Intelligent Human Systems Integration (IHSI 2024), Vol. 119, 2024, 137–143
https://doi.org/10.54941/ahfe1004477

Multimodal HCI: A Review of Computational Tools and Their Relevance to the Detection of Sexual Presence

Galaup Clément[1]**, Lama Séoud**[1]**, and Patrice Renaud**[2]

[1]Polytechnique Montréal, Montréal QC, Canada
[2]Université du Québec en Outaouais, St-Jérôme QC, Canada

ABSTRACT

Cybersexuality, referring to sexual interactions facilitated by or involving sexual technologies, for better or worse, is poised to play an increasingly significant role in people's lives. The psychophysiological states stemming from such interactions with sexual technologies, and especially virtual reality (VR) scenarios, is termed "sexual presence" (SP). This work aims to review the different methods used to analyse and algorithmically evaluate multimodal electroencephalography (EEG) -centric physiological signals through a multimodal human-computer interface (HCI) and to pinpoint those who prove relevant to the detection of SP. Multimodal HCI are defined as the processing of combined natural modalities with multimedia system or environment. Each modality engages different human capabilities (cognitive, sensory, motion, perceptual). These capabilities, in response to the multimedia environment, can be quantified through psychophysiological signals such as EEG, electrocardiography (ECG), skin conductance, skin temperature, respiration rate, eye gaze, head movements, to name only the most common. While existing surveys have focused on the specific use of EEG to analyse emotions or on the measurement techniques and methods that have been used to record psycho-physiological signals, this work reviews the computational tools, mostly using machine and deep learning, to process, analyse and combine various physiological signals in HCI. Papers published in the last 10 years, combining at least two psycho-physiological signals in an HCI system were collected and reviewed, regardless of the field of application. The focus was mostly on the methodological aspects such as signal synchronization and calibration, fusion approach, model architecture, learning strategy. We put an emphasis on the methods that can be used to detect a subject's condition in real time. At the light of this review, we can identify a research gap in terms of computational tools for multimodal data classification and prediction. This review will allow us to draw on existing work in other fields of application to address our specific application: to analyse EEG, oculometry and sexual plethysmography (penile for the men and vaginal for the women) signals together, using deep learning, to detect SP in subject immersed in an VR environment presenting sexual content.

Keywords: Physiological signal analysis, EEG analysis, Multimodal HCI

INTRODUCTION

The use of numerical technologies for sexual purposes, called Cybersexuality, has been on the rise for the last decades and will play a significant role in people's sexualities.

Sexual presence was defined as "a psychophysiological state of sexual arousal, including a subjective erotic perception, whose content and extent are determined by the interplay between individual psychological predispositions, idiosyncratic past experiences, and what is sexually afforded by a mediating technology" by (Renaud & Fontanesi) and a more complete will be available at (Brideau-Duquette & Renaud, 2023).

The detection and study of sexual presence can be used for psychoforensic purposes. It is possible to detect and assess sexual presence through the analysis of sexual signals like (Côté et al., 2021) did by using EEG and penile plethysmography to classify deviant sexual preferences. In this article, we are interested in reviewing tools that can be applied to perform a real-time assessment of sexual presence through the study of EEG, Eye movement data, genital plethysmography. However, the field of sexual presence detection is quite recent so we will look at articles assessing a psychological state through the analysis of physiological signals: emotion recognition and stress detection.

There are several reviews that have goals similar to the one we have. The review from (Rahman & Sarkar, 2021) focuses on all the methods used to classify emotion from the analysis of EEG signals. The paper from (Craik, 2019) reviews classifications that can be done by using deep learning on an EEG signal. However, both those reviews focused on the analysis of EEG while one of the goals of this article is to review the tools used to fuse multimodal physiological signals and analyse them.

While there exist reviews concerning multimodal HCI containing information on psycho-physiological analysis (Baig & Manolya, 2019) (Jaimes & Sebe, 2007) or more recently (Azofeifa & Noguez, 2022), our review focuses more in depth on the computational tools behind the analysis of signals from multimodal HCI.

METHODS

The field of sexual presence detection is quite recent and niche, and we have thus been obliged to review articles from other fields like emotion recognition and stress detection. The articles reviewed in this work may not be from the targeted field, but they have an objective similar to the one we want to achieve in the future: detecting and assessing a subjective psychological state through the analysis of physiological signals.

