Gender Effects on an EEG-Based Emotion Level Classification System

I. De La Pava[1(✉)], A. Álvarez[1], P. Herrera[2], G. Castellanos-Dominguez[3], and A. Orozco[1]

[1] Automatic Research Group, Faculty of Engineerings, Universidad Tecnológica de Pereira, Pereira, Colombia
{ide,andres.alvarez1,aaog}@utp.edu.co
[2] Psychiatry, Neuroscience and Community Group, School of Medicine, Universidad Tecnológica de Pereira, Pereira, Colombia
p.herrera@utp.edu.co
[3] Signal Processing and Recognition Group, Department of Electrical and Electronic Engineering, Universidad Nacional de Colombia, Manizales, Colombia
cgcastellanosd@unal.edu.co

Abstract. Emotion level classification systems based on features extracted from physiological signals have promising applications in human-computer interfaces. Moreover, there is increasing evidence that points to gender differences in the processing of emotional stimuli. However, such differences are commonly overlooked during the assessment and development of the systems in question. Here, we study gender differences in the performance of an emotion level classification system and its constituting elements, namely features extracted from electroencephalography (EEG) signals, and emotion level ratings in the Arousal/Valence (AV) dimensional space elicited from audiovisual stimuli. Obtained results show differences in the physiological and expressive responses of men and women, and in overall classification performance for the valence dimension.

Keywords: Emotion assessment · Electroencephalography · Gender differences

1 Introduction

The development of automatic systems for emotional response recognition arouses considerable interest because of their potential impact on the field of human-computer interfaces [9]. These systems aim to predict a subject's emotional response to a stimulus from audiovisual or physiological data. The emotional responses are coded using either a discrete representation (specific emotions such as happiness, fear, sadness, or anger) or a dimensional representation, that is, latent dimensions whose combination give rise to specific emotions. The most common dimensional representation being the Arousal/Valence (AV) emotional space, in which the arousal dimension places emotions in a range varying

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 810–819, 2019.
https://doi.org/10.1007/978-3-030-13469-3_94

from inactive or calm to active or excited, while the valence dimension does so in a range varying from negative or unpleasant to positive or pleasant [7]. Regarding the data needed to infer the emotional response, EEG has received increasing attention from affective computing researchers since it is a non-invasive, fast and relatively inexpensive neuroimaging technique with well-established connections to cognitive processes [1]. However, due to the complexity of the spectral and spatiotemporal relationships between EEG signals and emotional responses that need to be deciphered, the performance of EEG-based systems remains relatively low, especially when complex stimuli (e.g., music videos) are used for emotion elicitation. These hurdles have led to many feature extraction, feature selection and classification methods being explored to improve the performance of such systems [5,9,11].

Moreover, little attention has been paid to demographic characteristics that could impact the performance of emotional response recognition systems, such as gender or age. There is a growing body of evidence suggesting differences in the way men and women process emotional stimuli: men may rely more on the recall of past emotional experiences to evaluate new ones than women [1], unpleasant and high arousing stimuli may evoke greater electrophysiological responses in women relative to men [8], and reports on EEG patterns have shown stronger group coherence among women compared with men during emotion [12]. Despite this, a recent survey gathering research about emotion recognition from EEG signals over the past 9 years, highlights important concerns: 24% of the analyzed works do not specify the participants' gender and 68% are based on unbalanced samples in regard to the men-women ratio, being men overrepresented [1]. Therefore, it is relevant to study how gender differences affect the constituting components of emotional response recognition systems based on EEG, as well as their impact on those systems' overall performance.

In this work, we study the gender differences present in an EEG-based emotion level classification system. We do so at the level of the physiological responses measured by EEG and at the level of the subjective experience and/or expressive response associated with a provided stimulus. We also study whether these differences are reflected in the overall classification performance. To those ends, the EEG data is characterized through a differential entropy (DE) analysis, which has been shown to outperform other EEG characterization strategies in emotion level classification tasks [11]; while the expressive responses are coded as ratings in independent scales for valence and arousal. These features, along with gender class vectors, are used as inputs to simple K-nearest neighbor classifiers. The analyses are carried out on the publicly available Database for Emotion Analysis Using Physiological Signals (DEAP) [7]. Obtained results show a gender difference for the valence dimension in terms of classification performance. They also show that it is feasible to classify the subjects' gender from the DE features and from ratings in the AV emotional space.

