Studying Attention in Virtual Reality with Electroencephalography and Eye-Tracking

Victor Delvigne

IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

UMONS
University of Mons

Cotutelle IMT Nord Europe - Université de Mons

Thèse

pour obtenir le grade de :

Docteur de l'IMT Nord Europe
dans la spécialité
« Spécialité Informatique »

par

Victor Delvigne

Étude de l'attention en réalité virtuelle à l'aide de signaux électroencéphalographiques et du suivi du regard

Thèse soutenue le 04 Octobre 2022 devant le jury composé de :

Mme	Bernadette Dorizzi	Professeur émérite, Télécom SudParis	(Rapporteur)
M.	Benoit Macq	Professeur, Université Catholique de Louvain	(Rapporteur)
M.	Jonathan Weber	Maître de conférences HDR, Université de Haute-Alsace	(Examinateur)
M.	Olivier Debeir	Professeur, Université Libre de Bruxelles	(Examinateur)
Mme	Laurence Ris	Professeur, Université de Mons	(Examinateur)
M.	Jean-Philiipe Vandeborre	Professeur, IMT Nord Europe	(Examinateur)
M.	Thierry Dutoit	Professeur Université de Mons	(Co-Direct. Thèse)
M.	Hazem Wannous	Professeur IMT Nord Europe	(Direct. Thèse)

Laboratoire CRIStAL (UMR CNRS 9189) - ISIA Lab

CRIStAL
Centre de Recherche en Informatique, Signal et Automatique de Lille

ISIA Lab
INFORMATION, SIGNAL, ARTIFICIAL INTELLIGENCE

Studying Attention in Virtual Reality with Electroencephalography and Eye-Tracking

Victor Delvigne

victor.delvigne@umons.ac.be
victor.delvigne@imt-nord-europe.fr

Tuesday 4th October, 2022

A dissertation submitted to the Faculty of Engineering of the
University of Mons, for the degree of Doctor of Philosophy in Engineering Science and
to IMT Nord Europe for the degree of Doctor of Philosophy in Computer Science

Supervisor: Prof. Thierry Dutoit (Université de Mons)
Supervisor: Prof. Hazem Wannous (Institut Mines - Télécom Nord Europe)

Jury members

Dr. **Jonathan Weber** - Université de Haute Alsace, President
Prof. **Bernadette Dorizzi** - Telecom Sud Paris
Prof. **Benoit Macq** - Université Catholique de Louvain
Prof. **Olivier Debeir** - Université Libre de Bruxelles
Prof. **Thierry Dutoit** - Université de Mons, Supervisor
Prof. **Hazem Wannous** - IMT Nord Europe, Supervisor
Prof. **Laurence Ris** - Université de Mons, Co-supervisor
Prof. **Jean-Philippe Vandeborre** - IMT Nord Europe, Co-supervisor

Abstract

Attention Deficit Hyperactivity Disorder (ADHD) is the most prevalent psychiatric disease in childhood (5% of the total population). Several treatments and diagnoses have been designed, however, some of them present a high price, a low accuracy or unfavourable side effects. Neurofeedback training, a novel method to help in the reduction of attention-related symptoms consists of a real-time representation of the brain activity (in an understandable form) to teach the participants how to self-regulate their symptoms. To develop applications proposing neurofeedbacks, several aspects have been considered in this thesis.

First, participants' attention has to be objectively recorded and assessed based on several physiological signals. PhyDAA, a corpus of Electroencephalogram (EEG) and eye-tracking signals have been recorded in VERA a framework composed of novel Virtual Reality (VR) environments specially designed for this task. These environments have been jointly developed with specialists in the field.

Second, with the advancement of Artificial Intelligence (AI), several novel signal processing techniques have been designed for biomedical signals analysis. in order to estimate the attention state from physiological measurements.

Finally, a novel application reacting to participants' attention state based on the insights made above has been developed. The goal of the latter is to maximize the video game effects on the participant's attention. This application could help participants reduce attention-related disorders' symptoms by better detecting them.

Keywords: Electroencephalogram Signals Analysis, Brain Computer Interfaces (BCI), Attention Estimation, Deep Learning (DL)

Résumé

Le Trouble du déficit de l'Attention avec ou sans Hyperactivité (TDAH) est l'une des maladies psychiatriques les plus prévalentes auprès des enfants (5% de la population totale). Plusieurs traitements et diagnostics ont été conçus, cependant, certains d'entre eux présentent un prix élevé, une faible précision ou des effets secondaires défavorables. Une nouvelle méthode, le neurofeedback, consiste à représenter les signaux biomédicaux sous une forme compréhensible afin d'aider les participants à réduire leurs symptômes par eux-mêmes. Dans le cadre du développement de telles applications, différents axes de recherche ont été considérés au cours de cette thèse.

Premièrement, l'état d'attention des participants doit être mesuré et enregistré de manière objective à partir de signaux physiologiques. Pour ce faire, un corpus d'Électroencéphalogramme (EEG) et de mesure de la direction du regard a été enregistré dans des environnements virtuels spécialement dédiés à cette tache. Ces derniers ayant été développés en collaboration avec des spécialistes du domaine.

Deuxièment, grace aux récentes avancées en matière d'intelligence artificielle (IA), diverses nouvelles approches pour traiter les signaux biomédicaux ont été considérées. Ces avancées permettant au mieux d'estimer l'état d'attention à partir des mesures physiologiques.

Finalement, les découvertes faites au cours de cette thèse ont permis le développement d'une nouvelle application. Celle-ci a pour but de maximiser ses effets sur l'attention des participants le plus longtemps possible.

Mots Clés : Analysise des Signaux Électroencéphalographiques, Interface Cerveau-Machine (ICM), Estimation de l'Attention, Apprentissage Profond

to Lucie

Remerciements

Ces trois années de thèse furent riches en découvertes, émotions et défis. J'ai eu l'opportunité d'en apprendre énormément durant cette période que ce soit d'un point de vue scientifique ou personnelle. Les défis rencontrés au cours de cette thèse n'auraient pu être relevés sans le soutien de mon entourage, je tiens donc à les remercier pour leur aide et conseils. Cette thèse a été effectuée en co-supervision entre la Belgique et la France où j'ai eu l'occasion de rencontrer plusieurs personnes m'ayant permis de murir d'un point de vue professionnel ainsi qu'humain.

Tout d'abord, je tiens à remercier mes deux promoteurs: Thierry Dutoit et Hazem Wannous sans qui je n'aurais pas pu commencer cette thèse. Leurs conseils, confiances et temps m'ont permis de vivre cette expérience au mieux. Bien que leur emploi du temps ait toujours été particulièrement chargé, ils ont toujours pu prendre le temps pour m'aider dans mes travaux. En plus de mes deux promoteurs, cette thèse a également été co-supervisée par Laurence Ris et Jean-Philippe Vandeborre, je tiens donc également à les remercier pour leurs précieux conseils et leur guidance dans leurs domaines spécifiques. Bien qu'ayant un rôle de superviseur secondaire, j'ai pu compter sur leur soutien constant durant ces trois années.

Je souhaite également remercier les membres de mon jury de thèse qui ont consacré de leur temps pour permettre la meilleure évaluation de mes travaux ainsi que pour les précieux conseils et remarques.

Cette thèse a été rendue possible à l'aide des financements octroyés pour ces trois ans par l'Université de Mons et l'IMT Nord Europe. Sans ces derniers la science et plus particulièrement la recherche ne pourrait évoluer. Je remercie également l'Université de Mons ainsi que le réseau I-Site Université Lille Nord-

Europe pour les aides financières qu'ils m'ont octroyées dans le cadre de ma mobilité à l'Université de Central Florida aux États Unis.

Je remercie également mes collègues durant ces trois années du coté belge et français. Un merci spécial pour mes collègues avec qui j'ai partagé un bureau à Mons: Antoine, Bastien, Luca et Nathan; mais aussi à Lille: Antoine, Oriane, Sami, Théo et Vivien. Ces rencontres ont permis en plus des échanges scientifiques de passer divers moments de rire ou de décompression en dehors de la rigueur académique. Je souhaite également à remercier spécialement Nathalie pour son aide sans failles durant les différentes batailles qui ont constitué le cauchemard juridique de la cotutelle.

Enfin, cette thèse n'aurait pas pu voir le jour sans le soutien inconditionnel de mes proches, de ma famille et de mes amis. Merci à eux pour leur aide dans les bons comme dans les mauvais moments. C'est grace à vous tous qu'un nouveau Docteur Delvigne a vu le jour. Merci ! Spécialement, merci à Lucie pour son aide et son soutien dans mes moments de doute ou de crise (...surtout pendant les vacances...).

Finalement merci à vous qui parcourez ces lignes, bonne lecture.

Contents

List of acronyms

FFT	Fast Fourier Transform	20
DFT	Discrete Fourier Transform	113
MLP	Multi-Layer Perceptron	23
AI	Artificial Intelligence	3
SVM	Support-Vector Machine	22
RF	Random Forest	22
DT	Decision Tree	21
LR	Logistic Regression	58
ML	Machine-Learning	3
DL	Deep-Learning	3
HCI	Human Computer Interaction	5
HCI	Human Computer Interfaces	32
BCI	Brain Computer Interfaces	32
EEG	Electroencephalogram	3
EOG	Electrooculogram	16
EMG	Electromyogram	16
ECG	Electrocardiogram	33
GSR	Galvanic Skin Response	33
fNIRS	Functional Near-Infrared Spectroscopy	15
FFN	Feed Forward Network	94
MHA	Multi-Head Attention	92
RNN	Recurrent Neural Network	7
RNNs	Recurrent Neural Networks	24

Introduction

In a world where neurons and neural networks more often refer to their artificial twins than the nerve cell that inspired them, it has become clear that Artificial Intelligence (AI) and its use become the standard. Its advances have made possible the creation or the improvements of areas of research such as Machine-Learning (ML) or Deep Learning (DL) with, for instance, the come back of deep learning architectures for image classification, in 2012 [1,2] several years after their creation in 1998 [3].

Research in AI, as well as the resulting disciplines, are applied everywhere in today's world: images recognition, images generation, natural language processing (NLP), recommendation system, drug discovery, etc. However, one field where the use of AI techniques remains complex to address is the medical field. DL models have already been developed to predict disorder or to assess symptoms, but clear biases have been noticed [4]. Moreover, the considered approaches with several million or billions of parameters remain very often difficult to understand. For these reasons, the application of novel technologies and more especially DL in medicine has to be made carefully, by always keeping specialists in the experimental loop for instance. More generally, it is necessary to keep in mind that mathematical models are here to help but never to replace doctors and practicians. It is in this context that this thesis work has been performed.

DL models applied to biomedical signals compose a large part of the existing works in DL. The considered signals measured on humans or animals for this purpose can be various: movements, muscular activity, chemical composition, etc. One subset of the signals measures the electrical activity of the brain. In this context, one of the most often considered measurements is the Electroencephalogram (EEG). EEG signals measure the interaction between neurons on the scalp with several electrodes placed at different locations of the participant's head. EEGs are often considered in the case of attention

estimation due to several motivations: the measurement is non-invasive, has a low cost and the recorders can be easily moved/has not to be used in specific rooms. However, challenges remain: EEGs are signals difficult to analyse compared to images or sounds that are more easily understandable. Another stumbling block is their trend to be quickly affected by noise and artifacts from recording environments (e.g. power plants) or the subjects themself (e.g. eye blinks records on the useful signal).

The existing approaches to automatically processed EEGs are often based on ML or DL. These last goals are often to make predictions from EEG recordings during specific tasks. The existing methods can be separated into two categories: ML considering prior human knowledge in the signal processing pipeline and DL considering the model only to process the signals. The difference between the two methodologies is based on the use of feature extraction methods to reduce the dimension of the signals and extract the relevant information. Both of these approaches are considered in the literature for various reasons: feature extraction methods are based on analysis that are sometimes not generalizable or not applicable to a specific group of participants and considering models able to extract the relevant information automatically requires a large corpus of signals which is not necessarily the case. Moreover, DL models are more difficult to understand compared to more naïve approaches that make the results less understandable.

The goal of this research project is to study the relationship between the attention state during specific tasks promoting attention in Virtual Reality (VR) and physiological signals, e.g. the electrical activity of the brain or the direction of the sight. This objective is fulfilled by combining the expertise of several actors, i.e. engineers, psychologists and neuroscientists with the novel technologies, i.e. DL and VR.

Strategy

The original goal of this thesis is to design a Human Computer Interaction (HCI) system based on the joint measurement of EEG and eye-tracking in VR. It aims to enhance attention and reduce attention-related symptoms for healthy subjects between 18 and 30 years old.

For this purpose, a protocol aiming to maximize the effect of video games in VR on the attention state should have been designed. The design of the protocol would be based on the findings made during the PhD: an immersive environment in VR; signal processing with the help of novel techniques, e.g. AI and DL based approach.

The strategies to achieve our goals were divided into four main steps:

- Creating a VR benchmark promoting attention in the environment and allowing the recording of physiological signals.
- Collection of several signals with the above-mentioned benchmark to aggregate a dataset of sufficient size for DL based analysis [5]. The signals are recorded on participants between 18 and 30 years old.
- Development of a signal processing box considering the signal registered above to correctly estimate the attention state. The signal processing part studies novel methods for feature extraction/representation and signal classification.
- Application of knowledge learned above in the design of a proof of concept video game reacting with the attention state assessed with physiological recordings.

The strategy was to proceed with the steps above sequentially during the three years of the PhD[1].

[1] The initial goal of this thesis was to consider a specific sub-group of participants corresponding to children between 8 and 12 years old with inattentive and combined Attention Deficit Hyperactivity Disorder (ADHD) symptoms. However, due to covid limitations, it was decided to consider another more general group of participants, making easier the signals registration.

Contributions

Given the challenges and the plan, the original contributions of this thesis are listed below:

- Based on the limitations of the existing works for the joint recording of physiological signals and attention study in VR (non-realistic environments or limitations in the inspired life scene - e.g. virtual classroom), a novel framework VERA (Virtual Environments Recording Attention) to assess attention in VR has been designed. In this software, five different VR environments based on everyday life situations were proposed and in each of them, a task aiming to assess the participant's attention state has been designed. The framework has also been designed to record several physiological signals (including EEG, eye-tracking and head movements measurements) during the proceeding of the attention tasks. Moreover, to promote research in the field, it has been made freely available as open source software
- Given the novel framework promoting the simultaneous recording of physiological signals during attention tasks, we have created a dataset dedicated to attention assessment: PhyDAA (Physiological Dataset Assessing Attention). In the actual context where most of the proposed methods for signal processing are data-driven, this dataset could help for attention estimation. For this purpose, 32 healthy participants aged between 18 and 30 years old took part in a 15 minutes experiment during which they played a short video game in VR. Several signals including EEGs, eye-tracking and head position were recorded for later analysis of the relationship between attention state and physiological recordings.
- We propose a novel approach to process EEG signals with models initially dedicated to NLP: Transformer. The goal of this model is to estimate the attention state from EEG signals. The original Transformer being dedicated to text processing, a novel methodology to extract and represent features from EEG signals has also been developed. This analysis pipeline is based on a threefold representation of EEGs: Time (temporal evolution of the signals during the recording), Frequency (contribution among frequency bands) and Space (spatial contribution of each electrode on the participants' scalp). As well as this novel approach, other

architectures have also been considered: Variational Autoencoder (VAE), Recurrent Neural Network (RNN)-inspired models.

- In addition to the novel pipelines, novel models and training methodologies have been investigated for EEG signals processing. This includes novel approaches for multimodality fusion, i.e. how to combine several physiological signals including EEG recordings, eye-tracking or head movement for better estimation of the attention state. Another contribution in the context of DL is the development of self-supervised learning tasks for EEG signal processing, the goal of which is to design learning methodologies helping the model to process and understand the signal instead of only processing it for estimation purpose.
- We have developed an application proposing feedback based on the participants' attention in real-time estimated from the recording of EEG signals. The development of this application is based on the previously mentioned findings.

Organization of the Dissertation

- Chapter 1 gives the background of the thesis and defines theoretical notions to understand the next thesis chapters.
- Chapter 2 describes the environments designed with VERA for the assessment of attention in VR as well as the attention task.
- Chapter 3 presents the corpus of signal, PhyDAA, acquired in the VR environments.
- Chapter 4 introduces a novel approach for feature extraction and representation from EEG signals based on the Temporal, Frequency and Spatial evolution of the signal.
- Chapter 5 proposes methods to estimate attention from feature arrays based on DL approach. Moreover, DL models for EEGs processing in other tasks than attention estimation are described. Among the proposed architecture for attention estimation, one of the proposed model is based on the transformer that has proven its supremacy to automatically solve tasks, e.g. text or images analysis.

- Chapter 6 proposes three novel methods to process information from EEGs and other modalities including eye-tracking, head movements or respiratory signals in parallel.
- Chapter 7 describes the use of a novel learning approach in the context of EEG processing: self-supervised learning.
- A proof of concept application in the form of a video game giving feedback depending on the attention state of the participant is described in Appendix A.

Chapter 1

Thesis Background and Scope

Contents

This chapter is organized as follows: Section 1.1 gives the context in which this thesis was conducted, Section 1.2 presents the methods often used to assess attention. The last two sections focus on the use of physiological signals for attention estimation and the methodologies developed to process such signals.

1.1 Introduction

The attention state is a mental state during which a participant is considered focused on a specific task. The nature of this task can be expressed in various forms depending on the considered stimuli[1], e.g. visual or auditory. The attention state can be divided into two categories depending on the considered task:

- Sustained attention reflects a state during which the participant is focused on a specific and redundant task. Sustained attention measures the ability of the participant to keep a high attention state on the designed task without any distractors.
- Selective attention reflects the faculty of a participant to remain focused on a task and not be distracted by other stimuli. During the selective attention tasks, different kinds of stimuli may appear. It is asked to the participants to react to positive ones and not to react to or inhibit the negative ones.

1.1.1 Why assessing attention?

Attention assessments can be helpful for various purposes, e.g. attention disorders diagnosis, symptoms detection, marketing, etc. The existing methods to assess attention state can be classified into two categories depending on the assessment duration, i.e. the length of the measurement.

For the first category, assessing the general attention state can help classify people in specific clusters depending on their reaction to stimuli or sequences of stimuli. These clusters can represent subgroups with specific mental disorders or presenting particular symptoms.

[1] external stimulation composing a study to which a participant has to react or not

The second category aims at having an instantaneous assessment of the attention state. In this case, the goal is no longer to estimate the general participant behaviour but to measure the attention state at a specific time, e.g. the attention loss and gain regarding a baseline or the attention state under particular circumstances or situations.

1.1.2 Attention disorder

In the context of attention estimation, one of the most known disorders is the Attention Deficit Hyperactivity Disorder (ADHD). ADHD patients are often considered as people having difficulties focusing, but it remains much more complex. Three subtypes of ADHD have been defined by the DSM-V [6], depending on the symptoms encountered: predominantly inattentive (ADHD-I), hyperactive (ADHD-H) and combined (ADHD-C). Their distribution is not clearly defined; however, ADHD-I and ADHD-C seem more represented, with a slightly higher representation of the inattentive subgroups than ADHD-H [7]. Although the patients' symptoms may vary, all three subtypes present difficulties in planning, focusing and inhibiting perturbators. The degree of difficulty varies depending on the subgroup.

ADHD is mainly represented in childhood, where it is the most prevalent mental disorder within this age group. According to Raman et al. 2018 [8], approximately 5% of the children in the world are diagnosed via the DSM-V [6]. However, these figures may vary depending on the geographic region (e.g. North-America presenting a prevalence higher of 5.3% compared to UK [8]). Several questions can infer from these figures: are there much more children affected by this disorder in the world? Are there regions over or under-diagnosing?

In everyday life, ADHD is characterized by several drawbacks that affect the patients' life [9]: impairments in school, social life (with friends, classmates and parents), activities, etc. Different methods have been developed to reduce these symptoms: 1) Medication by daily intake of methylphenidate, but in addition to the controversial aspects that medicalizing children involves, it is not effective for 20 to 30% of the patients [10]; 2) Behavioural treatments to help the detection and reduction of the behaviours caused by the symptoms, but this method is more expensive and binding; 3) Neurofeedback training, a

method consisting of a real-time representation of the brain activity (in an understandable form) to teach how to self-regulate specific brain functions. This method is already used for the treatment of phobia [11], Autistic Spectrum Disorder [12], or Anxiety [13].

1.1.3 Use case for attention estimation

As mentioned above, the estimation of the attention state can be helpful for various applications. These applications can be summarized in different categories:

- Diagnosis of attention disorders: estimating attention under specific conditions can help in the context of disorders detection by comparing the results of the particular task.
- Symptoms assessment: Similar to diagnosis, tasks assessing specific symptoms can be designed to understand a disease's evolution better.
- Treatment: an approach could be to consider attention estimation to treat specific pathologies better. The goal is to help the participants better understand and detect their symptoms to reduce them by themself.
- Well-being, entertainment and other: this last category includes all use cases non-related to disorders. It can be applied in various fields, for instance, vigilance estimation during driving tasks. Other applications focus on using attention estimation to create a device to help with meditation tasks. Finally, the last area in which attention estimation can be helpful is for marketing by creating applications predicting the user behaviour in this context (e.g. early skipping or attention span in an advertisement).

1.2 Attention Assessment

After answering why and in which context it is helpful to know the attention state, this section focus on the available methods to assess attention. It is possible to classify the methods employed to estimate participant attention

into different categories: the general assessment and diagnosis, the neurophysiological assessments and the physiological recordings to measure attention.

1.2.1 Diagnosis and general assessment

One of the first approaches for assessing attention is based on the diagnosis (i.e. the subject has a specific pathology) and symptoms assessment (i.e. the subject present specific conditions or symptoms while not suffering from any disorders) of attention-related disorder.

This approach considers general questionnaires and diagnoses designed by specialists, i.e. (child) Psychiatrists or (Neuro-)Psychologists. The most important standard to establish a diagnosis is based on the Diagnosis and Statistical Mental Disorder (DSM-V) [6] that aimed at helping practicians to have the best definition of a given mental disorder. This manual also helps in the context of symptoms definition and quantification.

1.2.2 Neurophysiological assessment

The neurophysiological assessments correspond to tests validated on a large corpus of participants and aim at providing a discrete or continuous score representing the attention state of the participant. These tests represent the participants' ability to answer at or inhibit specific stimuli or sequence of stimuli. As previously said, neurophysiological assessments can be performed by assessing various senses, e.g. visual (reacting to an image appearing on a screen) or auditory (reacting as fast to a sound). These tests are and have to be performed by specialists.

On the other hand, more general questionnaires can be used to estimate attention. They correspond to observations made during the everyday life of the participants. It can be based, for instance, on their diet or school grades. In this case, the assessment must not necessarily be performed by practicians but can be proceeded by the participants' relatives or themselves [14].

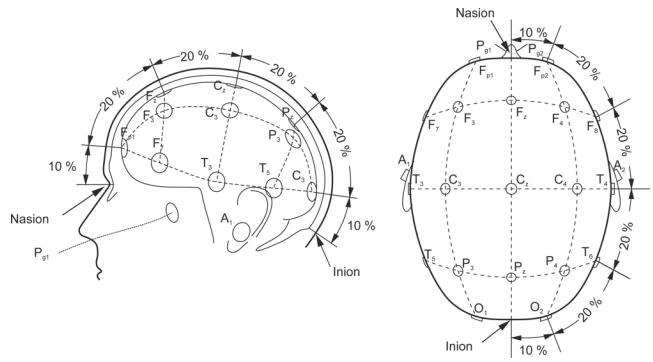


Figure 1.1. 10-20 Electrode placement. Electrode placement for the 10-20 configuration (taken from [15]).

1.3 Physiological Signals and Attention

Due to the large variety of signals, this section will focus on their description to correctly understand this thesis.

1.3.1 Electroencephalogram (EEG)

EEGs are signals representing the brain's electrical activity on the scalp. Information is transmitted in the brain through electrical impulses between neurons. The electrical activity measured can be caused by action potentials along the neurons themselves or by the electro-chemical interactions between neurons.

These signals are acquired with electrodes placed on the scalp following a specific configuration with a predefined location on the scalp. Figure 1.1 from [15] gives the representation of the the electrodes placement for one specific electrodes placement: the 10-20 disposition, the number of electrodes employed for this recording can vary depending on the chosen task. However, scientific works generally consider 8, 16, 32, 64, 128 or 256 electrode montages.

EEG signals are non-invasive and easy to use, their temporal resolution can vary depending on the resolution of the recorder: from 250 Hz to 2000 Hz respectively corresponding to a temporal resolution from 4 ms to 0.5 ms. The

precision with which these devices record the signals will vary depending on their prices, usage or experience of the experimenter (electrodes having to be correctly placed to register the cleanest possible signal). Moreover, the amplitude of the EEG signals are in a very low range of value between 5 µV and 250 µV.

External factors, sometimes uncontrollable, make the EEG signals challenging to register and analyze. Indeed, the amplifier can be affected by various artefacts and noise from several sources, e.g. the electrical power plant and the participant himself (e.g. muscular movements, heartbeats, etc.). For this reason, analyzing EEG signals remains a difficult task.

1.3.2 Functional Near-Infrared Spectroscopy (fNIRS)

The Functional Near-Infrared Spectroscopy (fNIRS) is another method to assess brain activity. In this case, the aim is not to measure the electrical activity but to measure the oxygenation of the blood cells in the brain. The variation of haemoglobin concentration is then measured via infrared, and the augmentation and decay of oxygenation are assessed. The temporal resolution of this approach is higher than EEG recordings, i.e. $\approx$ 100 ms - 200 ms. However, similarly to EEGs this method has a relatively low cost, more it is portable and easier to use compared to EEG (no need for conductive gel).

