Comic Readers’ Emotion Estimation Using Bio-signals by Supervised and Unsupervised Learnings

Hiroki Ono\textsuperscript{1a)}, Mutsumi Suganuma\textsuperscript{2}, and Wataru Kameyama\textsuperscript{2}

\textsuperscript{1}Department of Computer Science and Communications Engineering, Graduate School of Fundamental Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
\textsuperscript{2}Faculty of Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
\textsuperscript{a)} kshs_3579@akane.waseda.jp

Abstract: In this paper, we investigate estimating readers’ emotions during reading comics from bio-signals of brainwave, heartbeat, pupil diameter and gazing point. We collect the bio-signals from 11 subjects while they are reading comics, and ask them to answer a questionnaire after reading comics on raised emotions in seven categories including “pleased”, “excited”, “astonished”, “afraid”, “frustrated”, “sad” and “calm”. For the analysis, we use DNN (Deep Neural Network) as a supervised learning method as well as AE (Autoencoder) as an unsupervised one. The questionnaire responses are taken as the correct labels of the raised emotions for DNN, and used to evaluate the clustering results of them by AE. As the results of the analysis, we obtain a high F-score for each emotion estimation by DNN, and find several clusters dominantly including a particular emotion by AE. It suggests that comic readers’ emotions can be estimated from the bio-signals by both supervised and unsupervised learnings.

Keywords: Emotion estimation, bio-signal, comic, DNN, Autoencoder

Classification: Multimedia Systems for Communications

References

[1] Mizuki Matsubara, Olivier Augereau, Charles Lima Sanches, Koichi Kise, “Emotional arousal estimation while reading comics based on physiological signal analysis,” MANPU ’16, no.7, pp.1-4, December, 2016, DOI: 10.1145/3011549.3011556.

[2] J. A. Russell, “A circumplex model of affect,” Journal of Personality and Social Psychology, vol.39, no.6, pp.1161-1178, Dec. 1980.

[3] K. Aizawa, A. Fujimoto, T. Otsubo, T. Ogawa, Y. Matsui, K. Tsubota, H. Ikuta, “Building a Manga Dataset ‘Manga109’ with Annotations for Multimedia Applications,” IEEE Multimedia, 2020, DOI: 10.1109/MMUL.2020.2987895.
1 Introduction

In recent years, the demand for e-books has been on the rise, and comics account for a large share of the market. E-books of comics allow users to enjoy more works on a device, but it is becoming increasingly difficult to recommend comics that users want to read because of a vast amount of them. Most of the latest recommendation systems recommend comics based on the browsing histories of other readers who have similar ones to the target user. However, it is difficult to improve the recommendation accuracy only based on browsing history, because it does not include the reaction of readers to the comics. Therefore, to realize more accurate comic recommendation, we study a method to estimate the reaction of comic readers by using bio-signals to use it as new metadata. In this paper, we focus on raised emotions as a reaction of comic readers, and investigate estimating comic readers’ emotions from bio-signals.

There are many studies on estimating the users’ emotion to contents focused on video and music, but there are relatively few studies focusing on comics. [1] is one of those which uses bio-signals of comic readers, and classifies the arousal into high and low by SVM (Support Vector Machine). However, to improve the accuracy of comic recommendation, estimating only arousal is not enough, but valence should be estimated as well. Therefore, in this paper, we take seven emotions covering the two-dimensional space of arousal and valence proposed by [2], and try to estimate more detailed emotions of comic readers by using supervised and unsupervised learnings. For the supervised learning, we use DNN (Deep Neural Network), and the model is trained to use the questionnaire responses by subjects as correct labels. However, the questionnaire responses by subjects may not be always correct because of their hesitant to self-disclosure, or simply by forgetting, and so on. Therefore, we also use AE (Autoencoder) as the unsupervised learning to estimate the raised emotions, and compare the both results.