The articles reviewed have been chosen and analysed through the following criteria:

- The variety of physiological signals used and their natures.
- The preprocessing algorithms used on each of the physiological signals.
- The features extracted from each signal or from a combination of several signals. And whether those features extractions are applicable for real-time

- How were the signals combined into an object usable by machine learning algorithm.
- The machine learning algorithm used to determine the physiological state studied.

EXTRACTION OF FEATURES FROM PHYSIOLOGICAL SIGNALS

Signals Used for Emotion Recognition and Stress Detection

The fields of emotion recognition and stress detection use a plethora of physiological signals. Most papers on emotion recognition use preexisting dataset of volunteers watching audiovisual stimuli while their physiological data is being recorded such as the DEAP dataset created by (Koelstra et al., 2012) used in articles like (Tong et al., 2018) (Tripathi, Acharya, Sharma, Mittal, & Bhattacharya, 2017) or the seed V dataset used by (Guo, Zhou, Zhao, & Lu, 2019).

The physiological signals are recorded over a given window of time (one minute for the DEAP dataset (Koelstra et al., 2012)). These datasets and the articles reviewed used the following physiological signals:

- EEG recorded from electrodes (usually 32) placed on the head of the subject. The recorded raw signal passes through a low-pass filter and artefact due to reflex eye activity (blinking) are removed using the eye data recorded.
- Electrocardiogram (ECG) and electromyogram (EMG) are recorded via electrodes placed on the body and may be passed through a low pass filter.
- Blood volume pulse (BVP) recorded via photoplethysmography usually placed on a finger.
- Skin conductance or galvanised skin response (GSR) and Skin temperature recorded via a wearable device.
- Respiration activity recorded via a belt placed around the subject's chest.
- Information on pupil dilatation (PD), eye movement or eye images recorded via a camera.

Feature Extraction

EEG

Prior to feature extraction, some articles further divide the recorded signal into smaller time windows (around 6 seconds (Rahman & Sarkar, 2021) (Gümüşlü, Barkana, & Köse, 2020)).

The EEG signal is never used in its raw form, it is common to first divide it into five bands δ (1-4 Hz), θ (4-8 Hz), α (8-14 Hz), β (14-31 Hz), and γ (31-50 Hz) by using either a short-term Fourier transform (Guo, Zhou, Zhao, & Lu, 2019) or a wavelet transform (Gümüşlü, Barkana, & Köse, 2020).

Then from each of the frequency band (or only a few among the five) we can extract either the normalised rapport of energy (Tong, et al., 2018) and (Gümüşlü, Barkana, & Köse, 2020) or the differential entropy (Guo, Zhou, Zhao, & Lu, 2019).

The features are extracted from each electrode of the EEG and then put into a feature vector thus transforming a 32*N signal (assuming 32 electrodes with N the large number of samples acquired during the given window) into a 32*f signals where f is the desired number of features. According to (Rahman & Sarkar, 2021) this method of feature extraction is widely used in the field of emotion recognition.

The EEG signals are non-linear so some articles like (Liu & Sourina, 2014) state that using the Fourier transform to extract spectral features might not be the best solution but rather use fractal dimensions features extracted using Higuchi algorithm and high order crossing based features. However, while those methods seem to allow for the use of fewer electrodes which might be of interest for other applications, their classification results are not on par with more recent articles.

The feature vector can be 1 dimensional where each electrode is treated as independent but some articles like (Gümüşlü, Barkana, & Köse, 2020) give the extracted information a spatial value by creating an RGB image were each electrode is mapped to a pixel on an image where the position of the pixel is related to the position of the electrode on the cranium and its RGB values are the extracted energy of the α (8-14 Hz), β (14-31 Hz), and γ (31-50 Hz) frequency bands. While adding spatial information by formatting the signal into an image is an interesting idea and makes the resulting object appealing for a classification using a convolutional neural network (CNN), the values inside are still human-extracted and reduce a complex signal to just a few values. The lack of analysis of a raw EEG signal seems to be a research gap that as yet to be explored.

Other Physiological Signals

For most of the other physiological signals, the features extracted varies.

Those features are either time-related like the mean, average deviation, slope, skewness, and kurtosis over a defined window of time.

