T. Fukada, W. Huang, P. Janssen, K. Crolla, S. Alhadidi (eds.), *Learning, Adapting and Prototyping, Proceedings of the 23rd International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2018*, Paper 49 (Preprint).

AN EVOLVING BUILT ENVIRONMENT PROTOTYPE

A Prototype of Adaptive Built Environment Interacting with Electroencephalogram Supported by Reinforcement Learning.

TONGDA XU[1], DINGLU WANG[2], MINGYAN YANG[3], XIAOHUI YOU[4] and WEIXIN HUANG[5]
[1,2,3,4,5]*Tsinghua University*
[1]*x.tongda@ucl.ac.uk* [2]*Andrewwangdl@163.com*
[3]*mingyanyang@outlook.com* [4]*thuyouxiaohui13@126.com*
[5]*huangwx@tsinghua.edu.cn*

Keywords. Brain–computer interface; Reinforcement learning; Adaptive environment; Electroencephalogram; Mindfulness training.

Abstract. This paper proposes an environment prototype learning from people's Electroencephalogram (EEG) feedback in real-time. Instead of the widely adopted supervised learning method, a recently published affordable reinforcement learning model (PPO) is adopted to avoid bias from designers and to base the interaction on the subject and intelligent agent rather than between the designer and subject. In this way, development of interaction method towards a specific target is substantially accelerated. The target of this prototype is to keep the subject's alpha wave stable or decline, which indicated a more calming state, by intelligent decision of illumination state according to subject's EEG. The result is promising, a decent trained model could be gained within 500,000 steps facing this mid-complex environment. The target of keeping the alpha wave of subjects on a low or stable level purely by decision from computer agents is successfully reached.

1. INTRODUCTION

1.1. EEG BASED ENVIRONMENTAL INTERACTION

With the development of machine learning, the reliable analysis of electroencephalography (EEG) with classification has reached a practical level (Alarcao, 2017), e.g. 92.57% accuracy for four emotions; 94.86% for valence and 94.43 % for arousal (Lin, 2009). The reliable emotional benchmark from EEG data has been applied to various spatial-related studies and most of them are evaluation

studies with classification from Emotiv (Ramirez, 2012; Emotiv, 2017), covering issue on the influence of street space on people's emotion (Andreani, 2017; Mavros, 2012), the form of space and people's feeling inside (Shemesh, 2015) and the sound and emotive feedback (Dorothea, 2014).

1.2. REINFORCEMENT LEARNING AND INTERACTIVE ENVIRONMENT

While in our knowledge, most of the EEG spatial researches focus on evaluation instead of interaction since classification method is suitable for research while complex for space design. The method of design has to be proposed by designers, and effectiveness of which has to be assessed separately round by round for optimization.

Figure 1. left: the supervised learning; middle: the reinforcement learning(RL);.

Thus, the reinforcement learning (RL) should be applied in environmental interaction design with EEG instead of the widely adopted classification. The idea of RL was proposed in late 1979 (Sutton & Barto, 2016). The basic elements of a RL model are the environment and agent (Figure 1). Compared with supervised learning method, the policy is generated from agent instead of designers (Alarcao, 2017). The agent receives observation and reward from, and act on the environment automatically. The model is trained to maximize reward function, by which the result of action is evaluated whether desirable or not and changed to reach the target set. And after deep q-network (Mnih et al., 2015), physical- environment-based RL models are light and robust enough for designing environmental interaction.

1.3. STUDY GOAL

The aim of this research is to explore a simple, effective and computer-generated prototype of EEG mediated environmental interaction. For the first step, EEG alpha power and color (H,S,V) are picked as input and stimuli. Supported by a far more efficient agent-based search method, the prototype could avoid bias from designer, and substantially accelerate the design process.

