Journal of Theoretical and Applied Information Technology
31st January 2017. Vol.95. No.2
© 2005 - 2017 JATIT & LLS. All rights reserved·

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INTER-RATER RELIABILITY OF ACTUAL TAGGED EMOTION CATEGORIES VALIDATION USING COHEN'S KAPPA COEFFICIENT

[1]NOR RASHIDAH MD JUREMI, [2]*MOHD ASYRAF ZULKIFLEY, [3]AINI HUSSAIN, [4]WAN MIMI DIYANA WAN ZAKI

Department of Electrical, Electronic & Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.

E-mail: [1]norrashidahmdjuremi@siswa.ukm.edu.my, [2*]asyraf.zulkifley@ukm.edu.my, [3]draini@ukm.edu.my, [4]wmdiyana@ukm.edu.my

ABSTRACT

It is necessary to find the human inter-rater agreement in emotion recognition research especially when handling with publicly available database. This paper discusses the Cohen's Kappa coefficient technique to verify the actual tagged emotion categories for hybrid emotion model using music video as stimulus. This method has been done by finding the degree of inter-rater reliability between the five selected raters. As the results, the values of Cohen's Kappa coefficients are over 0.87 for four actual tagged emotion categories which are happy, relaxed, sad and angry. These values demonstrate that the degree of inter-rater agreement are excellent. The actual tagged emotion categories are selected based on the division of average value of arousal-valence rating.

Keywords: *Emotion recognition, Cohen's Kappa Coefficient.*

1. INTRODUCTION

Recently, emotion recognition is a new growing research field that researchers interested to explore especially in human-computer interaction (HCI) application. Apparently, human emotion can be recognized by speech [1], facial expression [2] and bodily gesture [3]. However, it is easy to hide the true human emotion. Therefore, many researchers have tried to identify the real inner emotion using bio-signal input such as electroencephalogram (EEG).

Various kinds of stimuli are used in EEG emotion recognition to evoke emotion such as emotional image [4], sound [5], music [6], movie [7] and music video [8]. Out of all these stimuli, music video has the greatest effect in evoking emotion [9] in a short time of period since it is a combination of visual and audio reactions. There is only one publicly available database to evoke EEG emotion based on music video stimuli, which is a dedicated database for emotion analysis using physiological signals (DEAP) [8]. Nevertheless, there are no actual tagged emotion categories for each of the music videos. Thus, validation needs to be performed to test the reliability towards actual tagged emotion categories each of music videos as used in DEAP database experiment.

2. LITERATURE REVIEW

Generally, there are two dominant models that widely used in tagged emotion categories toward the stimuli which are discrete and dimensional. Discrete model's philosophy falls as the idea that emotion can be classified into several categories that are mutually exclusive to each other. For example, Ekman [10] in 1971 first popularized this model by introducing six categories of emotion which are sad, anger, surprise, joy, fear and love. However, this model has limitation such that the total number of emotions are limited, therefore several others emotion such as bored, calm or excited cannot be identified. On the other hand, dimensional model stated that emotions can be mapped and occupied into dimension of space such as two dimensional Russell circumplex [11] ranges from unpleasant to pleasant state for valance axis and from deactivation to activation state for arousal axis. Participants need to rate ranging from 1 to 9 for both arousal and valence axis after experiencing the emotion as verbal self-assessment of emotion. Recently, a new model was introduced, namely as a hybrid discrete-dimensional [12-15] model to convert the dimensional model into discrete model. This model is more advance compared to the other models since this model can vary the number of emotion categories. The number of emotion categories depends on the division of the interval of rating, both arousal and valence axis. For example in Russell circumplex model, the number of emotions can be categorized to be either four [8], six [6, 13] or nine [14].

Variability tagged emotion categories exists when handling with the diversity of participants during data acquisition. For example, a participant might tag a particular music video as happy emotion but another participant might tag it as calm emotion. Therefore, the tagged emotion categories are not consistent. Due to this issue, validating the tagged emotion categories are crucial step, so that it can be used as actual tagged emotion categories (also known as ground truth of classes) particularly use in classification module. The output of emotion categories from the classification module need to be compared with the actual tagged emotion categories to calculate the performance measure of the system.

One of the most important benchmark database and publicly available is developed by Sander Koelstra et al. [8]. It comprises of 40 music videos tested on 32 participants. As in previous research [18], two experiments are conducted to investigate the reliability of publicly database, which are International affective picture system (IAPS) and International Affective Digitized Sound system (IADS) towards benchmark DEAP database. As a result, similar brain pattern for a particular emotion can be observed between the tested and benchmarked database. However, to the best of author's knowledge, there is no inter-rater reliability test has been devised to validate the reliability of actual tagged emotion categories towards DEAP database. Thus, the main objective of this paper is to validate the reliability of 40 music videos taken from DEAP database.

