the community of the University of Utrecht and clients of certain thesis support companies. The mean age was 25.84 (SD = 4.78). Three participants quit after 5 days or less. The reasons for quitting communicated to the experimental leader were not related to the study itself (namely illness and job requirements). We analyzed data from the remaining 16 participants. All participants were students at the time of the experiment and working on their thesis or other study projects, which is what they were asked to do in the laboratory offices. Participants were not in a dependent relationship with any of the people involved in the research. They received €150 upon completion of the study. All participants were fully briefed beforehand as to what was expected of them and what was recorded. They signed an informed consent form in accordance with the Declaration of Helsinki. This study was approved by the TNO Institutional Review Board (TCPE).

2.2 Materials

Figure 1 shows two participants during the study (picture taken and used with their permission). For each participant, a personal work area was available which consisted of a desk, an adaptable chair, a Dell Windows 7 laptop on a laptop stand, a mouse, a keyboard and a USB hub. The laptops were provided by TNO and equipped with software required for the experiment and a webcam. A maximum of 5 participants worked in one of three office rooms that were dedicated to this study.

Fig. 1. Study environment.

Heart rate was recorded using a MIO Fuse heart rate wrist wearable. At every heartbeat the device sent a heartrate value to the local data server together with the wearable ID.

Computer activity was recorded using Noldus uLog keylogger (research Edition version 3.3). This program logged various parameters of keyboard usage and mouse usage, of which the number of keypresses per minute, the error-ratio (number of error keypresses (backspace and delete) in proportion to the total number of keypresses) and the number of application switches per minute were assessed in this study. To protect the participant's privacy, typed strings of text and names or content of documents, emails and websites were not recorded.

For measuring the participant's facial features, a webcam snapshot was taken every minute. These images were later analyzed using Noldus FaceReader 7.0. to extract different basic emotions and expression components, so-called action units [16]. The output of this software are scores for 20 different action units and eight basic emotions. We examined action units dimpler, lip corner puller, lid tightener, brow lowerer, outer brow raiser and inner brow raiser since earlier research suggested that these may contain information as to emotion in an office situation [5, 15]. We also examined the basic emotion outputs for disgusted, scared, surprised, angry, sad, happy and neutral.

Every 15 min a pop-up screen with an affect grid [17] appeared on the participant's computer screen. Figure 2 shows the version we used (translated from Dutch to English). Participants were instructed to indicate their current emotion by clicking the appropriate location in the grid, and subsequently click the 'ok' button. They could also

click the pop-up away by clicking the 'ok' button immediately, but participants were asked to do this only when they really did not want to answer.

For investigating participants' overall experience as a participant in the study, including experienced obtrusiveness of the pop-up screen and privacy aspects, semi-structured one-to-one interviews were conducted at the end of the last working day. Open questions as well as Likert scale questions were used. The audio was recorded and notes were made by the experimental leader.

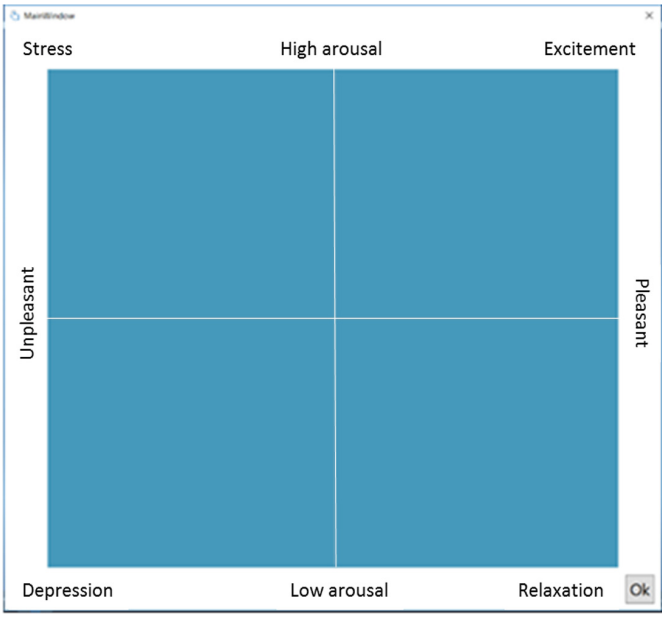

Fig. 2. Used pop-up affect grid (translated to English) for probing participants' current emotional state. Participants were instructed to indicate their current emotion by clicking the appropriate location in the grid, and subsequently click the 'ok' button.