2 Materials and Methods

Database. We use EEG and subjective data obtained from the Database for Emotion Analysis Using Physiological Signals (DEAP) [7]. This database holds EEG recordings obtained from 32 healthy subjects (15 females and 17 males of average age 25.4 years and 28.3 years, respectively) while performing 40 trials of an emotion elicitation experiment. In each experiment, the subjects were exposed to 1-min long music videos. Afterward, the participants rated the music videos on discrete 9-point scales for valence and arousal; where 1 and 9 represented the lowest and highest level of emotional elicitation in either dimension. The EEG data were acquired at a sampling rate of 512 Hz using a 32 channel BioSemi ActiveTwo system. The dataset underwent eye blink artifact removal via independent component analysis, frequency down-sampling to 128 Hz, and bandpass filtering from 4–45 Hz. Besides, the data were averaged to the common reference and segmented into trials lasting 63 s.

EEG Feature Extraction. We compute the differential entropy (DE) as follows: given the EEG data recorded from each subject $\{\boldsymbol{X}_n \in \mathbb{R}^{C\times M}\}_{n=1}^{N}$, where $C = 32$ is the number of channels, $M = 8064$ is the number of samples registered for each channel, and $N = 40$ is the number of trials or videos; we segment the last 10 s of each signal using a square window of 1280 points. This segmentation is performed under the premise that the subject's emotional response to the 1-min long music video should be more evident towards the end of the stimulus due to emotional reverberation [3]. Then, we compute the average power spectral density over each EEG rhythm (θ: 4–7 Hz, α: 8–13 Hz, β: 14–30 Hz, and γ: 31–45 Hz) using the Fast Fourier Transform of the segmented data $\boldsymbol{X}'_n \in \mathbb{R}^{C\times L}$, with $L = 1280$. The features are restricted to 4–45 Hz since the pre-processed version of the DEAP dataset is bandpass-filtered in that frequency range. Finally, we compute the DE as the logarithm of the power spectral density [11], obtaining for each subject a set of DE matrices $\{\boldsymbol{\zeta}_n \in \mathbb{R}^{C\times 4}\}_{n=1}^{N}$.

Gender Differences from DE Feature Sets. We concatenate the DE feature set so that the matrices $\boldsymbol{\zeta}_n$ are transformed into vectors $\boldsymbol{\zeta}'_n \in \mathbb{R}^{1\times(C\times 4)}$. Then, we stack the $N = 40$ row vectors $\boldsymbol{\zeta}'_n$, corresponding to each video to form a matrix $\boldsymbol{\Lambda}_i \in \mathbb{R}^{N\times(C\times 4)}$ that contains all DE features extracted from the i*th* subject. For each matrix $\boldsymbol{\Lambda}_i$, we assign a vector of gender labels $\boldsymbol{l}_i \in \{0,1\}^N$ (label "0" is assigned to men and "1" to women). Next, we set up a subject independent classification system with the aim of estimating the gender labels $\boldsymbol{l}_i$ from the DE features. We train a Euclidean distance-based K-nearest neighbor classifier using a 32-fold cross-validation setup. For the j*th* fold the features from subject j*th*, $\boldsymbol{\Lambda}_{i=j}$, are used as the testing set and the features from all other subjects, $\boldsymbol{\Lambda}_{i\neq j}$, are used as the training set. The classification is performed for all DE features, and for subsets of different EEG rhythms (θ, α, β, and γ) and cortical areas (frontal: Fp1, Fp2, AF3, AF4, F3, F4, F7, F8; central: FC1, FC2, FC5, FC6, C3, C4, CP1, CP2, CP5, CP6; parietal: P3, P4, P7, P8, PO3, PO4; temporal: T7, T8; and occipital: O1, O2) by selecting the appropriate columns of $\boldsymbol{\Lambda}_i$.

