Emotion Recognition by Point Process Characterization of Heartbeat Dynamics

Akshay Sujatha Ravindran, *Student Member, IEEE*, Sho Nakagome, *Student Member, IEEE*, Dilranjan S. Wickramasuriya, *Student Member, IEEE*, Jose L. Contreras-Vidal, *Fellow, IEEE* and Rose T. Faghih, *Member, IEEE*

***Abstract*— Recognizing human emotion from heartbeat information alone is a challenging but ongoing research area. Here, we utilize a point process model to characterize heartbeat dynamics and use it to extract instantaneous heart rate variability (HRV) features. These features are then fed into a convolutional neural network (CNN) to characterize different emotional states from small windows. On average, we achieved over 60% classification accuracy and as high as 77% in some subjects. This is comparable to other studies that use a combination of physiological signals as opposed to only HRV measures as done here. Informative features were identified for the different affective states. These findings enable the possibility of augmenting electrocardiogram or photoplethysmogram monitoring wearable devices with automated human emotion recognition capabilities for mental health applications. They also allow for the use of instantaneous estimation of HRV features to be used in combination with models that use other types of physiological signals for instantaneous emotion recognition.**

***Index Terms*—emotion, heart rate, point process modeling**

I. INTRODUCTION

Automated human emotion recognition has been an interesting and ongoing research area involving multidisciplinary expertise. It has been reported that over two million US citizens have been diagnosed with bipolar disorder [1]. Despite the prevalence, current practices used for assessing emotions are mainly conducted by means of basic questionnaires or are solely based on physician experience. Some of the commonly used emotional spaces are: the discrete emotion model proposed by Ekman [2] with six universal emotions (happiness, surprise, anger, disgust, sadness and fear), the two dimensional valence-arousal model by Russell [3], and the PAD (pleasure, arousal, and dominance) model [4] which incorporates an additional axis known as dominance.

In general, people tend to express their emotions through the tone of their voice, gestures, posture and facial expressions [5]. The usage of gesture, facial expression and speech-based emotion detection techniques are susceptible to social masking, as they can be easily modulated/suppressed by the subjects themselves [6]. This led to the popularity of emotion recognition techniques using physiological signals within the last decade, as they originate from autonomic nervous system (ANS) activity and hence cannot be triggered by volitional control [7].

Prior studies have shown that the fusion of multiple physiological signals provides better emotion recognition accuracy. Experimental evidence has demonstrated that the analysis of heart rate variability (HRV) in both the time and frequency domains can provide insight into changes associated with emotion processing [8], [9]. The ANS comprises of both the sympathetic and parasympathetic branches, both of which are innervated to the heart in the sinoatrial (SA) node, which is in charge of the heart's neuromodulation in response to sympathetic and vagal activities [10].

Many state-of-the-art classification rates in emotion recognition are now realized using deep learning models. However, most of them rely on electroencephalography (EEG), electrooculography (EOG) or electromyography (EMG) signals and do not employ heart rate (HR) as the former provides continuous recordings which could be used to train these models. The studies that use electrocardiogram (ECG) or photoplethysmography (PPG) typically extract features over the entire duration of the trial. Therefore, HR measures are typically avoided in continuous estimation problems with smaller windows since they do not estimate the instantaneous values of HR which can be obtained by modelling heartbeats as a point process, i.e. a sequence of binary events in continuous time.

Therefore in this study, we model heartbeats using a point process framework with parameters systematically chosen using maximum likelihood estimation (MLE) and the Bayesian information criterion (BIC) [11]. The model is used to extract instantaneous HRV features which are then fed into a convolutional neural network (CNN) to characterize different emotional states. By exploring the features learnt by the CNN, we also identify the most relevant features for classifying different emotional states. This could enable continuous emotion recognition using existing wearable devices that make use of PPG signals.