2.3 Procedure

Most participants worked for 10 working days from 9:00 AM until 5:00 PM at the study location. The participation days were planned as consecutively as possible. One participant requested to join for longer than the 10 required days because of the pleasant working environment - she participated for 16 days. Another participant participated for 11 days to compensate for being absent due to appointments away from the study location. Each morning participants logged in on the study software with their participant ID and personal password. Then the heart rate wearable was placed on the participant's wrist and the connection with the local data server was checked. While participants were working, unobtrusive measurements were gathered and they were asked to rate their emotion every 15 min. Short individual breaks were allowed. In

addition, each day at twelve o'clock all participants went to the canteen together for a lunch break. On their first participation day, participants received a short instruction about all measurements and they signed an informed consent form. Afterwards, two questionnaires were filled in (a personality questionnaire and a questionnaire on vitality: not analyzed here). At the end of each participation day, a short questionnaire was taken on global activities, emotion, stress and mental effort over the whole day (not analyzed here). On their last participation day, a one-to-one interview with the participant was conducted focused on their experience with the study obtrusiveness and privacy aspects.

2.4 Analysis

Some heart rate data were missing due to technical problems. This resulted in reduced datasets for four participants (6, 8, 9 days of data for three of the participants who had participated 10 days; and 14 days of data for the participant who joined the study for 16 days).

For each subjective rating, we determined an averaged value for each examined implicit measurement (heart rate, the thirteen facial expression features and four computer usage features) reflecting the 15 min interval preceding the time of rating. Spearman's correlation analyses were performed between these averaged values of each of the implicit measures, and valence and arousal scores on the other hand. This was done for each participant separately. The number of significant ($p < 0.05$) correlations as well as their direction of correlation were stored.

For classification, we trained linear SVM (Support Vector Machine) models for each participant to distinguish between the participant's highest and lowest 33% arousal scores, and between the participant's highest and lowest 33% valence scores. Included features were the averaged values for each examined implicit measurement (heart rate, facial expression features and computer usage features), except for the six action units since these did not appear to be informative from the correlation analyses. In order not to lose data points if there were missing data of one or some of the implicit measurements, missing values were replaced by mean values of the participant. Classification was performed using the Donders machine learning toolbox developed by [18] and implemented in the FieldTrip open source Matlab toolbox [19]. The features were standardized to have mean 0 and standard deviation 1 on the basis of data from the training set. We used 5-fold cross validation. For each participant, and each arousal and valence model, we determined whether classification was above chance using a binomial test. An alpha level of 5% was used.

3 Results

3.1 Ratings Affect Grid

Participants virtually never clicked the pop-up affect grid away without entering a location in the grid. Click locations in the affect grid suggested that none of the

participants had aimed to click the same location - variability in ratings of naturally occurring emotions during office work seemed enough for further analysis. Figure 3 shows the responses of three different participants. They are the participant with the most dense cloud, with the widest cloud, and one of the approximately 7 participants who used the arousal in a more continuous way than valence – i.e. these participants tended to describe their emotion as either pleasant or unpleasant.

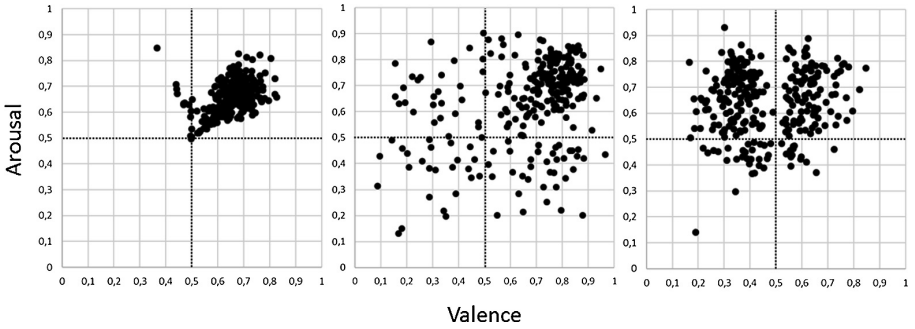

Fig. 3. Example data of three participants representing different types of response patterns. Each dot corresponds to one click in the affect grid – for each of the participants, all responses collected in the complete experiment are shown. See Fig. 2 for the lay-out of the grid as presented to the participants.

