DeepFlow: Detecting Optimal User Experience from Physiological Data Using Deep Neural Networks

Extended Abstract

Marco Maier
TAWNY
Munich, Germany
marco.maier@tawny.ai

Chadly Marouane
TAWNY
Munich, Germany
chadly.marouane@tawny.ai

Daniel Elsner
TAWNY
Munich, Germany
daniel.elsner@tawny.ai

ABSTRACT

The affective state called flow is described as a state of optimal experience, total immersion and high productivity. As an important metric for various scenarios ranging from (professional) sports to work environments to user experience evaluations, it is extensively studied using traditional questionnaires. In order to make flow measurement accessible for online, real-time environments, in this work, we present our preliminary findings towards automatically estimating a user's flow state based on physiological signals measured with a wearable device. We conducted a study of subjects playing the game Tetris in varying difficulty levels, leading to boredom, stress, and flow. Using a convolutional neural network, we achieve an accuracy of 70% in recognizing flow-inducing levels. In the future, we expect flow to be a potential reward signal for human-in-the-loop reinforcement learning systems.

KEYWORDS

affective computing; flow; socially intelligent agents; deep learning

ACM Reference Format:
Marco Maier, Chadly Marouane, and Daniel Elsner. 2019. DeepFlow: Detecting Optimal User Experience from Physiological Data Using Deep Neural Networks. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

The research field *Affective Computing* is dealing with recognizing, processing, interpreting, and simulating human affects and emotions [12, 16, 17]. With regard to the goal of recognizing emotions, typical approaches rely on various kinds of sensor data like images [13, 14], videos [1, 19], audio data [21], and physiological signals such as heart rate (HR) or electrodermal activity (EDA) [11, 15, 22].

Besides basic emotions such as *happy* or *sad*, other psychological models such as the *flow* theory [4] can be a valuable construct to assess a user's affective state. The state of flow is characterized by optimal experience, total immersion and high productivity, making it an interesting piece of information when assessing user experiences, from user interfaces to games to whole environments.

Traditionally, whether a subject experiences flow or not is determined through questionnaires [8, 10], which has the disadvantage of being only applicable after the actual occurrence and requires manual effort from the subject. In contrast, automatic flow recognition based on sensor data would be applicable unobtrusively and in real-time.

In this work, we propose a method to automatically measure flow using physiological signals from wrist-worn devices. The method is based on a convolutional neural network (CNN) architecture. For training data generation, we propose a study setup using the well-known game Tetris. We show the preliminary results of a small pilot study.

2 STUDY SETUP

For the data collection, we created a custom version of the game Tetris [20] as a mobile application. Tetris has already been used in similar studies and it has been found that depending on the difficulty of the game, users experience flow [2, 3, 6, 9]. The original game logic was modified so that there are only three different levels, i.e., *easy*, *normal*, and *hard*, in random order, each lasting 10 minutes, and independent from the player's performance. The difficulty of the three levels, i.e., the speed of the falling tetriminos, was set how we expected the game to lead to *boredom*, *flow*, and *stress* respectively. The recorded physiological data from each level was labeled accordingly.

Participants were selected so that they all had approximately the same skill level in the game. They were equipped with an Empatica E4 wrist-worn device [7] capturing physiological signals such as EDA, HR and HRV (heartrate variability). The E4 was worn on the participant's non-dominant hand. The smartphone (iPhone 5s) with the Tetris application was held in the other (dominant) hand.

We ran the following preliminary evaluations on a dataset from a small pilot study we conducted. There were 11 participants (3 female, 8 male) aged between 20 and 35. In total, we gathered 31 sessions, summing up to 15.5 hours of data. 4 participants played several sessions, 7 played only one session.

3 DATA AND PREPROCESSING

We used three streams of physiological signals from the E4: HR, HRV, and EDA. HR and EDA are provided by the E4 and were used in its raw form. With regard to HRV, the E4 provides the so-called RR-intervals, i.e., the time difference between consecutive heart beats, from which various HRV measures can be derived. EDA is sampled at 4 Hz while the HR values are provided at 1 Hz. RR intervals are not provided at regular intervals but when they occur.