II. METHODS

A. Data

The Database for Emotion Analysis using Physiological Signals (DEAP) is an open source data set [12] containing multimodal physiological signals – EEG, EMG, skin conductance, respiration, PPG and body temperature. The dataset was recorded from 32 healthy subjects (16 male; 16

A. S. Ravindran, S. Nakagome, D. S. Wickramasuriya, J. L. Contreras-Vidal and R. T. Faghih are with the Department of Electrical and Computer Engineering at the University of Houston, Houston, TX 77004 USA. (e-mail:{asujatharavindran, snakagome, dswickramasuriya, jlcontreras-vidal, rtfaghih}@uh.edu). This work was supported in part by NSF grant 1755780 – CRII: CPS: Wearable-Machine Interface Architectures. Correspondence should be addressed to senior author Rose T. Faghih.

female) while they watched a series of music videos meant to elicit different emotions [12]. Subjects were instructed to provide feedback on each music video they viewed and rate them separately on valence, arousal, dominance, liking and familiarity scales ranging from 1-9. Scores less than 5 were labelled as low and those greater than or equal to 5 as high.

In this study, we make use of the preprocessed data from the first 22 subjects to characterize emotions (valence, dominance and arousal) solely using PPG. The data from the remaining subjects were collected at a different study location and are not used here. Each of the signals were sampled at 128 Hz and segmented into the respective trials.

B. Preprocessing

The PPG signals were first downsampled to 64 Hz. They were then high-pass filtered at 0.5 Hz to remove drift and then low-pass filtered at 5 Hz to make the peaks prominent. Peak detection was used to detect the valleys of the signal and the differences between successive valleys were calculated as heartbeat intervals. These intervals calculated from the PPG are equivalent to the RR-intervals in an EKG (R-peaks are a prominent feature in an EKG signal and accompany ventricular contraction).

C. Point Process Modeling

We used the point process model in [13] for characterizing heartbeat dynamics. A point process characterization provides a mathematical basis for modeling physical activities such as heartbeats that have inhomogeneous Poisson arrival times. Assume K successive heartbeat occurrences at times u_k during the observation interval $(0,T]$ such that $0 < u_1 < u_2 < \cdots < u_K \leq T$. We can define the RR-intervals as $w_k = u_k - u_{k-1}$ and the history term as $H_k = \{u_k, w_k, w_{k-1}, \ldots, w_{k-p+1}\}$ where p is the model order (the number of history terms).

At time $t > u_k$, the inter-arrival time for the next heartbeat can be modeled using a History Dependent Inverse Gaussian (HDIG) density function:

$$f(t|H_k,\theta) = \left[\frac{\theta_{p+1}}{2\pi(t-u_k)^3}\right]^{\frac{1}{2}} \times \exp\left\{-\frac{1}{2}\frac{\theta_{p+1}[t-u_k-\mu(H_k,\theta)]^2}{\mu(H_k,\theta)^2(t-u_k)}\right\}, \quad (1)$$

where $\mu(H_k,\theta) = \theta_0 + \sum_{j=1}^{p}\theta_j w_{k-j+1} > 0$ and $\Theta = (\theta_0, \theta_1, \ldots, \theta_{p+1})$.

The model [13] is then used to estimate the instantaneous RR-interval at each discrete time bin $j\Delta$ for $j = 1, \ldots, J$, where $\Delta = 1/F_s$ seconds and F_s is the sampling rate of the physiological signals.

D. Model Selection

BIC is one of the commonly used criteria for model selection. We used MLE to calculate θ. MLE and BIC were computed iteratively by changing the values of p from 1 to 10 to determine the model order that gave the lowest BIC values.

$$BIC = -2\log(likelihood) + K \times \log(N) \quad (2)$$

where K is the number of parameters estimated and N is the sample size.

E. Features

We extracted the following features from the RR-intervals to characterize the HRV variations associated with different emotions in each of the trials. These features were selected as they have shown sensitivity towards emotion recognition in prior studies. The following instantaneous features were computed from the HDIG model:

1) Mean RR-interval (μ).
2) Variance of the RR-intervals (Var).
3) High Frequency Power (HFP): Power in 0.15 to 0.4 Hz.
4) Low Frequency Power (LFP): Power in 0.04 to 0.15 Hz.
5) Very Low Frequency Power (VLFP): Power $<$ 0.04 Hz.
6) Total Power (TP): The spectral power.

All the features were z-scored for each individual for the remainder of the analysis. A prior study had analyzed the best window size for emotion recognition in the DEAP dataset and had found that classification accuracy was higher when using windows which were 3-10 seconds long [14]. Therefore a window size of 5 seconds was selected in the study. Signals were then segmented into windows of 5 seconds for training the models.