3.2 Correlation Analyses

Figure 4 shows the number of participants for whom we found significant correlation coefficients when checking correlations between reported emotions and implicit recordings, separately for the different pairs of variables and for the direction of correlation (negative or positive).

By chance, about 1 significant correlation (5% of 16) is expected for each pair. In general, computer activity, overall face expression (but not the independent face action units) and heart rate seemed to be associated with the subjective ratings. The direction of correlation was not always equal across participants.

It is likely that the quality of the emotional ratings differs between participants. Some people cannot or will not report emotions very accurately, so we would expect that for them, no correlations will be found while for others it is possible. Indeed for some participants, only about 8% of the correlations were significant (i.e. around chance level) while for the best participant, almost 50% was. Leaving out correlations with face action units that seemed not informative in general, the highest score was even 70%.

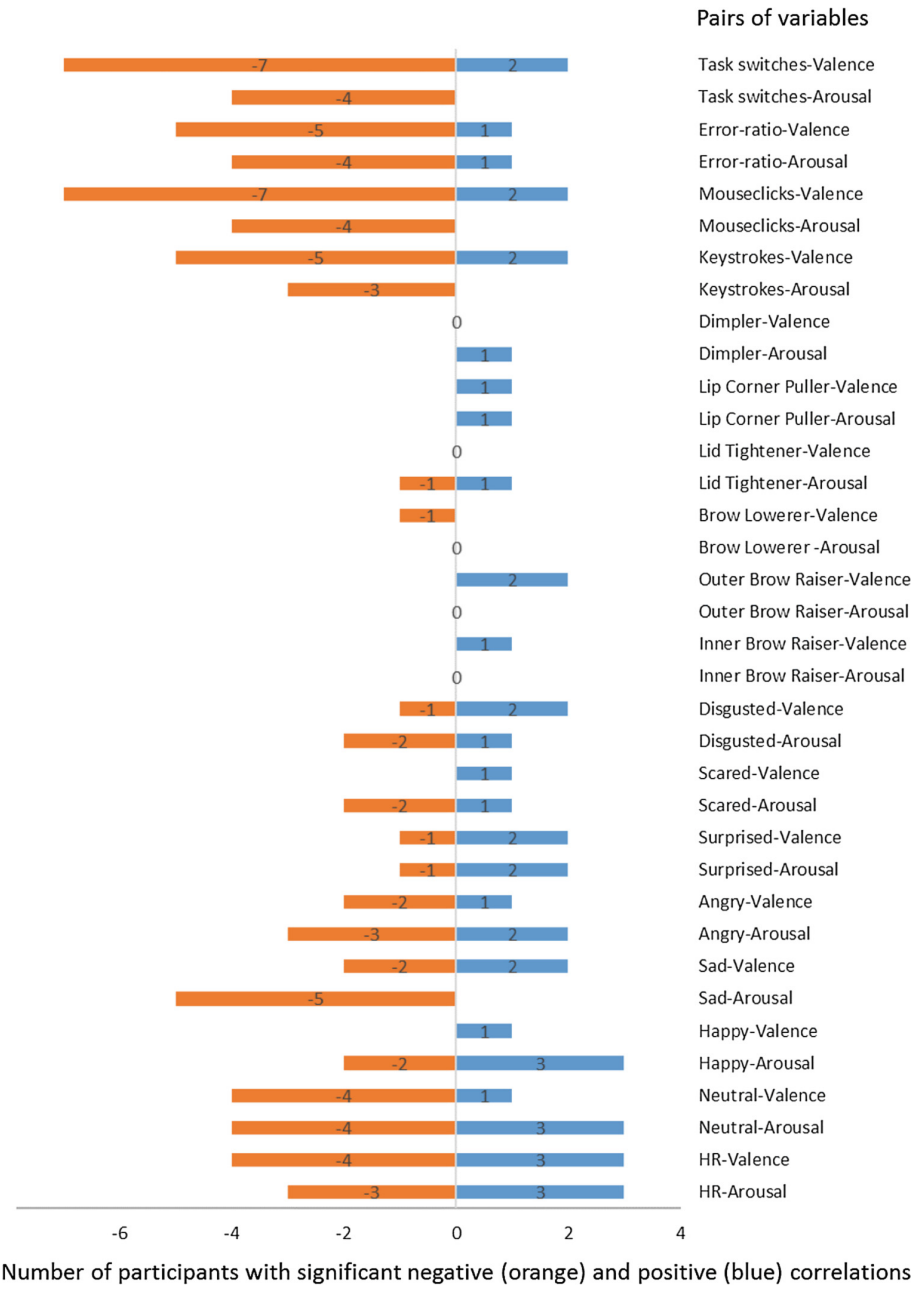

Fig. 4. Number of participants (out of 16) with significant correlations between rated emotional state and unobtrusive measures as indicated on the right (from top to bottom measures reflecting computer activity, facial expression and heart rate). Numbers are indicated as negative for negative correlations and positive for positive correlations. (Color figure online)

3.3 Classification

Classification accuracy averaged across accuracies per participant was 59% for valence (with 7 out of 16 participants showing significantly above chance classification of 50%) and 58% (with 5 out of 16 participants showing significantly above chance classification) for arousal.

3.4 Interviews

Participants generally liked the experiment (on a scale from 1 to 10, M = 7.3, SD = 1.1) and would participate again (M = 7.5, SD = 1.2). They reported that the affect grid was easy and quick to use, though participants also stated that at times it was challenging to indicate their 'true' emotional state. The affect grid was not experienced as interfering or bothering, except in a few cases when participants were very concentrated and wanted to finish a task. The unobtrusive recordings (Mio wrist band, web cam and computer activity logging) were indeed experienced as unobtrusive. Most participants thought that the recordings did not make them behave differently, they reported that awareness of being monitored decreased over time.