In order to align the RR intervals with the two other data streams, we calculated a common HRV measure called RMSSD (root mean

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

Accuracy [%]	Baseline	Leave-one-session-out	Leave-one-subject-out
boredom vs. not boredom	50.00	65.04	57.13
flow vs. not flow	50.00	70.37	69.55
stress vs. not stress	50.00	66.09	71.17
boredom vs. flow vs. stress	33.33	52.59	50.43

Table 1: Best mean test accuracies achieved in leave-one-session-out and leave-one-subject-out cross validation.

square of successive differences) [18]. The RMSSD measure is computed over windows of data and it is recommended to use a window size of at least 60 seconds [5]. Consequently, at each time step where an RR value was received, a window of size 60 seconds before this point in time was extracted and the RMSSD value was computed for that window. The sample times of the EDA series were used as a basis for the final time series. Both HR and RMSSD values were forward-filled to fit the 4 Hz sampling frequency of the EDA series. The result is an equidistant time series, sampled at 4 Hz, with EDA, HR and HRV (i.e., RMSSD) values at each time step.

In order to create the training and validation sets, we split each session in windows of n samples. The window interval slides forward one sample at a time, i.e., consecutive windows overlap by $n-1$ samples. For this work, we used 10 second windows, i.e., windows consisting of $n = 40$ samples, each containing three values. The window length of 10 seconds was chosen because preliminary tests showed that shorter windows do not allow to capture characteristic patterns.

4 TRAINING AND EVALUATION

Our approach is based on a convolutional neural network architecture. The network consists of four convolutional layers (32 filters, kernel size 3), connected through max pooling layers. After the convolutions, one fully connected layer (32 neurons) leads to a final dense layer with the number of neurons in accordance with the number of classes of the classification task and a softmax activation. Except for the last layer, we used ReLU activations for the layers. During training, dropout is applied after the convolutional (0.1) and dense (0.5) layers to prevent overfitting.

We evaluated three binary one-vs-all classification tasks and one task trying to distinguish between all three classes at the same time. When creating the training and validation data sets, examples were chosen in a balanced manner, i.e., for the binary tasks, only half of the examples for the negative class were randomly drawn from the available examples to keep an even split between the two classes. We trained and evaluated our model in two ways: leave-one-session-out cross validation and leave-one-subject-out cross validation, the latter only on subjects that had played only one session, thus, validating on a completely unseen subject in each iteration. Table 1 shows the results.

One can see that the examples associated with boredom are the hardest to get right. We suppose that the easy level leads to the highest diversity of feelings among the three levels, i.e., the very slow speed is sometimes perceived as relaxing, sometimes as stressful, and only sometimes as distinctively boring. All in all, the CNN model is able to differentiate between the three classes considerably more accurately than the baseline strategy.

As we have outlined before, the affective state of flow is often associated with high productivity or better performance. In the case at hand, the achieved score in the Tetris game can be interpreted as the user's productivity. Thus, we can apply our model to Tetris sessions in order to divide a session into intervals of boredom, flow and stress – this time without taking into account information about the actual game level! – and then observe how good the performance of the player is in the respective states. Players indeed performed best when the model has recognized the *flow* state (average of 2.59 points per 10-second window), second best when the player is estimated to be bored (2.04 points). In contrast, when our system recognizes the state of *stress*, players perform considerably worse, even obtaining negative scores during these phases (−0.50 points).

5 DISCUSSION AND FUTURE WORK

In general, the initial results of our approach seem promising. However, there are several surrounding conditions that have to be improved in future work.

The data set we used for training and evaluating our model is too small. We showed the first, preliminary results from a small pilot study, but clearly see the need to greatly increase the number of subjects. Furthermore, data was collected from a very homogeneous population (i.e., young and healthy subjects, biased towards males) which should be broadened in future iterations.

From a psychological perspective, it should be further verified if the affective states we are trying to induce with the different difficulty levels of the game really can be considered *boredom*, *stress* and *flow*. Even though our general setup is in accordance with previous studies examining flow and especially examining flow with the game Tetris, the exact variant of the game and the surrounding conditions have not been fully verified. Thus, combining our data collection process with psychology-validated flow questionnaires is advisable. On the other hand, we could observe that players perform best during time intervals our model classifies as flow, which could be regarded as an indicator for an actual flow experience.