Gender Differences from AV Rating Scales. For each emotion dimension, we build a matrix containing the ratings (in a 1 to 9 scale) given by the subjects to each video: $\boldsymbol{Y} \in \mathbb{R}^{S \times N}$, where $S = 32$ is the number of subjects. Then, a gender label is assigned to each row of $\boldsymbol{Y}$, obtaining a labels vector $\boldsymbol{\xi} \in \{0, 1\}^S$ (0 for men and 1 for women). Next, we employ a Euclidean distance-based K-nearest neighbor classifier to estimate the gender labels from $\boldsymbol{Y}$, following a leave-one-out cross-validation scheme. The gender classification is performed independently for valence and arousal.

Subject-Dependent Emotion Level Classification. We devise the emotion level assessment task as two binary classification problems, one for each dimension of the AV space. The rating scales for variance and arousal are divided into low (1 to 5) and high (5 to 9) levels, and given class labels -1 and 1, respectively. Thus, for each subject we have a set of matrices $\{\boldsymbol{\zeta}_n \in \mathbb{R}^{C \times 4}\}_{n=1}^{N}$ containing the DE features, and two labels vectors $\boldsymbol{\lambda} \in \{-1, 1\}^N$ (one for valence and one for arousal). Afterward, for each subject we train two Euclidean distance-based K-nearest neighbor classifiers, one for each emotional dimension, to estimate the emotion level labels from the DE features. We do so, following a leave-one-out cross-validation scheme.

For each of the above-described experiments, the number of nearest neighbors K of the classifier is selected through nested cross-validation from the set $K = \{1, 3, 5, 7, 9, 11\}$.

3 Results and Discussion

Gender Differences from DE Feature Sets: Table 1 shows the gender classification accuracy [%] per sample (video) discriminated by EEG band and cortical area. The highest accuracies are obtained for the parietal region across all frequency bands. Figure 1(a) presents the confusion matrix for classification from all frequency bands in the parietal region. Overall, the classification system is more apt at identifying male subjects than the female subjects. Figure 1(b) shows the confusion matrix obtained after estimating the subject's gender, not for each sample, but as the mode of its predicted sample labels, that is, the mode of the predicted labels of the $N = 40$ videos for each subject. The trends observed in Fig. 1(a) remain unchanged in Fig. 1(b), implying that the differences in classification performance between genders, probably cannot be attributed to variations in the recorded brain activity in response to specific stimuli (videos), but to more general differences among subjects. These differences are the result of larger variability in women's DE features as compared with men's. Figure 2(a) shows a Principal Component Analysis (PCA) based projection of the DE features (parietal region, all bands) into a 3D space. The projected features for men and women are represented as blue and red dots, respectively. Men's projected features are clustered together, while women's are more spread out, which translates into a higher variability in women's original DE features. As a consequence, the area where both groups overlap counts with a higher density of features belonging to men, which sheds light into the most common type of

Table 1. Average accuracy [%] per sample for gender classification from DE features.

	θ	α	β	γ	All
Frontal	60.6	45.7	41.9	48.1	50.2
Central	49.5	32.2	30.6	49.6	40.6
Parietal	76.9	68.9	69.6	61.3	75.1
Temporal	66.8	53.4	52.5	57.3	60.9
Occipital	56.3	59.6	55.8	52.4	61.2
All	65.0	44.0	55.0	52.3	61.4

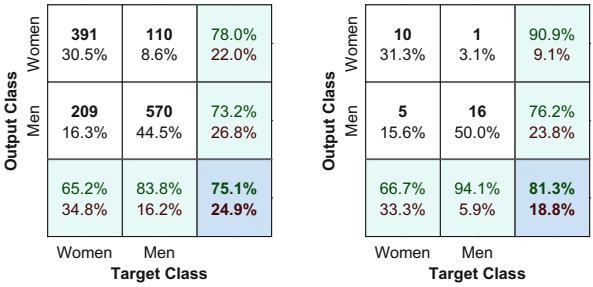

Fig. 1. Confusion matrices for gender classification from parietal DE features: (a) for each sample, (b) for each subject (mode of each subject's predicted sample labels).

classification error obtained (female subjects wrongly classified as males). This result seems to contradict previous evidence showing that women share more similar EEG patterns among them when emotions are evoked than men [12].