Opinions differed on wanting to use a potential application that could arise from research like this: 'an application that can provide insight in personal stress levels and generate advice based on this' (M = 5.0, SD = 2.4). Disadvantages that were mentioned were privacy issues, a lack of trust that it would work and the feeling that they are already aware of their stress level and thus would not need such an application. Privacy issues arose with some (not all) imagined practical applications: 15 out of 16 participants said to be willing to share data on a personal level with a doctor, 8 with their manager, and 5 with colleagues.

4 Discussion

The approach that we followed here to study associations between implicit, unobtrusive measures on the one hand and self-reported emotion in a natural, real life situation on the other hand, led to a strong adherence of participants. Despite the obligation to come to the recording facilities for 10 days, to respond to a pop-up every 15 min, and a low financial reward, only three out of nineteen participants did not finish the study and one participated longer than required. In the total study, it occurred only a few times that participants clicked the pop-up away without answering. As confirmed by the participants in the interviews, likely factors that contributed to this success are that the recorded implicit variables were indeed experienced as unobtrusive, that the emotion was probed with only one click (which was easy and quick, and not too annoying). Furthermore, participants valued the working space and the pre-scheduled working time to be devoted to their personal projects. Participants signed up completely voluntarily to the study, knowing about the requirements and measurements.

In a real-life office environment we found associations between heart rate, global facial expression, and computer activity on the one hand and currently experienced

valence and arousal on the other hand. The direction of the associations were not consistent across participants.