1.3.3 Eye-tracking signals

In the context of attention estimation, eye movements and other eyes-related signals can be helpful.

Among the available signals, gaze position or real-time recording of sight direction is employed to measure the exact location where sight is directed. This measure can be discrete, e.g. with the position of the observed object or point in the coordinate frame, or continuous with a probabilistic map giving the probability of a given region to be seen or observed. An example of this dual representation is given in Figure 1.2 from [16].

Another type of measurement among eye-tracking signals is eye vergence. It represents the focal properties of the lens composing the eyes (or "*cristallin*").

Finally, another insight measured is the pupils' size during specific tasks. The pupil is a hole placed at the centre of the eye. It can be seen as the diaphragm of a camera. It varies its size to regulate the light intensity.

1.3.4 Electromyogram (EMG)

Electromyogram (EMG) is a measurement of the electrical activity related to the movement of muscle fibres. Signals come from the synchronized muscular contraction of the fibres composing the studied muscle. This measurement can be made during rest state or during contractions. Practicians can perform the exams by placing a needle in the muscle or with surface electrodes.

The amplitude of surface EMG is higher than EEGs with a mean amplitude between 0.1 mV and 5 mV.

1.3.5 Electrooculogram (EOG)

Electrooculogram (EOG) is a specific case of EMG where the recorded muscles are those responsible for eye movement. This measurement is made by comparing the potential of the cornea (front of the eye) and the retina (back of the eye). The goal is thus to measure and study the variation of the induced dipole. By this means, it is possible to analyze the horizontal and vertical variation of this dipole and therefore deduce the movement of the eyes.

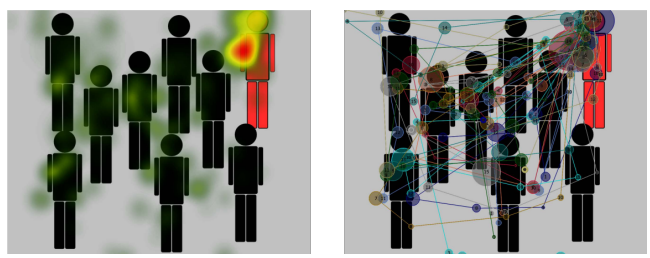


Figure 1.2. Eye tracking measurement. Example of measurement of continuous eye tracking signals (left) and discrete measurement of sight positions (right) (taken from [16].)

1.3.6 Limb and Head Positions

Another possible measurement is the position and orientation of the head and the limbs. This measurement helps to measure the motor activity at different locations of the body: e.g. wrist, hip, ankle, shoulders, head, etc. Various sensors can be used for this purpose, among which the accelerometers are the most often used due to their low size and weight combined with high precision. It allows a low price with a correct temporal resolution ≈ 5-10 ms.

1.4 Signal Processing Background

After listing all the available physiological signals in the context of attention estimation, this section aims at providing background on signal processing for 1-D signals, also mentioned as temporal series.

This section gives the different steps to process this uni-dimensional signal. First, feature extraction, which aims to reduce the dimension of the input signal. Second, the ML-based approach for signal processing that considers more basic and older methods. Finally, the DL-based methodology is a subtype of ML approach considering models with more parameters. Both ML and DL methodologies are data-driven, meaning that the models learn from statistical observations and not from defined equations.

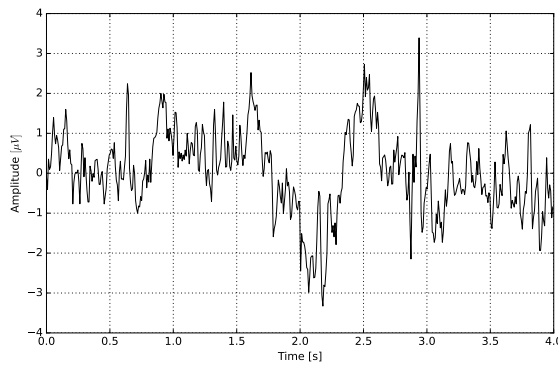


Figure 1.3. 1-D Signal. Example of 1-D signal (e.g. EEG from a single channel).

For clarity, we will consider a signal noted x, which could represent a recording at a given instant of EEG for a given electrode, as represented in Figure 1.3.

1.4.1 Feature Extraction

The first step in the signal processing pipeline is often to consider feature extraction from the signal to reduce its dimension and make it easier to be processed by computer and mathematical models. Another motivation behind the use of feature extraction methods is a low Signal-to-Noise Ratio (SNR) which means that the signal tends to be affected by noise or artefacts. Features vectors can be easier to process compared to noisy signals.

Among the existing approach to extracting features, three categories have been noted: statistical, time-based and frequency-based characteristics.

Statistical features

A naïve approach for extracting information can consider the statistical information of the signal x. These information can be based on the mean μ_x, the median med_x, the standard deviation σ_x, the skewness γ_x or the kurtosis κ_x of the signal x.

If we consider the signal x, composed of n samples, i.e. $x = [x_1, x_2, \ldots, x_n]$, these parameters are computed as:

$$\mu_x = \frac{1}{n}\sum_{i=0}^{n} x_i \tag{1.1}$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x_i - \mu_x)^2} \tag{1.2}$$

$$\gamma_x = \frac{\sum_{i=0}^{n}(x_i - \mu_x)^3}{(n-1)\sigma_x^3} \tag{1.3}$$

$$\kappa = \frac{\sum_{i=0}^{n}(x_i - \mu_x)^4}{(n-1)\sigma_x^4} \tag{1.4}$$

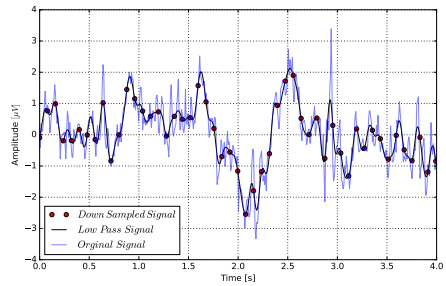


Figure 1.4. Down Sampled 1-D Signal. Example of temporal feature extraction of 1-D signal. In the figure the original signal is represented in black, the signal after low-pass filtering is represented in blue and the down-sampled signal in red dots. This constituting the temporal feature vector.

Temporal features

In addition to the statistical features, it is possible to re-express the signal in a shorter subspace considering temporal-based feature extraction.

A solution could be to consider the same statistical features for the first or second-order derivative, respectively x' and x''.

Another approach is to consider information related to the signal temporal evolution, with, for instance, the number of zero-crossing, i.e. the number of times the signal x has crossed the x-axis. A different method consists of considering a downsampled version of the signal after low-pass filtering. It aims to observe the signal's general trend as represented in Figure 1.4.

Frequency features

Another approach for feature extraction from temporal series is the study of the evolution of their oscillating frequency.

To evaluate the contribution of the signal in different frequency bands, the signal can be filtered, and the parameters mentioned earlier (i.e. statistical and temporal features) can be computed for each filtered signal. On the other hand, it is possible to re-express the signal as a weighted sum of harmonics oscillating at a different frequency with the help of dedicated algorithms, in-

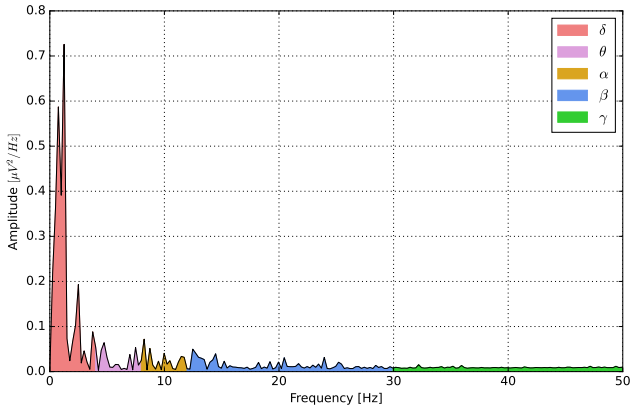


Figure 1.5. Power Spectral Density of 1-D Signal. Example of PSD of a signal between 0 Hz and 50 Hz. The figure represent the PSD with its contribution in each defined brain rhythm.

cluding Fast Fourier Transform (FFT). With the FFT, the signal x can thus be approximated by a periodic signal of period T, with x_T computed as :

$$x \approx x_T = a_0 + \sum_{i=1}^{N} a_i \sin(2\pi f_i t + \phi_i) \tag{1.5}$$

with a_i, f_i and ϕ_i respectively the amplitude, frequency and phase shift of the i^{th} harmonics.

After computing the signal FFT, it is possible to estimate its Power Spectral Density (PSD) representing the normalized contribution on the whole spectrum as represented in Figure 1.5.

Finally, the signal PSD can be directly used as a vector of reduced dimension compared to the signal x. However, in the context of brain signal processing, specific bands have been defined by specialists [17]. Each band is enhanced during specific behaviour or tasks:

- δ rhythm between 0 Hz and 4 Hz related to deep sleep state.
- θ rhythm between 4 Hz and 8 Hz related to the processing of information and memory.

- α rhythm between 8 Hz and 13 Hz related to relaxation with eyes closed.
- β rhythm between 16 Hz and 30 Hz related to open eyes and alert state.
- γ rhythm between 30 Hz and 50 Hz related to consciousness, stress state and motor movements.

In Figure 1.5 the contribution of each rhythm is represented. It is important to note that, although these rhythms have been validated for a large corpus of participants, they are not generalizable for everyone. It has been shown that a specific sub-group of participants presents different frequency bands limits or enhanced rhythm under different conditions [18]. These differences explain that frequency bands limits may vary with the considered literature [19].

1.4.2 Machine Learning

After representing the information in a shorter subspace, the signal processing block processes the input information, i.e. features extracted from the signal x, to analyze or estimate the input modality. This estimation is represented in various forms. For instance, it can be represented by the belonging of a specific class (e.g. symptomatic vs healthy) or a score (e.g. concentration state between 0-100 %).

Various approaches have been dedicated to this task. A non-exhaustive list of the existing models and their descriptions will be presented in this subsection.

Decision Tree (DT)

Decision Tree (DT) can be considered as a simple approach to processing input and making decisions based on answers to binary questions (e.g. is $\mu_x >$ given threshold) made one after another. The end of each tree ramification corresponds to a class corresponding to the answer to all the previous questions. The process is then repeated for all the signals composing the dataset, and the goal is to find the DT allowing correct classification for the whole database.

Figure 1.6 gives a schematic representation of a decision tree for flowers classification based on the iris dataset [20].

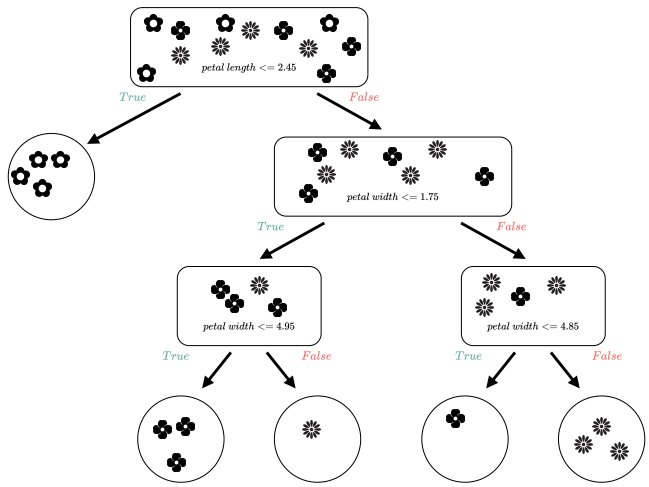


Figure 1.6. Schematic view of a decision tree. Example of DT in the case of flowers classification based on their petals and sepals size.

Random Forest (RF)

An existing method to improve DT is the Random Forest (RF). This method considers several DT constituting a forest. The process of estimating the feature vectors is similar to DT, except that it is repeated several times by considering several, i.e. a forest, instead of a single tree. The final decision consists of the mean decision of all the trees in the case of regression (i.e. estimation of a score) or the most occurrent estimation (i.e. estimation of a class).

Support Vector Machine (SVM)

Support-Vector Machine (SVM) are models that aim discriminate vectors with the help of hyperplans. Given a feature vector f from an input vector x, the goal is to estimate the class/score y with the help of the function:

$$y = w * f + w_0 \tag{1.6}$$

with w being a matrix with learnable parameters and w_0 a bias. During the parameters tuning, the goal is to define a hyperplane maximizing the frontier between classes. In the case of non-linearly separable data, several kernels can be used to transform the feature subspace into a linearly separable one, e.g. polynomial or Gaussian.

1.4.3 Deep Learning

The specificity of the approaches presented above lies in the relatively low amount of parameters they need to be efficient. Since the 2010s, a novel family of methods has known a considerable increase of interest, it constitutes a specific part of ML: Deep Learning (DL) models.

DL approaches aim to consider models composed of several parameters (several millions to billions) to help process information. Although various DL architectures have been created for several purposes during the last years, this section describes four approaches at the basis of the most used today's models.

Multi-Layer Perceptron (MLP)

The Multi-Layer Perceptron (MLP) is the most simple form of Artificial Neural Network (ANN). As its name suggests, MLPs are composed by several layers of perceptron, it corresponds to weighted sum followed by a non-linear function. The perceptron aims at modeling the relationship between several inputs, e.g. $x = [x_1,\ x_2,\ x_3]$, and an output o:

$$o = f(\sum_{i=1}^{3} w_i x_i + b_i) \tag{1.7}$$

with x_i, w_i and b_i representing respectively the input, learnable weight and bias. The difference with Equation 1.6 proposed for the SVM is the non-linearity function $f(.)$ aiming to replicate more complex non-linear functions.

Finally, from the perceptron above defined, it is possible to reproduce various non-linear functions by tuning the learnable weights and repeating the perceptron into several layers. These functions can model relationship from

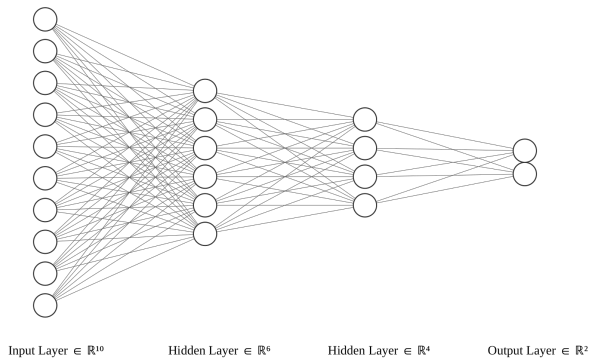


Figure 1.7. Schematic view of a MLP applied on a vector of dimension 10 and composed of three layers with output dimension equal to 6, 4 and 2.

observations, being thus able to model their relationship can help to make estimation from these observations.

Figure 1.7 represents a simple MLP with two hidden layers. The respective input size of each layer is 10, 6, 4 and 2.

Reccurrent Neural Network (RNN)

In addition to the simple MLP approach, architectures specially dedicated to sequential signals have also been designed: Recurrent Neural Networks (RNNs) [21].

Given a sequence of length n, i.e. $x = [x_1,\ x_2,\ x_3, \ldots,\ x_n]$, the aim of RNN is to process these sequences for estimation (e.g. by predicting a score or class) or to forecast the following element in the sequence x_{n+1}.

For this purpose, RNNs based architectures consider the sequence and process the information sequentially with the following equation:

$$h_i = f(Ux_i + Vh_{i-1} + b) \tag{1.8}$$

$$y_i = f(Wh_i) \tag{1.9}$$

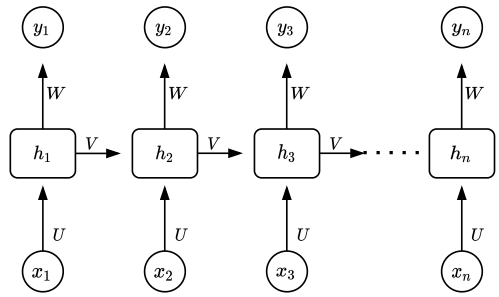


Figure 1.8. Schematic view of a RNN

with U, V and W being matrices with learnable parameters, x_i, h_i and y_i being the input, hidden and output of the i^{th} element of the sequence. Figure 1.8 gives an overview of the RNN with the hidden sequence being computed from the input and allowing for output computing.

In the context of EEG signals processing, the recurrence can be expressed toward the temporal (i.e. specific instant), frequency (i.e. location on the frequency spectrum) or spatial (i.e. electrode location) features.

In addition to the RNN, several variants have also been developed to improve the architecture by reducing its limitations (e.g. gradient vanishing or lost of memory). For instance, the most famous Long Short-Term Memory (LSTM) [22] and Gated Recurrent Unit (GRU) [23] models aiming to extract information more "far away" in the sequence.

Convolutional Neural Network (CNN)

The last family of neural networks introduced in this section is the Convolutional Neural Network (CNN). Initially dedicated to image processing, these networks are based on the succession of consecutive convolution operations.

An example of convolution is given in Figure 1.9. In this case, we consider a random image of dimension 9×9 and a 2-D convolution is applied to it with a kernel of dimension 3×3. In this case, a kernel with each parameter equal to $1/9$ is applied, the last consisting of mean filtering. However, other types of filtering can be applied depending on the value of each element composing

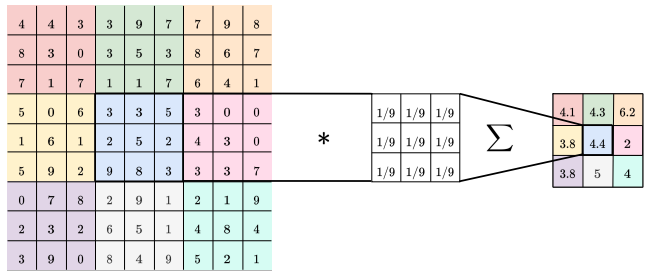


Figure 1.9. Simple convolution of an image of dimension 9×9 pixels with a single channel. The input and result of the convolution operation have the same color. The considered kernel for the convolution has a dimension of 3×3 and a stride of 3 (i.e. distance encountered by the kernel between two neighbors convolution).

the kernels. For this reason, the parameters composing each convolution can be learned for the CNN.

Several convolution layers are applied one after the other. The idea behind using these operations is that they help to extract the information at different levels of the input images (e.g. contours, textures or shape). Finally, the vectors deduced from successive convolutions can be used to process the information in a shorter subspace better.

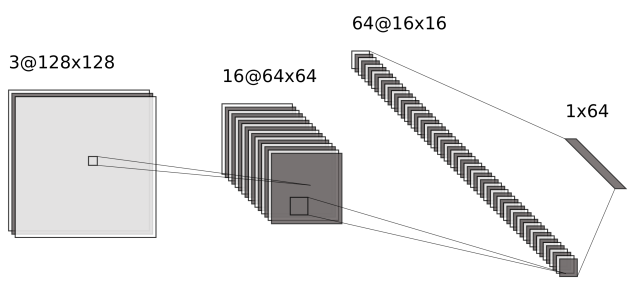


Figure 1.10. Schematic view of a CNN (drawn with [24]). Two convolutions layers are applied on the three channels input image with output channels equal to 16 and 64.

In the context of computer vision, CNNs have proven their supremacy on various tasks, including image classification, segmentation, Visual Question Answering (VQA) and other tasks with models specially developed for this purpose [25, 26].

Autoencoder (AE)

This architecture aims to process the information differently. In this case, the goal is no longer to estimate y from an input x but to encode the input in a shorter subspace. For this reason, the network learns to compress the information into a shorter subspace z and to reconstruct the input x from this representation in a shorter subspace.

During the training, the goal of the model is to find the functions $h(.)$ and $g(.)$ as:

$$z = h(x) \tag{1.10}$$

$$\hat{x} = g(z) \tag{1.11}$$

The two functions can be expressed by a succession of convolutions, recurrent operations or weighted sums, depending on the nature of the input.

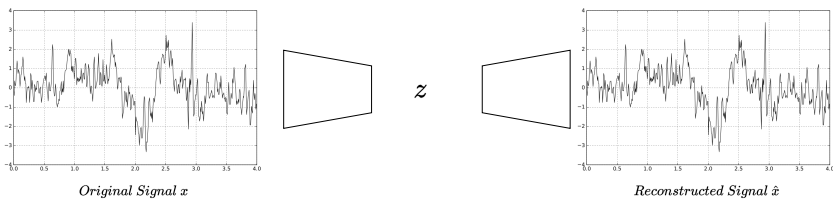


Figure 1.11. Schematic view of a Autoencoder (AE) with the input signal x is passed in the encoder part to have the compressed representation z. The reconstructed version of the original signal $\hat{x}$ is computed from z with the decoder.

1.5 Conclusion

This chapter aims to present the context and the basis for a correct understanding of the following chapters of this thesis.

The notion of attention and the psychological and medical aspects surrounding has been introduced. The physiological signals that can be analyzed for attention estimation have been listed and presented, as also their properties and origins. Finally, the chapter has described the signal processing methodology for 1-D signals. This last section has also introduced the notions of ML and DL models covering the data-driven approach from the simpler methods to the more complex. A more complete description of these models applied to EEG signals will be given in Chapter 5.

Chapter 2

Virtual Reality for Attention Assessment

Contents

This chapter is based on the following publication :

- "*VERA: Virtual Environments Recording Attention*", 8th IEEE International Conference on Serious Games and Applications for Health (SeGAH), 2020.

This chapter presents the existing approaches for assessing attention and the improvements proposed by VR in the field in Section 2.1. Then the VERA framework that aims to assess attention in VR and propose the joint recording of physiological signals will be shown in 2.2.

2.1 Related Work

The presented works aim to better assess specific attention states during predefined tasks to better understand the mechanism behind attention in the human brain. As explained, it exists various methods to evaluate participants' attention state. These methods can be applied during different ranges of time: from the measurement at specific times to the assessment or diagnosis of the symptoms made after a more extended period. Due to the medical concerns and lack of expertise in the medical field, we will not focus on diagnosing the attention-deficit-related disorder in this work. Practicians perform the diagnosis based on the DSM-V [6] constituting the guideline for symptoms and diagnosis of all the existing mental disorders.

2.1.1 Neurophysiological assessment

Neurophysiological assessments are tasks defined by specialists, i.e. psychologists, psychiatrists, and pedo-psychiatrists, that aim to assess a participant's attention state in a specific context. Various tests have been validated on large groups of participants and aim to evaluate multiple symptoms or behaviours.

In the context of attention estimation, three major tests are used to assess the specific behaviours: these tests have a standard procedure consisting of performing a task during which measurements are made, e.g. reaction time to stimuli appearance or rate of detected/omitted stimuli. An assessment based on the value of these measurements is deduced, e.g. measurement above/below a given threshold from predefined formulas. The three tasks are:

- Go/noGo task where the participant is asked to answer as fast as possible to a target, e.g. an *"x"* appearance on a screen, and respectively to inhibit the distractors (not corresponding to the target *"x"* in this case). This test aims to assess the reaction time to a stimulus as well as the

ability of the participant to inhibit specific stimuli as described in Dillon and Pizzagali [27]. An observation of attention state can be deduced by measuring the amount of correctly answered right stimuli, inhibited wrong stimuli and the time taken to answer. This task aims at assessing selective attention and the participant's inhibition ability.

- Conner's Continuous Performance Test (CPT) [28] is a similar task to the Go/noGo without the inhibition mechanism. The CPT consists of a task during which it is asked to answer as fast to a single stimulus appearing during a limited time without appearance of wrong stimuli. Attention is computed from the reaction time and the number of answered and missed stimuli. This task aims at assessing sustained attention during a long and repetitive task.
- The Wisconsin Card Sorting Test is a task during which four cards are presented to the participants, and they are asked to find the proper relationship between them. This assessment has already been used in research aiming to investigate ADHD symptoms as in Mullane and Corkum [29].

Although these tests have presented encouraging results for attention estimation at a specific instant, they are often made on a 2-D screen or on a paper format which can cause a loss of interest and biases in the results. For this reason, a novel trend has been to adopt these tests in VR environments to benefit from their advantages as presented by Bashiri et al. [30].

2.1.2 VR Environment for attention estimation

Attention assessments are often made in a paper format for questionnaires or on a 2-D screen for neurophysiological evaluations. However, for some years now, an increasing amount of neurophysiological tasks have been adapted in VR due to the advantages presented by Pollak et al. [31]: VR provides a higher ecological validity for various assessment (compared to 2-D screen) and in a more appreciated environment. Moreover, it brings larger freedom for the environment creation, greater safety, and more control for the distractors. For these reasons, the number of applications considering VR environments for applications aiming to assess attention has increased during the last decade.

Tan et al. [32] and Eom et al. [33] present in their work a VR environment with which they assess the attention state of participants. The goal of these VR environments is to help diagnose ADHD by considering the results of a specific task. Eom et al. [33] show a high correlation between the results of DSM-IV and VR-CPT defined in their work.

Another use of VR in the context of attention is the development of neurofeedback training aiming to reduce symptoms and increase perturbators' inhibition. As a reminder, neurofeedback training is a technique in which brain activity is displayed in a more understandable form (e.g. video games evolving with specific brain activation). Some recent works have presented neurofeedback in VR environments, as Blume et al. [14] detecting the inattention pattern with combined s and fNIRS recordings in a virtual classroom. Other works have shown different methods to detect ADHD patterns based on gaze direction or head position as described in Tan et al. 2019 [32].

