Estimation of affect scores accounting for user responsiveness

Hoang Nguyen (Grenoble Institute of Technology, France), *Serina Koyama (Okayama University), Zeynep Yucel (Okayama University), Akito Monden (Okayama University), Mariko Sasakura (Okayama University)

1. Introduction

Performance of interaction robots can be improved significantly, if they can adapt the services to the varying intrinsic, i.e. cognitive or emotional, state of users. Especially robots for dementia or autism threaphy may profit from such adaptations [1, 2, 3].

Estimation of user's state can be obtained by observing the individual user's behavioral or biological responses. Such observations enable tailoring the content to specific users. On the other hand, it is also possible to apply direct analysis of the stimuli in order to design a broad-gauge scenario for the general tendency of anticipated response.

In this respect, this study contrasts these the user specific and broad-gauge approaches for estimating users' state and provides a discussion on bounding performance rates. To that end, we expose human subjects to emotionally stimulating visual content over a prolonged period and try to build a model of the relation between the stimulus *intensity* and decline in user's responsiveness. We contrast this approach to direct estimation of emotional response, and particularly arousal, from the stimuli.

2. Background

Emotions are short-term changes in mental state arising as a reaction of human autonomic nervous system to various stimuli. Autonomic nervous system regulates other mostly unconscious functions of the human body. Therefore, observation of such psychological responses enables evaluation of changes in the emotional state.

2·1 Affect space

There is no de-facto categorization of emotions due to their subjectivity and cultural dependence. However, there are basically two fundamental ways of categorizing, namely discrete and dimensional.

The most prevailing discrete model belongs to Ekman, who defines six basic emotion categories as Anger, Disgust, Fear, Joy, Sadness and Surprise [4]. Ekman claims these categories are *universal* among all humans, which is highly debated.

As for dimensional approaches, the circumplex model is one of the most popular [5]. It represents emotions on two orthogonal axes, *valence* and *arousal*. Valence rates impressions (i.e. unpleasant to pleasant), whereas arousal rates intensity (i.e. passive to active) [6, 7, 8]. Here, we adopt circumplex model due to its continuous representation of emotions.

2·2 Physiological signals

Physiological signals are the readings taken from bodily processes of human beings, e.g. heart-beat or respiratory rate, skin conductance, or brain electrical activity. Being controlled by autonomic nervous system, they enable spontaneous observation of involuntary reactions to cognitive stimulation. In addition, they are not sensitive to cultural and social differences [9, 10], which makes them objective markers leading to an extensive deployment in affect analysis.

As physiological signal, we use electrodermal activity (EDA) due to its ease of recording. EDA is correlated with the activity of eccrine glands, which are found on nearly all skin locations and in highest concentration on hands [11, 12, 13]. From an anatomical point, EDA reflects two basic electrical properties of the skin: skin conductance level, also termed as *tonic level*, and skin conductance response, also termed as *phasic response*. EDA signal y can be decomposed as,

$$y = t + r + \epsilon, \tag{1}$$

where r and t stand for phasic response and tonic level, respectively, and ϵ is the noise. t changes slowly, whereas r is considered as rapidly changing small waves superimposed on t. Phasic response is sensitive to a wide range of factors such as stimulus novelty, intensity, or affective content. In this study, we focus on intensity of affective content (i.e. arousal).

3. Related work

Several studies investigate the use of physiological signals for estimating affect scores of stimuli. For instance, EDA and Electroencephalography (EEG), are often studied to estimate emotions induced by visual or acoustic stimuli. For instance, Gerdes et al. [14] studied brain activations against visual stimulus, where Trochidis et al. [15] studied the link between acoustic stimulus and EDA, respiration rate and blood volume pulse.

Physiological signals can be coupled with Neural Networks (NN) to predict an emotional state as well. Chanel et al. [16] and McFarland et al. [17] use NN to estimate subjects' emotional states from EEG. Kim et al [18] build a Deep NN (DNN) to predict affective levels of visual stimuli from color, foreground, background features. Peng et al. [19] develop NN to estimate affect levels of also acoustic stimuli.

In addition, EEG [20] and fNIRS [21] are used in interaction robotics to estimate user's attention

and/or stimulate his engagement or estimate several cognitive or affective states.

4. Experimental procedure

We adopt a discrete stimulus paradigm and present emotionally significant visual stimuli as a side show. Four subjects, one female and three male, aged between 21 and 37, participate in the experiments.

The stimuli (i.e. images) are selected from the Open Affective Standardized Image Set (OASIS), which consists of 900 images, labeled for valence and arousal by over 100 annotators. 200 images with mild valence and arousal are selected from OASIS to build the slide show (see Fig. 1), such that each image is displayed for 5 sec followed by a 5 sec reset period. As the subject watches the slide show, EDA is logged by two electrodes on the index and middle fingers' distal phalanges of the non-dominant hand. By this means, both the physiological signals and the affect level are represented quantitatively on respective continuous scales.