In this study, Cohen's Kappa coefficient method is proposed to measure the degree of agreement between the raters for multi-categorical items called as inter-rater reliability. Coefficient of Cohen Kappa was first introduced by Cohen [19] in 1960. This method has been widely utilized especially in speech recognition [20], biology [21], clinical [22] and others. Thus, Cohen's Kappa coefficient is used as a validation tool of actual tagged emotion categories in hybrid model.

3. METHODOLOGY

Overview of database: This study used the downloadable DEAP (a database for emotion analysis using physiological signals) database to investigate the inter-rater agreement of tagged emotion categories using music video as the emotional stimulus. Figure 1 shows the experiment protocol that conducted during the acquisition of physiological signal. The experiment consisted of two sections with a short break session between the two sections. Each section comprised of 20 trials that exhibit 20 music videos. Each trial started with the displaying the current number of trial on the screen. After 2s, the fixation cross presented on the center of LCD screen. After 5s, the particular music video played for the 60s of duration. After that, participant need to rate the arousal and valence values.

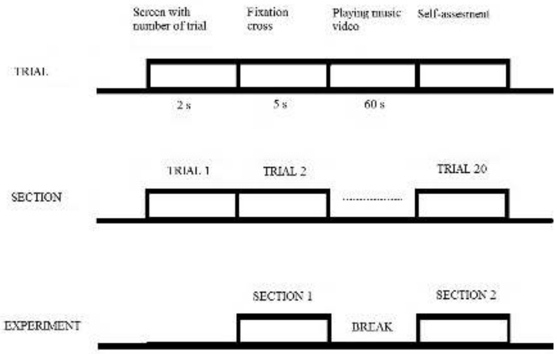

Figure 1: The Structure of Experiment during Evoking the Emotions.

Cohen's Kappa Coefficient Experiment: The tagged emotion categories of 40 music videos are not provided by the authors. However, the authors provided the list of YouTube links for each of music video. Furthermore, the self-assessment rating of the 32 participants for 40 music videos were also enclosed. Thus, four emotion categories elected as the actual tagged emotion categories by thresholding the middle value of arousal and valence rating, which is 5. The amount of actual tagged emotion categories depend on the suitability distribution of average values of arousal and valence rating (32 participants) towards 40 music videos as shown in figure 2. Based on the distribution, average rating values mostly lie on the middle line of arousal-axis, whereas the average values of rating are equally distributed for both high and low rating of valence-axis. Therefore, four actual tagged emotion categories are selected to avoid some of the categories that do not have sample data. Happy emotion represent the high arousal and high valence (HAHV) shown in the 1st quadrant, relaxed emotion represent the low arousal and high valence (LAHV) shown in the 2nd quadrant, sad emotion represent the low arousal and low valence (LALV) shown in the 3rd quadrant and angry emotion represent the high arousal and low valence (HALV) shown in the 4th quadrant. All the actual tagged emotion categories are illustrated in figure 3.

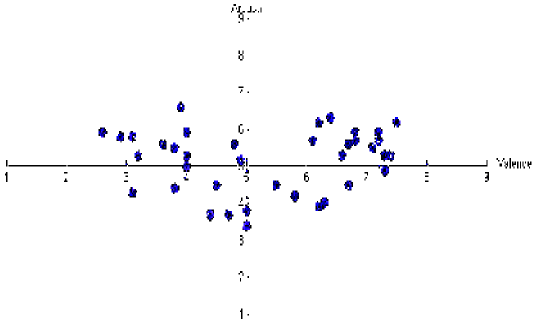

Figure 2: The Average Value For Valence And Arousal Rating For 40 Music Videos.