For heart rate, one may intuitively expect that heart rate increases with arousal. We did find a positive correlation for some individuals, but for others the correlation was negative. In fact, the literature shows that the relation between heart rate and arousal can go both ways. In our own studies, we found instances of positive relations, e.g. in social stress [20] and negative relations, e.g. when reading arousing sections in a novel [21]. The reason for this is probably that self-reported arousal can be associated with the body being prepared for action, cf. the defense reflex, or with a concentrated, focused state, cf. the orienting reflex where receptive and consolidating processes are facilitated [22]. With the defense system heart rate accelerations were found, while with the orienting system decelerations in heart rate were found [23].

For overall face expression, consistent directions would have been expected since the emotions covary with a certain level of valence and arousal. For instance, since anger is an emotion with low valence and high arousal [24], one would expect significant correlations between reported valence to be negative, and correlations with arousal to be positive. Such consistent patterns are not shown (Fig. 4). It might have been hard for the facial expression recognition algorithm to correctly interpret facial expressions in the office context, because facial expressions have a strong communicative function and in the present context, participants were not communicating with other individuals. One might therefore attribute the observed, across-participants, lack of coherence between facial expressions and subjective states to the fact that the studied context did not (or hardly) involve social interaction in the study. Note that even though the higher level interpretation of action units as performed by the face expression algorithm did not correctly predict the type of subjective state, it still provided information about valence and arousal as indicated by the number of participants that showed significant correlations (Fig. 4).

For computer activity the varying directions of correlations between participants were not unexpected. Some individuals are happy when they type a lot and are productive, for others it may be an indication that they have time pressure and feel negatively stressed.

The subjective rating plots, the correlation- and classification results suggest that some of the essential requirements to be able to proceed in the attempt to monitor emotion during real life office work are met: emotions under real life office working circumstances seem to vary enough, and can be reported well enough, in order to uncover relations with several unobtrusively measured variables on the level of an individual person. This opens the way for monitoring in real life that is more fine-grained than the frequency with which emotion is probed, which would be useful for research and application purposes. However, associations and classifications as reported here are still very modest or nonexistent for part of the participants.

There are several ways in which correlations and classification can be improved. Data from different (shorter) time intervals than the currently used 15 min preceding reported current emotion for all variables may better reflect current emotion. Context could be further specified by considering data per application (type) or use application (type) as a feature, or by adding adaptive learning of contexts [25]. Unreliable data (especially computer interaction data) could be excluded by using a criterion on the

percentage of the time that participants were actually at the computer during the used data interval. Future analyses should also focus on obtaining a deeper understanding of what is happening. When or for whom do we see which associations? Results of personality questionnaires or questionnaires on working habits and coping styles could be taken into account. Are the different tasks (applications) causing or hiding the associations? For instance, users will show a higher rate of key presses when using text processing applications than when using a web browser. When a user is more happy when browsing the web than using a text processing application, the application may be the underlying reason for key presses to be informative on mental state. On the other hand, while the number of keypresses may be informative about mental state in the word processing case, it may not be in the web browsing case; and the former association may be lost when data from all applications are taken together.

When sufficient classification accuracy can be reached, simulation of a real time situation would be of interest - training a model on successively incoming data and predict subsequent subjective rating. For future studies, it would be helpful and feasible to include eye tracker information (pupil size and blink features) as well as information about posture (which could also take the form of distance between the head and the webcam).

Acknowledgements. We would like to thank our TNO colleagues Bart Joosten and Thymen Wabeke for the technical setup; Leon Wiertz from Noldus for help with the face reader data; Jan-Willem Streefkerk (TNO) for project management. This publication was supported by the Dutch national program COMMIT (project P7 SWELL), and TNO Early Research Programs Making Sense of Big Data (Judith Dijk) and Human Enhancement.