F. Classification

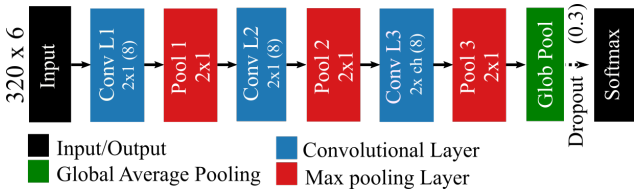

Fig. 1. CNN architecture used for emotion recognition. Inputs to the model were 5 second long HRV features and the output node corresponds to low/high class.

We used a deep learning model using the CNN architecture shown in Figure 1, which was implemented in Python 3.6 using Keras 2.1.5 wrapper with Tensorflow as the backend, to classify the HRV features based on the emotion scores (each for valence, arousal and dominance) from the windowed HRV signals. The CNN had 3 blocks of convolution-pooling layers with a Tanh activation following a batch normalization layer in each block. Each convolution layer had 8 nodes. Tanh non-linearity gave better performance compared to ReLu for the dataset. A global average pooling layer was used instead of a dense layer to reduce the number of parameters and a dropout layer (rate = 0.3) was used prior to the softmax layer to reduce the overfit. We used an Adam optimiser with a learning rate of 1e-5. The model was trained for a maximum of 200 epochs and a batch size of 32, with an early stopping condition to stop the training if the validation loss did not drop in 5 successive epochs. In

order to avoid overfitting to the class having more number of samples, the loss function was weighted during training alone, based on the ratio of samples in both the classes. This ensures that the model will pay attention to both classes equally. The hyper parameters and the architecture were selected heuristically based on assessing how incremental changes modified performance. Separate models of the same architecture were used for individual subjects for each of the emotion groups.

We also used a traditional machine learning model using ensemble learning implemented using the *fitcensemble* function in Matlab 2018b (MathWorks, Inc., Natick, Massachusetts, United States) to act as a baseline comparison. Bootstrap aggregation that bags tree-stump based weak learners were used for learning in the ensemble model. Multiple bootstrapped replicas were selected randomly from these data and these replicas were then used to grow decision trees. The average response of the prediction from all the trees gives the final prediction. The input to the ensemble were the mean values of the features in each of the 5 second windows.

The models were trained for each of the 3 emotional classes separately. For individual classifiers, 10% of trials from each of the low/high classes were kept for validation and the 80% trials that remained were used to train the models. A 5-fold cross-validation is performed such that different combination of trials are used in each of the folds.

G. Feature Learning

To identify the most relevant features for modelling different emotional states, we artificially corrupted individual features and assessed how the performance of the model got degraded. Signals were corrupted by randomly shuffling the features to lose the temporal congruity. This was followed by the addition of Gaussian noise with the mean and standard deviation of the original feature. The original pre-trained model was then tested on this data and the performance degradation was assessed for each feature. This was repeated for all subjects and the mean change for the top 50% of the subjects that gave the highest accuracy was averaged to estimate the feature importance.

III. Results

A. Goodness of Fit

We estimated the best order (3.1 ± 1.12 across subjects) of the HDIG model for each individual subject. These yielded the least BIC values. The selected order was used to fit the HDIG model for characterizing the heartbeat dynamics for each of the trials. An example showing the instantaneous RR-intervals estimated using the HDIG model for one subject is shown in Figure 2.

B. Classification

The 5-fold cross validated accuracy for the CNN and the ensemble learners is summarized in Table I. The mean classification accuracy across subjects was higher for valence and dominance compared to arousal. The CNN outperformed the Ensemble learning model when estimating emotion from the 5 second long windows. Also checking the F1 and recall score, we can see that by weighting the classes during training, the CNN has learned not to be biased towards any particular class as the classification scores are similar unlike that using the ensemble method.

