R-R vs GSR – An inter-domain study for arousal recognition

Martin Gjoreski
Department of Intelligent Systems,
Jožef Stefan Institute
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia
martin.gjoreski@ijs.si

Blagojce Mitrevski
Faculty of Computer Science and
Engineering
Skopje, R. Macedonia

Mitja Luštrek, Matjaž Gams
Department of Intelligent Systems,
Jožef Stefan Institute
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia

ABSTRACT

Arousal recognition is an important task in mobile health and human-computer interaction (HCI). In mobile health, it can contribute to timely detection and improved management of mental health, e.g., depression and bipolar disorders, and in HCI it can enhance user experience. However, which machine-learning (ML) methods and which input is most suitable for arousal recognition, are challenging and open research questions, which we analyze in this paper.

We present an inter-domain study for arousal recognition on six different datasets, recorded with twelve different hardware sensors from which we analyze galvanic skin response (GSR) data from GSR sensors and R-R data extracted from Electrocardiography (ECG) or blood volume pulse (BVP) sensors. The data belongs to 191 different subjects and sums up to 260 hours of labelled data. The six datasets are processed and translated into a common spectro-temporal space, and features are extracted and fed into ML algorithms to build models for arousal recognition. When one model is built for each dataset, it turns out that whether the R-R, GSR, or merged features yield the best results is domain (dataset) dependent. When all datasets are merged into one and used to train and evaluate the models, the R-R models slightly outperformed the GSR models.

Keywords

Arousal recognition; GSR; R-R; machine learning; health.

1. INTRODUCTION

The field of affective computing [1] has been introduced almost two decades ago and yet modeling affective states has remained a challenging task. Its importance is mainly reflected in the domain of human-computer interaction (HCI) and mobile health. In the HCI, it enables a more natural and emotionally intelligent interaction. In the mobile health, it contributes to the timely detection and management of emotional and mental disorders such as depression, bipolar disorders and posttraumatic stress disorder. For example, the cost of work-related depression in Europe, was estimated to €617 billion annually in 2013. The total was made up of costs resulting from absenteeism and presenteeism (€272 billion), loss of productivity (€242 billion), health care costs of €63 billion and social welfare costs in the form of disability benefit payments (€39 billion) [2].

Affective states are complex states that results in psychological and physiological changes that influence our behaving and thinking [17]. These psycho-physiological changes can be captured by a wearable device equipped with GSR, ECG or BVP sensor. For example, the emotional state of fear usually initiates rapid heartbeat, rapid breathing, sweating, and muscle tension, which are physiological signs that can be captured using wearables.

The affective states can be modeled using a discrete or a continuous approach. In the discrete approach, the affect (emotions) is represented as discrete and distinct state, i.e., anger, fear, sadness, happiness, boredom, disgust and neutral [15]. In the continuous approach, the emotions are represented in 2D or 3D space of activeness, valance and dominance [3]. Unlike the discrete approach, this model does not suffer from vague definitions and fuzzy boundaries, and has been widely used in affective studies [4] [5] [6]. The use of the same annotating model allows for an inter-study analysis.

In this study we examine arousal recognition from GSR and heart–related physiological data, captured via: chest-worn ECG and GSR sensors, finger-worn BVP sensor, and wrist-worn GSR sensor and pulse oximeter (PPG) sensor. The data belongs to six publicly available datasets for affect recognition, in which there are 191 different subjects (70 females) and nearly 260 hours of arousal-labelled data.

All of this introduces the problem of inter-domain learning, to which ML techniques are sensitive. To overcome this problem, we use preprocessing techniques to translate the datasets into a common spectro-temporal space of R-R and GSR data. After the preprocessing, R-R and GSR features are extracted and are fed into ML algorithms to build models for arousal recognition. Finally, the results between different experimental setups are compared, i.e., models that use only R-R features, models that used only GSR features and models that use both R-R and GSR features. This comparison is performed in a dataset-specific setup and merged setup where all datasets are merged in one. At the end, the experimental results are discussed and the study is concluded with remarks for further work.

RELATED WORK

Affect recognition is an established computer-science field, but one with many challenges remaining. There has been many studies confirming that affect recognition can be performed using speech analysis [8], video analysis [9], or physiological sensors in combination with ML. The majority of the methods that use physiological signals use data from ECG, electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), GSR, electrooculography (EOG) and/or BVP sensors.