2. METHOD

2.1. THE PROTOTYPE

2.1.1. The Prototype Setting

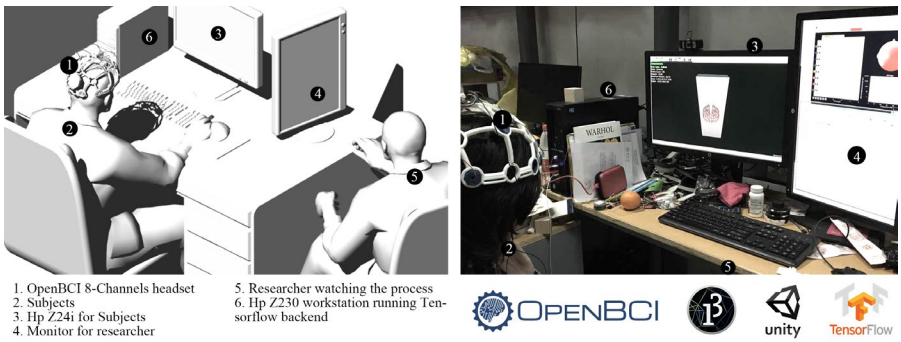

Figure 2. the prototype scene and environment setting.

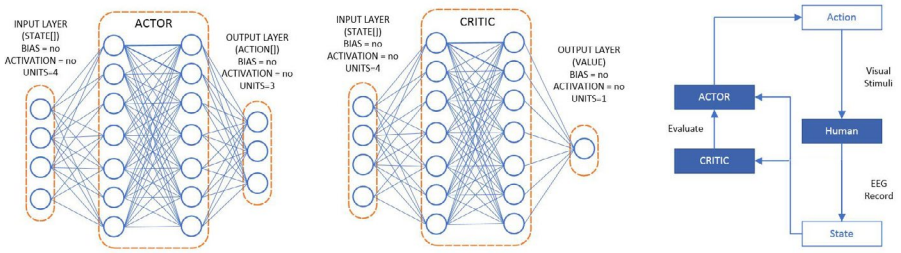

Figure 3. the Actor-Critic model and the training loop.

The prototype (Figure 2) is built in architecture studio, the device used to capture Electroencephalography (EEG) is OpenBCI-Cyton Biosensing Board (8-channels), with sampling rate @ 250HZ. The monitor subjects watch is HP-Z24i, with its color calibrated by Datacolor Spyder5 Elite, driven by HP workstation Z230. The machine learning backend is Tensorflow GPU r1.2, running on training environment with python 3.5.2., CUDA Toolkit 8.0 GA2, cuDNN v6.0 (April 27, 2017) and Windows 7.

The model used in this project is based on the open source Unity-ML agent (Juliani, 2017), which is built upon Proximal Policy Optimization model (Schulman, 2017), a simplified version of Trust Region Policy Optimization (Schulman, 2017). The core of the model is an Actor and Critic Neural Network (Konda, 2000) with 2 hidden layers. (Figure 3)

2.1.2. The Environment Configuration

For each loop an action array Vector3 would receive feedback from the model, adding to the existing Color (H,S,V). To avoid unnecessary jam, a part of reward function is set to punish an exceeding action which would bring the [Hue, Sat, Val] beyond [0f,1f]. (Juliani, 2017)

2.1.3. Reward Function

According to the precedent experiments (Juliani, 2017) and the target of mollifying the alpha wave, if the new alpha wave is 10% higher than the current, the punishment would be -1 with a reset, else the reward would be 0.1, targeting a successful loop.

```
if (action_h + Hue > 1.0f), reset = true, reward = −0.01f;
else { Hue = Hue + action_h };//Updating color
if ((alphaValue/oldalphaValue) > 1.1f ), reset = true, reward =
 ↪  −1f;
if (reset == false) { reward = 0.1f; }
```

According to the precedent experiments (Juliani, 2017) and the target of mollifying the alpha wave, if the new alpha wave is 10% higher than the current, the punishment would be -1 with a reset, else the reward would be 0.1, targeting a successful loop.

2.2. EXPERIMENT DESCRIPTION

2.2.1. the Pilot study and Subjects

In the pilot study of hyperparameters tuning, the experiment is done by 3 members from research group. The initial hyperparameters are copied from OpenAI Baselines (Dhariwal, 2017). The tuning process is assisted by 2 Ph.D. students majored in AI. The best hyperparameters are recorded and inherited in later experiments.

10 students majored in architecture are invited to participate as blind subjects, among which 60% are female, and 40% are male. The average age of subjects is 22.