2.1.3 Estimation of attention with physiological signals in VR

In addition to assessing attention with questionnaires or psychological assessments on 2-D screens or in immersive environments, it has been shown that it was also possible to evaluate attention state based on biomedical signals measurement in VR. The advance in the field and the creation of low-cost recording devices have led to an increase in research considering Brain Computer Interfaces (BCI). BCI is a sub-type of Human Computer Interfaces (HCI) that aims to link the brain with a computer. In BCI, the interfaces are controlled directly with brain signals. Moreover, it is possible to classify BCI into passive or active depending if the signals are directly controlling the interfaces in real-time (i.e. active) or if it is recorded and pre-processed outline and then processed through the interfaces delayed (i.e. passive).

The aims of BCI can be various: restore previously lost ability, improve, supplement, enhance or replace human capabilities. In the context of attention estimation, several works have been presented to estimate attention from EEGs and other physiological signals.

In their work, Blume et al. [14] propose an environment representing a VR classroom where attention is assessed based on a frequency-based feature from EEG signals. The computed feature corresponds to the ratio of two specific

frequency bands: θ and β. They demonstrate that using VR instead of the 2-D screen may increase the results of the tasks reflecting attention after numerous training.

In another field, Zheng and Lu [34], and Cao et al. [35] propose a driving car task in VR during which several signals were recorded and an attention score was assigned. In both works, they consider the use of EEG signals to assess the vigilance state. A strong correlation has been proven between these signals and the attention state. More, Zheng and Lu [34] also propose in their experiments the recordings of EOG to better assess the attention state.

Another domain where the use of entertainment recorders has been reported is for engagement score assessment during presentation or learning tasks. Similarly, an attention score is computed from EEG signals. Afterwards, participants receive feedback to tell them to focus again or to reward them in case of a high attention state, as in AttentivU [36]. Other works aim at studying the engagement of a large population and thus did not provide feedback as *Pay Attention!* [37] that assess public engagement during TEDx talks.

In addition to this work considering the use of EEG signals to assess attention state, other works with the same purpose considering other signals have also been reported.

García-Baos et al. [38] propose an attention assessment based on gaze information in their work. Varela Casal et al. [39] shows that biomarkers can be computed from eye-tracking signals for attention assessment. The attention state is estimated from the measurement of the fixation time to a target in the centre of the screen. The framework has been tested and shows an improvement in impulsivity and reaction time compared to a control group not taking part in the training.

Other biological signals, e.g. EMG, Electrocardiogram (ECG) or Galvanic Skin Response (GSR) may also be interesting biomarkers for attention estimation. The motivation behind using this type of biological signal is motivated by the encouraging results that they have presented for other tasks [40].

2.2 Neurophysiological Assessment in VR Environments

This section presents our framework VERA to assess attention in VR environments. This framework aims at assessing attention gain/loss from physiological signals, i.e. EEG, eye-related signals and head movement. This section presents the VR environments, VR tasks and the recorded signals during these tasks.

2.2.1 VR environments

In the context of this work, five VR environments have been designed. Their goal is to create a medium promoting attention, controllable and safe. Since these environments are not the first designed for this purpose [14], the aim was to create environments with the highest emotional comfort. For this reason, no school environment has been created to avoid the possible fears related [9]. The designed VR environments are represented in Figure 2.1.

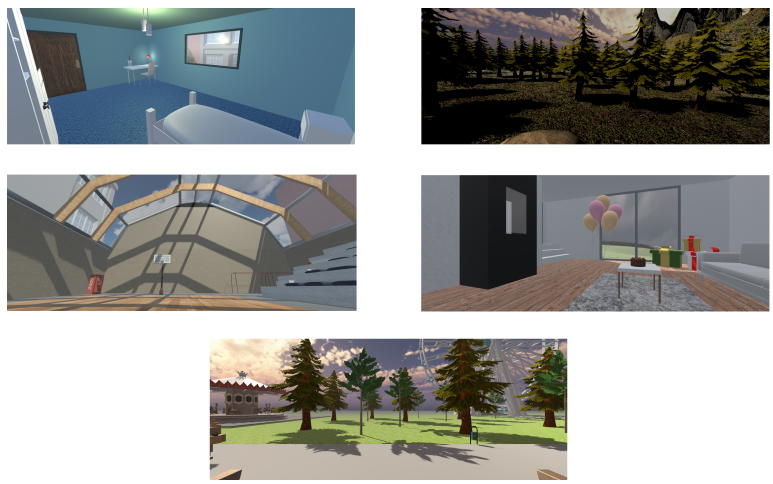


Figure 2.1. VERA - Virtual Environments. This figure gives a representation of the different VR environments that have been developed in the VERA (Virtual Environment Recording Attention) framework. Each of these environments describe lifespan scene of a child.

The framework is composed of five different environments promoting attention. The choice of creating separate environments is to allow the participant to choose the place he feels the most comfortable in. These five environments represent the lifespan scenes of a current child: a sports session in a gym hall; a bedroom; a walk in the forest; a birthday party in a living room, and a trip to an amusement park. For each scene, three different stimuli can appear at predefined times. All the perturbators are related to the environment, e.g. the basketballs in the gym hall or the squirrels in the forest. To measure the effectiveness of the visual, auditive and combined distractors on the physiological measurements, one stimulus of each type was considered as presented in Table 2.1.

The five proposed environments, i.e. bedroom, forest, gym hall, birthday party and amusement park, are displayed in Figure 2.1. In addition to the stimuli appearing during the relaxation task as listed in 2.1, other stimuli appear during the proceeding of the tasks promoting selective and sustained attention as explained in Subsection 2.2.2. Each of these last being related to the chosen environment:

- Basketball and football in the gym hall;
- Butterflies of different colours in the forest;
- Cats and dogs in the bedroom;
- Balloons of different colours in the living room;
- Animals in the amusement park.

2.2.2 VR tasks

Three tasks were designed in collaboration with cognitive psychology and neuropsychology specialists at the Faculty of Psychology of the University of Mons to assess participants' attention. Each of these exercises had a duration of five minutes. The choice of this duration was motivated by two factors: the duration had to suit with participants' profiles (i.e. a part of the population could have issues remaining still and calm during a too long time) and had to be sufficient for data-driven signal processing applications [5].

We considered in our framework three different tasks:

	Auditory	Visual	Audio-Visual
Bedroom	Knocking door	Blinking Light	Cars passing
Gym Hall	Fridge On/Off	Ball falling	Birds flying on roof
Forest	Bird singing	Butterfly	Squirrel running
Birthday	Fire cracking	Cat	Box falling
Amusement	Carousel	Balloon	Dog running

Table 2.1. Stimuli for each VR Environment. List of stimuli with their types for each environment.

- A relaxation task during which we asked the participants to be as calm as possible. The goal was to consider the mean baseline activity while resting. This task aims to begin the attention tasks with a specific attentional state and to have the best participant characterisation for the resting state. At random times during the relaxation task, three different perturbators related to the environment appear (one visual, one auditory and one combined) as listed in Table 2.1.
- A selective attention task. During this task, the participant was asked to look at a specific stimulus (considered positive) and to avoid looking at the other (considered negative). The stimuli corresponding to the environments are listed in the previous subsection. The difference between the wrong and the right stimulus can be based on its nature (different objects), texture or colour. The interstimulus interval (ISI) corresponding to the appearance time between a couple of stimuli follows a normal distribution of 3000 ms and a standard deviation of 250 ms, i.e. ISI $\in N(3000, 250)$. During the second task, a score representing the selective attention state for each trial was computed. This score was calculated with the following equation:

$$t_{task2}(k) = t_{target}(k) - t_{distractor}(k) \tag{2.1}$$

$$score_{task2}(k) = a * t_{task2}(k) + b \tag{2.2}$$

with $t_{target}(k)$ (resp. $t_{distractor}(k)$) being the amount of time during which the target (resp. the distractor stimulus) is looked during the trial k.

a and b are parameters that were experimentally computed to have the mean difference time corresponding to 50% and the lowest (resp. highest) time corresponding to 20% (resp. 80%). These parameters have been computed based on the participants dependent and independent mean attention score for the second task.

- A sustained attention task. It is asked to the participant to direct the sight in the direction of the stimulus. During this task, only the positive stimulus appears. The ISI for the third task also follows a normal distribution with a mean of 3000 ms and a standard deviation of 500 ms, i.e. ISI $\in N(3000, 500)$. From the recordings made during the third task, a score representing the sustained attention for each trial was computed using the following formula:

$$t_{task3}(k) = t_{elapsed}(k) \tag{2.3}$$

$$score_{task3}(k) = a * t_{task3}(k)^2 + b * t_{task3}(k) + c \tag{2.4}$$

with $t_{elapsed}(k)$ the time elapsed by the participants to move their gaze towards the appearing stimulus at trial k. a, b and c are experimentally chosen parameters that match the mean elapsed time with a score equal to 50%, the lowest (resp. highest) time to a score of 80% (resp. 20%). These parameters have been computed based on the participants dependent and independent mean attention score for the third task. During this task, a high attention state corresponds to a short duration, contrary to the previous task where the goal was to have the longest time as possible.

With the methodology presented above, assigning a score representing attention for each stimulus appearance or trial is possible. A threshold is then defined, and each trial with an attention score above 50% is considered as focused and below as distracted.

2.2.3 Recorded signals

In parallel to the task processing, it has been decided to jointly register several signals during the relaxation, selective and sustained attention task to investigate the evolution of physiological signals during attention processing. In the

Figure 2.2. Participant taking part to VERA Experimentaion.

actual context where more and more models are trained on large amounts and types of data, it has been decided to create a framework aiming to register:

- EEG recording the electrical brain activity during task promoting attention to investigate the mechanism behind attention in the brain.
- Gaze information including sight direction, pupils size and blinking instants. This recording is justified by the fact that eye-related information can be a good biomarker for attention estimation [39].
- Head movements are recording the linear and angular position of the head during the proceeding of the attention task. The recording is motivated by the fact that these signals have already proven their ability to help in attention estimation [32].

An example of a participant taking part in the experiment is shown in Figure 2.2. The participant is equipped with the EEG recorder and VR headset.

All the recorded signals have been segmented around the stimuli appearance (one second before and three seconds after) to extract information regarding the instant of interest to characterise the attention state. A more precise description of the signals acquisition will be made in Chapter 3.

2.3 Conclusion

In this chapter, we first introduced some existing approaches to assess attention and point out their drawbacks (i.e. loss of engagement due to the considered format). A list of the previous works considering the assessment of attention in VR has been presented.

We present a novel framework to assess attention state and jointly record physiological signals according to the observations made from previous works. The advantages of the proposed framework can be listed: 1) It allows assessment of both selective and sustained attention; 2) The participant can choose the environment in which the experimentation is made to be the most emotionally comfortable; 3) It allows the recordings of several physiological signals and is freely accessible online, in opposition to other environments that are licensed or paid.

Chapter 3

Datasets of Physiological Signals for Attention Assessment

Contents

This chapter is based on the following publication :

- "*PhyDAA: Physiological Dataset Assessing Attention*", in IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2022.

In this chapter, the methodology employed to create our corpus of physiological signals will be presented. In section 3.1, a review of the existing datasets proposing EEG signals during attention tasks will be presented, as well as other EEG corpuses during different types of assessments: epileptic seizure detection, emotion classification, motor imagery or sleeping stage classification. These datasets have also been employed in this thesis. Finally, after investigating the limitations induced by the existing works, our corpus of EEG PhyDAA will be presented in Section 3.2.

3.1 Related Work

There exist various datasets assessing multiple mental states, symptoms or specific behaviours from EEG signals. This section will first focus on a particular subset of these works: those used for assessing attention state. In the second subsection, we will also focus on other datasets composed of EEG signals for different paradigms and tasks unrelated to attention assessment.

3.1.1 Attention EEG Datasets

Cao et al. [35] propose in their work a dataset composed of EEG signals during a sustained attention task. The experimentation consists of a driving task in a simulated VR environment. The car's steering angle is randomly modified at specific times, and the time the participant takes to react is measured. Twenty-seven participants took part in the 90-minutes long experiments. The recordings have been acquired with a 32 channels biomedical EEG recorder. Each trial corresponds to a random modification of the car steering angle. Only the EEG raw, preprocessed signals and raw steering angle measurement are given with the dataset without pre-extracted features. No other modalities have been registered.

Zheng and Lu [34] propose a similar dataset studying the vigilance evolution during a driving task. Their work aims to predict drowsiness from neurophysiological signals: EEG and EOG. As in Cao et al. [35], the participants are asked to take part in a driving task. During this task, the participants wear eye-tracking glasses that help estimate a score representing their mental fatigue state. PERCLOS score [41] has been considered to assess the attention

state from eye-tracking signals. This score is considered a baseline for vigilance estimation [41] and is computed from the duration of eyes opened, closed or in movement as described in Equation A.2. No external stimuli were considered during the experimentation. Twenty-three participants took part in the approximately two hours long experiment. The signals have been segmented in fixed-duration windows, and a label corresponding to the drowsiness state has been assigned. In addition to the raw signals from EEG, EOG and eye-related features, preprocessed signals and frequency-based features are provided in the dataset.

Except for to these two datasets, no other corpus that assesses the selective or sustained attention state has been found in the literature. For this reason, we consider creating our own coprus of EEG signals with the help of the VERA framework presented in Chapter 2.

3.1.2 Other EEG Datasets

In addition to the datasets assessing attention, other EEG signal corpuses have been considered for the experimentations presented in this thesis. It is possible to divide the EEG datasets into different categories depending on the studied paradigm.

1) Seizure detection is an important application of EEG signal processing. This has resulted in a large amount of datasets freely available online [42, 43]. The motivation behind the consideration of EEGs for seizure detection is justified by the relative simplicity of the task: seizures are temporal events that are easier to visualize compared to oscillatory events. However, it is crucial to remember that even if seizures seem easy to visually identify, epiletic seizures detection remains a difficult task.

2) Estimating emotion from various modalities is a trending goal in today's research world. Emotion can be estimated from texts, images, videos or voices. In this context, several works have envisaged the estimation of emotional states from EEG signals with sometimes other modalities: EMG, EOG or eye-tracking signals.

3) Motor imagery is an application that aims to estimate imagined or real motor movements from EEG signals. The considered motor movement corresponds to movements of upper/lower limbs, tongue or fingers. Moreover,

Paradigm	Dataset	# Participant	Other Modalities
Seizure Detection	TUH [42]	> 50	/
	CHB-MIT [43]	22	/
Emotion Estimation	DEAP [44]	32	EMG, EOG, videos
	SEED [45]	15	/
	SEED-IV [46]	15	Eye Movement
Motor Imagery	BCI IV 2a [47]	9	/
Sleep Stage	Kemp et al. [48]	22	/
Driving Vigilance	Cao et al. [35]	27	/
	SEED-VIG [34]	23	EOG

Table 3.1. List of considered EEG datasets in this thesis. It is presented by paradigm, number of electrodes, number of participants and additional modalities.

various applications have shown the feasibility of these applications and have helped many people suffering from paralysis or amputated limbs.

4) Sleep monitoring is another field where the use of EEGs is often considered. EEG signals can help in the context of sleeping stage classification, e.g. detecting rapid eye movement (REM) and non-REM sleep, to help treating sleep-related disorders, e.g. sleep apnea detection.

Table 3.1 proposes a review of a part of the existing EEG datasets for the above-presented paradigms. As seen, the corpus varies from one paradigm to another. For instance, emotion-related datasets often propose other signals/modalities, e.g. EMG, EOG, videos or sound. Another interesting aspect of the presented datasets is their relative size: seizure and sleep stage-related EEG corpus present a larger size than other paradigms. As their result from systematice recordings in medical environments for diagnosis and not in the specific context of a research project, paper or challenge, as is the case for other paradigms.

3.2 Data collection

Since no corpus was available for tasks promoting selective and sustained attention, we decided to create our new dataset. In this section, we present the data collected with the help of VR environments described in Chapter 2. It first describes the recording condition, then the datasets and finally setup for further classification are presented.

The proposed dataset is composed of physiological signals: EEGs, eye-tracking and head movements. It is available on the Zenodo platform[1] [49].

3.2.1 Recording conditions

EEG signals were recorded using the Brainvision actiCHamp biomedical EEG recorder[2]. The eye-tracking signals (i.e. gaze direction, blinking time and pupil diameter) were recorded with the eye-tracker located in the VR headset and the head position with the accelerometer also located on the HTC Vive Pro Eye VR headset[3]. The recording setup used to record PhyDAA corpus is shown in Figure 2.2.

The brain vision recorder software has been used to record EEGs at 500 Hz. Due to the initial incompatibility of this software with an external recorder, the stimuli appearance time in the VR environments has been automaticly annotated with keyboard input simulation.

We have three types of signals in the dataset: 1) Signals recorded with the sensors located in the HTC Vive Pro Eye VR headset, concatenated in a `.txt` file at a sampling frequency of 5 Hz; 2) Raw EEG signals recorded with the biomedical EEG recorder provided in the original brain vision format `.vhdr`; 3) The pre-processed and segmented signals as well as pre-extracted features (see Chapter 4 for further information on feature extraction) that are provided in NumPy [50] format `.npy`.

[1] https://zenodo.org/record/4558990
[2] https://brainvision.com/products/actichamp-pluss
[3] https://www.vive.com/eu/product/vive-pro-eye/overview

3.2.2 Dataset description

Thirty-two participants (11 females and 21 males) aged between 18 and 30 took part in the approximately 15 minutes experiment. Each participant was in healthy condition and did not present any neurophysiological disorder or trouble related to the use of VR (e.g. cybersickness [51]). Figure 2.2 presents a representation of the experimental procedure.

Due to the place occupied by the VR headset, it was decided to adapt the original position of the electrodes that could have been wrongly impacted by the VR headset as shown in Figure 3.1 (P3 → AF3; Pz → FCz; P4 → AF4). EEG signals have been recorded with an adapted cap of 32 electrodes following the 10/20 placement [52]. Moreover, an automatic bandpass filtering was applied to the signal between 0.5 and 140 Hz during the registration.

The physiological signals recorded by the HTC Vive VR headset are:

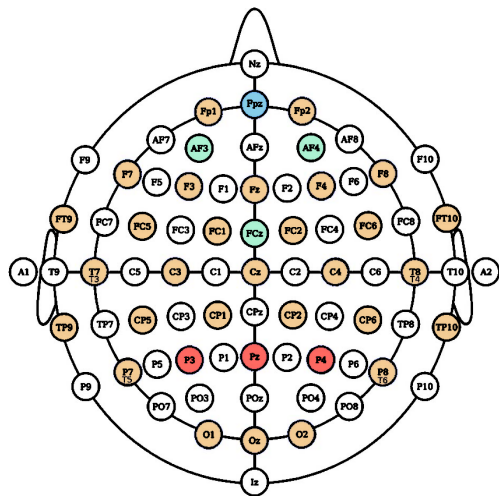


Figure 3.1. 10-20 Electrodes placement adapted to VERA. Electrode placement for the 10-20 configuration adapted to VERA: in blue the ground, in red the electrodes from the original placement, in green the electrodes from the adapted version and in orange the position in common.

- Position and orientation of the eyes in the VR coordinate frame at a given time;
- Position and orientation of the considered stimuli represented as a game object at a given time;
- Position and orientation of the head in the VR coordinate frame;
- Information from the eye state, i.e. eye blinking state (opened vs closed) and eye pupils diameter for both eyes.

The above-described signals have been cut into segments of 4 s (one second before the stimulus appearance and three after). This duration was chosen to investigate the information from the signal variations around the stimuli appearance.

Each signal recorded from the VR headset has been sampled at a frequency of 5 Hz. Moreover, to focus on variation of signals instead of on their instantaneous value, it has been chosen to consider the first-order derivative of the signals.

The attention score has been computed following the methodology from VERA framework presented in subsection 2.2.2. As mentioned in Chapter 2, eye tracking signals have been recorded for both the labels (with the instantaneous eye-direction) and the raw signals (with EEG and other physiological signals). A corresponding label representing the attention state was assigned for each signal segment (composed of EEG and physiological signals). These labels correspond to the binary attention state during the sustained and selective attention task as described in Chapter 2.

The description of the signals composing the dataset is provided in Table 3.2. In addition to the raw signals provided with the dataset, it was decided to provide already extracted feature maps from EEG signals, these last representing respectively the time and frequency-based characteristics of EEG signals. A method based on automatic feature extraction from autoencoder: latent space representation, is also provided in the dataset. Further information about the feature extraction methods is provided in Chapter 4.

Modality	Extract Feature	File Name	Dimension
EEG	Raw EEG Signal	`raw_eeg.npy`	$[3946 \times 2001 \times 31]$
	Power Spectral Density (PSD)	`freq_band.npy`	$[3946 \times 5 \times 31]$
	Hjorth parameters	`hjorth.npy`	$[3946 \times 3 \times 31]$
	Latent-space representation	`AE.npy`	$[3946 \times 16]$
Eye-tracking	Variation pupils diameter [mm]	`phy_sig.npy`	$[3946 \times 1]$
Eye-tracking	Head acceleration (linear and angular)	`phy_sig.npy`	$[3946 \times 6]$

Table 3.2. List of the files in the PhyDAA dataset on Zenodo.

3.2.3 Classification tasks setup

To compare signals corresponding to a high/low attention state, it is necessary to compute a label representing this feature. For this purpose, from the attention label representing the attention state mentioned above, a threshold has been deduced from the mean attention score of each participant separately and the global mean attention score. If the score of a trial was above (resp. below) the threshold, the trial corresponds to a shorter (resp. longer) reaction time and therefore to a high (resp. low) attention state.

A binary label has been computed for each trial constituting the dataset, and the goal of the proposed task was to estimate this label from (neuro)physiological signals. The motivation behind this choice is justified by the fact that it was desired to evaluate participants' attention gain/loss during a task promoting attention and not a continuous score. Estimating the exact score was considered useless in this context.

As mentioned above, the final dataset comprises a set of signal segments with the corresponding attention binary label (focused vs distracted). In addition to this score, a vector provides participant information. This table informs on the considered participant for a given signal segment. This information will help to measure the approach's ability to generalize well or to be trained on a smaller subset of signals.

There are two protocols to evaluate a model's ability to estimate participants' related signals: subject-independent (also known as Leave-One-Subject-Out (LOSO)) and subject-dependent cross-validation. The first protocol consists in training the model with all the participant signals except one used for

the validation. This step is repeated for all the study's subjects, and a mean cross-validation accuracy and its standard deviation are computed. The benefit of this cross-validation method is to measure the model's ability to generalize its knowledge to participants it has not been trained, i.e. it has never met. However, if the amount of participants is large, it leads to a large amount of training repetition time compared to the original K-fold cross-validation, and then the training/validation is more resources and time-consuming. With the second approach, the model is trained and validated with the same participant following a regular K-fold cross-validation. The process is repeated for each participant, and the mean of cross-validation accuracy and its standard deviation is computed. The advantage of this method is that it gives a good insight into the model's ability to estimate with fewer signals. However, an approach based on DL models generally needs a large amount of data. If the study is not containing enough trials per participant, it thus may lead to overfitting issues.

Another interesting approach would be to investigate the effect of the chosen task. As mentioned, in this dataset, we consider two different tasks promoting attention: selective and sustained tasks. They aim to investigate the ability of the model to extract information from a task to transpose it into the other. It is also possible to assess the training and evaluation with task-dependent or task-independent cross-validation.

3.3 Conclusion

In this chapter, the existing datasets used for assessing attention have first been presented. Then a non-exhaustive list of the existing datasets composed of EEG signals during other tasks was given. After analyzing the current works, several limitations have been noted, in addition to the lack of an existing dataset assessing selective and sustained attention in VR.

For this reason, we present a novel dataset PhyDAA to assess attention state with the help of the framework already described in Chapter 2. The advantages of the proposed dataset are various: 1) Proposing EEG signals during relaxation, selective and sustained attention tasks; 2) A joint record of EEG and physiological signals to assess their contribution; 3) A binary label related to the attention loss and gain during the mentioned tasks.

Chapter 4

Feature Extraction and Representation

Contents

EEGs are signals of high dimension that can be represented in $\mathbb{R}^{n_{electrodes} \times time}$ with $n_{electrodes}$ being the number of electrodes and $time$ the signal duration in samples. Being able to design architectures to compress them has been a challenging task for several years now. For this purpose, feature extraction methods are often considered. These are often based on the signals' time, frequency or space properties [17].

Moreover, EEG signals are sets of 1-D temporal series. However, other representations can be considered for the set of features extracted or the raw signal from EEGs. For instance, it is possible to consider embeddings (from AE as explained in Chapter 1) or image-based representation.

In this chapter, Section 4.1 will first focus on the existing work for EEG feature extraction as well as our proposed methodology. In Section 4.2, a similar analysis will be performed for feature representation.

4.1 Feature Extraction Methods

Representing signals with high dimensions in shorter subspace or finding a solution to the curse of dimensionality has been a stumbling block in ML works for many years. In this section, the existing works to solve this problem in the case of EEG processing, as well as our proposed method, will be presented.

4.1.1 Related work

Due to the increasing interest in BCI and ML applied to the medical field, EEG signal processing has often been studied in recent works. In most cases, EEG signal analysis follows a predefined pipeline separated into different steps, as presented in Lotte et al. [17]: 1) Signal acquisition during which the EEGs are recorded in a specific context; 2) Signal preprocessing, where the noise and artefacts are removed from the signals, this step is sometimes merged with the previous one or not considered; 3) Feature extraction, where the most relevant information is extracted from the signal; 4) Feature classification or regression, during which the feature vectors are used to classify the signals or assign them a specific score.