Figure 2(b) shows the normalized difference of the average DE between male and female subjects. As expected, the larger differences are found in the parietal and temporoparietal region, especially in the right hemisphere. It is this difference in the DE, and thus in EEG power in those regions, which accounts for the ability of the proposed classifier to discriminate between male and female subjects. These results are in agreement with previous studies that identified gender differences in parietal, temporoparietal and occipital regions in the δ, θ and β bands [4,6]. However, the cited studies found those differences during simple visual stimulation and meditation tasks. Which implies that the gender differences observed here may not be related to the emotional response to the audiovisual stimuli, but be an epiphenomenon of intrinsic gender differences such us thicker cortical gray matter and increased neural process in women, or skull thicknesses variability [6]. To determine if the performance of our DE-based gender classification system depends on an emotional response, we evaluated the average classification performance for each video and contrasted those results against the distribution of the videos in the AV dimensional space generated by the DEAP subjects' ratings shown in Fig. 2(c). The results of the carried out

analysis are presented graphically in Fig. 2(d). We did not find any significant differences in gender classification when the subject was exposed to stimuli of different emotional content, according to the four quadrants of the AV space. Therefore, the results discussed in this section point to the existence of gender differences in the DE features extracted from EEG data. However, we fail to directly link those differences with the subjects' emotional response to the audiovisual stimuli.

Gender Differences from AV Rating Scales: Figure 2(e) shows the distribution of the average AV ratings discriminated by gender. A simple visual inspection reveals gender differences in the self-reported emotional estates elicited by each video, implying that besides the gender differences in the subject's physiological responses, there are differences in the subjective experience and/or expressive response associated with the stimuli. In the following, we attempt to analyze these differences independently for each emotional dimension and exploit them to identify a subject's gender from his/her ratings. Figure 3(a) presents the results of performing gender classification from the self-reported

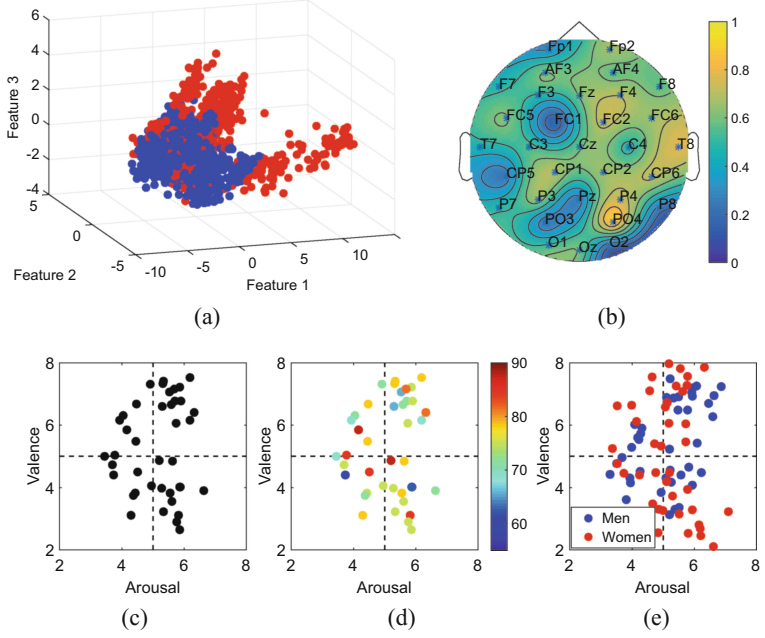

Fig. 2. (a) Gender differences in a space generated by PCA analysis over the DE features (parietal region, all bands; men - blue dots, women - red dots). (b) Normalized difference of the average DE between male and female subjects. (c) Average ratings for each video in the AV space for all subjects. (d) Gender classification accuracy for each video distributed on the AV space according to the subject ratings. (e) Average ratings for each video in the AV space discriminated by gender. (Color figure online)