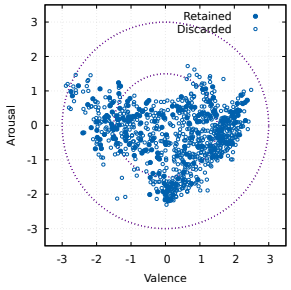

Fig.1 Distribution of valence and arousal labels for OASIS. The 200 images used the experiment are presented with solid circles.

5. Estimation of arousal accounting for responsiveness

In order to interpret EDA, a number of markers are used such as latency, amplitude, rise time etc [12]. For estimating emotional response, the best marker is suggested to be the amplitude of peaks in r [22]. Therefore, as a first step, the EDA signal is decomposed to obtain r. For obtaining the terms in Eq. 1, we use cvxEDA, a convex optimization approach proposed by Greco et al. [22], who model tonic component as a linear combination of cubic spline functions B, together with an offset with a linear term C:

$$t = Bl + Cd,$$

where l and d are the coefficient matrices. On the other hand, r is modeled with a Bateman function,

$$h(\tau) = \left(\exp(-\tau/\tau_0) - \exp(-\tau/\tau_1)\right) u(\tau), \qquad (2)$$

where u is the unit step. Eq.2 is represented with an autoregressive moving average model in frequency domain enabling convex optimization.

After decomposing y into its components, we focus on the amplitude of the peaks P in r, which are known to be sensitive to the arousal of stimuli [22]. We model the relation between P and the annotated arousal levels A accounting for the time elapsed from the start of user's task as in Eq. 3. Namely, P is modeled as a function of arousal score A and display time τ of the stimulus. P is subject to an exponential rate of decay with τ:

$$P(A,t) = A \exp\left(\alpha\tau + \beta\right) + C, \qquad (3)$$

where α, β and C are the parameters to be calibrated.

6. Performance Tests and Metrics

In order to test whether the responsiveness in EDA is attenuated according to the model in Eq. 3 (due to a possible fatigue etc), we adopt a reverse estimation strategy. That is, we assume that the model in Eq. 3 works efficiently and we estimate an arousal level A' for each image, given the amplitude of peaks P and the display time τ,

$$A' = \frac{P - C}{\exp(\alpha\tau + \beta)}.$$

This reverse estimation strategy enables an objective evaluation of performance. In other words, since the arousal level A of the stimulus (i.e. image) is provided as ground truth in OASIS, A' can be compared to A using one of the conventional metrics.

Sign agreement metric (SAGR) is the most common performance measure in affect analysis [23]. SAGR considers an estimation $\hat{\Theta}_{ij}$ as successful, if its sign equals the sign of ground truth Θ_{ij}. Namely,

$$SAGR = \frac{1}{n}\sum_{i=1}^{n} \delta(sign(\hat{\Theta}_i), sign(\Theta_i)), \qquad (4)$$

where δ is the Kronecker delta function. Obviously, SAGR is a nonlinear metric. To have a better insight into performance, we also present Root Mean Squared Error (RMSE) values in performance evaluation.

7. Comparison to reference methods

We consider two reference methods for comparing the proposed approach. Since sensation depends to a large extend on personality, it is not uncommon for human coders to have disagreements. Therefore, for estimating the level of disagreement inherent in the ground truth, we carry out a Monte Carlo (MC) simulation in Sec. 7·1. In addition, we develop a NN based estimator in Sec. 7·2 to provide a comparison with a broad-gauge approach as described in Sec. 1.

7·1 Ecological agreement with MC simulation

Here we use directly the statistics provided with the ground truth of OASIS. Namely, each image in OASIS, I_i, is evaluated by $N_i > 100$ coders. Instead

of each label $\{A_{ij}, j \in [1, N_i]\}$, OASIS provides mean, μ_i, and standard deviation, σ_i for each image I_i.

The optimization procedure described in Sec. 5., considers μ_i as the true of arousal score of I_i. In order to evaluate the disagreement reflected by σ_i, we take a MC standpoint. Namely, we compare m pairs of annotations sets, both randomly drawn from the ground truth distribution, under the assumption that the corresponding levels of arousal $\{A_{ij}\}$ for a given image I_i come from a normal distribution such that $A_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$. For each pair of annotation sets, we compute the corresponding SAGR as in Eq. 4 and average it over the m pairs, as well as RMSE.

7·2 Learning affect scores using NN

The second reference method uses a DNN architecture, which is a compositional model, building potentially more complex information from primitives in a layered manner. Although DNNs are used to estimate affective content of a stimulus [14], [24], they are trained with data from a large number of subjects and thus are often not specific to users.