Signal pre-processing

The preprocessing step consists of preparing the signals for further analysis: computing the classes/scores from the observations, removing the noise and artefacts, segmenting the signals, etc. Signal denoising is not a mandatory step, as shown in Lawhern et al. [53]. However, if an artefacts removal policy is considered, it often consists of bandpass filtering to remove the continuous contributions and artefacts that occur at specific frequency bands (e.g. electrical noise oscillating at 50 Hz). Moving average filter can also be applied to remove residual noise components. Another approach used to remove artefacts (e.g. ocular) is the use of Independent Component Analysis (ICA) to investigate the external contributions in the signal corresponding to eye movements and removing them [54, 55].

After preprocessing the signals, several feature extraction methods specially designed for BCI applications have been developed. These methods can be classified into two categories: Frequency and Time-based approaches.

Frequency features

Frequency features can be extracted by studying the contributions of the signal in the frequency spectrum. The most commonly considered approach is the PSD representing the signal power distribution in different frequency bands. In the context of EEG processing, the spectral information is mainly used to characterise the power of the signals in different frequency bands. For instance, in the case of ADHD assessment, the θ - wave [$\sim 5 - 7$ Hz] and β - wave [$\sim 15 - 30$ Hz] powers are often studied [56].

There are different methodologies to extract spectral features, but the most commonly considered is FFT. A large part of the current work in the context of attention assessment considers spectral features [14, 57, 58]. Similar methods have also been considered for other tasks, e.g. vigilance and emotion estimation [34, 59]. However, an important aspect that has to be taken into account is the correct definition of the frequency bands that can cause issues in the case of erroneous definition [18]. As explained previously, the frequency band limit may vary from one participant group to another, which may lead to wrong frequency band definition and incorrect feature extraction.

Temporal Features

One application in which temporal features are often used is in the context of Evoked Related Potential (ERP) analysis. ERPs consist of specific patterns occurring in brain signals recording during particular tasks, e.g. appearance of visual stimuli. EEG temporal variation can be quantified and used as a feature for signals classification. A temporal feature generally consists of a downsampled signal after low pass filtering. This feature extraction method aims to represent the signal's general trend, as shown in Figure 1.4. Although these methods have been first designed for ERP analysis, it is also possible to consider them for other tasks, e.g. sleep stage analysis or seizure detection.

Another approach often considered in EEG processing consists in studying the mean of the signals for a given task for all the participants to investigate general trends and insights among the studied groups [17].

Other feature extraction algorithms can also be applied to EEG signals. These may express spatial information (e.g. electrodes position or activated brain regions) or features representing the signal's disorder with chaos theory-based feature extraction methods (e.g. Higuchi fractal dimension [60]).

4.1.2 Proposed Approach

In this context, it was thought to investigate previous approaches to extract information from 1-D signals and analyse their results for feature extraction from EEGs.

Differential Entropy

Differential Entropy (DE) is a measurement directly correlated with the variance of the signal's amplitude. The method has presented encouraging results for emotion estimation from EEGs [46, 59]. Given a temporal series x, DE is computed as:

$$DE = \frac{1}{2} * \log 10 \left(e * \sigma(x)^2\right) \tag{4.1}$$

with e the Euler's number and $\sigma(x)$ the standard deviation of the considered temporal series.

Fisher Information

Fisher Information (FI) parameter is a measurement evaluating the quantity of information contained in a set of measurements, for instance EEG signals. FI measurement has already been considered for EEG signal processing [61, 62]. The method employed to compute the FI parameters is the one proposed in Py-EEG [63]. Given a temporal serie $x = [x_1, x_2, \ldots, x_N]$, FI is computed from the singular values σ of the delayed vector y:

$$\begin{aligned} y(i) =& [x_i, x_{i+\tau}, x_{i+2\tau}, \ldots, x_{i+(d_E-1)\tau}] \\ Y =& [y(1), y(2), \ldots, y(N-(d_E-1)\tau)] \\ FI =& \sum_{i=1}^{M-1} \frac{(\sigma_{i+1} - \sigma_i)^2}{\sigma_i} \end{aligned} \tag{4.2}$$

with τ the delay, d_E the embedding dimension and $\sigma_1, \sigma_2, \ldots \sigma_{d_E}$ the singular values computed from the singular value decomposition of the matrix Y. The singular value decomposition consists of a specific matrix factorization where a rectangular matrix is decomposed into two orthogonal matrices and a diagonal matrix. The diagonal matrix's coefficients are the original matrix's singular values. The delay and embedding dimension are two parameters that can be tuned with the considered matrix, in the original implementation $\tau = 2$ a,d $d_E = 20$.

Hjorth Parameters

Hjorth parameters are temporal-based features representing the signal's activity, complexity and mobility. These parameters are computed from the evolution of the first and second-order derivatives of the signal. This feature extraction method has been considered in EEG signal processing for emotion

and attention estimation [40, 49] and is computed as:

$$\begin{aligned} Activity &= \sigma(x)^2 \\ Mobility &= \sqrt{\frac{\sigma(x')^2}{\sigma(x)^2}} \\ Complexity &= \frac{Mobility(x')}{Mobility(x)} \end{aligned} \tag{4.3}$$

with x' being the first-order signal's derivative. The signal derivative is computed with numerical differentiation by computing the differences between consecutive samples.

Petrosian Fractal Dimension

Petrosian Fractal Dimension (PFD) is a measurement of the signal disorder, due to the high variability of EEGs, methods from chaos theory as fractal dimensions have been widely used for to process these signals, e.g. for drowsiness detection [64]. It is computed as:

$$PFD = \frac{\log N}{\log N + \log(\frac{N}{N+0.4N_{sign}})} \tag{4.4}$$

with N being the signal's length and N_{sign} the signal's sign changes, i.e. the number of crossings of the x-axis during the signal's duration.

Teager Energy

Teager Energy (TE) operator [65] is an operator extracting the information from signals' shape [66]. TE is computed from the evolution of the signal and its first and second order derivatives. In this thesis, TE operator has been applied on EEG and is computed as presented in the original work:

$$TE = \frac{1}{N}\sum_{i=0}^{N-1}(x_i'^2 - x_i * x_i'')^2 \tag{4.5}$$

With x' being the first order signal's derivative and x'' the second order signal's derivative. Both derivatives are computed with numerical differentiation method as for the hjorth parameters.

Other Feature Extraction Methods

Other feature can be considered but have not been studied in detail in this thesis. However, due to no-convergence, bad feature extraction and poor results, we will not enter into details with the latter: Lyapunov exponents [67], Discrete Wavelet Transform (DWT) [68], Statistical parameters (i.e. mean, standard deviation, kurtosis and skew), Principal Component Analysis (PCA), Empirical Mode Decomposition (EMD) [69] and Higuchi Fractal Dimension (HFD) [70].

4.2 Representation of Information

As mentioned, EEG and the corresponding feature maps can be represented under various forms depending on the information willing to proceed: graphs, images, temporal series, arrays, etc. Other works also propose considering the raw signals directly in the considered DL architecture. However, this approach will be more precisely presented in Chapter 7.

4.2.1 Related Work

This section presents some existing methodologies to represent features extracted from EEG.

Arrays

Arrays representation of EEG is the most straightforward approach for representing feature matrices. It represents the feature maps with the same dimensionality as raw EEGs. The corresponding representation is in three dimensions, each corresponding to the evolution among trials (i.e. measurement instant), electrodes and number of features.

The processing of these arrays can be made with different approaches depending on the considered methods that will be investigated in Chapter 5.

The simpler way to consider these arrays is to flatten them to obtain a 1-D vector easily processed by many mathematical or ML models, e.g. Logistic Regression (LR), DT, RF, SVM, etc. Although these approaches present a correct baseline for EEG features maps processing, they did not consider the specificity of EEG signals, i.e. time, frequency or space evolution. By flattening the vector, the information is mixed regardless of temporal position in the sequence, frequency evolution or electrode placement.

For this reason, methods dedicated explicitly to processing sequences or series have been designed, it is based on RNN. This family of models extracts and analyses the recurrence in these types of signals. The recurrence can be expressed with the temporal or spectral evolution of the signals or the spatial feature of EEG signal (i.e. where is the information located on the scalp). To process the signal regarding these specificities (i.e. time, frequency or space), the feature maps have to be re-arranged on the chosen axis, i.e. temporal, spectral or spatial.

Images

Another existing method to represent EEG feature maps is the image-based representation. This method considers EEG feature matrices as images where each non-empty pixel corresponds to the value of the feature matrix for the given electrodes. The pixel can be organised by keeping the neighbour electrodes as neighbour pixels as shown at the center of Figure 4.1. However, this methodology does not consider the distance that can separate electrodes, i.e. two neighbours electrodes are neighbours pixels regardless of their distance [71].

An improved method aims to take into account the positions of the electrodes by considering a more complex methodology: First, assigning to each electrode in 3-D coordinates the corresponding feature vectors; Second, projecting the information from 3-D representation to a 2-D discrete representation with azimuthal projection, which gives a discrete image; Third, creating a continuous image with bicubic interpolation [72,73]. The resulting image-based represen-

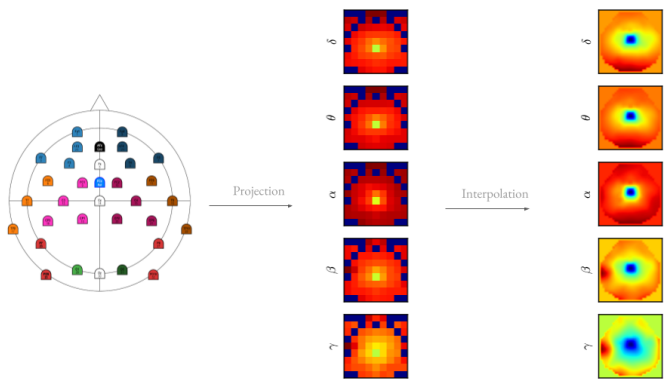


Figure 4.1. Image-based representation of EEGs. Employed methodology to construct images-based representation of feature maps from EEGs for each frequency bands. At the left the representation in arrays, at the center the naïve representation with re-organised pixels and at the right the interpolated images.

tation can be processed with images dedicated DL models, e.g. CNN and is represented at the right of Figure 4.1.

4.2.2 Multi-dimensional representation of EEG

First, let's consider an EEG dataset as a set of segmented signals $X^r = [S_1^r, S_2^r, \ldots, S_C^r] \in \mathbb{R}^{C \times T}$ with C and T representing respectively the amount of electrodes on EEG recorder and the length of the signal.

Among the current work aiming to predict attention state from EEGs, it has been noted that spectral information plays an essential role in the estimation of attention [34,49]. For this purpose, it has been thought to filter the signals into five frequency ranges whose intervals have been defined in previous studies [19]:

- δ-rhythm$\in$ [0.5, 4] Hz
- θ-rhythm$\in$ [4, 8] Hz
- α-rhythm$\in$ [8, 13] Hz
- β-rhythm$\in$ [13, 30] Hz

- γ-rhythm$\in$ [30, 50] Hz

As mentioned in previous works, each of these bands is enhanced during specific tasks or behaviour, e.g. $\delta-$rhythm is enhanced during the deep sleep state [74] or $\alpha-$rhythm for attention during specific task eye closed [18]. Finally, the preprocessed EEG dataset is re-expressed after band filtering as a set of signal $X_f^r = [S_{1,f}^r, S_{2,f}^r, \ldots, S_{C,f}^r] \in \mathbb{R}^{Freq \times C \times T}$ with $S_{i,j}^r$ being the EEG segment of i-th channel and j-th frequency band and $Freq$ being the amount of considered frequency bands. The bandpass filtering applied on each segment consists of a FIR filter with a Hanning window of 1-second length and the cut-off frequencies corresponding to the boundaries of each frequency band.

It was reported that a wrong spectral band definition might lead to behaviour misclassification in the context of attention. Moreover, the spectral bands' boundaries may change from one participant to another [18]. To avoid these issues, it has been decided to adapt it by comparing the maximum of the participant dependent PSD and the mean for participant independent PSD. For this purpose, the mean PSD for each participant separately has been computed and compared to the global mean PSD, the difference between the frequency of each maximum has been computed and considered as the frequency shift occuring between participants.

Another filtering strategy has also been envisaged to avoid the drawbacks of fixed frequency bands. This strategy consists of filtering the EEG signal with regular boundaries between 0 Hz and the cut-off frequency at 50 Hz. This method splits the frequency spectrum into frequency windows of the same size.

After separating EEGs into spectral contributions, EEG segments are separated into time windows. The goal of this segmentation is to capture the information from the signal's temporal evolution during the processing task. It has been proven that specific patterns occur in EEGs during the sight of a stimulus [53]. From this insight, the novel signal representation is $X_f^t = [S_{1,f}^t, S_{2,f}^t, \ldots, S_{C,f}^t] \in \mathbb{R}^{F \times C \times T' \times \frac{T}{T'}}$ with T' being the amount of temporal window to consider in the segmentation. Thus, the length of the segment after the division into windows is $\frac{T}{T'}$.

Finally, from X_f^t, it is possible to extract features helping to express the signal information in a shorter subspace. All the proposed feature extraction

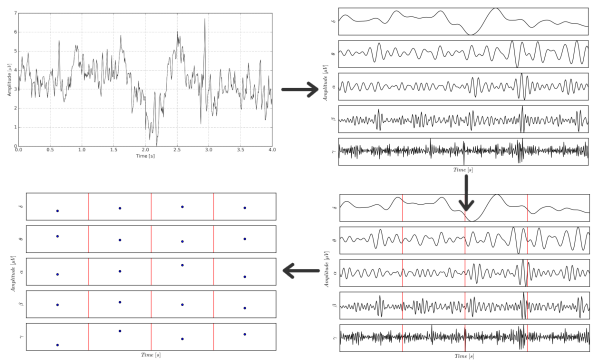


Figure 4.2. Mutli-dimensional representation of EEG. At the top left the original signal, the top right the set of signals after frequency filtering, at the bottom right the temporal slicing into windows and at the bottom left the final representation with the feature extracted for each frequency band and temporal window. For this figure, five frequency bands and four temporal windows have been considered.

approaches have been considered to represent EEG segments in a shorter subspace and will be evaluated in Chapter 5 to investigate the best feature extraction approach.

With the feature matrice $F \in \mathbb{R}^{Freq \times C \times T' \times n_{feat}}$, it is possible to consider its representation as a sequence in three dimensions: frequency, time and space. The spectral direction considers the signal evolution among the considered spectral bands. The temporal dimension expresses the time-based evolution of each feature vector. The spatial direction represents the electrodes-based relationship between feature information and depends on the order in which the electrodes are sorted.

Practically, this representation can be expressed considering each of these dimensions after transposing and merging axes. From the original feature matrice, it is possible to generate the following arrays:

- $EEG|_{frequency} \in \mathbb{R}^{Freq \times n_{feat-freq}}$ a set of feature vectors representing the sequential information within the considered spectral bands.
- $EEG|_{temporal} \in \mathbb{R}^{T' \times n_{feat-temp}}$ a set of feature vectors representing the sequential information within the considered temporal bands, i.e. the segment position in the trial.

- $EEG|_{spatial} \in \mathbb{R}^{C \times n_{feat-spat}}$ a set of feature vectors representing the sequential information within the considered spatial position, i.e. the considered electrode.

The feature dimension for each of these three representations is deduced from the reshape dimension of the feature matrice F. These dimensions can vary with the number of temporal windows, frequency bands or considered electrodes.

One of the main interests of this representation is that it allows having a sequential representation of information considering each stream (i.e. spectral, temporal and spatial) separately. Moreover, this representation allows to avoid losing information, unlike the image-based representation that considers an interpolation of a feature map. More, image-based EEGs of conventional dimension $[32 \times 32]$ is created from low dimension feature vector $\in \mathbb{R}^C$ that could lead to a bias induced by the interpolation.

4.3 Conclusion

This chapter lists the existing methods to extract and represent features from raw EEG signals. After the literature review, it was noted that the current approaches to express and represent these signals in a shorter subspace present some drawbacks: they consider solely the specificity of the signals (i.e. frequency, time or space) and without any merging strategies. Moreover, most existing works consider PSD based features that could present issues, e.g. variation of the frequency bands between participants.

To address all these concerns, it has been thought to consider a novel method to segment the signals into three streams, each corresponding to one of the signal specificity. After considering this novel representation of EEG segments, various feature extractions have been applied to each segment. The approach proposed in this chapter will help to evaluate each stream's effect and feature extraction for the attention estimation purpose in Chapter 5. This sequential representation of information will be considered for architectures specially dedicated to this representation of information, i.e. RNN and transformers. The experiments with this representation of information are given in the Subsection 5.3.1.

Chapter 5

Deep Learning and EEG signals

Contents

This chapter is based on the following publication :

- "*Spatio-Temporal Analysis of Transformer based Architecture for Attention Estimation from EEG*", IEEE International Conference on Pattern Recognition (ICPR), 2022.
- "*A Saliency based Feature Fusion Model for EEG Emotion Estimation*", IEEE International Engineering in Medicine and Biology Conference (EMBC), 2022.
- "*Attention Estimation in Virtual Reality with EEG based Image Regression*", IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), 2020.

For over a decade, research in AI and especially DL have known an increase of interest. The novel DL models have become standard in many fields. For instance, computer vision, natural language processing, speech, etc. However, another area in which their use remains more elusive is biomedical signals and, particularly, for EEG processing.

In this context, this chapter will first discuss the existing works investigating the use of DL with EEG signals in general in Section 5.1. For clarity, the current works will be classified according to their architecture. The proposed novel architecture dedicated to EEG processing in different contexts and tasks (i.e. estimation of attention, assessment of emotion and study of visual saliency maps from EEGs) will be presented in Section 5.2. Finally, the results will be discussed in Section 5.3.

5.1 Related Work

This section presents a literature review of some existing works proposing DL models for EEG signals processing. The context of attention estimation from EEG signals being limited, related works covering similar analyses with other paradigms and tasks will also be presented, although they are not specially dedicated to attention estimation.

5.1.1 Multilayer Perceptron (MLP)

The Multi-Layer Perceptron (MLP) remains the most straightforward method among the existing DL models. It consists of the succession of linear combinations of inputs followed by a non-linearity as explained in Chapter 1. Although this naïve approach is often considered as a baseline for comparison with the more complex DL approach, some works have considered architecture exclusively composed by MLP [75–78].

One of the explanations for the decline in research projects considering MLP is that it does not allow the adaptation to signals' specificities, e.g. MLP are not specially dedicated to images, sounds or graph signals processing. MLP can be considered a mathematical function that processes the information regardless

of its nature. For this reason, other DL approaches specially dedicated to the processed modality have been developed.

However, it should be noted that it is also possible after modifications to adapt MLP to the nature of the considered signals. Recent works have shown encouraging results using exclusively MLP architecture images processing [79].

5.1.2 Convolutional Neural Networks (CNN)

In the context of EEG processing, Convolutional Neural Network (CNN) can be considered to model the relationship between electrodes or extract the signal's temporal evolution.

The first approach consider an image-based representation of EEGs (as explained in Chapter 4) and process them with architecture initially dedicated to image processing [71–73, 80]. CNN can thus process the image representation with(out) interpolation.

The second approach considering CNN aims at extracting temporal information from signals. EEGs being set of time series, it is thus possible to use CNN based architecture to extract the relevant information from these signals with the help of convolution operations on the temporal axis. Several architectures composed by convolution layers have been designed to overcome this challenge like, Shallow ConvNet [81] considering 1-D convolution successively on the time and spaces (i.e. electrodes) axes, EEGNet [53] proposing a pointwise convolution to mix the feature maps from differents electrodes and StagerNet [82] considering convolution only on the temporal axis.

Finally, the specificities of each of these architectures are based on the choice made to construct the network: considered layers, skipped connections, combinations with other architectures, and learning methodologies. This aspect makes the network more or less robust for solving tasks.

5.1.3 Reccurent Neural Networks (RNN)

Recurrent Neural Network (RNN), including Long Short-Term Memory (LSTM) and GRU, are often considered to process the sequential information in EEG recordings. They correspond to a neural network family processing the input

information as a sequence of input vectors. RNN has been initially designed to process the recurrence in modalities. For instance, in text for natural language processing, they have provided encouraging results in text translation or sentiment analysis. In addition to their results with these modalities, RNN also appear to be a good approach for EEG processing with various task such as vigilance estimation [83,84], emotion estimation [40,59] or motor imagery [72].

The explanation for the use of this type of architecture can be various. For instance, Li et al. [59] consider RNN to manage the spatial information in EEG signals. Their work considers two recurrent streams over all the electrodes: one vertical and one horizontal, to model the relationship between EEG electrodes.

Another aspect that can be covered is the evolution of the signals over time. Bashivan et al. [72] proposed in their work, a model processing the recurrence of EEG segments. Each EEG trial is segmented into sub-segments, and each of them is processed sequentially as an input cell of the RNN.

Although these methods present several advantages, they still present drawbacks.

First, they process information in a unidirectional pathway. This means it is necessary to organize the signal's information on one single axis to process it. Therefore, it is not possible to process information in more than one direction at the same time. However, EEG electrodes are challenging to organize in 1-D.

Second, they process the information recurrently, which makes the relationship between non-neighbours (or faraway) electrodes more difficult to process. Third, the gradient in RNN is propagated over many stages that may cause an issue for the gradient descent [85] during the training. This is due to the partial derivative's self-multiplication that causes the gradient's exploding or vanishing.

5.1.4 Autoencoder (AE)

Autoencoder (AE) composes a network family aiming to encode information into a shorter subspace. During the training, the network's goal is to recreate the input modality from a representation in a smaller subspace as closely as possible. This information compression is acquired by updating the weights to

minimize the mean reconstruction error rate, i.e. mean squared error, between the input and output.

The aims of the AE is to project the signal in a subspace with less dimensions than the input. This representation in shorter subspace is more easily classifiable by ML models, e.g. LR, DT, RF, SVM, or handled by DL models.

AE are used nowadays in various fields, for instance in vigilance estimation [86], epileptic seizure detection [87] or representation of EEGs for better interpretability [88].

5.1.5 Graph Neural Network (GNN)

Graph Neural Network (GNN) are network considering EEG feature maps as graph, i.e. feature values of each electrodes are assigned to each vertices and their edges are proportional to the inter-electrodes distances in cartesian or geodesic space. They have proven their supremacy in various fields such as emotion estimation [89], seizure/abnormal signals detection [90] or ERP detection [91] from EEGs.

GNN can be useful to model several aspects of EEG signals. They can thus model the temporal evolution of the signal in each electrode, study the spatial interaction of electrodes regarding their distance, and investigate the functional neural connectivity that studies the brain regions' synchronization during specific tasks.

5.1.6 Generative Adversarial Neural Network (GAN)

Generative Adversarial Neural Network (GAN) is a family of neural networks where two networks (the generator and the discriminator) are trained in an adversarial manner: the discriminator aims at detecting if a given modality has been artificially created by the generator or corresponds to the ground truth, and the generator tries to fool the discriminator by developing modalities very close to reality [92].

GAN have already been used for generating images representing thoughts and/or dreams [93, 94]. Although this research field is still under develop-

ment, the authors have high hopes that one day, it will be possible to visualize our thoughts or dreams.

Although these approaches present encouraging results, it is essential to consider that visualizing thoughts from biomedical signals is a very challenging task [95] even if some research teams express the opposite in media [96].

5.2 Methodology

This section will describe the DL-based models developed for EEG signal processing. As for the previous section, proposed approaches have been organized depending on their architecture. The proposed methods have been applied to solve various tasks from EEGs: attention estimation, emotion estimation and visual saliency estimation.

5.2.1 Hierarchical RNN

An H-RNN has been developed to model the spatial relationship between electrodes. The methodology is inspired by previous works considering different stages of RNN to predict action from skeleton-based signals [97]. The skeleton hierarchical method aims to process the information at a different levels, from the general body information to the interactions toward the limbs, to the position of each joint.

A similar approach has been considered to process EEG feature maps in this context and is represented in Figure 5.1. The spatial analysis focused on the different levels of information: from brain hemispheres to electrodes region to specific electrodes relationship between neighbors.

From the given feature matrix $x^{feat} \in \mathbb{R}^{c \times n}$ with c the number of electrodes and n the feature dimension, it is first possible to split it into two sub-matrices depending on the considered hemispheres, i.e. left or right. Then, both matrices can be separated into i sub-matrices to represent the contribution for each electrode region (i.e. frontal, occipital, temporal, parietal). These matrices are $x_i^{l,feat}$ and $x_i^{r,feat}$ representing the feature matrices for each hemisphere (i.e. left l and right r) and each region i.

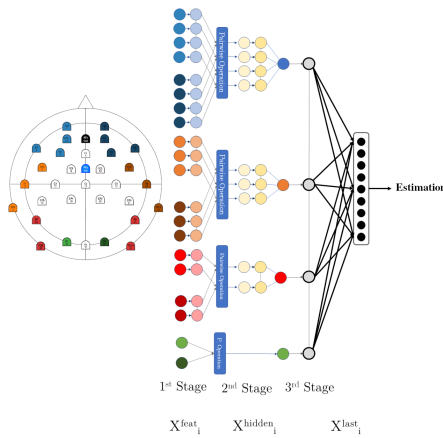


Figure 5.1. Hierarchical-Recurrent Neural Network (H-RNN) architecture overview. Three stages of RNN, the first stage capturing information in the brain region, the second taking into account the relationship between the two hemispheres and the third considering the relationship between the regions. After the RNN stages, a fully-connected network was used to make the estimation.