2 Experiment

2.1 Devices and Comics

As the bio-signals, we use brainwave, RRI (R-wave R-wave Interval of heartbeats), pupil diameter and gazing point. The brainwaves are measured by EMOTIV EPOC+ with 14 channels of AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 and AF4 in the international 10-20 system with 128[Hz] sampling, RRI is measured by the heartbeat sensor of Polar H10 with V800, and the pupil diameters and gazing points are measured by the eye-tracker of Tobii X60 with 60[Hz]
Five comics are selected from MANGA109 [3][4] dataset, that are expected to raise a lot of emotions during the reading. The titles of the comics are: “Jo-volley” (pp.1-25), “Hisoka Returns!” (pp.1-26), “Akkera Kanjinchou” (pp.1-25), “Eien No Wiz” (pp.2-39), and “Kuroido Ganka” (pp.1-33).

2.2 Procedure
The experiment was conducted on 11 subjects including 8 males and 3 females (μ=21.8, σ=0.386). The subjects wear EPOC+ and Polar H10 and sit on a chair in the experimental space, where Tobii X60 was set. The comics were shown on 23-inch display (Nanao FS2332), and a keyboard was used to turn the pages. After calibrating the eye-tracker, we started to measure the bio-signals, and asked the subjects to read the comics one by one shown on the display. The subjects read the next page at their own pace, but was not allowed to go back to the previous page. After reading each comic, the subjects answer a questionnaire on the PC, where they indicate which emotions raised in which frames during the reading. The subjects choose one emotion per a frame, if any occurred, from the following seven emotions: “pleased”, “excited”, “astonished”, “afraid”, “frustrated”, “sad”, and “calm”. This procedure was repeated for the five comics.

3 Dataset
The time-series of the data measured in the experiment are preprocessed as follows.

Taking the output of EMOTIVPRO software using the 0.5-second window with 0.25-second slide, theta wave (4-8 [Hz]), alpha wave (8-12 [Hz]), low-beta wave (12-16 [Hz]), high-beta wave (16-25 [Hz]) and gamma wave (25-45 [Hz]) are individually averaged for the 14 channels to obtain the data every 0.25 seconds. After that, we calculate the ratio of beta over alpha and the ratio of gamma over alpha for each channel, where the sum of the low-beta and high-beta waves is taken as the beta wave. On each frequency band, the left-right difference and the left-right ratio are also calculated from the symmetrical electrodes of EPOC+.

The RRI is linearly interpolated so that data less than 0.5 seconds or greater than 1.5 seconds are removed as outliers, and the data are resampled at 0.25 seconds. After that, the change rate of RRI is calculated by dividing each RRI value per 0.25 seconds by the mean value of the RRI of the whole.

For the pupil diameters, first, we remove the pupil diameter fluctuations caused by the luminance change of the comic images by applying the pupillary light reflect compensation [5]. In this process, the luminance change of the minimum and maximum values in the square wave form with the sequence of 0.2 [Hz], 0.71 [Hz], 1.25 [Hz], 1.87 [Hz], 2.14 [Hz], 2.5 [Hz] and 3.0[Hz], one cycle for each, is displayed to subjects 5 times before reading comics. The obtained pupil diameters at 60 [Hz] are linearly interpolated to complement the missing data of less than 5 seconds, and the averages of right- and left-pupil diameters are calculated. Then, the DNN as in [5] is used to obtain the pupillary light reflex model to predict the pupil diameters only caused by the luminance. After that, the average luminance value of the comic image within a circular region with 100-
pixel radius centered on the gazing point is fed to the model to obtain the predicted pupil diameter variation, where the gazing points are set to the midpoints of the right- and left-eye gazing points. After applying the same linear interpolation and the averaging to the measured pupil diameters while reading comics, the compensated pupil diameters are calculated by subtracting the predicted values produced by the model from the measured values. Finally, the compensated ones are resampled and averaged with 0.25-second window and 0.25-second slide. And the same resampling and averaging are applied to the gazing points, too.

After the above preprocessing, 144-dimensional data are taken as the explanatory variables in the analysis, as detailed in the following:

- 140 dimensions of brainwave: $4 \times 14 \times 14$ channel $+ 2 \times 14 \times 14$ channel $+ 4 \times 7$ symmetric-channel $+ 4 \times 7$ symmetric-channel
- 1 dimension of RRI change rate.
- 3 dimensions of compensated pupil diameter: diameter at the current time, and these after 1 and 2 seconds.