They can also be frequency related for BVP, EMG and ECG where the ratio of Low over High frequencies, heart, and respiration rates, means and standard deviations of interbeat intervals are the most used features. For the Galvanised Skin Response (or skin conductance), most paper extract the number, average intensity, rising time, energy of responses as well as the mean and standard deviation of the signal (Barreto, Zhai, & Adjouadi, 2007) (Can, Chalabianloo, Ekiz, & Ersoy, 2019) (Tripathi, Acharya, Sharma, Mittal, & Bhattacharya, 2017) (Gümüşlü, Barkana, & Köse, 2020) (Tong, et al., 2018).

More recent works focused on learned features from the "raw" data like (Gümüşlü, Barkana, & Köse, 2020) that uses the temporal signal directly in their deep learning algorithm. And (Guo, Zhou, Zhao, & Lu, 2019) extract high level features from signals like the eye image by using CNN and LSTM.

Application to Real Time

All the algorithms focus on collecting data during a certain time window and extracting features from that data. If the window used is short enough (around 6 seconds (Gümüşlü, Barkana, & Köse, 2020), (Rahman & Sarkar, 2021)) and if the feature extraction techniques aren't too time consuming.

The application of the same methods to a sliding time window (for example the last 6 seconds of the record).

SIGNALS FUSION AND CLASSIFICATION ALGORITHMS

Most of the reviewed articles use a features level fusion: they extract different features from each signal individually and then concatenate them into a one-dimensional feature vector. Once such a feature vector is created, it will be labelled with the psychological state associated with the given time window (either an emotion classification, a valence and arousal value or a stress level for the articles related to stress detection).

Once such a dataset is created, it is usually divided into training, validation and testing datasets which are used to train different machine learning algorithms. The algorithms used differs according to the articles and having a wide variety of algorithms was an important factor when selecting the reviewed algorithm.

For stress detection, (Barreto, Zhai, & Adjouadi, 2007) analysed BVP, GSR and skin temperature using Bayes Naives Classification, Decision Trees, and Support Vector Machine (SVM) with respective accuracies of 79%, 89% and 90%. While (Can, Chalabianloo, Ekiz, & Ersoy, 2019) tested Principal Component Analysis (PCA) with Linear Discriminant Analysis (LDA), PCA with radial SVM, k-Nearest Neighbours (kNN), Logistic Regression, Random forest and Multilayer Perceptron (MLP) with respective accuracies of 82.3%, 82.3%, 80,4%, 90.2%, 86,27% and 92.2%. using heart rates, GSR and movement data.

For emotion recognition using valence and arousal scales on the DEAP dataset using two classes, (Tong, et al., 2018) used Logistic Regression and AdaBoost on BVP and EEG signals and obtained respectively 61% and 66% accuracy for arousal and 67% and 69% accuracy for valence while (Tripathi, Acharya, Sharma, Mittal, & Bhattacharya, 2017) managed to attain 82% and 74% accuracy for valence and arousal using a Convolution Neural Network architecture.

Using the SeedV dataset (Guo, Zhou, Zhao, & Lu, 2019) used EEG and features extracted from eye images by deep learning to classify 5 emotions using a SVM with an accuracy of 73.9%.

Using their own dataset recording BVP, GSR, ECG, EMG and Respiration (Haag, Goronzy, Schaich, & Wiliams, 2004) managed to obtain accuracy rates of 96.6% and 89.9% for recognition of emotion arousal and valence.

(Gümüşlü, Barkana, & Köse, 2020) used a gradient boosting machine (GBM) on physiological data (BVP, GSR and skin temperature) or features extracted from EEG to classify between unpleasant, neutral, or pleasant emotions with accuracies of 95% and 75%.

Another possibility from using a feature level fusion followed by a machine learning algorithm is using deep learning algorithms such as a combination of CNN or a Bimodal deep auto encoder (BDAE) to extract high level representations of features. Using a BDAE, (Guo, Zhou, Zhao, & Lu, 2019) managed to up their accuracy to 76% in differentiating between five classes of emotions.

(Gümüşlü, Barkana, & Köse, 2020) used a custom architecture based on two CNN that took in input either the raw data from one or several physiological signals or an image like feature vector constructed from the EEG by extracting frequency-based features. After each of the CNN reduced their input to an appropriate size, their results are concatenated, and some other layers are added to the neural network. While the results obtained by fusing signals of different physiological sources with EEG and using a CNN are not as good as the results, they obtained by using GBM on an extracted feature vector. It can be justified by the small number of participants they worked with which resulted in an insufficient dataset to successfully train deep learning algorithm.