This feature array is then passed through three-stage of RNN. The hidden state from the different RNN stages can be formulated as explained in Equation 1.9. A schematic representation of the model is given in the top-centre of Figure 5.2. As seen in the figure, three stacks of RNN (corresponding to neighbors, electrodes regions and hemispheres) have been successively applied to the EEG feature maps, and a MLP layer estimates the hidden representation at the third stage of RNN.

5.2.2 Saliency based RNN-CNN

Given the approach presented above, an improvement is to consider the combination of different models and feature representations. For this purpose, a novel approach was envisaged considering the fusion of CNN with H-RNN. The advantage of this method is that it combines two representations of EEG feature maps, i.e. arrays and images-based representation. An illustration of this dual processing approach is summarised in Figure 5.2.

From the results provided by the methods presented above, a combination of these last has been considered to improve the classification accuracy. An approach based on saliency analysis has been considered to promote the synchronized learning of both networks and help transfer learning between modules. This method consists of finding the most important electrodes to estimate a saliency image from H-RNN model. This "saliency" information is then transfered to the image-based network. The feature vector representing Saliency from the H-RNN is computed as:

$$Saliency = \left|\left| \frac{\partial \text{ Estimation}}{\partial \text{ Class}} \right|\right| \tag{5.1}$$

with delta the partial derivative operator representing the variation of the estimation with regard to the considered class. The *Saliency* is represented as a normalized score giving the elements for which the gradient is the highest. These vectors measure the importance of each electrode's feature to make the estimation. From the saliency vector, an image representation is considered and used to weigh the image representation of the feature vector as shown at the top-right of Figure 5.2. This process aims to concentrate the learning

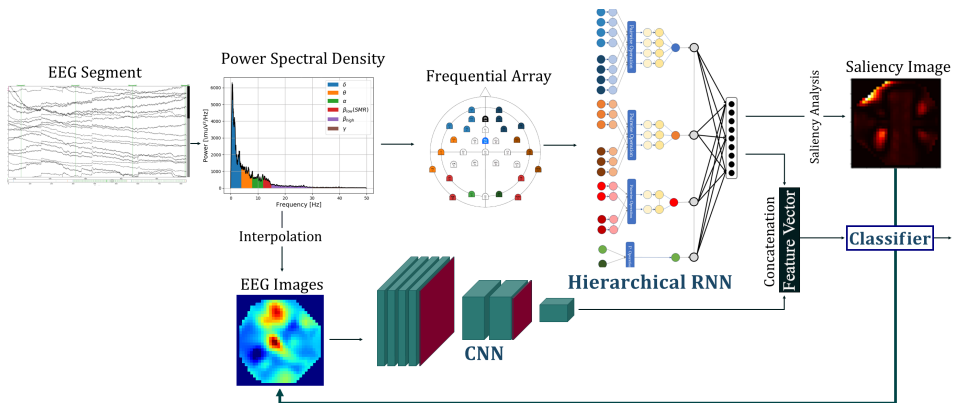


Figure 5.2. H-RNN and CNN architecture overview. Overview of the framework composed of the feature extraction and represenation (left part), the H-RNN (top part), the CNN (bottom part) and the emotion classification (right part). The saliency maps extracted by the H-RNN are used to weight input images of the CNN.

around the most crucial region of the EEGs for both representations. The opposite process consisting of guiding the learning of H-RNN based on CNN has not been considered for two reasons: 1) The process to compute the saliency feature vector from the image presents a high computing cost; 2) Results from H-RNN were better than those guided by the CNN.

5.2.3 Transformer-based Model

Another innovative method to process the information from EEG feature maps is to consider the novel methods based on transformer architecture initially presented by Vaswani et al. in 2017 [98]. This architecture family was initially dedicated to Natural Language Processing by helping to understand and generate texts. Transformers have proven their supremacy in this field [99, 100]. However, as seen in recent years, transformer-based models have also presented great result for other tasks not specially related to text processing, e.g. computer vision [101], speech analysis [102] or VQA [103].

In this context, it was envisaged to use this sequential model to process the EEG feature maps. In addition to their encouraging results in various fields, their use is also motivated by their ability to solve the recurrence issues of RNN. As explained by Merkx et al. 2020 [104], transformers can process sequential information in a non-recurrent manner.

Given the multi-dimension representation of EEGs in the three dimensions (i.e. time, frequency and space), it was envisaged to consider a transformer-based approach to process this specific information representation. An adapted version of the encoder layers from the transformer architecture [98] has been considered. This architecture comprises different parts, each responsible for a specific processing step. If we consider the input feature matrix $EEG|_i$ representing the sequential information in one of the three dimensions i, the estimated class $\hat{y}$ is computed after the following steps:

Embedding

The embedding aims to have a continuous representation of the feature in a vector of lower dimension. In machine translation problems, this embedding consists of assigning to a word a vector that is more easily handled by the

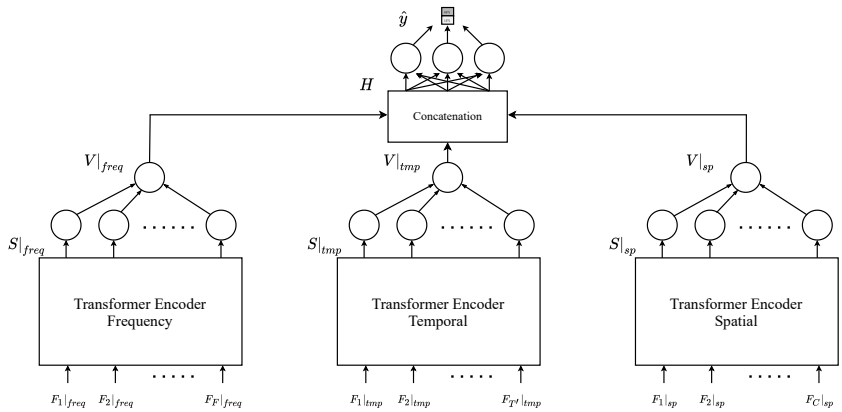


Figure 5.3. Transformer EEG architecture overview. The three representations of the EEG features arrays are passed to the correspond transformer encoder stream. Then the outputs are decoded and concatenated to create an hidden vectors passed to a fully-connected networks to estimate the attention state.

DL model. In this thesis, the embedding step has been replaced by a linear operator expressed as:

$$E = F\, W_{embed} \tag{5.2}$$

with E the embedded feature vectors $\in \mathbb{R}^{SeqLen \times EmbDim}$, F the feature matrix $\in \mathbb{R}^{SeqLen \times FeatDim}$ and W_{embed} a learnable transformation matrice $\in \mathbb{R}^{n_{feat} \times EmbDim}$, $SeqLen$, $EmbDim$ and $FeatDim$ being three parameters representing respectively the sequence length (proportional to the temporal windows and frequency bands length or amount of considered electrodes), the choosen size for the embbed vector and the dimension of the considered feature extraction method.

Positional encoding

Position encoding allows adding information about the element position in the sequence. The motivation to consider this step is justified by the fact that the self-attention based network did not recurrently process information, i.e. no

information on the elements' position is provided in self-attention network. This issue is solved with a trick consisting of adding a sinusoidal function (sine for even position and cosine for odd positions) oscillating at different frequencies proportional to the embedded features vectors positions [98]. With this step, the model can handle the element position in the sequence.

In the context of EEG signal processing, it has been decided to adopt positional encoding for the spatial transformer encoder (as seen in Figure 5.3) to the modality: the considered frequencies for the sinusoidal function being proportional to the electrode distances instead of the position in the array. The original position encoding has been kept for spectral and temporal transformer encoders.

Transformer encoder

The transformer encoder aims to create a novel representation of the embedded sequence. This module is divided into two parts, each considering residual connections [26] and normalization layers. The first block is composed of the multi-head attention layer. With the help of this layer, the transformer can find the relationship between the sequential information, not necessarily neighbours. For instance, in the case of spatial information of EEGs, it considers the relationship between one electrode on the frontal lobes in the right hemisphere and the parietal on the left hemisphere; in the case of machine translation, it considers the relationship between the first and fifth word of a sentence. This architecture applied the self-attention mechanism [98]. This mechanism consists of the multiplication of three learnable matrices, the query Q, key K and value V. The three matrices are computed from linear operations of the set of encoded feature vectors E and are processed as:

$$Attention(Q, K, V) = softmax\left(\frac{Q\ K^T}{\sqrt{d_k}}\right) V \tag{5.3}$$

With d_k, the dimension of each matrix corresponds to the sequence length. The second block composing the transformer encoder is a MLP applied on each sequence element separately and identically. The resulting sequence is written $S \in \mathbb{R}^{SeqLen \times FeedDim}$ with $FeedDim$ corresponding to the output size of the MLP.

From this continuous representation of the features, the original aim of the transformer in NLP is to estimate the next feature vector in the sequence. This process is part of predicting the next element of the sequence. The prediction is made with the decoder layer comparing the combination of the hidden representation and the shifted original sequence. In our case, the goal is to estimate a class, not a sequence. This step has been modified to suit our task consisting of estimation from EEG feature maps. The estimation part consists of first a MLP applied separately on each vector S_i composing the sequence, and then computing another representation V_i followed by a concatenation of the hidden representation of the sequence. The succession of operation is represented at the top of Figure 5.3. The resulting hidden arrays called H and calculated as:

$$V_i = relu(S_i\ W_{decoder} + b_{decoder}) \tag{5.4}$$

$$H = [V_1, V_2, \dots V_{SeqLen}] \tag{5.5}$$

with $W_{decoder}$ and $b_{decoder}$ being responsively the learnable weight and bias of the MLP decoder network. The hidden representation is finally passed through a dense layer to estimate the class/score suiting with the input sequence.

5.2.4 Variational Autoencoder

A novel approach proposed by this thesis to process raw EEG signals consists of considering VAE to compress and extract the relevant features from these signals. These feature vectors can thus be mapped with another modality with which it is studied. This approach is divided into a two steps:

- In a first time, an unsupervised training of two VAE. The first VAE processing EEG signals, the second considering the unsupervised learning with the other modality to map with EEGs. During this step, the training consists to compress information from EEGs and the other modality individually.
- In the second time, merging half of both networks, i.e. the encoding part of EEG VAE with the decoding part of VAE for the other modality.

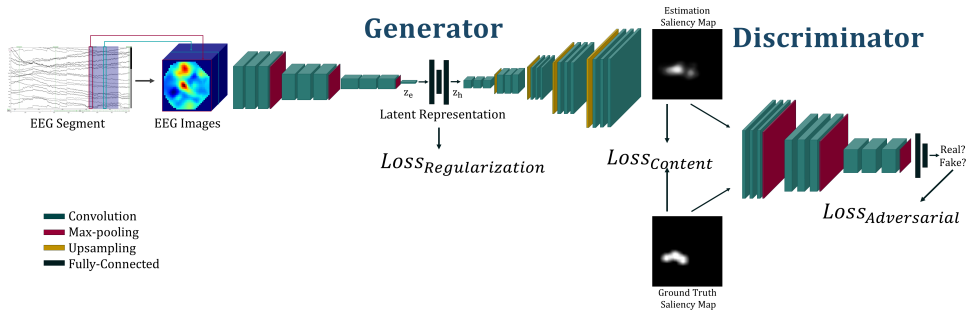


Figure 5.4. VAE architecture overview. At the left the encoding of the EEGs, to its mapping with the embedding of the saliency maps with the decoder. And The discriminator fulling the generator (composed by the encoder and decoder) at the right.

Given the registered signals from our experimentation, a typical use case was to study the relationship between visual saliency maps and EEG signals. Visual saliency maps represent the probability of an area in a visual scene to attract the participant's visual attention. Concretely, a visual saliency map is a single-channel image with each pixel representing the probability between 0 and 1 to be observed [105]. The idea was to create a model to estimate visual saliency maps from EEG signals. The final model is then composed of the encoding part of the EEG VAE and the decoding part of the saliency VAE. Figure 5.4 represents the model. This model can be divided into two sub-models, each aiming to encode/decode specific information.

Autoencoding saliency maps

From raw eye-tracker recordings, it is possible to create a visual saliency map representing the area of attention in an image of one channel with values between 0 and 1 representing the degree of visual attention on specific pixels and their neighbours. It can also be considered as a probability for a given pixel to be watched or not.

During the experimentation, sight direction has been jointly recorded with EEGs with the help of the eye-tracker located in the VR headset as shown in Figure 2.2. The recordings have been separated into trials corresponding to

a specific time. Then, the discrete eye-tracking measurements were projected on 2-D images (one per trial) as explained by Salvucci et al. [106].

The considered network architecture is based on the ResNet proposed by He et al. [26]. We consider for the encoding part four stacks of ResNet layer, each composed of three convolutional layers and batch normalization, each stage except the being separated by a max-pooling operation. Given the visual saliency images, a VAE has been trained to represent them in the corresponding latent space. A similar approach has been considered for the decoding part with upsampling layer instead of a max-pooling operation. The padding has been adapted to ensure that the output size matches the input.

The goal of this network is double: 1) Recreating an image as faithful as possible to the original saliency map via a representation in a shorter latent space; 2) Computing a continuous and complete latent space representation and therefore not favouring one dimension among others.

After some experimental tests, it has been observed that the VAE tends to overfit after a certain amount of epochs. A data augmentation policy has been considered to handle this issue and build a more robust network. The data augmentation process had to be well designed to keep the physical behaviour behind the visual saliency map. For this reason, we have considered for each training sample of each batch from the training set a random horizontal flip with a probability of 0.5 and a random vertical and horizontal translation between -5 and +5 pixels. This method helped to generate a wider range of visual saliency maps with a lower error rate between the initial image and its reconstruction with a better representation of the latent space.

Autoencoding EEG

EEG signals can be represented as two-dimension time series, the first corresponding to time evolution and the second to the considered electrodes. Bashivan [72] proposes to create EEG-images from spectral feature maps based on the electrodes' spatial location. The process to construct our EEG-images was inspired by their work, except that the process was directly applied to the pre-processed signal and repeated on each time sample.

This map representation of EEGs allows keeping the spatial (in the first and second dimension) and temporal relationship between samples. It enables the

consideration of squared shape kernels, unlike the older approaches considering uni-dimensional kernels for feature extraction from EEGs [53]. Moreover, this methodology is better suited for CNN than the array representation of EEGs.

In a similar way to the saliency-VAE, the EEG images have been passed through a VAE to reduce the EEGs dimension and to represent them in a continuous and completed subspace. For this purpose, the VAE has been trained with the images-based EEG.

A similar methodology for the saliency map has been considered to construct the most robust network possible. To that end, a random signal following a gaussian distribution of zero mean and standard deviation = 0.05 has been added to EEG images to increase the model stability and to promote a better understanding of the difference between noise and EEGs. In addition, some pixels composing the EEG images have been supposed to remain equal to zero, these groups of pixels corresponding to the region of the space where no electrodes were placed. A checking was set up to verify that those regions of the images remained equal to zero.

Latent space mapping

From the representation of the saliency map and EEGs in their corresponding shorter subspace, the possibility of mapping the two distributions has been investigated.

As mentioned above, previous works have already investigated the GAN architecture for EEG processing [93, 94]. However, previous works consider array-based representation. A novel architecture adapted to EEG-images representation was then designed.

From both hidden representations, this model aims to estimate a saliency map from EEGs without considering a one-to-one correspondence between modalities. To solve this issue, a GAN has been considered with a generator aiming to recreate the latent representation from EEGs. This generator is combined with a discriminator seeking to distinguish the generated images from the accurate saliency maps. Similar to existing approaches [92], noise following a normal centred distribution (i.e. $\mu = 0$ and $\sigma = 1$) has been concatenated to the latent vector at the centre of the generator. This concatenation aims to guide the generator for the saliency map generation.

The overall architecture of the networks to translate EEG space into image space is represented in Figure 5.4. As seen, the generator consists of the concatenation of the encoding part of EEG VAE and decoding part of Saliency VAE through a generator composed of Fully-Connected (FC) layers.

5.3 Experiments

Let us present the experimental evolution of the different methods developed in this chapter on various datasets. This section shows the results depending on the considered task: attention, emotion and visual saliency estimation. During the experiments, it was noted that some approaches did not converge for some tasks, this being explained by the datasets specificity (e.g. number of electrodes, duration of the segments or tasks). For clarity, only the best results will be presented in this section.

5.3.1 Attention estimation

This subsection discusses the attention state estimation from EEGs and presents a comparison with some existing methods.

For this purpose, two datasets proposing EEG signals with the corresponding attention state will be considered: the PhyDAA datasets recorded in the context of this thesis and the corpus presented by Cao et al. [35]. A more detailed description of both datasets is given in Chapter 3.

To evaluate the proposed methodology, the architecture has been trained and validated with two training methodologies aiming to assess the model's ability to generalize: 1) Subject-Independent (or LOSO) classification, where the model is trained with all the participant signals except one that is used for the validation. The step is repeated for all subjects, and a mean cross-validation accuracy and its standard deviation are computed. The benefit of this method is to measure the model's ability to generalize its knowledge to never met participants; 2) Subject-dependent classification where the model is trained and validated with the same participant following a regular 5-fold cross-validation, the process is repeated for each participant and the mean, and standard deviation of cross-validation accuracy is computed. The advantage of this method

is that it gives a good insight into the model's ability to make estimations with fewer signals.

It was also thought to compare the different methodologies to estimate attention from feature matrices constructed from EEG signals. Among all the existing ML models, four have been kept:

- Traditional ML models: RF and SVM based classifier to define a baseline result for attention estimation.
- H-RNN classifier considering the multi-level representation of information for spatial features from EEGs combined with regular RNN for the time and frequency stream.
- CNN classifier considering ResNet [26] inspired architecture to process images-based representation of EEG feature maps.
- Transformer classifier with the threefold representation of feature maps combined at the output as represented in Figure 5.3.

Approach	Driving EEG [35] ACC/STD [%]	PhyDAA [49] ACC/STD [%]
SVM	68.09/9.55	64.61/9.22
RF	67.81/10.17	61.55/9.79
H-RNN	72.12/8.27	70.86/9.82
ResNet	62.07/6.20	66.82/5.21
Transformer	**74.41/9.27**	**77.24/6.11**
TCA + LR [107]	72.70/9.42	-
MIDA [108]	73.01/9.17	-
Graph Network [49]	-	72.41/5.51

Table 5.1. Classification performance for attention estimation with participant independent protocol. Results above the bold line are obtained from our model experiments.

As seen in Tables 5.1 and 5.2, results obtained by the transformer-based approach present the highest accuracy compared to other baseline approaches for both datasets, which demonstrate the proposed framework's ability to estimate attention from EEG signals. Moreover, it demonstrates the efficiency

of the self-attention based models architecture in processing sequential signals as EEGs for attention assessment.

Furthermore, results from previous works have also been compared to evaluate the developed architectures. For the first dataset, the previous works consider transfer learning approach to increase cross-subject accuracy. These last are based on Transfer Component Analysis (TCA) or Maximum Independence Domain Adaptation (MIDA) with traditional ML architecture. The consideration of naïve ML approaches may cause an accuracy decay compared to the more complex methods. As seen in Tables 5.1 and 5.2 the results from our experiments from traditional ML approaches, i.e. SVM and RF, are lower than other DL methods. It makes us think that considering a more complex training methodology, including transfer learning, may increase the transformer accuracy, although its accuracy already outperforms the state-of-the-art methods.

The best results from the related works for the second dataset are based on GNN. Its architecture includes a graph convolution and a pooling operation conserving the most discriminant nodes (i.e. electrodes) by removing those that can be represented by neighbors electrodes. Unlike the transformer, GNN only considers the spatial stream to estimate attention from EEG. This single representation may explain the lower results.

As mentioned in Section 4, different feature extraction methods and segmentation parameters have been considered. We have investigated the corresponding cross-validation accuracy for each combination. In Figure 5.5, the feature ex-

Approach	Driving EEG [35] ACC/STD [%]	PhyDAA [49] ACC/STD [%]
SVM	76.07/9.65	70.82/13.25
RF	75.60/8.76	75.63/12.89
H-RNN	80.03/8.09	79.64/10.55
ResNet	75.96/8.98	70.39/6.91
Transformer	**83.31/6.71**	**85.04/7.56**
MLP [78]	81.32/6.02	-
Graph Network [49]	-	77.34/10.24

Table 5.2. Classification performance for attention estimation with participant dependent protocol. Results above the bold line are obtained from our model experiments.

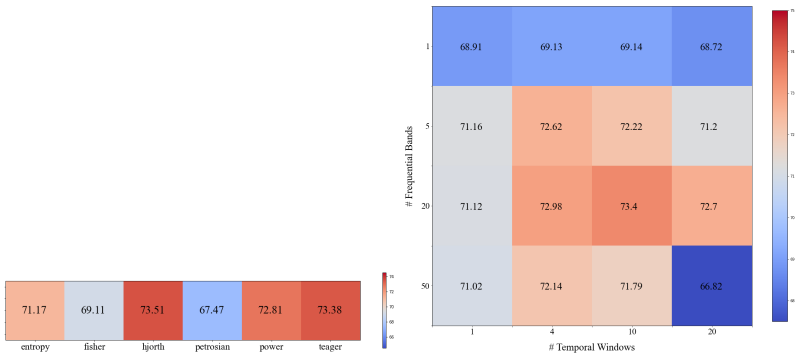


Figure 5.5. Mean cross-validation accuracy as a function of the feature extraction methods (at the left); the amount of temporal windows and frequential bands (at the right).

traction methods present cross-validation accuracy around similar range of value $\approx 70\%$. As shown, Hjorth, TE, PSD and DE present the best results. Moreover, the two best feature extraction methods: Hjorth parameters and TE based operator, consider both the signals' derivative that corroborating the fact that the derivative play an important role in the attention estimation from EEGs.

As seen in Figure 5.5, the amount of both time windows and frequency bands play an essential role in attention estimation. For both the number of temporal windows and frequential bands, a too small number of time windows/frequency bands leads to a decrease in accuracy. This decay can be caused by the difficulty of representing the evolution of the brain activity during the stimuli appearance or among the spectrum. Moreover, a large number can decrease accuracy due to high dimension or overfitting issues.

The medium values present the higher results in temporal and spectral parameters. More precisely, better results are proposed for regularly cut bands (i.e. with 20 frequency bands) compared to pre-defined bands (i.e., five frequency bands). This insight can be explained by the fact that some populations do not present the same band limits as the pre-defined [18, 109].

5.3.2 Emotion Estimation

Estimating emotion from EEG is a task often considered for the design of BCI. In this context, exploiting the previously described models for emotion estimation was considered. After experimentation, it has been noted that the saliency-based H-RNN-CNN presents better results than other approaches.

The experimentation has been made on four public datasets considering EEG signals during task promoting specific emotion: SEED-IV [46], SEED [45], DEAP [44] and MPED [40]. Each of these datasets were composed resepectively of four classes for SEED-IV [46], i.e. happy, sad, neutral and fear, three classes for SEED [45], i.e. positive, neutral and negative, four classes for DEAP [44], i.e. positive and negative dominance and arousal, and seven classes for MPED [40], i.e. joy, funny, anger, disgust, fear, sad and neutral.

The consideration of several datasets was motivated by the desire to present a general model instead of a finely tuned approach working only in a specific context.

The LOSO cross-validation accuracy has been chosen to assess the model's ability to generalize to previously unmet participants. EEG signals being very person-specific [95], a significant gap is often noted for the cross-validation accuracy between participant dependent and participant-independent. Nevertheless, BCI applications are supposed to be directly used on the participant in real-life, i.e. their signals are not used during the training of the DL model.

In Table 5.3, a comparison of the results of our approach is presented with state-of-the-art methods for emotion estimation from EEGs. Our approach presents the best results for some datasets and remains on the same scale

Dataset	SEED-IV	SEED	DEAP	MPED
Saliency	**74.42/4.8**	84.11/2.9	**78.47/4.9**	**32/4.7**
BiHDM [59]	69.03/8.6	**85.40/7.5**	-	28.27/4.9
RGNN [89]	73.84/8.1	85.30/6.7	-	-
RODAN [110]	60.75/10.4	-	56.60/3.5	-

Table 5.3. Classification performance for emotion estimation. Results above the bold line are obtained from our model experiments.

for others. This proved our approach's ability to estimate emotion in various cases. Although our approach may seem slightly better than previous works, it is important to note that the proposed approach can achieve a comparable subject-independent cross-validation performance as previous works on four datasets. Similar results have not been observed in the literature. As said, our work aimed to propose an easily generalizable approach instead of fine-tuning our results only on one dataset.

5.3.3 Visual Saliency Estimation

Another model presented in this chapter is the VAE. The task dedicated to this model was to estimate visual saliency maps from EEG signal. The signals from PhyDAA (presented in Chapter 3) were considered to solve this task. To this end, the EEG and raw eye-tracking segments were processed for each trial, and the saliency maps were computed from the raw eye-tracking signals as presented by Salvucci et al. [106] and shown in Figure 1.2. The goal of this model is to estimate a visual saliency map from an EEG segment as shown in Figure 5.4.

Due to the specificity of the task, no works considering the same modality, i.e. EEG, to estimate visual saliency maps were found in the literature. It was not possible to have a fair comparison of our approach with previous work. For this reason, the effect of the discriminator has been investigated by comparing the results for the estimation with the whole model vs the model without the discriminator. The considered metrics to quantitatively evaluate the saliency maps estimation were:

- The area under the curve (AUC) represents the area under the Receiver Operating Characteristic (ROC) curve. In the case of visual saliency estimation, the AUC has been adapted to suit the problematic by considering a changing threshold for class estimation from a value between 0 and 1 (corresponding to the saliency value). This adapted AUC is sometimes also called *AUC-Judd* [111].
- The Normalized Scanpath Saliency (NSS) is a straightforward method to evaluate the model's ability to predict the visual attention map. It measures the distance between the normalized around 0 ground-truth saliency map, and the model estimation [112].