References

1. Byrne, E.A., Parasuraman, R.: Psychophysiology and adaptive automation. Biol. Psychol. **42**(3), 249–268 (1996)
2. Baddeley, J.L., Pennebaker, J.W., Beevers, C.G.: Everyday social behavior during a major depressive episode. Soc. Psychol. Personal. Sci. **4**(4), 445–452 (2013)
3. Platt, T., Hofmann, J., Ruch, W., Proyer, R.T.: Duchenne display responses towards sixteen enjoyable emotions: individual differences between no and fear of being laughed at. Motiv. Emotion **37**(4), 776–786 (2013)
4. Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M.A., Kraaij, W.: The SWELL knowledge work dataset for stress and user modeling research. In: Proceedings of the 16th ACM International Conference on Multimodal Interaction, pp. 291–298 (2014)
5. Koldijk, S., Neerincx, M.A., Kraaij, W.: Detecting work stress in offices by combining unobtrusive sensors. IEEE Trans. Affect. Comput. (2016)
6. Hogervorst, M.A., Brouwer, A.-M., van Erp, J.B.F.: Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. Front. Neurosci. **8**, 322 (2014)
7. Okada, Y., Yoto, T.Y., Suzuki, T., Sakuragawa, S., Sugiura, T.: Wearable ECG recorder with acceleration sensors for monitoring daily stress: office work simulation study. In: Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE Engineering in Medicine and Biology Society, Annual Conference, pp. 4718–4721 (2013)

8. Alberdi, A., Aztiria, A., Basarab, A.: Towards an automatic early stress recognition system for office environments based on multimodal measurements: a review. J. Biomed. Inform. **59**, 49–75 (2016)
9. Brouwer, A.-M., Zander, T.O., van Erp, J.B.F., Korteling, J.E., Bronkhorst, A.W.: Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. Front. Neurosci. **9**, 136 (2015)
10. van Os, J., Delespaul, P., Barge, D., Bakker, R.P.: Testing an mHealth momentary assessment routine outcome monitoring application: a focus on restoration of daily life positive mood states. PLoS ONE **9**(12), e115254 (2014)
11. Mark, G., Iqbal, S.T., Czerwinski, M., Johns, P.: Bored mondays and focused afternoons: the rhythm of attention and online activity in the workplace. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3025–3034 (2014)
12. Trull, T.J., Ebner-Priemer, U.: Ambulatory assessment. Annu. Rev. Clin. Psychol. **9**, 151–176 (2013)
13. Moskowitz, D.S., Young, S.N.: Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. J. Psychiatry Neurosci. **31**(1), 13–20 (2006)
14. Kreibig, S.D.: Autonomic nervous system activity in emotion: a review. Biol. Psychol. **84**, 394–421 (2010)
15. Craig, S.D., D'Mello, S., Witherspoon, A., Graesser, A.: Emote aloud during learning with AutoTutor: applying the Facial Action Coding System to cognitive–affective states during learning. Cogn. Emot. **22**(5), 777–788 (2008)
16. Ekman, P., Friesen, W.V., Hager, J.: The Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action. Consulting Psychologists Press, Inc., Palo Alto (1978)
17. Russel, J.A., Weiss, A., Mendelsohn, G.A.: Affect grid: a single-item scale of pleasure and arousal. J. Pers. Soc. Psychol. **57**(3), 493–502 (1989)
18. Van Gerven, M., Bahramisharif, A., Farquhar, J., Heskes, T.: Donders Machine Learning Toolbox (DMLT) for matlab from https://github.com/distrep/DMLT, version 26 June 2013 (2013)
19. Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M.: FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. (2011). Article ID 156869
20. Brouwer, A.M., Hogervorst, M.A.: A new paradigm to induce mental stress: the Sing-a-Song Stress Test (SSST). Front. Neurosci. **8**, 224 (2014)
21. Brouwer, A.M., Hogervorst, M.A., Reuderink, B., van der Werf, Y., van Erp, J.B.F.: Physiological signals distinguish between reading emotional and non-emotional sections in a novel. Brain-Comput. Interfaces **2**(2–3), 76–89 (2015)
22. Sokolov, E.N.: Higher nervous functions: the orienting reflex. Annu. Rev. Physiol. **25**(1), 545–580 (1963)
23. Graham, F.K., Clifton, R.K.: Heart-rate change as a component of the orienting response. Psychol. Bull. **65**(5), 305 (1966)
24. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev. Psychopathol. **17**(3), 715–734 (2005)
25. Sappelli, M., Verberne, S., Kraaij, W.: Adapting the interactive activation model for context recognition and identification. ACM Trans. Interact. Intell. Syst. **6**(3) (2016). Article 22