All in all, the positive initial results open up several possibilities for future work. In addition to improving the data set and tuning the model, we see a lot of potential in transferring the general approach to other, similar tasks, especially typical tasks of an office job.

More clearly scoped to the field of AI research, we are especially interested in using automatic flow detection as a feedback mechanism in human-in-the-loop reinforcement learning. Socially intelligent agents could benefit from the information about this affective state by incorporating it as a reward signal for their behavior.

REFERENCES

[1] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2015. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 43–55.

[2] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. 2008. Boredom, Engagement and Anxiety As Indicators for Adaptation to Difficulty in Games. In *Proceedings of the 12th International Conference on Entertainment and Media in the Ubiquitous Era (MindTrek '08)*. ACM, New York, NY, USA, 13–17. https://doi.org/10.1145/1457199.1457203

[3] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. 2011. Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 41, 6 (2011), 1052–1063.

[4] M Csikszentmihalyi. 1990. *Flow. The Psychology of Optimal Experience.* New York (HarperPerennial).

[5] Michael R Esco and Andrew A Flatt. 2014. Ultra-short-term heart rate variability indexes at rest and post-exercise in athletes: evaluating the agreement with accepted recommendations. *Journal of sports science & medicine* 13, 3 (2014), 535.

[6] László Harmat, Örjan de Manzano, Töres Theorell, Lennart Högman, Håkan Fischer, and Fredrik Ullén. 2015. Physiological correlates of the flow experience during computer game playing. *International Journal of Psychophysiology* 97, 1 (2015), 1–7.

[7] Empatica Inc. 2018. Real-time physiological signals | E4 EDA/GSR sensor. https://www.empatica.com/research/e4/. Accessed: 2018-09-03.

[8] Susan A Jackson and Herbert W Marsh. 1996. Development and validation of a scale to measure optimal experience: The Flow State Scale. *Journal of sport and exercise psychology* 18, 1 (1996), 17–35.

[9] Johannes Keller, Herbert Bless, Frederik Blomann, and Dieter Kleinböhl. 2011. Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *Journal of Experimental Social Psychology* 47, 4 (2011), 849–852.

[10] J Matias Kivikangas et al. 2006. *Psychophysiology of flow experience: An explorative study.* Ph.D. Dissertation. Helsingfors universitet.

[11] Hindra Kurniawan, Alexandr V Maslov, and Mykola Pechenizkiy. 2013. Stress detection from speech and galvanic skin response signals. In *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*. IEEE, 209–214.

[12] CL Lisetti. 1998. Affective computing.

[13] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2017. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. (2017), 18. https://doi.org/10.1109/TAFFC.2017.2740923 arXiv:arXiv:1708.03985

[14] Ali Mollahosseini, Behzad Hasani, Michelle J Salvador, Hojjat Abdollahi, David Chan, and Mohammad H Mahoor. 2016. Facial expression recognition from world wild web. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 58–65.

[15] Lennart Nacke and Craig A Lindley. 2008. Flow and immersion in first-person shooters: measuring the player's gameplay experience. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share*. ACM, 81–88.

[16] Rosalind W Picard. 1999. Affective Computing for HCI.. In *HCI (1)*. Citeseer, 829–833.

[17] Rosalind W Picard. 2003. Affective computing: challenges. *International Journal of Human-Computer Studies* 59, 1-2 (2003), 55–64.

[18] Fred Shaffer and JP Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in public health* 5 (2017), 258.

[19] Shangfei Wang and Qiang Ji. 2015. Video affective content analysis: a survey of state of the art methods. *IEEE Transactions on Affective Computing* 6, 4 (May 2015), 410–430.

[20] Wikipedia. 2018. Tetris. https://en.wikipedia.org/wiki/Tetris. Accessed: 2018-09-03.

[21] Min Xu, L-T Chia, and Jesse Jin. 2005. Affective content analysis in comedy and horror videos by audio emotional event detection. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 4–pp.

[22] Jing Zhai and Armando Barreto. 2006. Stress detection in computer users based on digital signal processing of noninvasive physiological variables. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. IEEE, 1355–1358.