Approach	AUC	NSS	CC
VAE with Discriminator	**0.697**	**1.9869**	**0.383**
VAE without Discriminator	0.574	1.6891	0.251

Table 5.4. Estimation performance for visual saliency estimation.

- The Pearson's Correlation Coefficient (CC) is a linear correlation coefficient measuring the correlation between the ground truth and model estimation distributions [113].

The results in Table 5.4 show that the proposed approaches present encouraging results for the AUC score representing the classification ability of the model to estimate a pixel as being seen or not, i.e. modelling the participant's visual attention. To illustrate, an AUC of 0.5 corresponds to a random classification with a model without knowledge. The closer the AUC is to 1.0, the better the model performs. In this study, we can consider the classification ability acceptable, meaning that our model cannot perform well in every case but can already distinguish specific patterns. However, the goal of our model is to detect a small visual attention region, predicted map without any salient pixel could have presented results with high AUC. We must consider other metrics to evaluate our model's ability to estimate saliency maps from EEG.

For this reason, the NSS and correlation factors have also been studied. These corroborate the insights given by AUC score that our model performs well for some configurations but cannot estimate the saliency map for every case. It is crucial to keep in mind that the goal of this approach is not to improve an existing work already stated but to discuss the possible relationship between brain activity and visual saliency maps.

The discriminator's goal is indirectly to force the generator to create images as similar as possible to the visual saliency map generated with eye-tracking signals. This phase is achieved through a competitive training process between the generator and the discriminator. In Table 5.4, we note that the adversarial model, i.e. with the discriminator, presents better results for the three metrics than the approach composed only from the generator. It seems that in addition

to promoting the generation of faithful maps, adversarial learning could also help to make a better estimation.

5.4 Conclusion

The existing DL techniques to process EEGs have been presented in this chapter. Due to the small number of works considering attention estimation, approaches considering DL for EEG processing for other tasks have also been presented. Although DL is an active field of research in the context of EEG signals processing, it was noted that a few architectures were specially designed for EEGs, i.e. by considering all the signal specificities, including time, frequency and space properties, and not copying existing ones dedicated to image or text processing.

For this reason, it was decided to present three of our contributions in the field of EEG signals processing, including 1) A Hierarchical-Recurrent Neural Network (H-RNN) architecture for processing spatial information. This architecture can be combined with an image-based EEG to consider a dual representation of information; 2) A transformer architecture to consider spatial, spectral and temporal information; 3) A Variational Autoencoder (VAE) to process EEG signals based on spatial and temporal information.

Even though several approaches have been presented in this chapter, two aspects have still to be considered in the context of DL-based EEG processing.

First, as mentioned previously, physiological signals have been co-registered jointly with EEGs. For this reason, Chapter 6 will focus on a novel approach to merge EEG signals with physiological signals.

Second, no specific approach has been considered for the training of DL models, they are trained in a supervised (i.e. the goal during the training is to make the model able to estimate input by reproducing examples) or unsupervised (i.e. the goal during the training is to find specific patterns/organization of information in the datasets) manner. For this reason, Chapter 7 will investigate a novel approach to represent EEGs in a shorter subspace by training the model in a self-supervised manner, i.e. by teaching the model how to process information instead of how to reproduce behaviours (e.g. classifying, doing regression or clustering the signals).

Chapter 6

Fusion of EEG and Physiological Signals

Contents

When constructing the PhyDAA corpus, physiological signals have been recorded in addition to EEG signals during the task promoting attention in VR in addition to EEG signals. This chapter will first focus on the existing works to combine those types of signals in Section 6.2. From these insights, three novel methods for fusion modalities are presented in Section 6.3. They have all been trained and validated on three public datasets. Due to the few works considering modality fusion in the context of attention estimation, this chapter will focus on fusion policy for EEG and other signals, but not specially dedicated to attention estimation.

6.1 Introduction

Several public datasets propose joint recordings of EEG signals and other modalities. The additional modalities can be of various natures and vary depending on the considered tasks perfomed during the dataset acquisition. A naïve approach could be to consider each separately by designing one model for each modality and training them individually. In addition to presenting better results, a fusion-based approach would also reduce the model size and training time. Indeed, instead of considering n models for the n different modalities, the n signals are processed jointly in the same model.

For this reason, several techniques in the DL field have emerged. They aim to focus on the processing of several modalities. The existing methods to combine modalities can be classified into three categories depending on the location of the fusion block: Early fusion, Middle fusion (or DL feature fusion) and late fusion. The location of the modality combination in the architecture characterizes the fusion methods: early fusion consists of fusing the input directly after the preprocessing steps, middle fusion consists of combining the feature vectors computed by DL model, and late fusion consists of combining the prediction for each standalone modality, this method can also be seen as a voting estimator. A representation of these three methods is shown in Figure 6.1.

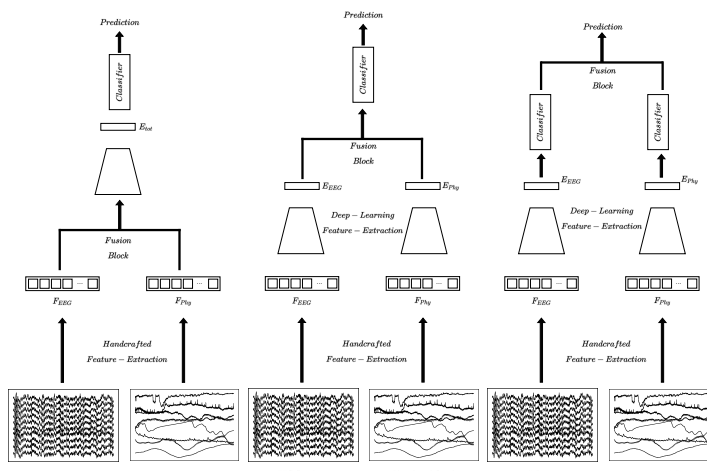
6.2 Related Work

This section reviews the existing works proposing a method to fuse EEG signals with one or more physiological modalities.

6.2.1 Early fusion

In the early fusion, the signals from different modalities are combined and passed jointly in the ML models. It is important to note that with this fusion method, handcrafted feature extraction methods have been considered. These features could represent the signals' time or frequency properties [34, 46, 49]. The fusion part mostly consists of a concatenation of all feature vectors [34, 84, 114, 115].

The concatenated feature vectors can then be passed in more or less complex architectures such as SVM [34], MLP [115], CNN [114] or a cascade of LSTM layers and Capsule Network [84]. The improvements of the proposed methods



(a) Early fusion (b) Middle fusion (c) Late fusion

Figure 6.1. Summary of the modality fusion methods in the context of EEG and external physiological signals.

can thus come from the specificity of the considered DL architecture: Cimtay et al. [114] consider the InceptionResnet architecture [116] that have shown encouraging results in computer vision-related tasks. Zhang et al. [84] adapte the original capsule networks [117] introducing spatiality information in image processing to EEGs.

Another possible analysis consists of investigating the aspects of the modalities that affect the model performances. For instance, Zheng et al. [34] propose a comparison of the fusion between EOG signals and EEG signals from different brain regions: posterior, temporal or forehead. They prove that merging information from temporal EEG regions and EOG presents the best results for vigilance estimation.

6.2.2 Deep learning feature fusion

Another category consists of fusing the extracted features by the DL models from each modality. Handcrafted features are passed to their corresponding networks in this case, and the resulting embeddings are merged. Although the considered approaches to combine these vectors can be various, concatenation remains one of the most used methods to combine the embedded vectors [46, 73,86,118,119]. However, other methods are also presented in the state-of-the-art. Zhang et al. [120] give a global fusion layer with regularization merging the embedded vectors from each modality, and Liu et al. [121] propose a canonical correlation analysis to rectify and regularize the embedded vectors from each modality.

Similarly, the handcrafted features are extracted from the raw signals before passing them through their corresponding DL model for input fusion. Several motivations explain this methodology: EEG and the considered physiological signals are easily affected by noise. To ensure the correct proceeding, features are extracted from raw signals to keep only the relevant information from them and to have a better representation. Moreover, a large part of the available datasets for EEG signal processing has a relatively small size regarding the high dimension of the signals, which makes using DL models more complex to be deemed with high dimension signals.

Although lots of existing approaches are considering concatenation as feature fusion, they may still differ based on the considered DL model and its complex-

ity: Restricted Boltzmann Machine (RBM) [46], CNN [73], MLP [120,121] and AE networks, although most DL methods consider an AE-based architecture to merge the encoded information from several modalities [86, 118, 119].

6.2.3 Late fusion

In addition to the proposed methods, few works present late fusion architectures, where each modality is used separately to make the estimation, and the decisions of each model (i.e. logits) are weighted to take the final decision. Lu et al. [122] consider this method for emotion recognition from EEG and eye-tracking signals. Although their feature fusion approach present better result than the processing of signals solely, other more complex approaches based on DL fusion have presented better results [46].

6.2.4 Other fusion methods

A less envisaged method consists in considering raw signals both for EEGs and the other modalities and combining the feature vectors extracted by each corresponding DL model [123]. As mentioned above, the motivation for using pre-extracted features from EEGs in the context of modality fusion is explained by the relatively small size of the available datasets presenting EEG signals with other modalities and the low signal-noise ratio that presents EEG signals. This issue is even more represented in the context of multimodal signal acquisition.

6.3 Proposed Methodology

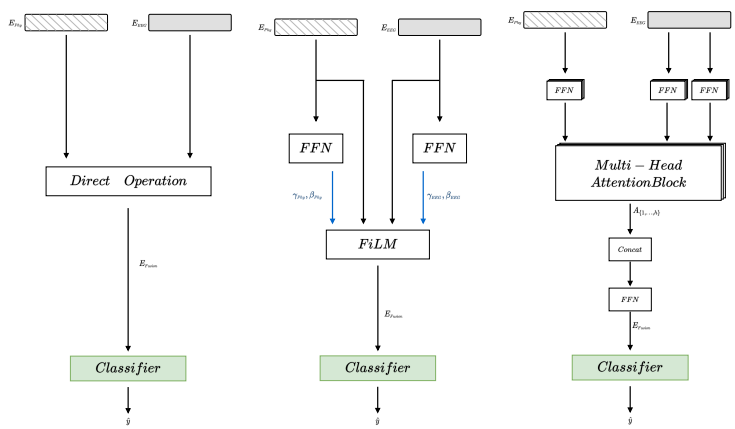
In this section, we present three novel methods to fuse information from two biomedical modalities: EEG and a set of physiological signals depending on the considered dataset. From our point of view, fusing raw signals or manually extracting features from different modalities/sources makes less sense due to their specificity. Moreover, many recent works considering DL-based fusion have shown encouraging results.

For simplicity and understanding, it was decided to consider the same notation in the entirety of this chapter. The pre-extracted features from EEG and Physiological signals (i.e. EMG, EOG, etc depending on the considered dataset) are respectively written X_{EEG} and X_{Phy}. Given the considered methodology, both vectors are passed in a DL pipeline to have an embedded representation of the input easier to fuse. The corresponding embedded vectors are respectively written E_{EEG} and E_{Phy} for both set of modalities. The two embedded vectors are passed through the fusion layers. The resulting fused vectors are written E_{fusion}, and finally, the corresponding estimation, i.e. classes or regressed score, is written $\hat{y}$. These notations are also shown in the center of Figure 6.1: middle fusion.

The methods we propose to combine the embedded vectors are Basic approaches, Feature-wise Linear Modulation (FiLM) layers and Multi-Head Attention (MHA) based fusion are represented in Figure 6.2.

6.3.1 Basic Approaches

Given both embedded vectors, the first considered methodology to merge the information from sets of modalities proposed in the state-of-the-art consists



(a) Early fusion (b) Middle fusion (c) Late fusion

Figure 6.2. Overview of the proposed methods for biomedical modality fusion.

of using *direct operations*. Among these, one of the most considered method is concatenation, whether for input fusion [34,84,114,115] or DL based fusion [46,73,86,118,119].

However, other straightforward methods can also be considered in addition to this baseline approach. The addition and inner product are simple operations that consider the interaction between pairs of elements composing both latent spaces [59]. These two basics approaches for merging embeddings are listed in Equations 6.1 and 6.2. Although these methods can express this one-to-one relationship, they present a drawback due to the fact that no interaction occurs between two elements of each embedded vector which are not located at the same position. A solution to overcome this drawback would be to consider the outer product. However, the resulting dimension of the fusion embedded vector could be drastically high: if $E_{EEG} \in \mathbb{R}^{EEG-emb}$ and $E_{Phy} \in \mathbb{R}^{Phy-emb}$, the resulting dimension of the fused embedded vector $E_{fusion} \in \mathbb{R}^{EEG-emb*Phy-emb}$. With dimensions similar to previous works, i.e. $\text{EEG}_{emb} = \text{Phy}_{emb} = 64$, the resulting embedded vector would have a dimension of 4096.

To avoid a huge dimentional representation, Gao et al. [124] propose a novel method in the context of image processing: Compact Bilinear Pooling (CBP), aiming to reduce the computational burden while speeding up computations. This method is based on the finding that an operation on two vectors' outer product can be re-expressed as the convolution of both operations applied on each vector separately [125]. In the original work, the authors introduce an FFT in the equation to benefit from the convolution theorem, which states that *"the Fourier transform of a convolution of two signals is the pointwise products of their Fourier transforms"*. It is possible to get rid of the convolution between vectors and re-express the equation with their products [124] as seen in Eq. 6.3.

Finally, the resulting direct operations are summarized:

$$E_{add} = E_{EEG} + E_{Phy} \tag{6.1}$$

$$E_{mult} = E_{EEG} \odot E_{Phy} \tag{6.2}$$

$$E_{CBP} = \text{FFT}^{-1}\left(\text{FFT}(E_{EEG}) \odot \text{FFT}(E_{Phy})\right) \tag{6.3}$$

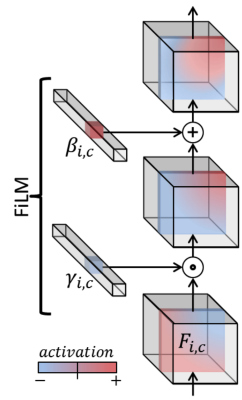


Figure 6.3. Representation of the FiLM layer as defined in the original paper [126]. The two parameters γ and β are computed from the activations maps from the modality to condition the neural network with (taken from [126]).

with E_j being the resulting embedded vectors for the method j and $\odot$ being the inner product. In our case, the fused modalities include EEG and external signals.

6.3.2 Feature-wise linear modulation

Perez et al. [126] proposed a novel approach for merging information from several modalities. This approach modulates the network during the learning phase through Feature-wise Linear Modulation (FiLM) layers. Instead of considering a direct operation between both embedded vectors, the proposed method aims to condition the learning of each neural network processing the embedded vectors. The idea is here to pass each embedded vector in a Feed Forward Network (FFN) and to condition the learning of DL feature extractor with parameters computed from the other. The general conditioning process is represented in Figure 6.3 and its application to the fusion between EEG and physiological signals is represented in the center of Figure 6.2.

The mentioned FFN corresponds to a linear transformation or a multiplication by a parametrized matrix followed by adding a bias, followed by a non-linearity

follows it. The corresponding relationship can be written as:

$$\mathrm{FFN}(x) = f(x * W + b) \tag{6.4}$$

with W being the parametrized matrix or weight, b the bias and $f(.)$ the non-linear function.

Concretely the idea is to estimate two parameters from each embedded modality computed as:

$$\gamma_i = f_\gamma(E_i) \qquad \beta_i = f_\beta(E_i) \tag{6.5}$$

with i being the considered modality, γ and β two parameters computed with the functions $f_\gamma(.)$ and $f_\beta(.)$. In this work, the considered function to estimate these parameters is a FFN that automatically estimates both parameters.

Given the parameters computed above, the FiLM layer applied the following operation on each embedded vector:

$$E_i^{FiLM} = \gamma_j \odot E_i + \beta_j \tag{6.6}$$
$$E_j^{FiLM} = \gamma_i \odot E_j + \beta_i \tag{6.7}$$

This approach aims to promote the joint processing of both neural networks by modulating their learning process.

Finally the proposed approach for FiLM based fusion between E_{EEG} and E_{Phy} can be expressed as:

$$\gamma_{EEG} = f_\gamma(E_{EEG}) \qquad \beta_{EEG} = f_\beta(E_{EEG}) \tag{6.8}$$
$$\gamma_{Phy} = f_\gamma(E_{Phy}) \qquad \beta_{Phy} = f_\beta(E_{Phy}) \tag{6.9}$$
$$E_{Film} = \gamma_{Phy} \odot E_{EEG} + \beta_{Phy} + \gamma_{EEG} \odot E_{Phy} + \beta_{EEG} \tag{6.10}$$

Given the encouraging results proposed by FiLM based neural network conditioning in various fields, e.g. question answering from images [126] or audiovisual feature fusion [127], it has been decided to investigate its use for biomedical signals fusion.

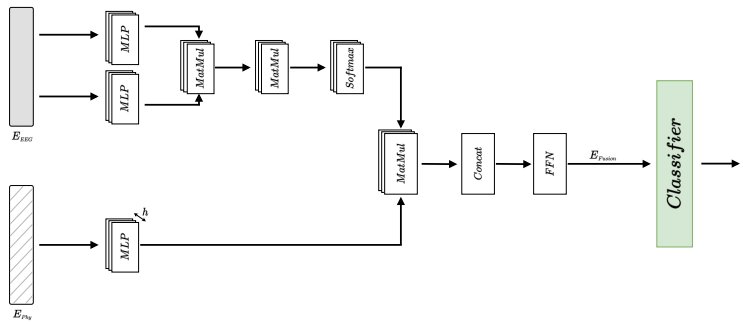


Figure 6.4. Representation of the MHA based fusion between EEGs and physiological signals. From both DL feature vectors, the attention mechanism is repeated h times representing the number of heads.

6.3.3 Multi-Head attention fusion

As explained in Subection 5.2.3, for five years now, a new family of neural networks called transformers or attention-based networks [98] has known an increase in interest, and have been used in many applications. They have proven their supremacy to solve many types of tasks including NLP [99], image processing [101], speech recognition [128] and even all of them simultaneously [129].

The main idea behind a transformer is based on the concept of attention mechanism. This mechanism has been created to overcome the drawbacks of previous works based on RNN. RNNs are considering sequential data recurrently, i.e. by considering the last entry when processing the current one, the processing (i.e. computing an hidden representation by taking into account the current entry and the hidden representation of the previous entry) is repeated for each element composing the sequence as shown in Figure 1.8. It is to overcome these drawbacks that transformers have been created. Finally, this approach initially dedicated to NLP has been adapted for estimation tasks with other modalities.

The attention mechanism can be expressed mathematically with the equations 5.3 from [98].

Instead of performing the attention mechanism once, we propose considering a parallelization of the attention function h times in different projected dimensions. The resulting MHA mechanism is represented in Figure 6.4 and mathematically expressed as:

$$\begin{aligned} \text{MHA}(Q, K, V) &= \text{concat}(head_1, \ldots, head_h)W_0 \\ \text{with } head_i &= Attention(QW_i^Q, KW_i^K, VW_iV) \end{aligned} \tag{6.11}$$

With W_i^j being the parameter matrice or weight of the i^{th} head and j^{th} input matrix and h being the number of heads.

Recent works have also shown that it is possible to consider these architectures for modality combination or fusion. Some of them have presented encouraging results in the combination of images and LiDAR view [130], visual, acoustic and linguistic modalities for emotion recognition and sentiment analysis [131], or audio and visual modalities for emotion estimation [132, 133].

For this thesis, Equation 5.3 and 6.11 defined above have been re-expressed to suit the paradigm studied. The resulting equations are:

$$\begin{aligned} \text{MHA}(E_{EEG}, E_{Phy}) &= \text{concat}(A_1, \ldots, A_h)W_0 \\ \text{with } A_i &= Attention(E_{EEG}W_i^{EEG}, E_{EEG}W_i^{EEG}, E_{Phy}W_i^{Phy}) \end{aligned} \tag{6.12}$$

As seen in Eq. 6.12, the modalities are combined with the help of the attention mechanism.

In addition to providing encouraging results in other fields, the use of MHA based fusion is motivated by its ability to capture the relationship between sequential signals with the help of the attention mechanism. Compared to existing methods based on RNNs, the strength of this approach is that transformers take into account the information regardless of its position in the sequence. In contrast, conventional approaches capture only the recurrence between consecutive elements. Therefore, elements are processed differently based on their relative/mutual distance in the sequence (two elements *far* from each other will be processed differently than two *neighbour* elements with RNNs).

6.4 Results

6.4.1 Experimental details

In the experiments, three different datasets have been used. These datasets are composed of EEG signals recorded with a specific experimental procedure following particular conditions. In addition, each dataset is composed of physiological recordings representing other aspects of the experimentation.

The differences between datasets consist of the number of considered electrodes during the recording, the nature of the additional physiological signals, and the proceeding task during which the signals have been recorded. The considered datasets are:

- DEAP [44] assessing emotion during the recording of EEG, EOG, EMG, GSR.
- SEED-VIG [34] assessing vigilance during the driving task with eye-tracking signals, EEG and EOG signals are recorded during the study.
- PhyDAA [49] assessing attention in VR based on eye-tracking signals. The recording procedure is described in Chapter 2. This dataset proposed the joint recording of EEG and physiological signals consisting of heading movement and eye-tracking signals.

These datasets are described more in detail in Chapter 3.

Due to the number of experiments and datasets considered for our experiment, we decided to present the results for each dataset separately. For clarity, the same nomenclature and acronyms will be kept as in this sections. EEG and Phy correspond to estimation from only EEG signals and physiological signals. Among the direct operation, Concat, CBP and Mult respectively correspond to concatenation, compact bilinear pooling and multiplication of the embedded vectors. Finally, FiLM and MHA correspond to the feature-wise linear modulation and multi-head attention-based fusion.

6.4.2 Attention estimation

An interesting task on which our method can be tested is that of attention estimation. The datasets proposing this task consider EEGs with other modal-

ities registered during task promoting attention. The goal is to estimate the attention class, i.e. focused vs distracted. For this purpose, we consider the PhyDAA dataset [49] proposing EEG and physiological signals (corresponding to eye-tracking and head position) to estimate attentional state during a task promoting selective and sustained attention state in VR. In this section, the ability of the model to predict a high (resp. low) attention state will be measured, and subject-dependent and independent cross-validation as explained in Subsection 3.2.3 accuracy will be used to measure the model's ability to combine modalities efficiently.

Table 6.1 provides the results of attention estimation from fused modalities. As seen in the table, the proposed methods to fuse modalities present higher results than those considering modalities separately. Moreover, the results of the proposed methods to combine information from several modalities present better results than existing works in the state-of-the-art.

The first column of Table 6.1 presents the cross-validation accuracy for the participant dependent study. As seen, the proposed methods' results present higher results than previous approaches based on GNN with a pooling opera-

	Accuracy Dep.	Accuracy Ind.
GNN [49]	77.3/10.3	72.4/5.51
CNN [73]	73.9/9.8	70.4/6.9
EEG	74.1/7.6	71.8/7.1
Phy	72.1/8.6	61.8/6.9
Concat	81.1/8.8	72.4/4.6
CBP	78.3/8.3	71.2/5.6
Mult	78.2/8.4	71.2/5.5
FiLM	82.2/8.3	73.6/6.8
MHA	**84.6/6.9**	**76.9/4.9**

Table 6.1. Classification performance of the different methods considering participant dependent and independent cross-validation accuracy on the PhyDAA dataset [49]. The results above the bold line correspond to the state-of-the-art models. The experimentation is performed for each participant so the each metric is represented by its mean and standard deviation.

tion to keep only the relevant nodes [49] and a CNN [73]. However, among the proposed methodologies to fuse information, MHA is the approach presenting the best results for attention estimation.

The second column of Table 6.1 presents the cross-validation accuracy for the participant independent analysis. A similar observation can be inferred for the subject independent study: the proposed approaches present higher results than the existing methods, and the attention-based method for feature fusion also acquire the best results.

Finally, the results above corroborate that MHA based fusion correctly combines information from different modalities. Moreover, this method can also predict attention from a smaller set of signals, i.e. subject dependent cross-validation, and extrapolate information from never met subjects, i.e. subject independent cross-validation.

6.4.3 Vigilance estimation

The second dataset for multimodal biomedical signal processing is SEED-VIG [34] which is composed of EEG and EOG signals. This dataset studies the vigilance state during a driving task, promoting sustained attention. The task has been segmented into four seconds length segments, and for each of them, a vigilance score between 0 and 1 corresponding to awake to asleep has been computed. It corresponds to the PERCLOS score computed from eye-tracking signals. The studied paradigm in this context corresponds to a regression, the estimation being a score and not a class. Although the goal is to predict the nearest possible score representing the vigilance state, the original paper [34] also propose to consider the specific threshold to express this problem as a classification.