valence ratings. Unlike the results obtained using the DE features, the classification system is more apt at identifying women than men, with class accuracies of 80.0% and 64.7%, respectively. The possibility of identifying a subject's gender from the valence ratings is accounted for by gender differences in psychological responses to low and high valence stimuli [10], which are at least partly observable in Fig. 2(e). Figure 3(b) shows the Euclidean distance between the valence scale self-reported ratings for all subjects ordered by gender. The distances among women's valence ratings are smaller than among men's, which is reflected in an area of small distances and little variability at the bottom right corner of the plot, explaining the results shown in Fig. 3(a). Figure 3(c) presents the confusion matrix resulting of performing gender classification from the self-reported arousal ratings. Contrary to the valence ratings, the arousal ratings do not allow to carry out gender classification successfully. At this point, it is worth noting the opposing trends described so far, regarding the intra-gender variability exhibited by the valence ratings and the DE features computed from the EEG signals. Men's DE features have less variability than women's, while for the valence ratings the opposite is true. This result poses a challenge to emotion level classification systems based on that information because the features from which the emotional responses are being inferred present large variability in similar emotional responses and vice-versa.

Subject-Dependent Emotion Level Classification: Given the gender differences in the EEG patterns and in the valence ratings found in the previous sections, it follows that the algorithms of emotion level classification should perform differently in the valence dimension for male and female subjects. For our subject-dependent binary emotional dimension level set-up, we obtain average classification accuracies of $60.0 \pm 9.7\%$ and $61.1 \pm 12.5\%$ for valence and arousal, respectively. These accuracies are in the same range as those of recent works that deal with the problem of emotion assessment from EEG and test their methods

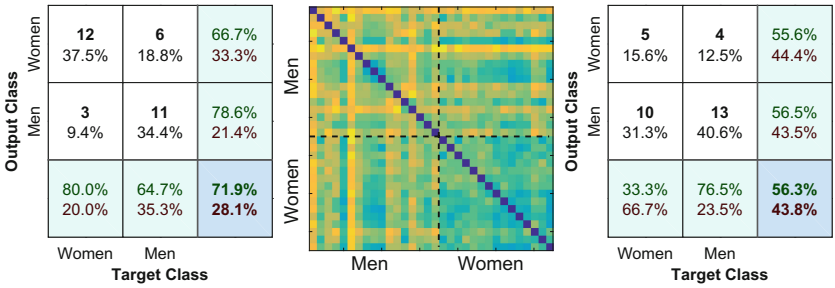

Fig. 3. (a) Confusion matrix for gender classification using the valence ratings. (b) Euclidean distance between the valence scale self-assessment ratings for all subjects ordered by gender. (c) Confusion matrix for gender classification using the arousal ratings.

in the DEAP database. A comparison with such methods is presented in Table 2. However, it has been noted that for emotional level discrimination in the DEAP database the classification accuracy can be misleading [9] because of the class imbalances in the valence and arousal ratings. The area under the Receiver Operating Characteristic (ROC) curve is a fairer way of assessing the performance of the classifier. Figure 4(a) shows the average accuracies and areas under the ROC curve per subject for valence classification. All subjects with high areas under the ROC curve also have high classification accuracies, while the contrary is not true. Finally, Fig. 4(b) shows the boxplots, displaying with the distribution of the areas under the ROC curve for all subjects, discriminated by gender for valence. As seen, there is a difference in classification performance between men and women in the valence emotional dimension, with women displaying an average area ROC of 0.58 ± 0.11 versus men's 0.52 ± 0.11. For arousal, average areas ROC of 0.50 ± 0.06 for women and 0.52 ± 0.10 for men are observed.

Table 2. Emotion level classification accuracy [%] for all subjects in DEAP.