CONCLUSION

The current methods used in the field of emotion recognition and stress detection use similar signals and their acquisition and analysis don't seem make their methods unapplicable to the domain of sexual presence detection. The use of a computationally efficient on a short-timed window is compatible with real time analysis.

In addition to the use of the techniques mentioned in the articles reviewed. We identified several research gaps. We didn't come across any articles that used raw signals for the EEG and the extraction of features from the signal, while backed up by field specific knowledge, seems to simplify the signal to much for its use in deep learning models. This can be explained by size of EEG signal, but it might be interesting to analyse the EEG signals in its entirety and not through a subdivision in bands and their respective energies/entropies.

About signal fusion several techniques are aborded, with feature level fusion still remaining as the go-to method; but also, mid-level with a high-level feature extraction (Guo, Zhou, Zhao, & Lu, 2019)before a classifier and end level with connected layers after classifiers trained on each signal (Gümüşlü, Barkana, & Köse, 2020). However, it's a given that the accuracy results obtained by the articles can't be directly compared as their aims and methods of classification differs. All of the fusion methods will have to be tested on datasets relevant to sexual presence detection.

While most of the 'classic' machine learning algorithms were used in the articles found, we found no use of more recent architectures (Transformers, GAN, …).

In our case where we want to distinguish between a neutral psychological state and a state of sexual presence, such algorithms might prove useful and be the objects of future papers.

REFERENCES

(n.d.).

Azofeifa, J. D., & Noguez, J. (2022). Systematic Review of Multimodal Human-Computer Interaction. *informatics*.

Baig, M. Z., & Manolya, K. (2019). A Survey on Psycho-Physiological Analysis &. *Multimodal technologies and interactions*.

Barreto, A., Zhai, J., & Adjouadi, M. (2007). Non-intrusive Physiological Monitoring for Automated Stress Detection in HCI. *Human computer Interaction.*

Brideau-Duquette, M., & Renaud, P. (2023). Sexual presence a brief introduction. In *Encyclopedia of Sexual Psychology and Behavior.*

Can, Y. S., Chalabianloo, N., Ekiz, D., & Ersoy, C. (2019). Continuous Stress Detection Using Wearable Sensors. *Sensors*.

Côté, S. S.-P., Paquette, G. R., Neveu, S.-M., Chartier, S., Labbé, D. R., & Renaud, P. (2021). Combining electroencephalography with plethysmography for classification of deviant sexual preferences. *International Workshop on Biometrics and Forensics*. Rome.

Craik, A. (2019). Deep learning for electroencephalogram (EEG). *Journal of Neural Engineering*.

Gümüşlü, E., Barkana, D. E., & Köse, H. (2020). Emotion Recognition using EEG and Physiological Data for Robot-Assisted Rehabilitation Systems. *International Conference on Multimodal Interaction.*

Guo, J.-J., Zhou, R., Zhao, L.-M., & Lu, B.-L. (2019). Multimodal Emotion Recognition from Eye Image, Eye Movement and EEG using deep neural network. *Engineering in Medecine and Biology Conference.*

Haag, A., Goronzy, S., Schaich, P., & Wiliams, J. (2004). Emotion Recognition Using Bio-sensors. *Affective Dialogue Systems, Tutorial and Research Workshop, ADS*, (pp. 48–60).

Jaimes, A., & Sebe, N. (2007). Multimodal human–computer interaction: A survey. *Computer Vision and Image understanding*.

Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., ... Patras, I. (2012). DEAP: A Database for Emotion Analysis. *Transactions on Affective Computing*.

Liu, Y., & Sourina, O. (2014). Real-time Subject-dependent EEG-based Emotion Recognition algorithm. In *Transactions on Computational Science*.

Rahman, M., & Sarkar, A. (2021). Recognition of human emotions using EEG signals: A review. *Computers in biology and medecine*.

Renaud, P., & Fontanesi, L. (n.d.). Sexual presence: Toward a model inspired by evolutionary psychology. *New Ideas in psychology*.

Tong, Z., Chen, X., He, Z., Tong, K., Fang, Z., & Wang, X. (2018). Emotion recognition based on photoplethysmogram and electroencephalogram. *International Conference on Computer Software & Applications.*

Tripathi, S., Acharya, S., Sharma, R., Mittal, S., & Bhattacharya, S. (2017). Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on the DEAP Dataset. *Conference on Innovative Applications.*