Table 6.2 and 6.3 present the results obtained by our model to predict the vigilance score from EEG and EOG recordings separately and with the five propose approaches to merge information. Since the problem is now a regression problem. The considered metrics have been adapted: Root Mean Squared Error (RMSE) and Correlation Coefficient (CORR) have been considered to evaluate the prediction performance. As seen in both tables, the proposed approaches perform well and obtain results compared to the state-of-the-art

	CORR	RMSE
MAE [118]	0.85/0.06	0.094/0.030
SVR [34]	0.85/0.09	0.133/0.017
AE [86]	0.89/0.14	0.08/0.04
CapsAtt [84]	**0.98/0.01**	**0.030/0.010**
EEG	0.87/0.04	0.2/0.002
Phy	0.73/0.18	0.14/0.017
Concat	0.95/0.03	0.06/0.002
CBP	0.92/0.04	0.07/0.003
Mult	0.93/0.04	0.07/0.003
FiLM	0.95/0.03	0.063/0.002
MHA	0.97/0.03	0.062/0.001

Table 6.2. Vigilance prediction performance of the different methods considering participant dependent cross-validation on SEED-VIG dataset [34]. The results above the bold line correspond to the state-of-the-art models. The experimentation is performed for each participant so the each metric is represented by its mean and standard deviation.

methods for the subject-dependent and independent cross-validation. Moreover, the proposed fusion methods also present better results than signals taken separately.

	CORR	RMSE
ADDA [115]	0.84/0.13	0.141/0.051
CapsAtt [84]	0.88/0.11	**0.109/0.070**
EEG	0.69/0.15	0.682/0.021
Phy	0.67/0.18	0.694/0.025
Concat	0.73/0.12	0.247/0.023
CBP	0.91/0.06	0.221/0.023
Mult	0.93/0.04	0.23/0.024
FiLM	0.73/0.18	0.22/0.031
MHA	**0.95/0.03**	0.2/0.025

Table 6.3. Vigilance prediction performance of the different methods considering participant independent cross-validation on the SEED-VIG dataset [34]. The results above the bold line correspond to the state-of-the-art models. The experimentation is performed for each participant so the each metric is represented by its mean and standard deviation.

The subject-dependent estimation results are presented in Table 6.2. The proposed approach based on MHA presents a lower error rate represented by the RMSE compared to other proposed approaches. The previous works are constituted by Multimodal AE [118], Support Vector Regression [34], AE [86] and Capsule Attention Networks [84] presenting the best results for vigilance estimation on SEED-VIG dataset. Although the model performs well with a low error rate, a slightly lower CORR is shown for our best approach compared with previous methods. However, their standard deviation makes the two cross-validation correlations overlap, which makes these improvements non-significant. Given the results of Table 6.2, it is possible to affirm that the proposed methods can estimate vigilance from a single participant given its EEG, EOG or both. Moreover, the proposed approaches present non-negligible improvements compared to single modality methods.

Table 6.3 presents the results for the participant-independent cross-validation metrics. The results obtained during the participant-independent validation are lower due to the relatively more complex task than a subject-dependent study and the necessity of considering inter-subject variability in the estimation process. In this context, the results from two previous works are presented: Adversarial Discriminative Domain Adaptation for AE [115], and Capsule Attention Networks [84]. The best results for the subject-dependent analysis are also acquired for the second approach. The lower amount of work considering subject-independent results is explained by the higher computational cost that this cross-validation method presents: with this methodology, the training is repeated for each participant, and the resulting training set is $n_{subject} - 1$ times bigger than for the subject dependent analysis, with $n_{subject}$ being the number of the participant taking part to the study. Table 6.3 shows that the results obtained by the proposed approaches outperform the results for estimation from a single modality. Moreover, MHA based method presents the best results outperforming other experiments. These results prove that most proposed approaches can take the relevant information from participants' signals and make estimations from unseen participants.

As mentioned above, it is possible to re-express the regression as a classification task. As noted in the dataset's original paper: a score below 0.35 can be considered awake, between 0.35 and 0.7 as tired and above 0.7 as drowsy. Thus by considering these two thresholds, it is possible to re-express the regression problem as a classification task. The thresholds have been applied to each

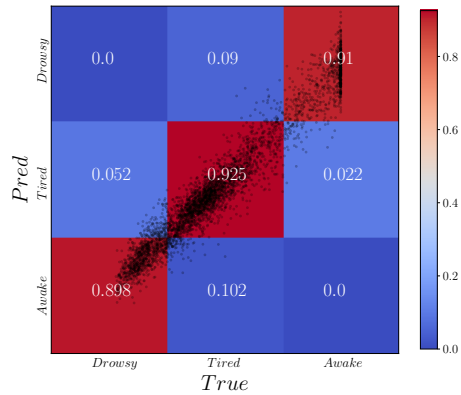


Figure 6.5. Confusion matrix regression/Classification. Confusion matrix from the regressed predictions, black dots represent the discrete prediction and ground truth values between 0 and 1. Boxes represent the density of the predicted and truth classes inferred from regression.

method, and the resulting accuracy reflects the same behaviour as the RMSE and CORR: a higher accuracy is denoted for merging methods compared to single modality methods. Moreover, the best results are obtained with MHA for the subject (in)dependent cross-validation with an accuracy of 91.48/5.33 (64.29/8.32). Figure 6.5 gives the dual representation of information with dots representing the discrete estimation and the boxes representing the predicted and ground-truth classes, i.e. awake, tired and drowsy. One of the advantages, in addition to good predictions, is that wrong forecasts remain near ground truth, i.e. no predictions of a drowsy state were made from awake signals. Therefore, in addition to giving accurate predictions, the errors made by the model remain very low.

6.4.4 Emotion estimation

The emotional task consists of estimating an emotional state from physiological signals. Moreover, emotion can be represented into a three axis coordinate frame, each representing a metric of the emotional state: arousal (i.e. excitation to disinterest), valence (i.e. pleasant to unpleasant) and dominance. In this context, it was decided to measure our approach's ability to estimate a high (resp. low) score for each metric (i.e. low versus high arousal, valence or

	valence	arousal	dominance	mean
Inception [114]	86.6	84.7	-	-
SAE [119]	-	65.9	78.1	-
DCCA [121]	-	84.3	-	85.6
EEG	79.4/23.5	82.2/23.9	84.9/22.1	82.1/23.3
Phy	78.4/22.9	83.6/26.6	84.4/24.6	88.1/20.3
Concat	82.5/23.9	88.5/18.4	84.1/9.9	88.4/18.9
CBP	87.2/21.7	87.1/19.2	88.1/22.8	87.4/21.3
Mult	88.1/18.9	83.3/20.8	84.1/21.7	86.1/19.8
FiLM	87.8/21.5	87.9/18.4	**88.7/20.9**	84.8/23.7
MHA	**90.6/15.6**	**91.6/21.7**	86.8/23.2	**89.2/23.2**

Table 6.4. Classification performance of the different methods considering participant-dependent cross-validation accuracy on the DEAP dataset [44]. The results above the bold line correspond to the state-of-the-art models. The experimentation is performed for each participant so the each metric is represented by its mean and standard deviation.

dominance) and a mean score in the case where all the metrics are estimated simultaneously.

As seen in Table 6.4 and 6.5, the proposed methods to merge information from different modalities present higher results compared to considering each modality separately. Moreover, the results presented by the model proposed are generally higher than the ones proposed by the existing approach. This insight is observed for both participant-dependent and independent cross-validation.

Table 6.4 presents the results for the participant-dependent training. The best valence, arousal, and mean results are obtained by the MHA approach corresponding to the attention-based fusion. The second best results are obtained by the FiLM based approach. Although the methods result in a similar ranges of values, these generally perform at least equal or better compared to approaches proposed by the state-of-the-art. A recurring issue is that previous works mainly focus on estimating one or two metrics from the DEAP dataset. For this reason, it was decided to present results from several earlier works to present a fair comparison with state-of-the-art methods: InceptionResnet model [114] estimating valence and arousal; Stacked AE [119] presenting feature fusion for arousal and dominance estimation; the Deep Canonical Correlation Analysis based approach that proposed the arousal and general emotion

	valence	arousal	dominance	mean
Inception [114]	-	-	-	32.5/15.6
DFKM [120]	64.5	63.1	-	-
MSD [134]	-		-	77.6
EEG	77.3/5.5	79.2/9.9	78.6/7.3	78.4/7.8
Phy	74.8/3.2	75.3/5.2	77.4/8.4	78.5/6.1
Concat	79.6/5.3	80.9/6.5	79.9/9.9	80.1/7.5
CBP	79.9/7.5	78.5/5.6	80.5/10.6	79.7/8.2
Mult	79.9/5.1	81.5/6.5	81.1/9.4	80.9/7.3
FiLM	78.6/5.1	79.2/7.3	79.8/9.5	79.2/7.5
MHA	79.9/4.1	81.2/6.9	81.8/8.9	80.9/7.2

Table 6.5. Classification performance of the different methods considering participant independent cross-validation accuracy on the DEAP dataset [44]. The results above the bold line correspond to the state-of-the-art models. The experimentation is performed for each participant so the each metric is represented by its mean and standard deviation.

estimation from both signals. The results of Table 6.4 corroborates that most of the proposed approaches perform well for feature fusion in the context of subject-dependent studies.

Table 6.5 proposes a similar analysis for the subject-independent study. Participant-independent estimation is generally more complex than participant-dependent due to the variability in biomedical signals with the chosen participant. However, results also outperform previous works. Similarly, the best results are proposed by MHA. However, the proposed methods do not provide the score for all the metrics of DEAP dataset. For this reason, several previous works have been considered: InceptionResnet model [114], Deep Fusion Kernel Machine [120] and Multi-Stage Decision Method [134]. The results of Table 6.5 also corroborate that most of the proposed approaches perform well in generalizing for participants' signals they have never met.

In addition to the raw results presented by Tables 6.4 and 6.5, a comparative study can also be made to investigate the differences between the proposed approaches. Similarly, for the model-based comparison, the study has been performed for the estimation of valence, arousal and dominance and the mean estimation. This experiment is also made for participant-dependent and independent cross-validation.

As seen in Table 6.4, the unimodal approach, i.e. considering only physiological signals or EEG signals, present lower results compared to the merged approaches. Although the results remain pretty good, no trend is noted from the standard deviation, which does not bring out a specific model stability compared to others. It confirms the assumption that considering several modalities, each reflecting particular behaviours instead of a single one, helps forecast this specific behaviour. Moreover, the MHA feature fusion method seems to be an interesting approach for feature fusion, given its relatively high performance compared to other methods.

From Table 6.5, more detailed observations can be made. The hypothesis regarding the signals separately or merged is likewise confirmed. However, where a trend was observed for participants-dependent analysis, it is impossible to identify a methodology outperforming other approaches. The greater difficulty of participant-independent analysis can explain this difference with an intra-subject validation. No comparison is also possible for the model stability, the standard deviation of each model overlapping each other. However, a lower standard deviation is observed for the participant-independent methods (four times lower). This insight could be explained by the fact that more signals have been considered for this analysis compared to subject-dependent analysis.

6.5 Conclusion

In this chapter, we first presented the existing approaches and the gaps for multimodal analysis of neurophysiological signals. From the observations, three novel methods have been designed: Compact Bilinear Pooling (CBP), Feature-wise Linear Modulation (FiLM) and Multi-Head Attention (MHA).

The three methods have been trained and validated on three public datasets, and the results obtained corroborate the fact that considering a more complex approach to fuse EEGs with other signals can be benefic for various tasks: attention, vigilance and emotion estimation. Finally, the MHA based method provides the best results for all the tasks and the cross-validation methods. MHA also outperforms the results of previous works.

Chapter 7

Self-Supervised Approach for EEG Feature Reduction

Contents

As mentioned in the previous chapters, EEGs are rarely processed directly in DL models. Moreover, several approaches to extract information from EEG signals have been developed. These feature vectors are deduced from the time, frequency or space properties. Although these approaches suit various applications, the evolution of DL techniques has demonstrated that handcrafted feature extraction from modalities presents several limitations that can be overcome with the use of DL pipelines. For this reason, it has been decided to investigate the design of such approaches with a specific learning methodology.

This chapter first introduces the concept of Self-Supervised Learning (SSL) in Section 7.1. The related works considering SSL for EEG processing are presented in Section 7.2. Based on the limitations of previous works, the proposed methodology for SSL processing is presented in 7.3 with the results of the preliminary study in Section 7.4.

7.1 Introduction

For some years now, DL algorithms have proven their supremacy to solve tasks automatically and sometimes even better than humans. The scope of these last is huge: computer vision, natural language processing (NLP), speech recognition, image generation or biomedical signal processing. Even if these algorithms present encouraging results, several concerns remain: how to ensure that the model is learning and if so, is the learning based on relevant and unbiased information? [4, 135] Moreover, is it appropriate to consider supervised methods to train these models (i.e. by showing a large number of examples to the model to make it reproduce them), while it is not the case in the nature for animals and humans babies?

In this context, we decided to consider the use of recent methods to train algorithms to promote a better understanding of information. This new learning approach, called Self-Supervised Learning (SSL), aims to put forward the understanding of the information by the model prior to make it reproduce example.

Another important aspect of this approach is the separation of the learning into two steps: the pretext and downstream tasks. The pretext task is the first step where some basic knowledge is learned with SSL algorithms. The

tasks can take various forms: solving a jigsaw puzzle from images [136], estimating rotation from images or the relative position of patches [137]. Novel methodologies propose to evaluate clusters to which the vector representation create by DL architecture can belong, these clusters being computed in previous steps and automatically updated [128, 138]. After considering the pre-trained pipeline resulting in the pretext task, it is possible to re-use the part aiming to extract features from the modalities during the downstream task. The pipeline is re-used to proceed to a supervised learning task. The aggregated knowledge from the pretext tasks can thus be re-used to improve the learning during the downstream task.

Other pretext tasks often developed consist of processing to an attraction or repeal mechanism between feature vectors [139, 140], these vectors are automatically computed from a DL model called backbone. This attraction (resp. repeal) mechanism consists of teaching the model to create feature vectors with a mathematical high (resp. low) similarity. The goal of this task is to *attract* two embedded vectors from the same images, video or sound after different modifications (e.g. rotation or cropping) and to *repeal* the ones being the transformation of other inputs. This problem can be re-expressed as an optimization problem where the goal is to decrease a contrastive loss between two transformations of a given modality x respectively z_1 and z_2 computed as [139]:

$$l(z_1, z_2) = -log\left(\frac{\exp(sim(z_1, z_2))}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq 1]} \exp(sim(z_1, z_k)/\tau)}\right) \tag{7.1}$$

with $sim(z_i, z_j)$ being a similarity function, e.g. the cosine similarity function shown in Equation 7.2. The Function $\mathbb{1}_{[k\neq i]}$ is an indicator function equal to 1 for each index except the i^{th}, and τ is the temperature parameter controlling the penalty of the negatives samples [141] fixed to 0.1 in several implementations.

$$sim(z_1, z_2) = \frac{{z_1}^t z_2}{||z_1||||z_2||} \tag{7.2}$$

7.2 Related Work

Self-Supervised Learning (SSL) is a recent concept initially dedicated to processing images-based modalities. However, recent works have emerged and present the use of SSL for EEG processing.

Banville et al. [142] propose a novel approach for processing of EEGs based on SSL algorithms. Their methodology considers pretext tasks aiming to predict if a temporal window is temporally close (i.e. the difference between both recording times is lower than a predefined threshold) or far (i.e. time difference is higher than the given threshold). Their approach assumes that the signals' representations should evolve slowly with time [143]. The pretraining has been performed on the TUH Abnormal Dataset [144] composed of more than a year of EEG recordings. The resulting trained backbone for automatic EEG feature extraction is then re-used to predict sleeping stage from EEGs [48]. Their approach presents promising results and proofs that SSL methods can be applied for EEG signal processing, especially to create a general feature extraction backbone. However, it can be seen as naïve due the fact that it is only processing the temporal evolution of EEG signals (frequency properties and electrode location are not taken into account during the training).

In 2019, Baevski et al. [102] presented wav2vec 2.0, a model based on the transformer to process speech signals. The motivation behind this consideration is the fact that speech signals can be expressed, like EEGs, as temporal series of unique dimensions (the only difference being that EEG signals can be recorded on several regions of the brain considering them as a set of time series instead of a unique one). In addition to this new architecture, an innovative training methodology is also proposed. It implements a contrastive training between feature vectors extracted from transformer with masked part of the signal. The advantage of wav2vec pointed out by the authors is that one hour of label speech data can be enough to fine-tune the model and beat any state-of-the-art methods. From this promising approach, Kostas et al. [145] adapted this model and its specific pretraining methodology to EEG signal. They kept the wav2vec backbone [102] and fine-tuned it with the TUEG dataset [144]. Their method has been tested on the same downstream task for sleeping-stages estimation [48] and other public EEG dataset [47]. Their approach corroborates that SSL learning can be leveraged to create a general backbone for EEG feature extraction. Although the presented results are very encouraging, this

method is not considering other signal properties than its temporal evolution. Moreover, even if EEG signals are represented as 1-D vectors, they present differences from speech signals: they have different spectral properties and are more difficult to separate into discrete segments (e.g. the word in NLP or speech units in speech signals).

In this context, it was essential to consider the pros and cons of creating a method to extract features from raw signals based on SSL. The novel approach processes the properties and specific aspects of the EEGs.

7.3 Methodology

In this section, we proposee our approach for automatically extracting feature vectors from EEG signals.

7.3.1 Architecture

Due to the lack of baseline networks to extract features from raw EEGs, unlike to well-known image dedicated networks, e.g. ResNet [26]. We propose a general DL approach to extract feature vectors from EEGs.

To consider the most generalist and simple architecture, it has been decided to design a VGG-liked architecture [25] inspired from previous works [81,102,142]. This architecture is considering 1-D convolutions layers applied sequentially on each EEGs segments resulting to one feature vector for each EEG segment of one electrode. Moreover, to promote the use of this model with the various datasets, the convolution among the dimension of the electrode have been removed from this architecture to keep a more general model.

As mentioned, the main idea is to create a novel backbone that automatically extracts a feature vector from raw EEG signals, this feature vector corresponding to the resulted embedding from our VGG architecture after the SSL pretext task.

7.3.2 Pretexts tasks

At first, we consider pretext tasks to train the feature extractor or backbone to process EEG signals. Recent works [139, 140] have shown that applying first a pretext task presents better results for the knowledge transfer during the downstream task (i.e. classification or regression).

In this context, instead of pretraining the architecture described above in a supervised manner, we decided to teach the backbone how to analyze EEG signals instead. This was done by designing tasks based on the EEG signal' properties, i.e., their temporal evolution, frequency-based properties and electrode location.

In the transfer learning context, this step aims at helping the knowledge transfer to classify signals from another dataset.

Time-domain task

Given an EEG signal x, a jigsaw version of this signal can be created. The jigsaw EEG $x_{swapped}$ corresponds to a swapped version of the original signal x in the temporal domain. To create $x_{swapped}$, the original signal x is first augmented with data augmentation methods s(presented in Subsection 7.3.3) and segmented into $n_{windows}$ (of duration equal to 0.8 s). The corresponding windows are then randomly sorted and concatenated to create the temporally swapped signal. An overlap has been considered for shared signals between windows to keep information between temporal windows.

The time-related task predicts the original sequence from the temporally transformed signal. Concretely, this task sorts the windows that been randomly shuffled in the correct order. This task can also be considered a classification task consisting of predicting the proper order among the $n_{windows}!$. Concretely, if the model estimates a wrong prediction with a part of the correct information (i.e. partially predicting the sequence), it is computed as similar to a random prediction without any predicted windows at the right position. It is explained by the fact that during the learning process, the network can only predict right or wrong sequence, and not a partially right sequence. The drawback with this method is that it only promoted the accurate forecast versus wrong and did not differentiate among the errors that can appear. For this

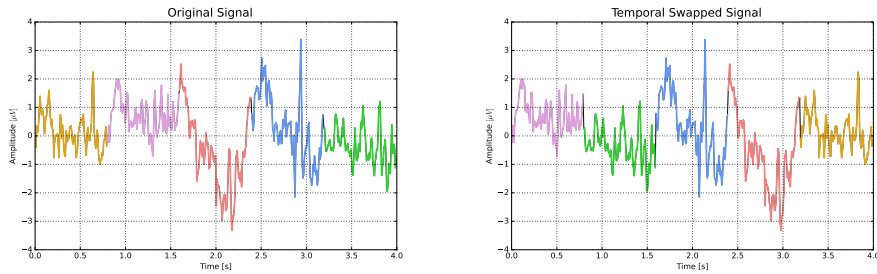


Figure 7.1. Time-domain task for SSL. On the left, the original signal is divided into time windows on the right, the windows are shuffled in random positions.

reason, the similarity between the predicted order and the order of the true windows has been considered, and a similarity loss has also been implemented. The loss corresponding to the temporal task can thus be summarized as:

$$Loss_{temporal}(y, \hat{y}) = Loss_{classification}(y, \hat{y}) + Loss_{similarity}(y, \hat{y}) \tag{7.3}$$

with $Loss_{classification}$ a classical cross-enropy, $Loss_{similarity}$ equal to the inverse of the cosine similarity, y and $\hat{y}$ the ground-truth and predicted sequence order.

Frequency-domain task

After the temporally modified EEG signal $x_{swapped}$, we decide to also take into account the spectral information located in EEG signal x. Given this signal x, it is possible to compute the PSD representing the signal contributions in each frequency range. The corresponding periodogram P_f is computed from the squared absolute value of Discrete Fourier Transform (DFT) for each frequency f_f, i.e. $P_f = |\text{DFT}(x)|^2$ at the frequency f_f. Finally, with this information, a harmonic reconstruction of the signal x is computed as:

$$x_{harm}(t) = \sum_{i}^{N} P_i * sin(2\pi f_i t) \tag{7.4}$$

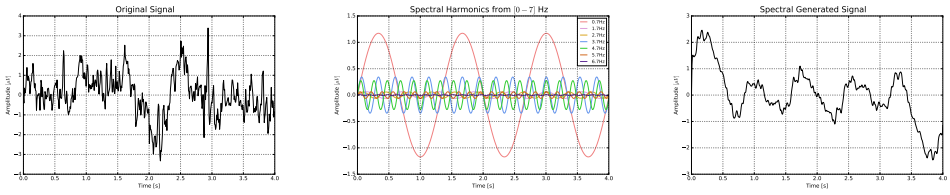


Figure 7.2. Frequency Task for SSL. Representation of the pipeline for the frequency-related task: on the left the original signal, in the centre its decomposition into harmonics and on the right the resulting harmonic reconstruction.

After computing the x_{harm}, the goal of the spectral task is to proceed to the attraction between x and its harmonic representation x_{harm}. The corresponding signal aims at representing its contribution to the spectral domain. The repeal mechanism was applied between the harmonics and different signals with the similar process (in this case the denominator of Equation 7.1 was smaller than the numerator that makes high value for the contrastive loss). To proceed with this task, the contrastive loss presented in Equation 7.1 has been adapted to suit the signals.

Spatial-domain task

In addition to the time and frequency information located among EEG signals, it is possible to process electrodes' positions on the scalp as a source of information for EEG signal processing. In previous work, several methods considering this type of information by selecting the relevant channels or making spatial-based filtering have presented promising results [17].

The idea was to adapt the advances proposed by the SSL approach in computer vision [139, 140] to EEG signal processing. As mentioned, these approaches perform an attraction and repeal mechanism with similar elements after modification or augmentation of the information. Two elements can thus be considered similar if they were the transformed version of the same image in the case of computer vision task. A repeal mechanism is made between patches from different images. In the context of EEG, we designed a similar approach by attracting the embedded vectors from the same EEG trial (i.e.

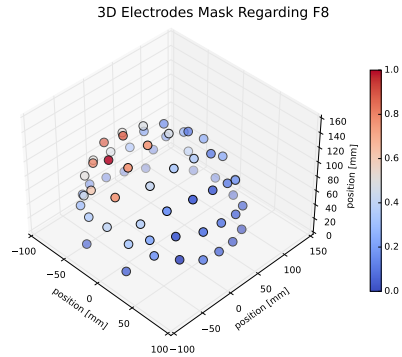


Figure 7.3. 3-D representation of mask with respect to the F8 electrode for spatially weighted loss for the Spatial Task. Reder (resp. bluer) are electrodes, higher (resp. lower) is the attraction/repeal mechanism during the training between the feature vectors from these electrodes and F8.

signals recorded at the same time but from different electrodes), respectively repealing embedded vectors from different EEG trials.

Given two EEG signals x_1, x_2 of dimensions $[n_{electrodes} \times s_{duration}]$ with $n_{electrodes}$ and $s_{duration}$ being the amount of electrodes and signal length in samples, it is possible to construct two augmented version of these EEG signals $\tilde{x}_1, \tilde{x}_2$. After passing the signals into the same backbone, the signals can be re-expressed as two embedded vectors z_1 and z_2 of dimensions $[n_{electrodes} \times d_{embedded}]$ with $d_{embedded}$ being the dimension of the resulting embedded vector. The embedded vectors can also be considered as a set of signals, one for each channel: $z_i = [z_i^1, z_i^2, \ldots, z_i^{n_e electrodes}]$.

From the above-described signal representation, an attraction and repeal mechanism is applied with the contrastive loss function described in Equation 7.1. However, the formula has been slightly modified to introduce information from the relative position of electrodes. The equation remains the same except that the indicator $\mathbb{1}_{[k \neq i]}$ is replaced by a scalar matrix $G = G_{i,j}$ multiplying the $\exp(sim(z_i, z_j)/\tau)$ term:

$$l(z_1, z_2) = -log\left(\frac{\exp(sim(z_1, z_2))}{\sum_{k=1}^{2N} G_{ik} \exp(sim(z_1, z_k)/\tau)}\right) \tag{7.5}$$

The elements of the scalar matrix elements are in the range $\in [0, 1]$, depending on the normalized distance between the signal from the i^{th} and j^{th} electrode. A 3D representation of this matrice is given in Figure 7.3.