2.2.2. Visual Stimuli

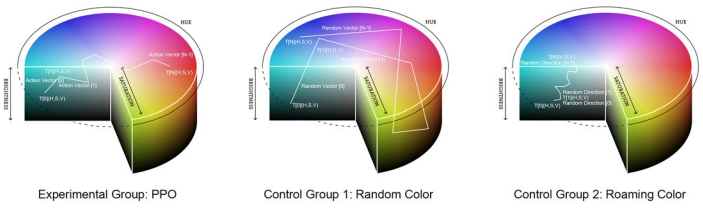

Figure 4. the search method of Experiment Group, Control Group 1-Random Color, Control Group 2-Roaming Color.

To eliminate other possible factors including fatigue, uneasiness and stimuli of random color, 2 control groups (Figure 4) are set with everything same as the experimental group despite the visual stimuli. The first control group watch random color, while the second watch roaming color, which means module of vector (H[T-1]-H[T], S[T-1]-S[T], V[T-1]-V[T]) was set uniformly.

2.2.3. Experiment Process

The common method used in video-based emotional stimulation (Valenzi, S, 2014; M. Murugappan, 2013) is adopted. The experiment is single-blinded and subjects are asked to sit before a screen with OpenBCI wear. The screen condition is monitored and adjusted by Datacolor Spyder5, ensuring the precision of color.

All the subjects use the same OpenBCI hat adjusted to their head size. Then the subjects are asked to focus on the screen playing consecutive pure color images generated by PPO (experiment group) or other (control group) until the experiment ends, which is around 40 minutes. The training data is recorded and exported into TensorFLow model and TensorBoard summary files for analysis.

3. RESULT

3.1. PILOT STUDY: BEST TRAINING STEPS AND TUNING

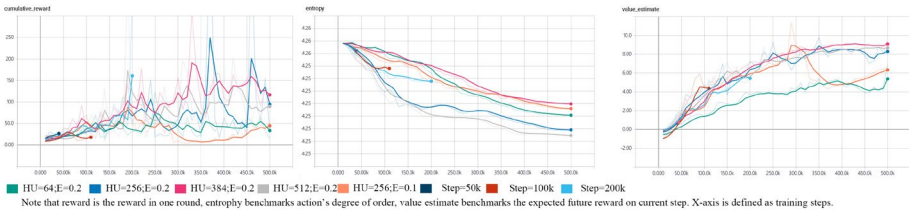

Figure 5. the reward, entropy and value-estimated of the pilot study.

Table 1. the t-test comparison of reward between 0-50k steps and 450-500k steps in tuning.

Hyperparameters	0-50k Mean	450-50k Mean	Sig.	95% Confidence Interval of the Difference	
				Lower	Upper
Hidden Units=64; Epsilon=0.2	16.3190924	42.03521786	0.05	0.80987004	50.62238
Hidden Units=256; Epsilon=0.2	18.6999662	127.1963913	0.00	64.32437534	152.6684748
Hidden Units=384; Epsilon=0.2	17.5128353	159.162178	0.13	-64.29230435	347.5909898
Hidden Units=512; Epsilon=0.2	13.8719385	82.21812058	0.00	49.12540774	87.56695638
Hidden Units=256; Epsilon=0.1	21.1898155	44.65787582	0.00	14.39013931	32.5459813

In best-training-step experiment, the reward and value estimate continue to increase at 50k, 100k, 200k steps (Figure 5), while reach a ceiling at around 400k steps, which means the model converges locally.

In tuning experiment (Figure 5) all the rewards increase with fluctuation. When compared the reward of 0-50k steps and 450-500k, the reward of Trial 2 (Hidden Units = 256; Epsilon = 0.2) witnesses a stable and significant increase. (Table 1) Trial 1 (Hidden Units = 384; Epsilon = 0.2) works better but oscillates too much. In later experiment, the training step and hyperparameters are inherited.

3.2. CUMULATIVE REWARD

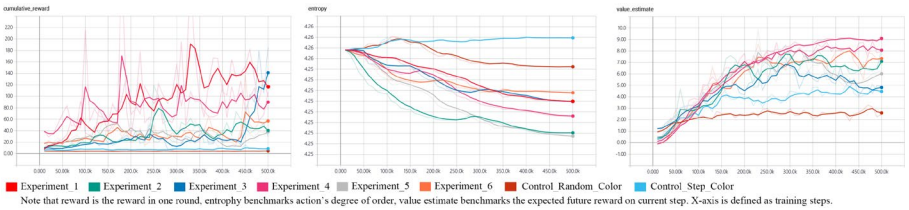

Figure 6. the reward, entropy and value-estimated of the experiment and compare group.