In general, the methods based on EEG data outperform the methods based on other data [4] [5], probably due to the fact the EEG provides a more direct channel to one’s mind. However, even though EEG achieves the best results, it is not applicable in normal everyday life. In contrast, affect recognition from R-R intervals or GSR data, is much more unobtrusive since this data can be extracted from ECG sensors, BVP sensors, PPG or GSR sensors, most of which can be found in a wrist device (e.g., Empatica [10] and Microsoft Band [11]). Regarding the typical ML approaches for affect recognition, Iacoviello et al. have combined discrete wavelet transformation, principal component analysis and support vector machine (SVM) to build a hybrid classification framework using EEG [12]. Khezri et al. used EEG combined with GSR to recognize six basic emotions via K-nearest neighbors (KNN) classifiers [13]. Verma et al. [14] developed an ensemble approach using EEG, electromyography (EMG), ECG, GSR, and EOG. Mehmood and Lee used independent component analysis to extract emotional indicators from EEG, EMG, GSR, ECG, and (effective refractory period) ERP [15]. Mikuckas et al. [16] presented a HCI system for emotional state recognition that uses spectro-temporal analysis only on R-R signals. More specifically, they focused on recognizing stressful states by means of the heart rate variability (HRV) analysis.

However, a clear comparison between ML methods for affect recognition from unobtrusively captured sensor data (e.g., R-R vs. GSR data) has not been presented yet, since most of these studies focused on only one dataset and a combination of the sensor data, aiming towards the highest performance and disregarding the obtrusiveness of the system. In this work, we analyze which ML algorithms in combination with which data type (either R-R intervals or GSR) would yield best performance across six different datasets (domains) for arousal recognition.

2. DATA

The data belongs to six publicly available datasets for affect recognition: Ascertain, Deap, Driving workload dataset, Cognitive load dataset, Mahnob, and Amigos. Overall, nearly 260 hours of arousal-labelled data that belong to 191 subjects. The Table 1 presents the number of subjects per dataset, the mean age, number of trials per subject, mean duration of each trial, duration of data per subject - in seconds, and overall duration.

Table 1. Experimental data summary.

					Duration in seconds		
Dataset	**Subjects**	**Females**	**Mean age**	**Trials**	**µ trial**	**Per subject**	**Overall data**
Ascertain	58	21	31	36	80	2880	167040
DEAP	32	16	26,9	40	60	2400	76800
Driving	10	3	35,6	1	1800	1800	18000
Cognitive	21	0	28	2	2400	4800	100800
Mahnob	30	17	26	40	80	3200	96000
Amgos	40	13	28	16	86	1376	55040
Overall	191	70	29,25	135	884,0	15080	458640

The four datasets, Ascertain, Deap, Mahnob and Amigos, were already labelled with the subjective arousal level. One difference between these datasets was the arousal scale used for annotating. For example, the Ascertain dataset used 7-point arousal scale, whereas the Deap dataset used 9-point arousal scale (1 is very low, and 9 is very high). From the both scales, we split the labels in the middle, which is the same split used in the original studies. Similar step was performed for the Mahnob dataset. The two datasets, Driving workload and Cognitive load, did not contain labels for subjective arousal level. The Driving workload dataset was labelled with subjective ratings for a workload during driving session. For this dataset, we presume that increased workload corresponds to increased arousal. Thus, we used the workload ratings as an arousal ratings. The split for high arousal was put on 60%. Similarly, the cognitive load dataset was labelled for subjective stress level during stress inducing cognitive load tasks (mathematical equations). The subjective scale was from 0 to 4 (no stress, low, medium and high stress). We put the limit for high arousal on 2 (medium stress).

3. METHODS

3.1 Pre-processing and feature extraction

3.1.1 R-R data

The preprocessing is essential, since it allows merging of the six different datasets. For the heart-related data, it translates the physiological signals (ECG or BVP) to R-R intervals and performs temporal and spectral analysis. First, a peak detection algorithm is applied to detect the R-R peaks. Next, temporal analysis, i.e., calculating the time distance between the detected peaks, detects the R-R intervals. Once the R-R intervals are detected they can be analyzed as a time-series. First, each R-R signal is filtered using median filter. After the median filter, person specific winsorization is performed with the threshold parameter of 3 to remove outlier R-R intervals. From the filtered R-R signals, periodogram is calculated using the Lomb-Scargle algorithm developed by Lomb and further analyzed by Scargle. The Lomb-Scargle algorithm is used for spectral analysis of unequally spaced data (as are the R-R intervals). Finally, the following HRV features were calculated from the time and spectral representation of the R-R signals: meanHR, meanRR, sdnn, sdsd, rmssd, pnn20, pnn50, sd1, sd2, sd1/sd2, lf, hf, lf/hf [32].