Convolutional NNs (CNN) have recently become ubiquitous in many computer vision fields. In this study, an existing image classification CNN named *Inception*, which is trained with 14 million images from Imagenet database, is used as a basis for developing a transfer learning architecture [25, 26]. Since our emotional stimuli is visual, we consider Inception to be a suitable starting point for adapting it to estimate affect scores via transfer learning. Namely, the final classification layer is replaced with a bunch of new layers, which are trained with the images from OASIS to estimate their arousal scores.

In particular, two convolutional layers are added for extracting affective features. Subsequently, a flatten layer is added to turn the multi-dimensional feature table into a column of features that are fed to a fully connected (FC) layer followed by a dropout layer that will randomly set 50% of the values to 0 so as to prevent over-fitting. Finally, a last FC layer with a one dimensional output is added with a sigmoid activation function. This layer estimates the affective score of an input image with a continuous value between [0,1]. This value is either rescaled to $[1, 7]$ in case of RMSE cost function or considered as positive or negative, in case of SAGR cost function.

8. Evaluation of performance

For testing the proposed method, EDA signal is divided into 3 minute long sections, corresponding to batches of 18 images. For each image in a batch, the amplitude of peaks P is computed using cvxEDA. A subset of 9 images is randomly chosen and corresponding P values are used to calibrate Eq. 3 by minimizing the squared error. Using the resulting $\{\alpha, \beta, C\}$, the arousal score A' of the remaining subset is estimated. This random selection process is repeated 100 times for all batches and all subjects. The estimation performance in represented in terms of the mean of SAGR and RMSE in Table 1. Since most images

Table 1 Comparison of performance

	SAGR (%)	RMSE
MC-200	41.7 ± 0.01	2.21 ± 0.01
MC-900	43.6 ± 0.01	2.19 ± 0.01
CNN-200	55.5 ± 0.16	1.76 ± 0.10
CNN-900	62.6 ± 0.01	1.98 ± 0.04
Proposed	53.6 ± 0.01	2.5 ± 0.10

used in the experiment have mild arousal scores, i.e. in $[-2, 2]$, (see Fig. 1), even a small error in A' may change its sign and degrade the performance in terms of SAGR easily. For this reason, RMSE is useful for interpreting the error. Therefore, MC and CNN are tested using (i) the 900 images in entire OASIS data set (MC-900 and CNN-900) and (ii) the 200 images which are used as visual stimuli in our experiments (MC-200 and CNN-900), in order to reflect the impact of arousal range of stimuli.

Table 1 shows that the ecological agreement is virtually the same in different arousal ranges (MC-200 and MC-900) in terms of both SAGR and RMSE. This is due to the high standard deviation of the ground truth scores (i.e. some people are extremely sensitive or insensitive to the same stimuli).

The effect of arousal range is obvious on CNN-200. Although the mean SAGR (55.5) seems to improve, it has a high standard deviation (0.16). Most tests achieve around 40% and few tests achieve about 90%, which increase the mean performance misleadingly, and prove the over-fitting issue. Therefore, also the low RMSE (1.76) can be explained by the learning of the input range rather than real affective features. Thus, the higher SAGR and lower RMSE values are misleading in interpreting performance of CNN-200.

The proposed method achieves an SAGR of 53.6% with a very small standard deviation, and estimates the *person-specific arousal* better and more stably than CNN-200. The RMSE is quite high since output range of Eq. 3 is not bounded. Therefore, it estimates the sign of arousal score correctly more often but its amount is overestimated. However, since affect analysis considers the correct estimation of sign to be more important than absolute error, a higher SAGR is desired rather than a lower RMSE.

In case of CNN-900, SAGR is found to be 62.6 ± 0.01 and RMSE is 1.98 ± 0.04. Obviously, larger amount of training samples improve the performance of CNN considerably. However, we consider it unfair to compare these values to the proposed method, since our performance is evaluated using a small set (200 images), but is one of the future research directions to diversify the affect range of the stimuli.

In addition, we searched for studies estimating

arousal scores of OASIS but we could not come across any such studies. Actually, most studies use the images as stimuli and analyze their impact on various physiological signals. In this sense, perhaps the most relevant study to our is by Hu et al [27], who achieve around 40% success rate using image features and offer to integrate image tags to improve recognition of affect qualities.

9. Conclusion

This study proposes a model for user responsiveness using EDA signal. Since there is no standard metric or ground truth for responsiveness, we take a reverse strategy and estimate arousal scores accounting for the decay in responsiveness. The proposed approach is found to give a more stable estimations of arousal score than CNN-200. In addition, although CNN-900 has better performance rates, a direct comparison is not fair, since the input sets are different. Therefore, as a future work, we suggest to (i) diversity the affect range of the stimuli and (ii) embed the model of responsiveness into an NN architecture.