Table 2. the t-test comparison of reward between 0-50k steps and 450-500k steps of

Subject	0-50k Mean	450-50k Mean	Sig.	95% Confidence Interval of the Difference	
				Lower	Upper
Experiment_1	44.7766415	88.88543554	0.01	12.24482579	75.97276223
Experiment_2	13.4832205	40.79454079	0.00	16.21144368	38.41119694
Experiment_3	18.6999662	127.1963913	0.00	64.32437534	152.6684748
Experiment_4	18.4327291	52.203936	0.00	25.02001318	42.52240053
Experiment_5	18.3190727	36.35396156	0.00	8.75080647	27.31897121
Experiment_6	8.67793531	50.07698365	0.03	5.092693117	77.70540356
Control_Random_Color	3.7920248	4.009628439	0.19	-0.132167107	0.567374378
Control_Roaming_Color	6.40332823	8.588227463	0.00	0.99178359	3.37801488

All the experiment groups witness significant increase in reward, value estimate, and stable decrease in entropy, which indicates a successful and stable training (Figure 6). Concerning the reward increase of 0-50k steps and 450-500k (Table 2), the experiment group is significantly higher than comparison group.

3.3. ELECTROENCEPHALOGRAPHY

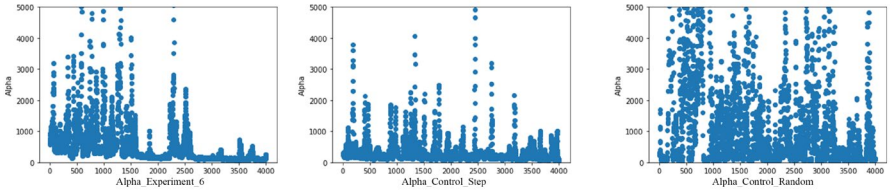

Figure 7. the comparison of Alpha power between experiment, step and random in the experiment process.

Table 3. the comparison of Alpha power mean and standard deviation between 0-50k steps and 450-500k steps of experiment_6, roaming and random.

Variable	Experiment_6		Control_Step		Control_Random	
	0-50k	450-500k	0-50k	450-500k	0-50k	450-500k
Alpha_Mean	757.55	137.13	407.17	233.02	2268.62	519.29
Alpha_Std	509.29	73.9	468.13	221.23	3889.22	872.11
Beta_Mean (Std)	770.01(424.34)	163.31(62.44)	327.02(366.82)	229.68(199.94)	2048.31(3631.47)	519.58(910.03)
Delta_Mean (Std)	13812.42(14413.34)	1220.79(956.77)	5810.33(8764.14)	3014.13(4673.44)	12314.84(22609.37)	3187.01(5376.04)
Gamma_Mean (Std)	2161.44(948.93)	1184.33(507.63)	3933.92(761.69)	3903.94(646.13)	2735.01(3848.60)	1627.08(1455.75)
Theta_Mean (Std)	923.24(798.62)	173.23(59.66)	533.12(767.98)	265.15(277.62)	2176.38(3753.47)	489.53(762.87)

Three participants are selected from different groups with alpha change. The deviation of alpha wave is kept lower in experiment group in training than control group (roaming and random) in later period of viewing visual stimuli (Table 3).

4. DISCUSSION

4.1. BEST TRAINING STEPS AND HYPERPARAMETERS

The current training step is set as 500,000 as the model's reward and value estimate stop to climb at 450k steps (around 30 minutes). This number is reasonable as OpenAI set the standard training steps as 1,000,000 when training MuJoCo environment with PPO. (Schulman, 2017) For more complex tasks such as humanoid running and steering, OpenAI adopts 50,000,000 to 100,000,000 steps for benchmark (Schulman, 2017). While 50millionsteps would take 55.6 consecutive hours, not practical for human experiment.

Due to the pilot study results, the best hidden unit number is around 256 per hidden layer. Further detailed experiment could be conducted to exact this number. Epsilon benchmarks how cautious the model is in exploring, and for now 0.2 fits the model best.