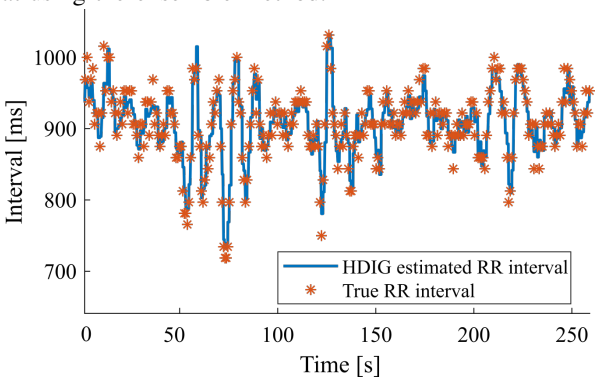

Fig. 2. An example showing both the original RR-intervals and the estimated instantaneous intervals using the HDIG model for a continuous recording from one subject.

TABLE I

Cross Validation accuracy of classifiers. The mean/max is computed across subjects; Ens: Ensemble Model

Metrics	Valence (%)		Dominance (%)		Arousal (%)	
	CNN	Ens	CNN	Ens	CNN	Ens
Max Accuracy	72.9	65.9	77.5	71	77.4	56.6
Mean Accuracy	61.9	54.4	62.6	51.1	60.1	49.3
F1 Score	59.7	52.1	60.7	44.9	57.8	40.2
Precision	63.6	54.4	63.9	51.2	62.3	49.3
Recall	61.9	52	62.6	42.5	60.1	36.2

Figure 3a shows the distribution of the cross validation accuracies for the CNN model for each of the classes of emotional states.

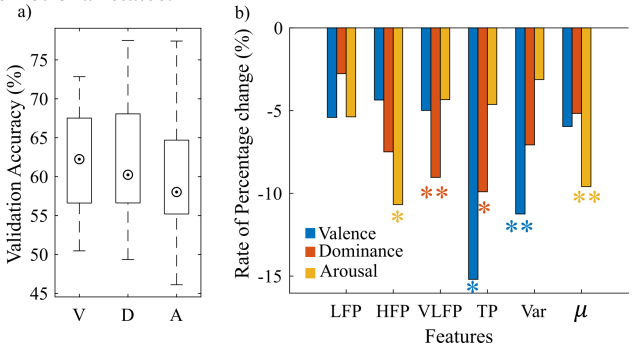

Fig. 3. a) Distribution of the cross validation accuracies across subjects (CNN model); V: valence, D: dominance, A: arousal b) Most important features identified by the CNN model; *: Most important feature, **: Second most important feature.

C. Feature Learning

Figure 3b shows an example of the most informative features for each of the three conditions. For valence, variance was the most important feature followed by TP. For dominance, TP and VLFP were identified to be the most important and for the arousal class, the HFP followed by the mean RR were the most relevant features recognized by the CNN.

IV. Discussion

In the original DEAP paper, Koelstra *et al.* [12] reported an overall F1 score of 60.8% for valence and 53.3% for arousal by using features from all the peripheral signals (skin conductance, PPG, EMG, temperature and respiration). In this study we achieved a comparable or better accuracy (59.7 % for valence and 57.8% for arousal) solely by using heartbeat dynamics and that too from a smaller window size. The study by Candra *et al.* [14] which compared variable windows sizes for emotion recognition using the DEAP dataset obtained a maximum classification accuracy of 65.03% for valence and 65.33% for arousal which is slightly higher than what we obtained (61.9 % for valence and 60.1% for arousal). This is expected as they used EEG and not HR, which is expected to have more discriminatory information. In the original paper [12], they found a high correlation between dominance and valence ratings which could be one of the reasons for the similar performances in both valence and dominance and similar to them we had a higher accuracy in classifying valence.

Mean HR is expected to change with varying levels of arousal scales since we tend to have a higher HR when we are angry or fearful. For instance, studies have shown that arousal state induces a higher rate of respiration [15]. HFP is strongly associated with the respiratory cycle driven modulations of the SA nodes. The stretch of the SA node in association with atrial pressure induced by respiratory modulations could change the HF-HRV as a result of change in mean HR. This could be why HFP was identified as the most important feature by the CNN for arousal state detection. Mean HR and power in high frequency bands are also shown to be modulated in response to different arousal levels in prior studies as well [16]. Previous studies have also shown that variation in band power and the standard deviation/variance of HR are important features to discriminate differences in valence conditions as well [16], [17].