Since there is a delay in pupil diameter variation in response to stimuli, we use pupil diameter of the current time together with these after 1 and 2 seconds.

For the explained variable, we use questionnaire responses by the subjects for the supervised learning. In the case of unsupervised one, the responses are used to evaluate the clustering results. The matching between the questionnaire responses and the bio-signals is done by using the coordinates of the frames in the annotation of MANGA109 and the calculated gazing points in the preprocess.

Then, a dataset is created by mapping the bio-signals to the corresponding raised emotions.

### 4 Analysis
We use DNN for the supervised learning, and AE for the unsupervised one with applying k-means clustering to the latency space.

**Table 1.** The Execution Parameters of DNN and AE

|                      | DNN                          | AE                      |
|----------------------|------------------------------|-------------------------|
| **Layer structure**  | 144, 250, 180, 100, 50, 7    | 144, 432, 144, 36, 144, 432, 144 |
| **Optimization algorithm** | RMSprop                      | Adam                    |
| **Loss function**    | MSE                          | MSLE                    |
| **Activation function** | ReLU                         | tanh                    |
| **Batch size**       | 64                           | 64                      |
| **Epochs**           | 300                          | 250                     |
| **Regularization**   | L2 regularization            |                         |
We use Keras with Tensorflow backend for DNN and AE, and scikit-learn for k-means clustering. The number of clusters in k-means clustering is set to 35, referring to the Gap statistic. The parameters of DNN and AE are shown in Table I. For DNN, stratified 5-fold cross validation on each subject is conducted to level off the variation of each dataset. For AE, the dataset is randomly divided into 5, and then 5-fold cross validation is applied to make 5 models for each subject. And in order to remove the initial value dependency of k-means, we also apply k-means 10 times per a model.

5 Results and Discussions

Fig. 1 shows the relationship between the F-score of each emotion estimation and the percentage of the emotion labels for each subject in the DNN analysis. From the figure, the overall accuracy is high, although the estimation accuracy varies depending on the subjects and emotions. In addition, “frustrated” seems easy to be estimated from bio-signals because high F-scores are obtained with one exception even though the number of the questionnaire responses is small.

![Fig. 1. Relationship between the Percentage of Emotion Labels and the F-score.](image)

In the AE analysis to check the clusters dominantly possessing a certain emotion, first, we calculate the first quartile on the number of data in all clusters, and ignore the clusters of less than that as noise. Then, we select the clusters possessing a certain emotion more than or equal to 50 percent from the extracted clusters as the characteristic ones with that emotion. After repeating the above operation 10 times per a model, we check whether the characteristic cluster appears at least 7 times out of 10 clustering results. If such a cluster is found in all of the 5 models, we consider that the AE analysis can successfully extract the characteristic clusters corresponding to subjects’ emotions. Fig. 2 shows the clustering results of the AE analysis meeting the above conditions for Subject 2 and 10, showing how many data corresponding to each emotion are clustered. In the case of Subject 2 from Fig. 2 (a), the following clusters are found as representing a certain emotion: Cluster 19 as “astonished”, Cluster 1, 4 and 13 as “afraid”, Cluster 20 and 21 as “sad”, and Cluster 14 as “calm”. In the case of Subject 10 from Fig. 2 (b), Cluster 14 and 24 are found as representing “sad”.
By applying the same analysis to all the clusters of each subject, the clusters dominantly possessing a certain emotion of “astonished”, “afraid” and “calm” are observed in several subjects. It suggests that such emotions can be estimated even by using AE as an unsupervised learning method.

6 Conclusion
In this paper, we estimate comic readers’ emotions while reading comics by using bio-signals aiming for a comic recommendation system. By using DNN as a supervised learning, about 70-90 percent estimation accuracy is achieved for all the emotions. By using AE as an unsupervised learning with applying k-means clustering to the latency space, some clusters are found to dominantly include some specific emotions, which shows the possibility to estimate the emotions even by using an unsupervised method.

However, in both analysis methods, the accuracy varies depending on the subjects, and there are some emotions with low classification accuracy, especially in the AE analysis. In the future, we will improve the accuracy by collecting more data and optimizing the parameters in the analysis.