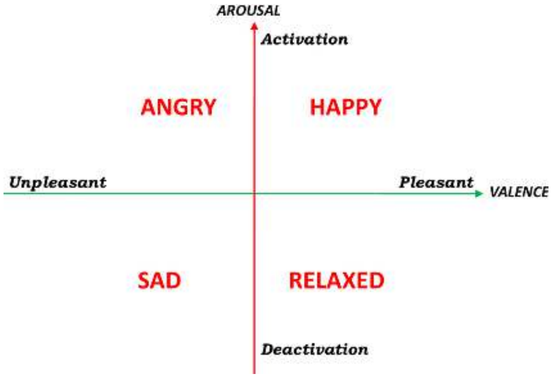

Figure 3: Four Category Of Emotions

In this study, the raters consisted of 3 women and 2 men to rate the actual tagged emotion categories for each of 40 music videos. All tested music videos were first downloaded and then they are played on a computer screen after brightness adjustment has been done for comfortability of the rater's eyes. Playback order of the music videos is similar for each of the raters. After watching the videos, each participant needs to choose one of the actual tagged emotion categories either happy, relaxed, sad or happy to explain the best emotion that characterizes a particular music video. Later, the proportion values of actual tagged emotion categories are computed and then compiled into the contingency table. Basically, Cohen's Kappa coefficient is a method to measure the level of agreement for qualitative items (categories) between two raters. Nevertheless, when multiple raters are involved, the combination technique is used to select the pair of raters. Specifically, the total number of combination $\begin{pmatrix} p \\ q \end{pmatrix}$ is calculated as in equation 1 where q is the number of ways which is two and p is the total number of raters which is five.

$$\begin{pmatrix} p \\ q \end{pmatrix} = \frac{p!}{q!(p-q)!} \qquad (1)$$

Table 1 illustrates the typical contingency table to access the agreement of k number of actual tagged emotion categories between two raters. For $i,j = 1,2,\ldots,k$ each of p_{ij} value represents the proportion value of actual tagged emotion categories that rater 1 classified a music video as category i and rater 2 classified a music video as category j. The diagonal elements, p_{ii} in the table 1 represent the proportion values of both raters agreed with the actual tagged emotion categories.

Table 1: Two Raters Proportion with k Categories.

Rater 1	Rater 2				Total
	1	2	…	k	
1	p_{11}	p_{12}	…	p_{1k}	p_{1+}
2	p_{21}	P_{22}	…	p_{2k}	p_{2+}
…	…	…	…	…	…
k	P_{k1}	P_{k2}	…	p_{kk}	p_{k+}
Total	P_{+1}	P_{+2}	…	P_{+k}	1

Therefore, the coefficient of Cohen Kappa is calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \qquad (2)$$

where p_o is the proportion of both raters agreed for the actual tagged emotion categories and p_e is the overall proportion of chance-expected agreement. The value of p_o and p_e are calculated using equation (3) and (4).

$$p_o = \sum_{i=1}^{k} p_{ii} \qquad (3)$$

$$p_e = \sum_{i=1}^{k} p_{i+}p_{+i} \qquad (4)$$

The value of κ varies from 1 to -1 depending on the extent of agreement among the raters. Perfect agreement of κ which is 1 is rarely achieved, however the value closed to 1 indicates that the degree of agreement is excellent. Otherwise, when the value closed to -1 shows that the raters are highly disagreed with the actual tagged emotion categories. The degree of agreement based on Cohen's Kappa coefficient are listed in table 2.

Table 2: The Degree of Agreement Based on Cohen's Kappa Coefficient.

κ	Degree of Agreement
$0.8 < \kappa \leq 1.0$	Excellent
$0.6 < \kappa \leq 0.8$	Good
$0.4 < \kappa \leq 0.6$	Fair
$\kappa \leq 0.4$	Weak

4. RESULTS AND DISCUSSION

In this study, the actual tagged emotion categories of 40 music videos from DEAP database are validated by five raters. The actual tagged emotion category for each of the music videos is selected based on average value of arousal and valence ratings. R^{a-b} represent the Cohen's Kappa coefficient where $a,b = \{1,2,3,4,5\}$. For instance, the symbol of R^{1-2} describes the Cohen's Kappa coefficient for rater 1 and 2. The value of Cohen's Kappa coefficient computed for all possible pairs of raters. Hence there are ten pairs of the rater combinations which are R^{1-2}, R^{1-3}, R^{1-4}, R^{1-5}, R^{2-3}, R^{2-4}, R^{2-5}, R^{3-4}, R^{3-5} and R^{4-5}. Thus, the value of Cohen's Kappa coefficient for each pair was plotted as shown in figure 4. In general observation, the degree of Cohen's Kappa coefficient for all pairs are excellent, ranging from 0.8655 to 0.9664.

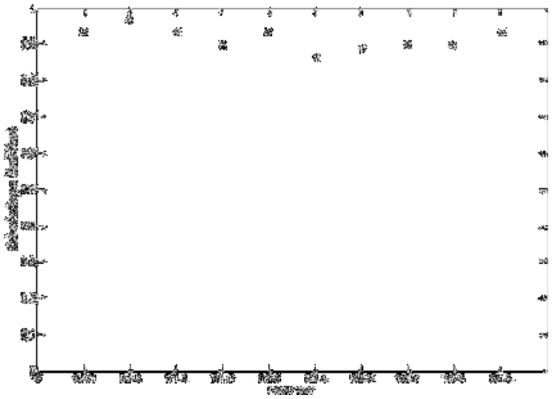

Figure 4: Value Of K For Rater Pairs.