3.1.2 GSR data

For merging the GSR data, several problems were addressed. Each dataset is recorded with different GSR hardware, thus the data can be presented in different units and different scales. To address this problem, each GSR signal was converted to µS (micro Siemens). Next, to address the inter-participant variability of the signal, person-specific min-max normalization was performed, i.e., each signal was translated between 0 and 1 using person specific winsorized minimum and maximum values. The winsorization parameter was set to 3. Finally, the GSR signal was filtered using lowpass filter with a cut-off frequency of 1HZ.

The filtered GSR signal was used to calculate the following GSR features: mean, standard deviation, quartiles (25th and 75th), quartile deviation, derivative of the signal, sum of the signal, number of responses in the signal, rate of responses in the signal, sum of the responses, sum of positive derivative, proportion of positive derivative, derivative of the tonic component of the signal, difference between the tonic component and the overall signal[24].

3.2 Machine learning

After the feature extraction, every data entry consists of 16 R-R features and 14 GSR features which can be input for typical ML algorithms. Models were built using seven different ML algorithms: Random Forest, Support Vector Machine, Gradient Boosting Classifier, and AdaBoost Classifier, KNN Classifier, Gaussian Naive Bayes and Decision Tree Classifier. The algorithms were used as implemented in the Scikitlearn, the Python ML library. For each algorithm, randomized search on hyper parameters was performed on the training data using 2-fold validation.

4. EXPERIMENTAL RESULTS

Two types of experiments were performed, dataset specific experiments, and experiments with merged datasets. The evaluation was performed using trial-specific 10-fold cross-validation, i.e., the data segments that belong to one trial (e.g., one affective stimuli), can either belong only to the training set or only to the test set, thus there was no overlapping between the training and test data.

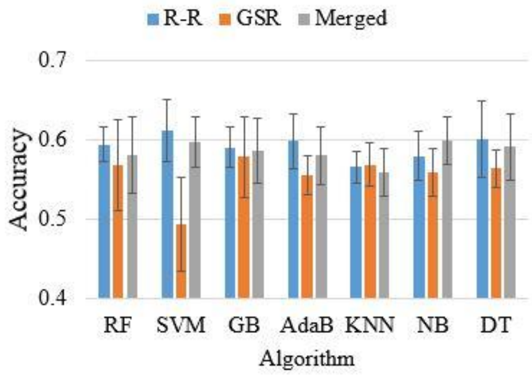
4.1 Dataset specific

The results for the dataset specific experiments are presented in Table 2. The first column represents the ML algorithm, the second column represents the features used as input to the algorithm (R-R, GSR or Merged - M) and the rest of the columns represent the dataset which is used for training and evaluation using the trial-dependent 10-fold cross-validation. We report the mean accuracy ± the standard evaluation for the 10 folds. For each dataset, the best performing model(s) is(are) marked with green. For example, on the Ascertain and the Driving workload dataset, the best performing algorithm is the SVM, on the Deap dataset the best performing algorithm is the RF, on the Cognitive Load and the Mahnob datasets the best performing is the NB, and on the Amigos dataset the best performing is the AdaBoost algorithm.

When we compare which input (R-R features, GSR features or Merged-M) provide better accuracy, on two datasets (the Asceratin and the Driving workload) the results are the same, on the Deap dataset, the R-R features provide better results, on the Cognitive Load dataset the highest accuracy is achieved both for the GSR and the Merged features, on the Mahnob dataset the GSR features provide best accuracy and on the Amigos dataset the Merged features.

4.2 Merged datasets

For these experiments, all datasets were merged into one, and the trial-dependent 10-fold cross-validation was used to evaluate the ML models. The results are presented in Figure 2. The results show that the models that use the R-R intervals as input, consistently outperform the models that use GSR features as input.



5. CONCLUSION AND DISCUSSION

We presented a study in….

The results on the dataset specific setup showed that, out of the ML algorithms tested, none yields the best performance on all datasets. In addition to that, a clear conclusion cannot be made whether the R-R, GSR or the Merged features yield the best results – this is domain (dataset) dependent.

Table 2. Dataset specific experimental results. Mean accuracy ± stdDev for trial-specific 10-fold cross validation. The best performing models per dataset are marked with green.