Acknowledgement

This research was supported by JSPS KAKENHI Grant Number 18K18168.

References

[1] A. Tapus, C. Tapus, and M. J. Mataric, "The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia," in *ICORR*, pp. 924–929, IEEE, 2009.

[2] A. Ramachandran, C.-M. Huang, and B. Scassellati, "Give me a break!: Personalized timing strategies to promote learning in robot-child tutoring," in *HRI*, pp. 146–155, ACM, 2017.

[3] K. Wada, T. Shibata, T. Asada, and T. Musha, "Robot therapy for prevention of dementia at home," *Journal of Robotics and Mechatronics*, vol. 19, no. 6, p. 691, 2007.

[4] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[5] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[6] G. Ball and J. Breese, "Modeling the emotional state of computer users," in *Workshop on Personality and Emotion in User Modelling*, 1999.

[7] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm, "Looking at pictures: Affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, no. 3, pp. 261–273, 1993.

[8] P. J. Lang, "The emotion probe: Studies of motivation and attention.," *American psychologist*, vol. 50, no. 5, p. 372, 1995.

[9] P. Rani, N. Sarkar, and C. Liu, "Maintaining optimal challenge in computer games through real-time physiological feedback," in *HCI*, vol. 58, pp. 22–27, 2005.

[10] J. H. Westerink, E. L. Van Den Broek, M. H. Schut, J. Van Herk, and K. Tuinenbreijer, "Computing emotion awareness through galvanic skin response and facial electromyography," in *Probing experience*, pp. 149–162, Springer, 2008.

[11] J. Marieke van Dooren, G.-J. de Vries, and J. H. Janssen, "Emotional sweating across the body: Comparing 16 different skin conductance measurement locations," *Physiology & behavior*, vol. 106, no. 2, pp. 298–304, 2012.

[12] M. E. Dawson, A. M. Schell, and D. L. Filion, "The electrodermal system," *Handbook of psychophysiology*, vol. 2, pp. 200–223, 2007.

[13] B. Figner, R. O. Murphy, *et al.*, "Using skin conductance in judgment and decision making research," *A handbook of process tracing methods for decision research*, pp. 163–184, 2011.

[14] A. Gerdes, M. J. Wieser, A. Mühlberger, P. Weyers, G. W. Alpers, M. M. Plichta, F. Breuer, and P. Pauli, "Brain activations to emotional pictures are differentially associated with valence and arousal ratings," *Frontiers in human neuroscience*, vol. 4, p. 175, 2010.

[15] K. Trochidis, D. Sears, D.-L. Trân, and S. McAdams, "Psychophysiological measures of emotional response to romantic orchestral music and their musical and acoustic correlates," in *From Sounds to Music and Emotions*, Springer, Jan. 2013.

[16] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals," in *MRCS 2006*, pp. 530–537, Springer, 2006.

[17] D. J. McFarland, M. A. Parvaz, W. A. Sarnacki, R. Z. Goldstein, and J. R. Wolpaw, "Prediction of subjective ratings of emotional pictures by eeg features," *Journal of neural engineering*, vol. 14, no. 1, p. 016009, 2016.

[18] H.-R. Kim, Y.-S. Kim, S. J. Kim, and I.-K. Lee, "Building emotional machines: Recognizing image emotions through deep neural networks," *IEEE TMM*, 2018.

[19] S. Peng, L. Zhang, Y. Ban, M. Fang, and S. Winkler, "A deep network for arousal-valence emotion prediction with acoustic-visual cues," *ArXiv e-prints*, May 2018.

[20] D. Szafir and B. Mutlu, "Pay attention!: designing adaptive agents that monitor and improve user engagement," in *SIGCHI*, pp. 11–20, ACM, 2012.

[21] M. Strait and M. Scheutz, "What we can and cannot (yet) do with functional near infrared spectroscopy," *Frontiers in neuroscience*, vol. 8, p. 117, 2014.

[22] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxEDA: A convex optimization approach to electrodermal activity processing," *IEEE T Bio-Med Eng*, vol. 63, no. 4, pp. 797–804, 2016.

[23] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *arXiv preprint*, 2017.

[24] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE STSP*, vol. 11, no. 8, pp. 1301–1309, 2017.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierrchical image database," in *CVPR*, pp. 248–255, Ieee, 2009.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, pp. 2818–2826, 2016.

[27] A. Hu and S. Flaxman, "Multimodal sentiment analysis to explore the structure of emotions," *arXiv preprint arXiv:1805.10205*, 2018.