The average value of Cohen's Kappa coefficient, $\overline{\kappa_a}$ is calculated as in equation (5).

$$\overline{\kappa_a} = \frac{\sum_{a=1}^{n} \kappa_a}{n} \qquad (5)$$

where n is 4 which is the total number of a pairs.

Generally, the $\overline{\kappa_a}$ results shown in figure 5 demonstrate that the degree of agreement are excellent, ranging from 0.9045 to 0.9322.

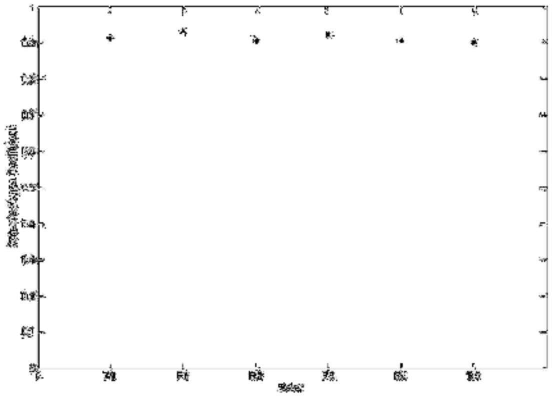

Figure 5: Average Value of Individual Rater.

The distribution of Cohen's Kappa coefficient for each individual rater was illustrated in figure (6). Each rater returned a small variance of Cohen's Kappa coefficient, which indicates that the values of Cohen's Kappa coefficients are nearly consistent and closed to the mean value.

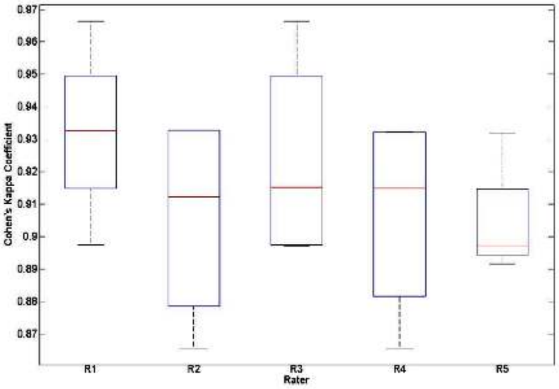

Figure 6: The Distribution of Cohen's Kappa Coefficient for Five Raters.

The main limitation of this study is due to limited choice of emotions in which the raters are forced to choose to tag the emotion categories; either happy, relax, sad or angry. In addition, Cohen's Kappa method is one of the inter-rater reliability tests that concerns on the experience of multiple raters in a large number of samples, which is not ideal for this study. Various raters might have different experience and interpretation of each of the music videos that lead to the misclassified of the actual tagged emotion categories. Nevertheless, these findings which is high value of Cohen's Kappa coefficient proved that the raters highly agreed with the actual tagged emotion categories of 40 music videos based on the average values of arousal and valence rating.

5. CONCLUSION

In conclusion, the Cohen's Kappa coefficient has been successfully applied in this study to validate the actual tagged emotion categories for hybrid emotion model using music video as stimulus. From the Cohen's Kappa coefficient results, DEAP database shows high reliability and can be used as benchmark database for investigating relationship between emotion recognition and the music video as stimulus. In future, the actual tagged emotion categories for 40 music videos obtained in this study are suitable to be applied in emotion classification module using machine learning techniques such as artificial neural network, support vector machine and others.

6. ACKNOWLEDGMENT

This work was funded by Ministry of Science and Technology Malaysia (06-01-02-SF1005) and Universiti Kebangsaan Malaysia (GUP-2015-053).

REFRENCES:

[1] T. Kostoulas, I. Mporas, O. Kocsis, T. Ganchev, N. Katsaounos, J. J. Santamaria, S Jimenez-Murcia, F. Fernandez-Aranda and N. Fakotakis, "Affective Speech Interface in Serious Games for Supporting Therapy of Mental Disorders", *Expert Systems with Applications*, 39, 2012, pp. 11072-11079.

[2] H. Ali, M. Hariharan, S. Yaacob and A. H. Adom, "Facial Emotion Recognition using Empirical Mode Decomposition", *Expert Systems with Applications*, 42, 2015, pp. 1261-1277.