Algorithm	Features	Dataset					
		Ascertain	Deap	D. Workload	Cog. Load	Mahnob	Amigos
RF	R-R	0.655 ± 0.07	0.556 ± 0.03	0.785 ± 0.24	0.739 ± 0.13	0.580 ± 0.11	0.536 ± 0.06
	GSR	0.638 ± 0.06	0.503 ± 0.04	0.780 ± 0.24	0.763 ± 0.12	0.611 ± 0.07	0.473 ± 0.11
	M	0.653 ± 0.05	0.540 ± 0.04	0.785 ± 0.25	0.755 ± 0.13	0.611 ± 0.10	0.559 ± 0.10
SVM	R-R	0.664 ± 0.07	0.536 ± 0.05	0.795 ± 0.26	0.717 ± 0.21	0.623 ± 0.15	0.521 ± 0.24
	GSR	0.664 ± 0.07	0.525 ± 0.05	0.795 ± 0.26	0.712 ± 0.20	0.588 ± 0.10	0.470 ± 0.12
	M	0.664 ± 0.07	0.513 ± 0.03	0.795 ± 0.26	0.691 ± 0.18	0.623 ± 0.15	0.506 ± 0.13
GB	R-R	0.649 ± 0.07	0.554 ± 0.03	0.785 ± 0.20	0.736 ± 0.15	0.578 ± 0.11	0.543 ± 0.06
	GSR	0.642 ± 0.05	0.500 ± 0.04	0.800 ± 0.21	0.743 ± 0.12	0.609 ± 0.08	0.527 ± 0.09
	M	0.644 ± 0.05	0.533 ± 0.03	0.755 ± 0.23	0.761 ± 0.15	0.609 ± 0.11	0.542 ± 0.09
AdaB	R-R	0.658 ± 0.06	0.532 ± 0.02	0.750 ± 0.23	0.718 ± 0.13	0.580 ± 0.09	0.531 ± 0.07
	GSR	0.633 ± 0.05	0.485 ± 0.03	0.750 ± 0.22	0.740 ± 0.13	0.589 ± 0.08	0.514 ± 0.09
	M	0.623 ± 0.05	0.526 ± 0.03	0.755 ± 0.22	0.766 ± 0.16	0.610 ± 0.08	0.560 ± 0.08
KNN	R-R	0.625 ± 0.05	0.509 ± 0.02	0.710 ± 0.19	0.715 ± 0.13	0.582 ± 0.07	0.509 ± 0.05
	GSR	0.590 ± 0.06	0.496 ± 0.04	0.795 ± 0.26	0.772 ± 0.09	0.605 ± 0.06	0.533 ± 0.08
	M	0.600 ± 0.05	0.490 ± 0.02	0.750 ± 0.23	0.770 ± 0.13	0.601 ± 0.09	0.533 ± 0.06
NB	R-R	0.654 ± 0.07	0.537 ± 0.04	0.735 ± 0.15	0.748 ± 0.15	0.574 ± 0.06	0.485 ± 0.09
	GSR	0.602 ± 0.04	0.537 ± 0.05	0.540 ± 0.22	0.803 ± 0.09	0.624 ± 0.07	0.454 ± 0.10
	M	0.591 ± 0.04	0.535 ± 0.06	0.665 ± 0.17	0.804 ± 0.12	0.592 ± 0.06	0.486 ± 0.09
DT	R-R	0.664 ± 0.07	0.519 ± 0.05	0.685 ± 0.17	0.736 ± 0.15	0.597 ± 0.09	0.505 ± 0.06
	GSR	0.640 ± 0.05	0.542 ± 0.05	0.765 ± 0.22	0.734 ± 0.08	0.583 ± 0.09	0.483 ± 0.11
	M	0.650 ± 0.05	0.524 ± 0.04	0.615 ± 0.22	0.704 ± 0.09	0.581 ± 0.13	0.551 ± 0.09

On the merged dataset experiments, the R-R models slightly outperformed the GSR models. This might be due to: (i) having more R-R features that GSR; (ii) having R-R features in frequency domain but no GSR features in frequency domain; (iii) the method for merging the data from the heart-related sensors providing more consistent features across datasets due to less noise in the ECG, BVP data.

In future, we plan to investigate intelligent combinations of ML models in order to gain accuracy. In addition to that, we plan to investigate more advanced techniques such as deep neural networks and transfer learning, which might be able to learn general models that will be able to generalize across different domains. Finally, once we find the best performing scenario, we will generalize the method for arousal recognition to method for valence recognition and method for discrete emotion recognition.