In conclusion, an HDIG model optimized using MLE and BIC for each subject was used to characterize RR-intervals and generate instantaneous HRV features. We were then able to distinguish between high and low levels of valence, dominance and arousal using only heartbeat dynamics based on a CNN model yielding similar/better classification scores to that of the decoder used in the original paper that used all physiological signals. The most important features learnt by these models were also identified.

V. Limitations and Future Work

There exists the possibility that the model accuracy might have suffered due to the use of subjective ratings in the study, as it could be prone to subjective bias and rating inexperience. The features were not optimized for the ensemble learner which would have resulted in sub-optimal performance for that model. However, that model was used to serve as a baseline alone. Also, the window length for the classification was selected based on a prior study that used EEG and that might not be the best window size for HR. Due to the computation overhead this was not optimized.

Future work would involve incorporating EEG features for improving classification accuracy. Examining physiological signal changes in additional scenarios (e.g. art, dance and drama) that evoke different emotions could be yet another direction of research.

References

[1] R. C. Kessler, K. A. McGonagle, S. Zhao, C. B. Nelson, M. Hughes, S. Eshleman, H.-U. Wittchen, and K. S. Kendler, "Lifetime and 12-month prevalence of dsm-iii-r psychiatric disorders in the united states: results from the national comorbidity survey," *Archives of general psychiatry*, vol. 51, no. 1, pp. 8–19, 1994.

[2] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

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

[4] A. Mehrabian, "Framework for a comprehensive description and measurement of emotional states." *Genetic, social, and general psychology monographs*, 1995.

[5] T. Bänziger, D. Grandjean, and K. R. Scherer, "Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (mert)." *Emotion*, vol. 9, no. 5, p. 691, 2009.

[6] P. Rani, C. Liu, N. Sarkar, and E. Vanman, "An empirical study of machine learning techniques for affect recognition in human–robot interaction," *Pattern Analysis and Applications*, vol. 9, no. 1, pp. 58–69, 2006.

[7] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on*. IEEE, 2011, pp. 410–415.

[8] M. T. Valderas, J. Bolea, P. Laguna, M. Vallverdú, and R. Bailón, "Human emotion recognition using heart rate variability analysis with spectral bands based on respiration," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 6134–6137.

[9] R. D. Lane, K. McRae, E. M. Reiman, K. Chen, G. L. Ahern, and J. F. Thayer, "Neural correlates of heart rate variability during emotion," *Neuroimage*, vol. 44, no. 1, pp. 213–222, 2009.

[10] K. Sunagawa, T. Kawada, and T. Nakahara, "Dynamic nonlinear vago-sympathetic interaction in regulating heart rate," *Heart and vessels*, vol. 13, no. 4, pp. 157–174, 1998.

[11] E. Wit, E. v. d. Heuvel, and J.-W. Romeijn, "all models are wrong...: an introduction to model uncertainty," *Statistica Neerlandica*, vol. 66, no. 3, pp. 217–236, 2012.

[12] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[13] R. Barbieri, E. C. Matten, A. A. Alabi, and E. N. Brown, "A point-process model of human heartbeat intervals: new definitions of heart rate and heart rate variability," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 288, no. 1, pp. H424–H435, 2005.

[14] H. Candra, M. Yuwono, R. Chai, A. Handojoseno, I. Elamvazuthi, H. T. Nguyen, and S. Su, "Investigation of window size in classification of eeg-emotion signal with wavelet entropy and support vector machine," in *2015 37th Annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 7250–7253.

[15] I. Homma and Y. Masaoka, "Breathing rhythms and emotions," *Experimental physiology*, vol. 93, no. 9, pp. 1011–1021, 2008.

[16] C. Godin, F. Prost-Boucle, A. Campagne, S. Charbonnier, S. Bonnet, and A. Vidal, "Selection of the most relevant physiological features for classifying emotion," *Emotion*, vol. 40, p. 20, 2015.

[17] S. Rezaei, S. Moharreri, N. J. Dabanloo, and S. Parvaneh, "Evaluating valence level of pictures stimuli in heart rate variability response," in *2015 Computing in Cardiology Conference (CinC)*. IEEE, 2015, pp. 1057–1060.