[3] L. Hinzman and S. D. Kelly, "Effects of Emotional Body Language on Rapid Out-group Judgments", Journal of Experimental Social Psychology, 49, 2013, pp. 152-155.

[4] Y. Liu and O. Sourina, "EEG Databases for Emotion Recognition", *Proceedings of International Conference on Cyberworlds*, October 21-23, 2013, pp. 302-309.

[5] O. Sourina and Y. Liu, "A Fractal-Based Algorithm of Emotion Recognition from EEG Using Arousal-Valence Model", *Proceedings of the International Conference on BioInspired*

Systems and Signal Processing, January 26-29, 2011, pp. 209-214.

[6] Y. Liu, O. Sourina, and M. K. Nguyen, “Real-Time EEG-Based Emotion Recognition for Music Therapy”, *Journal on Multimodal User Interfaces*, Vol. 5, No.1, 2012, pp. 27-35.

[7] D. Nie, X.-W. Wang, L.-C. Shi, and B.-L. Lu, “EEG-Based Emotion Recognition During Watching Movies”, *Proceedings of the International IEEE EMBS Conference on Neural Engineering*, 2011, pp. 667-670.

[8] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, “DEAP: A Database for Emotion Analysis using Physiological Signals”, *IEEE Transactions on Affective Computing*, Vol. 3, No. 1, 2012, pp. 18-31.

[9] R. Parke, E. Chew, and C. Kyriakakis, “Quantitative and Visual Analysis of the Impact of Music on Perceived Emotion of Film”, *Computers in Entertainment*, 2007, Vol. 5, No. 3, pp. 1-21.

[10] P. Ekman, “An Argument for Basic Emotions”, *Cognition and Emotion*, 1992, Vol. 6, No. 3, pp. 169-200.

[11] J. Russell, “A Circumplex Model of Affect”, *Journal of Personality and Social Psychology*, 1980, Vol. 39, pp. 1161–1178.

[12] L. F. Barrett, “Solving The Emotion Paradox: Categorization and the Experience of Emotion”, *Personality and Social Psychology Review*, 2006, Vol. 10, No. 1, pp. 20-46.

[13] I. C. Christie and B. H. Friedman, “Autonomic Specificity of Discrete Emotion and Dimensions of Affective Space: A Multivariate Approach”, *International Journal of Psychophysiology*, 2004, Vol. 51, pp. 143-153.

[14] K. S. Kassam, A. R. Markey, V. L. Cherkassky, G. Loewenstein and M. A. Just, “Identifying Emotions on the Basis of Neural Activation”, PloS one, 8(6), 2013, pp. 1-12.

[15] J. Russell, “Core Affect and the Psychological Construction of Emotion”, *Psychological Review*, 2003, Vol. 110, No. 1, pp. 145-172.

[16] Y. Song, S. Dixon and M. Pearce, “Evaluation of Musical Features for Emotion Classification”, *Proceeding of International Society for Music Information Retrieval*, 2012, pp. 523-528.

[17] Y. Liu, O. Sourina, and M. K. Nguyen, “Real-Time EEG-Based Emotion Recognition and Its Applications”, *Transactions on Computational Science XII*, Vol. 6670, 2011, pp. 256-277.

[18] Y. Liu and O. Sourina, “EEG databases for emotion recognition”, *Proceedings of International Conference on Cyberworlds*, 2013, pp. 302-309.

[19] J. L. Fleiss, B. Levin, and M. C. Paik, “Statistical Methods for Rates and Proportions”, New Jersey: *John Wiley*, 2003, pp. 604-605.

[20] M. H. M. Zaman, M. M. Mustafa and A. Hussain, “Inter-rater Reliability of Accessing the Intelligibility of Band-limited Transformed Speech Using Nonlinear Frequency Compression”, *International Conference on Computer, Communications, and Control Technology*, September 2-4, 2014, pp. 126-129.

[21] S. A. Fattahi, Z. Othman and Z. A. Othman, “New Approach for Imbalanced Biological Dataset Classification”, *Journal of Theoretical and Applied Information Technology*, Vol. 72, No.1, 2015, pp. 40-57.

[22] N. Wongpakaran, T. Wongpakaran, D. Wedding and K. L Gwet, “A Comparison of Cohen’s Kappa and Gwet’s AC1 When Calculating Inter-Rater Reliability Coefficients: A Study Conducted with Personality Disorder Samples”, *BMC Medical Research Methodology*, Vol. 13, No.61, 2013, pp. 1-7.