6. REFERENCES

1. R. Picard. Affective Computing. Cambridge, MA: MIT Press, 1997.
2. Depression cost: http://ec.europa.eu/health//sites/health/files/mental_health/docs/matrix_economic_analysis_mh_promotion_en.pdf, [Accessed 27.03.2017].
3. J. A. Russell. A circumplex model of affect. Journal of Personality and Social Psychology, 1980.
4. R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, N Sebe. ASCERTAIN: Emotion and Personality Recognition using Commercial Sensors. IEEE Transactions on Affective Computing. 2016.
5. S. Koelstra, C. Muehl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras. DEAP: A Database for Emotion Analysis using Physiological Signals (PDF). IEEE Transaction on Affective Computing, 2012.
6. M.K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras. Nicu Sebe. DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses. IEEE Transactions on Affective Computing, 2015.
7. N.R. Lomb. Least-squares frequency analysis of unequally spaced data. Astrophysics and Space Science, vol 39, pp. 447-462, 1976
8. hwG. Trigeorgis et al.Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.
9. I. Abdic, L. Fridman, D. McDuff, E. Marchi, B. Reimer, Schuller, B. Driver Frustration Detection From Audio and Video. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'16), 2016.
10. M. Garbarino, M. Lai, D. Bender, R. W. Picard, S. Tognett. Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. 4th International Conference on Wireless Mobile Communication and Healthcare, pp. 3-6, 2014.
11. Microsoft band. https://www.microsoft.com/microsoft-band/en-us
12. D. Iacovielloa, A. Petraccab, M. Spezialettib, G. Placidib. A real-time classification algorithm for EEG-based BCI driven by self-induced emotions. Computer Methods and Programs in Biomedicine, 2015.
13. M. Khezria, M.Firoozabadib, A. R. Sharafata. Reliable emotion recognition system based on dynamic adaptive fusion of forehead biopotentials and physiological signals.
14. G. K. Verma, U. S. Tiwary. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. NeuroImage, 2014.
15. R. M. Mehmooda, H. J. Leea. A novel feature extraction method based on late positive potential for emotion recognition in human brain signal patterns. Computers & Electrical Engineering, 2016.
16. A. Mikuckas, I. Mikuckiene, A. Venckauskas, E. Kazanavicius2, R. Lukas2, I. Plauska. Emotion Recognition in Human Computer Interaction Systems. Elektronika Ir Elektrotechnika, Reserarch Journal, Kaunas University of Technology, 2014.
17. Wei Liu, Wei-Long Zheng, Bao-Liang Lu. Multimodal Emotion Recognition Using Multimodal Deep Learning. Online. Available at: https://arxiv.org/abs/1602.08225, 2016.
18. W-L. Zheng, B-L Lu. A multimodal approach to estimating vigilance using EEG and forehead EOG. Journal of Neural Engineering, 2017.
19. P. Bashivan, I. Rish, M.Yeasin, N. Codella. Learning Representations From Eeg With Deep Recurrent-Convolutional Neural Networks. Online. Available at: https://arxiv.org/abs/1511.06448.
20. Z. Yin, M. Zhao, Y. Wang, J. Yang, J. Zhang. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. Comput Methods Programs Biomed. 2017.
21. H.P. Martínez, Y. Bengio, G. N. Yannakakis. Learning Deep Physiological Models of Affect. IEEE Computational intelligence magazine, 2013.
22. K.Weiss, T. M. Khoshgoftaar, D. Wang. A survey of transfer learning. Journal of Big Data, 2016.
23. S. Schneegass, B. Pfleging, N. Broy, A. Schmidt, Frederik Heinrich. A Data Set of Real World Driving to Assess Driver Workload. 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2013.
24. M. Gjoreski, M. Luštrek, M. Gams, H. Gjoreski. Monitoring stress with a wrist device using context. Journal of Biomedical Informatics, 2017, in press.
25. M. Gjoreski, H. Gjoreski, M. Luštrek, M. Gams. Continuous stress detection using a wrist device: in laboratory and real life. ACM Conf. on Ubiquitous Computing, Workshop on mentalhealth, pp. 1185-1193, 2016.
26. M. Soleymani, T.Pun. A Multimodal Database for Affect Recognition and Implicit Tagging, IEEE Transactions On Affective Computing, 2012.
27. L. H. Negri. Peak detection algorithm. Python Implementation. Online. Available at: http://pythonhosted.org/PeakUtils/.
28. M. Wu, PhD thesis. Michigan State University; 2006. Trimmed and Winsorized Eestimators.
29. J.D. Scargle. Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal, vol 263, pp. 835-853, 1982.
30. D. P. Kingma, J. Ba. Adam: A Method for Stochastic Optimization, http://arxiv.org/abs/1412.6980, 2014.
31. Tensorflow. Online. Available at: https://www.tensorflow.org/
32. R. Castaldoa, P. Melillob, U. Bracalec, M. Casertaa,c, M. Triassic, L. Pecchiaa. Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis. Biomedical Signal Processing and Control. 2015.
33. Scikit-learn, Python machine-learning library http://scikit-learn.org/dev/_downloads/scikit-learn-docs.pdf
34. L.J.P, van der Maaten., G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research. 9: 2579–2605, 2008.