4.2. EFFECTIVENESS

The training is effective. The agent learns how to mollify the EEG alpha power by color stimulation successfully. A significant increase of reward (sig.<0.05 in t test) is witnessed in experimental group. Although the roaming color mollifies people's emotion significantly, the effectiveness is far less than control group. Taking the Lower Interval (-3.32) of the most effective control group and the Upper Interval (-8.78) of the most ineffective experiment group, the distance is still more than twice (2.64).

In experiment groups, the entropy declines slowly but steadily, which means the training of the model is successful (Juliani, 2017). The model is converging and learning from environment. While in no control groups the entropy declines steadily, which means although the reward increases, the model is not learning.

The change in Alpha wave indicates possible effect of interaction in moderating deviations in Alpha. The alpha power of the experiment group also decreases in the process, implying possible effect of calming (Chiang, Li, & Jane, 2017).

The result of this study is limited to individual case study. Further experiment and analysis with the influence of balanced individual differences are required to

draw detailed comparison in EEG between experiment groups and control groups.

4.3. POSSIBLE BETTER METHODS

Due to the subject's limit, a continuous training period like several days is impractical. An alternative off-policy model for non-discrete control like Deep Deterministic Policy Gradients (DDPG) would probably work, since it could gather data from many different participants, and collect them together to learn a single policy (Juliani, 2017). Another possibility is to rewrite the code of training environment and rephrase it into discrete state, and adopt Deep Q-learning for training.

References

Alarcao, S. M. and Fonseca, M. J.: 2017, Emotions Recognition Using EEG Signals: A Survey., *IEEE Transactions on Affective Computing*.

Andreani, S. and Sayegh, A.: 2017, Augmented Urban Experiences: Technologically Enhanced Design Research Methods for Revealing Hidden Qualities of the Built Environment, *Proceedings of the 37th Annual Conference of the Association for Computer Aided Design in Architecture*, 82-91.

Chiang, Y.C., Li, D. and Jane, H.A.: 2017, Wild or tended nature? the effects of landscape location and vegetation density on physiological and psychological responses, *Landscape & Urban Planning*, **167**, 72-83.

Juliani, A. and Pierre, V.: 2017, "Github" . Available from Open Source Repository<https://github.com/Unity-Technologies/ml-agents> (accessed Oct 2017).

Kalogianni, D. and Coyne, R.: 2014, Thinking about sound and space, Recording people, *eCAADe 2014 Volume 2*.

Konda, V. R. and Tsitsiklis, J. N.: 2000, Actor-critic algorithms, *Advances in neural information processing systems*, pp. 1008-1014.

Lin, Y.P., Wang, C.H., Wu, T.L., Jeng, S.K. and Chen, J.H.: 2009, EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine, *Acoustics, Speech and Signal Processing*, 489-492.

Mavros, P., Coyne, R. and Roe, J.: 2012, Engaging the Brain-Implications of mobile EEG for spatial representation, *eCAADe 30 Volume 2 User Participation in Design*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G. and Petersen, S.: 2015, Human-level control through deep reinforcement learning, *Nature*, **518**, 529-533.

Murugappan, M. and Murugappan, S.: 2013, Human emotion recognition through short time Electroencephalogram (EEG) signals using Fast Fourier Transform (FFT), *2013 IEEE 9th International Colloquium on Signal Processing and its Applications*, 289-294.

Ramirez, R. and Vamvakousis, Z.: 2012, Detecting Emotion from EEG Signals Using the Emotive Epoc Device., *International Conference on Brain Informatics*, Berlin, Heidelberg., 175-184.

Schulman, J., Dhariwal, F., Radford, P. and Klimov, O.: 2017, Proximal policy optimization algorithms, *arXiv preprint arXiv:1707.06347.*.

Shemesh, A., Bar, M. and Grobman, Y.J.: 2015, Space and human perception: exploring our reaction to different geometries of space. In Emerging Experience in Past, Present and Future of Digital Architecture., *The 20th International Conference of the Association for Computer-Aided Architectural Design Research in Asia*.

Sutton, R.S. and Barto, A.G.: 2016, *Reinforcement Learning: An Introduction*, The MIT Press, Cambridge, Massachusetts.

Valenzi, S., Islam, T., Jurica, P. and Cichocki, A.: 2014, Individual Classification of Emotions Using EEG, *Journal of Biomedical Science and Engineering*, **7**, 604-620.