Approach	Arousal	Valence
Koelstra et al. [7]	62.0	57.6
Gupta et al. [5]	65.0	60.0
Padilla-Buritica et al. [9]	52.8	58.6
Arnau-González et al. [2]	67.7	69.6
This work	**61.1**	**60.0**

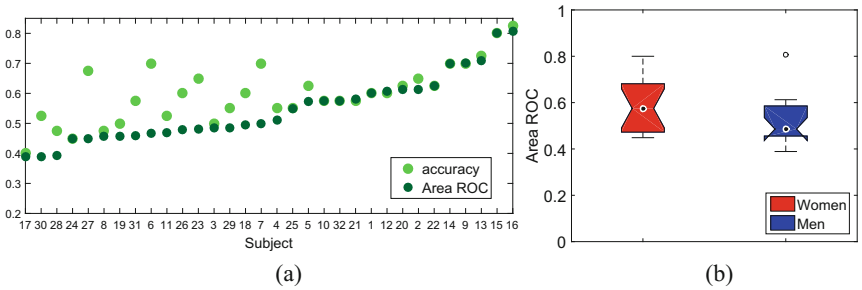

Fig. 4. (a) Average accuracies and areas under the ROC curve per subject for valence classification. The subjects are ordered according to increasing area ROC. (b) Area under the ROC curve for all subjects discriminated by gender for the valence dimension.

4 Conclusions

In this work, we study the differences between male and female subjects during an emotion elicitation experiment using EEG and behavioral measures. We assess the effects of such differences on an emotion level classification system and its components. We have adopted two different perspectives: regarding the subject's physiological responses, as measured by EEG, and their subjective experiences, as measured by rating scales in the AV dimensional space. Our results show differences between men and women in DE features extracted from EEG signals of the parietal region across all frequency bands, higher DE feature variability among female subjects, and higher variability in male subjects' ratings in the valence dimension. Our results also show a gender difference for the valence dimension in the overall performance of our emotion level classifier. Therefore, gender is a relevant factor to take into account during the development and assessment of systems that aim to automatically classify emotional responses, at least those elicited by audiovisual stimuli. Our future work will focus on the development of emotional response recognition strategies based on EEG that integrate demographic information. In particular, subject-independent emotion level classification systems that exploit gender differences to improve classification performance.

Acknowledgments. This work was supported by projects 1110-744-55778 and 6-18-1 funded by Colciencias and Universidad Tecnológica de Pereira, respectively. Author I. De La Pava was supported by the program "Doctorado Nacional en Empresa - Convoctoria 758 de 2016", also funded by Colciencias.

References

1. Alarcao, S.M., et al.: Emotions recognition using EEG signals: a survey. IEEE Trans. Affect. Comput. (2017)
2. Arnau-González, P., et al.: Fusing highly dimensional energy and connectivity features to identify affective states from EEG signals. Neurocomputing **244**, 81–89 (2017)
3. Droit-Volet, S., et al.: Emotion and time perception: effects of film-induced mood. Front. Integr. Neurosci. **5**, 33 (2011)
4. Güntekin, B., et al.: Brain oscillations are highly influenced by gender differences. Int. J. Psychophysiol. **65**(3), 294–299 (2007)
5. Gupta, R., et al.: Relevance vector classifier decision fusion and EEG graph-theoretic features for automatic affective state characterization. Neurocomputing **174**, 875–884 (2016)
6. Hashemi, A., et al.: Characterizing population EEG dynamics throughout adulthood. ENeuro **3**(6), ENEURO-0275 (2016)
7. Koelstra, S., et al.: DEAP: a database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**(1), 18–31 (2012)
8. Lithari, C., et al.: Are females more responsive to emotional stimuli? A neurophysiological study across arousal and valence dimensions. Brain Topogr. **23**(1), 27–40 (2010)

9. Padilla-Buritica, J.I., et al.: Emotion discrimination using spatially compact regions of interest extracted from imaging EEG activity. Front. Comput. Neurosci. **10**, 55 (2016)
10. Rukavina, S., et al.: Affective computing and the impact of gender and age. PloS one **11**(3), e0150584 (2016)
11. Zheng, W.L., et al.: Identifying stable patterns over time for emotion recognition from EEG. IEEE Trans. Affect. Comput. (2017)
12. Zhu, J.-Y., Zheng, W.-L., Lu, B.-L.: Cross-subject and cross-gender emotion classification from EEG. In: Jaffray, D.A. (ed.) World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada. IP, vol. 51, pp. 1188–1191. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19387-8_288