Concretely, with the channel weighted loss, the signals from various electrodes belonging to the same EEG trial are attracted proportionally to the distance between the electrodes, i.e. two spatially close electrodes will be more attracted than two electrodes with a higher distance separating them. The process is repeated and reversed for the repeal mechanism: two electrodes' signals belonging to different EEG segments will be repealed proportionally to the distance separating them. In other words, two signals from the same electrodes but different EEG trials will be more repealed compared to two electrodes' signals from electrodes further from each other.

7.3.3 Dataset augmentation

During both pretext and downstream tasks, we decided to consider a data augmentation policy to help the network generalize and be more robust against overfitting issues. This process aims at proposing to the network a different version of EEG signals with the same properties.

Three tasks have been applied in this work, each of them considering one of the specific aspects of EEG signals: temporal (among time), frequency (among the spectral bands) and spatial (among the location of the electrodes on the scalp). In this context, it has been necessary to design data augmentation methods that will not impact the properties of the signals regarding the task. Indeed, we decided to consider two categories of augmentation: time conservative augmentation keeping the temporal properties of the signal and frequency conservative keeping the frequency-related properties.

Time conservative data augmentations are listed as:

- Local Mean Assignation: assigning each elements of a signal's segment randomly choosen, the mean value of the signal in this window.
- Signal inversion: consists of multiplying the signals by -1.
- Random crops: consists of cropping the signal at a ratio of 80% randomly on the signal. This step consists to only keep a segment of length equal to 80% of the original signal length and to remove the border.

- Noise addition: consists of adding snoise following a gaussian distribution with a zero mean and standard deviation ratio between signal and noise equal to 100.

Frequency conservative data augmentation is similar to previously proposed methods described above with two specificities:

- Local Mean Assignation: assigning each elements of a signal's segment randomly choosen, the mean value of the signal in this window.
- Signal inversion: consists of multiplying the signals by -1.
- Random crops: consists of cropping the signal at a ratio of 80% randomly on the signal. This step consists to only keep a segment of length equal to 80% of the original signal length and to remove the border.
- Noise multiplication: consists of adding snoise following a gaussian distribution with a zero mean and standard deviation ratio between signal and noise equal to 100.
- Random swapping: consists to the same process than for the time-domain task except that it is used as a data augmentation with a number of windows equal to 2.

7.3.4 Downstream task

To verify the performances of the backbone trained with SSL on the TUEG Dataset [144], we envisaged to re-use the pre-trained backbone to perform a classification task. The classification has been tested on three public datasets considering motor imagery [47, 146] and sleeping stage classification [48].

The datasets used for the experimentation present the following characteristics:

- BCI competition 2008 - Graz dataset A [47] is a challenge from 2008 consisting to estimate the motor movements from EEG signals. The dataset includes four classes: movement of right, left hand, tongue and both feet.
- MMI [146] is another dataset for motor imagery movement. The dataset is separated into three classes depending on the task: 1) resting state; 2) left, right or both fists; 3) left, right or both feet.

- Sleep-EDF [48] is a dataset that aims to classify the sleeping patterns or hypnograms. The dataset is composed of EEG and Physiological recordings, i.e. EOG and chin EMG. However, the backbone being pre-trained with EEG signals we decided to only keep the EEG signals.

The choice of these datasets is motivated by studying several aspects: the effect of dataset size, electrode amount/placement and behaviour to estimate. All of these aspects also have to be considered by keeping in mind that the dataset is suitable and comparable for automatic estimation with raw signals, a large amount of datasets only considering pre-extracted features [46, 49].

7.4 Results

Table 7.1 presents the experimental results for the downstream task in various contexts depending on the considered datasets and tasks performed during the pretraining.

The first insight made from the observation of Table 7.1 is the improvements of the results for the downstream tasks when combined with a pretext task. This observation is repeated for each dataset and pretext task. This insight can be explained by the fact that pretraining a model, even with another dataset, can be helpful for better extraction of information from raw EEG.

Dataset	Supervised Standalone	Dual Approach			
		Temporal	Spectral	Spatial	Mixed
BCI [47] *4 – classes*	32.81/1.26	41.54/3.7	42.82/4.1	39.50/5.1	43.96/5.2
MMI [146] *3 – classes*	66.92/5.68	70.66/7.04	72.54/2.5	68.03/7.68	74.42/4.9
Sleeping-EDF [48] *5 – classes*	38.59/8.34	42.80/6.62	43.37/6.39	40.72/6.08	45.28/6.04

Table 7.1. Mean/standard deviation of the cross-validation accuracies for the downstream (ie classification) task for different datasets and with different pretext tasks.

In this experiment, we demonstrate that transferring knowledge from one EEG dataset to another is possible. Moreover, it can also be generalized for the dataset with other paradigms, tasks or acquisition procedures. This aspect could help for various applications given many datasets with several hours of recording that remain unused due to their poor labels.

In addition to this general observation, Table 7.1 also shows that aconsidering the combination of all the pretext tasks instead of one in particular helps in the processing of the downstream task. This can be explained by the large spectrum of information covered by these tasks, the later analyzing the temporal trend of the signal, its contribution to the frequency spectrum and the spatial organization of the electrodes. However, a trend stands out for each task: the spectral task seems to provide the best results, followed by the temporal and the spatial task presenting the worst results (but still higher than the supervised learning standalone). This aspect can be explained by the fact that more information is located in the spectral evolution of the signal than in another property, i.e. temporal or spatial. This is explained by the fact that the considered datasets present tasks inducing oscillatory patterns in the EEG signals. Moreover, it can explain that extracting spatial information from EEGs remains a complex task due to the volume conductions effect.

Finally, another important aspect of this work is that it is generalizable for several tasks. As shown in Table 7.1, the proposed approach presents improvements when considering the downstream task for various paradigms. It can therefore be applied to multiple cases.

7.5 Conclusion

In this chapter, the notion of Self-Supervised Learning (SSL) has first been introduced. Then the advances presented in several DL techniques in recent years have been listed, after which the existing models considering SSL for EEG signals processing have been presented.

From the observations, SSL-EEG approaches to design models for EEG dimension reduction has been developed. Although this novel method is not able to replace all the handcrafted feature extraction approaches for EEG signals processing, it has shown encouraging results, and several insights can be made from this first study: 1) SSL promotes better results than a supervised learning standalone with raw signals; 2) Training a DL model with SSL method helps the model to better process the signals based on their properties; 3) SSL allows the transfer learning between datasets and tasks.

Conclusion

Assessing attention in various contexts with the help of physiological signals could have several benefits in various fields. For instance, to help in the diagnosis and symptoms assessment of attention-related disorders, for vigilance estimation during specific task (e.g. driving), in entertainment, but also in many other applications. Undoubtedly, being able to assess attention with the help of novel DL-based methods could allow for improvements in many people's lives.

In addition to the medical aspect, artificial intelligence has known a revolution during the 2010s with the innovations provided by DL. These innovations led to advances in various fields, including in signal processing and lately in biomedical signal processing.

It is in the above-described circumstances that this thesis was born. As initially defined, the research project aimed at applying AI for studying attention in VR with the help of physiological signals, including EEG and eye-tracking signals.

Concretely, it was possible to divide the project into four milestones:

- Designing VR environments promoting various levels of attention, e.g. focused or distracted.
- Acquiring a corpus of EEG and eye-tracking signals in VR during attention-related tasks.
- Characterizing attention state from physiological signals with the help of novel DL methods.
- Developing a proof of concept video-game reacting with attention state.

These milestones have taken place sequentially during the three years of this PhD from September 2019 to August 2022. As explained in the introduction, the initial goal was to study the attention state of specific subgroups of par-

ticipants, i.e. children between 8-12 years old with specific ADHD symptoms. However, due to COVID restrictions for enrolling children in or experiments, the thesis has been redirected to study a more general group of healthy participants aged between 18 and 30.

Contributions

As mentioned, four milestones have been noticed for the accomplishment of this thesis. In this context, the main contributions of this PhD work are related to the abeovementioned milsones and can be summerized as follows.

Virtual Environments Recording Attention (VERA) [147] - Although VR has been often used for research, most available VR environments for attention assessment consist in VR classroom setups. In our research, we have considered the creation of VR environments representing everyday life scenarios. Two neurophysiological tasks designed by specialists have been adapted to VR to assess attention. The resulting environments in this context are freely accessible online to promote research in the field[1].

Physiological Dataset Assessing Attention (PhyDAA) [49] - It exists a large variety of datasets providing EEG signals during various tasks. In this thesis, a corpus of EEG and eye-tracking signals have been registered for the the assessment of attentional state in VR. This novel corpus aims at solving the lack of existing works providing physiological signals recorded in VR during tasks promoting attention. Several advances are proposed by this novel corpus: physiological signals assigned to an attention score; a dataset recorded in VR; the joint measurement of EEGs with a biomedical recorder and eye-tracking signals. The signals recorded in VR for the 32 participants are freely available online (but subject to EULA)[2].

Attention estimation from EEG - four majors advances for DL-based analysis of EEG signals have been presented in this thesis [73, 148–150].

[1] https://github.com/VDelv/VERA
[2] https://zenodo.org/record/4558990

A novel approach to represent EEG segments by considering their frequency, time and space properties has been designed. An analysis of the most critical aspects of the signals has also been proposed.

Several DL models specially dedicated to EEG signals or their resulting feature maps have been proposed: Hierarchical-Recurrent Neural Network (H-RNN), combined H-RNN-CNN with saliency-based fusion [148] and Transformer [149].

Three novel approaches have also been proposed to fuse EEG and physiological signals. The existing approach processed these signals separately or considered naïve based methods to merge the information. Novel methods based on Compact Bilinear Pooling (CBP), Feature-wise Linear Modulation (FiLM) and Multi-Head Attention (MHA) have been developed.

Finally, experimentations using novel learning techniques specially dedicated to EEGs have been performed. This novel technique aims to learn the model how to process the signals instead of naïvely reproducing examples to make classification or regression. The latter is based on Self-Supervised Learning (SSL).

An application with attention assessment - a demo appplication in which attention assessment could be useful has been developed, with the help of the insights made for EEG signals processing. We hope that further works will be made on the prototype to deploy it.

The contributions mentioned above have led to several publications (see Appendix B) to promote the works in the scientific community and to confront our-self with peers.

Perspectives

From the contributions of this PhD, several further directions can be inferred. Among the existing further works, we propose five directions for future works:

- Study of attention for other groups: as previously mentioned, the initial goal of this thesis was to work with pathological participants. It would be interesting to consider attention estimation and study from physiological

signals of children with ADHD. Signal processing pipeline presented in this work can be kept.

- Transfering insights and knowledge to other paradigms: based on the findings made during this PhD, it could be interesting to study their implementation for different tasks or paradigms, for instance, by considering our DL approaches for the processing EEG recordings for other detection or classification of symptoms and diseases (e.g. Alzheimer epileptic seizure detection, or sleeping stage classification).
- Pursuing the analysis of SSL for EEGs: SSL being an increaslingly hot research topic, it could be interesting to do further works to test novel SSL tasks and/or adapt recent results for 1-D signal analysis [128].
- Validation of the proof of concept: from the prototype video game presented in this thesis, it could be interesting to consider a more extensive study to analyze its effect on participant attention state. The video game could be a great help for the regulation of attention loss.
- Considering a general DL pipeline for EEG processing: as it is the case for other modalities, e.g. BERT for text translation [99], or ResNet for image processing [26], it could be interesting to work on novel architectures or improve the ones proposed in this thesis to suit with EEG in various contexts, i.e. signal duration or the number of channels. This perspective aims to create a novel and general pipeline for the automatic analysis of EEG signals.

Appendix A

A Proof of Concept: Attention Rythm Video Game

To consider a concrete application for the research presented in this thesis, we envisaged developing a HCI interface reacting to the participant's attention. The goal of this application would be to propose a video game responding to the attentional state through rewards (resp. motivations) during high (resp. low) engagement.

In this chapter, the previous application evolving with attention state will be presented in Section A.2. Then the followed process to create our proof of concept will be presented in Section A.3. Finally, further works and perspectives related to this prototype will be proposed in Section A.4.

A.1 Introduction

Creating an application able to estimate and/or evolve with the participant's attention state is a very challenging task. An application able to measure attention state could help in various fields:

- Entertainment by designing tools that can be deployed for everyone and everywhere. These applications could help to drive a drone, for meditation or improve attention-related skills in a non-medical environment.
- Symptoms assessment & diagnosis of attention-related disorder with the help of physiological measurements. As seen in previous works, BCI have already been applied to assess and detect specific symptoms. It can be useful for the help of diagnosis or treatment of particular disorders.

- Marketing and Neuromarketing applications aiming to better target needs with the help of physiological signals. BCI has also been designed in this context with applications, for instance, predicting premature video skipping from EEG signals.

After working on the signal analysis pipeline aiming to estimate attention state from EEGs, developing an application to export and mobilize the knowledge gained during this thesis is proposed.

An essential point in the development of BCI is dedicated to neurofeedback. As previously explained, biofeedbacks are applications displaying physiological signals in a more understandable form, e.g. a bubble with its size evolving with the heart rate [151]. Neurofeedback is a specific category of biofeedbacks where the considered measurements represent the brain activity.

In the context of attention estimation from EEGs, the development of neurofeedback evolving with the attention state could have benefits. It can reduce symptoms of specific disorders without a daily intake of medication which is a stumbling block, especially for children.

Thus, for all of these motivations, we thought to consider developing an application able to assess and react to the participants' attentional state.

A.2 Related Work

This section presents some related works proposing a prototype to increase or assess attention. The common point of the methods is that they all consider EEG as physiological signals. The resulting BCI can be considered passive or active depending on whether the signals are processed online and if feedbacks are given after the processing of the task.

In their work, Kosmyna and Maes [36] propose AttentivU an active BCI to estimate attention from peripherical EEG. The attention assessment is based on an engagement score E computed as:

$$E = \frac{\beta}{\alpha + \theta} \tag{A.1}$$

with β, α and θ corresponding to the signal mean power spectral density in the chosen spectral bands as described in Figure 1.5. As a reminder, frequency bands are related to specific behaviour in EEGs, e.g. δ is increased during sleeping or *beta* during motor movements. During the experiment, the participants are asked to watch a video on several general subjects (DNA, Bitcoins and Neural Networks), and a stimulus (corresponding to vibration in a scarf) occurs during the loss of engagement. The study considers 48 participants split into three groups: 1) Feedback related to engagement score; 2) Random feedback; 3) No feedback. The study shows that participants who receive the correct feedback present a higher mean score on the questionnaire following the videos than other groups.

Zheng et al. [34] present a vigilance estimator during driving task based on joint EEG and EOG recording. The goal of the prototype is to record signals to retrieve an engagement score based on the PERCLOS index computed as:

$$PERCLOS = \frac{blink + CLOS}{blink + fixation + saccade + CLOS} \tag{A.2}$$

with $blink$, $CLOS$, $fixation$ and $saccade$ respectively representing the duration of blinking time, closed eyes, fixation and ocular saccades. At the time of writing this thesis, the authors have only presented a passive BCI with an outline analysis of the signals. Twenty participants took part in this experiment.

Pay Attention is another passive BCI proposed by Szafir and Mutlu [37]. This project aims to assess audience engagement during TEDx talks, more especially, to investigate the effect of speaker parameters on the score (e.g. gesture or volume). The goal is to tune speaker behaviours based on the audience's loss/gain of attention. The score is computed based on frequential band contribution as described in the AttentivU equation A.1.

Yan et al. [152] proposed a similar study to the one proposed for Pay Attention adapted for an art performance. The engagement is similarly computed with Equation A.1. The 3D theatre scene is adapted from this score to enhance the audience's engagement with dance movements, lighting effects or state machinery.

Another project proposed by Huang et al. [153] is FOCUS. This BCI aims to help children learn to read. During the reading task, virtual book parameters

evolved to reward attention gain and encourage during the drop in engagement. Similarly, the concentration is computed with the frequency band features as explained in Equation A.1.

Another application where the engagement score can be used is for drone pilot engagement as proposed by Pham et al. [154]. In their work, they present an active BCI displaying the engagement during the driving task. The goal of the application is to help the pilot to focus during the task with the feedback. A study was led with 10 participants into two groups with attention scores displayed or not. From the participants' point of view, the attention score display helps for the in-game performance, however, no more analysis has been made.

Finally, a BCI working with attention estimation was proposed by Libert and Van Hulle [155]. The passive BCI aims at predicting premature video skipping in advertising. Moreover, an emotion study is also performed to estimate the valence and arousal of the video. Four participants took part in the research.

Although the proposed works present promising innovations in various fields, they present two major drawbacks:

- The considered method to estimate attention state from EEGs is based on the engagement score formula proposed by Equation A.1 that may seem naïve. It has already been considered inefficient in various cases due to variation in the frequency limits or the public not being concerned by this method for attention estimation [18, 58, 109].
- The used in some of the proposed works of entertainment recorder can be affected by noise. Moreover, it has been reported that considering the pre-extracted features from this recorder may induce a bias in the experiment.

A.3 PoC Description

The idea of the proposed application is to create a video game reacting to the participant's attentional state. The general pipeline of this application is summarized in Figure A.1. The system is composed of different parts interacting with each other:

- The participant whose attention is assessed.
- The EEG recorder registers the signals during the task.
- The Biomedical Signal Analysis part takes the signals and estimates the attention state from the raw signals.
- The application or video game gives the score of that attention's related video games and feedback depending on the attention state estimated during the previous step. Moreover, distractors are also shown to evaluate participant inhibition.

For clarity, the following section will be separated into the major steps of the application.

A.3.1 Signal Acquisition

The first step concerns the acquisition of the EEG signals during the tasks. The signals are registered with a hybrid recorder presenting the advantages of entertainment recorders, i.e. mobility and ease of use while avoiding their drawbacks, i.e. low signal-noise ratio and artefacts appearances. The chosen recorder for this purpose is the Unicorn Hybrid Black.

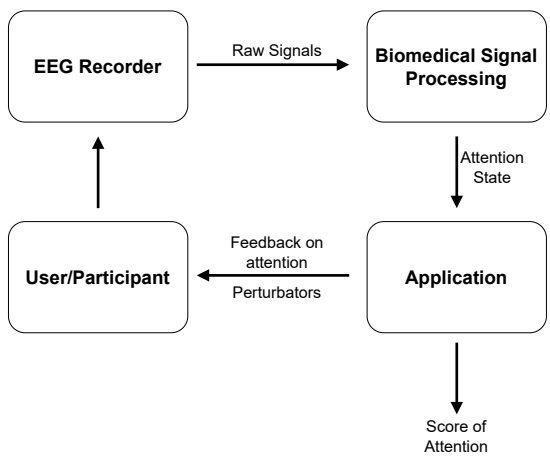


Figure A.1. Attention Rythm Pipeline

Due to synchronization issues and to break free from the constructor limitations, it has been considered not to use the API sold with the recorder. This software being effective only for this specific recorder model, it was envisaged to develop a more general solution.

The proposed solution is to stream the signals from EEG recorder through a local acquisition server, this server being made with OpenViBE software [156]. This methodology allows access to the signals whenever needed by listening at the predefined port while considering open-source software.

A.3.2 Signal Analysis

After the first step, the signal acquired during the task is directly accessible. It is extracted as arrays with the python interface of Lab Streaming Layer (LSL) with pylsl library [157]. The choice of this format is motivated by its ease of processing in python language.

After being able to read in real-time the signal stream, it is analyzed to estimate an attention state, i.e. focused vs distracted. During this step, the knowledge encountered in the previous chapters has been deployed.

First, the features have been extracted following the specific feature representation as explained in Chapter 4. With this method, the EEG raw arrays are represented in 3D matrices, with each dimension representing the specificity of the EEG signals, i.e. time, frequency and space. Then, after considering this approach for feature representation, DE is extracted for each element composing the array.

A transformer based methods to the ones presented in Chapter 5 has been considered for the representation and classification of EEGs. The method employed to estimate attention state from feature maps is based on the approach presenting the best results for attention estimation in Chapter 5.

Finally, with the abovementioned methodology, an attention score is computed for each EEG segment. This information is also streamed locally via UDP sockets on a specific port [158].

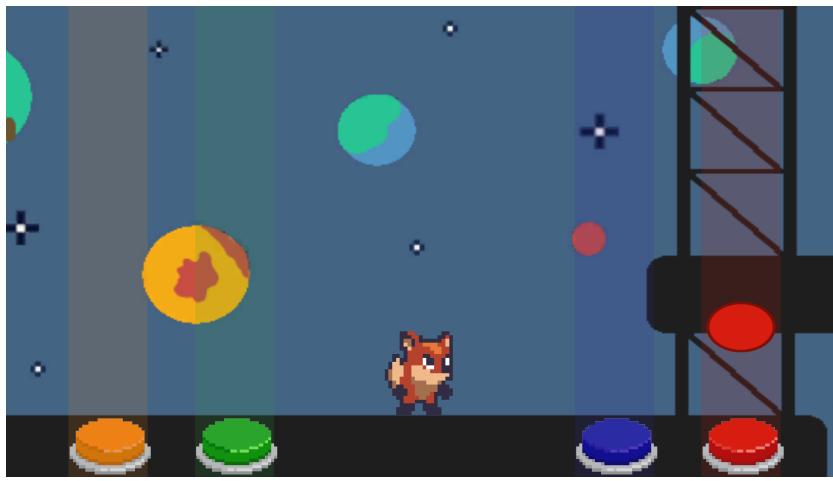


Figure A.2. Representation of the Attention Rythm Video-Game

A.3.3 Game evolution

In this context, an application assessing attention in a portable and easy-to-use manner has been considered. This application aims to provide a medium to evaluate attention loss and gain to help the participant better detect these specific behaviours.

The video game has been designed with Unity game engine and represents an environment in space. The game asks the participants to reproduce sequences by pushing specific keys on the keyboard. Four different configurations of keys have been set to let the participants place their hands at the location they feel the most comfortable in.

A representation of the game is given in Figure A.2 with the four different buttons represented in different colours. During the game, the number of keys correctly pressed is recorded jointly with the number of omissions and wrong input. The time taken to process the information has been measured, i.e. the time between the appearance of the tile and the button pressed. Moreover, to reduce the learning effect during the task, i.e. the participant's ability to play better during the game's evolution, the difficulty is progressively increased by decreasing the time between consecutive tiles occurrence. From all this

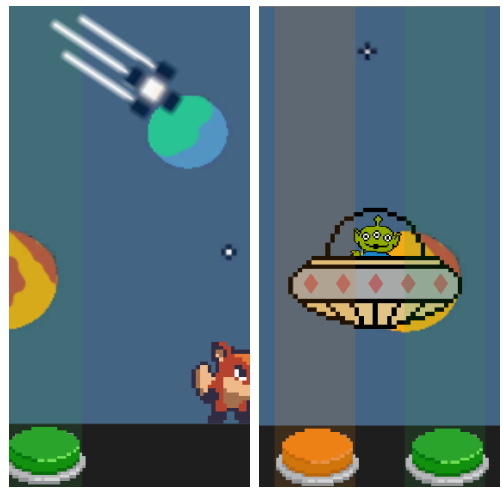
information, a game-related attention score is computed, giving a good insight into the degree of game completion.

In parallel to the task and tile appearance, it has been chosen to consider the appearance of perturbators as shown in Figure A.3. Two types of distractors have been considered, both appearing randomly during the game and being related to the environment. The first perturbator corresponds to a shooting star blinking and crossing the screen. The shooting star aims at assessing the inhibition during a short time distraction. The second perturbator is a ufo flying randomly and at a lower speed. It lands at a random location (out of the playing area) and remains there for 10 seconds. The second perturbator aims at assessing the inhibition during more persistent external stimuli. Considering these two external stimuli is motivated by the fact that it helps cover a larger span of attention and helps in recording attention loss compared to more repetitive tasks where the loss of engagement can be more challenging to detect.

The attention state computed with EEGs has not been considered yet. However, the aim of this project and, more generally this thesis was to use the physiological signals to help participants reduce the symptoms related to attention loss. This aspect could make this video game helpful to increase the participant's general attention state by training with the video game. Moreover, it has been proven that using physiological signals was helpful in this context [14, 18]. For these reasons, we decided to consider the attention observation from EEGs to corroborate this measurement.

The combination of information has been made by considering the attention state computed in the previous section. This attention state is sent from the python script to the Unity video game through UDP socket [158]. This attention state is then stored and registered several times to observe and study the attention trend as made in previous works [36, 147]. Finally, if a high (res. low) attention state is measured during a significant duration, the participant is considered attentive (resp. distracted), and thus he is rewarded (resp. helped). The rewards and help have not been designed yet but will be considered in further works.

The main idea of this prototype is to ensure that considering positive feedback could help improve the attention state.



(a) Star (b) UFO

Figure A.3. Video-Game Perturbators

A.4 Conclusion and Discussion

We present in this chapter an innovative video game to study and help improve attention state. Although the application remains a prototype, we wish to improve it to become an actual application usable by everyone and everywhere (especially by participants with attention-related disorders and/or symptoms at home).

In the future, we plan to conduct a study on a large public to investigate the effect of the feedback based on attention estimation from EEG measurement as it has been made for previous works [36]. This study will help validate the prototype by comparing the feedback from EEG recordings with random feedback or no feedback at all.

For the following years, we hope that EEG acquisition combined with novel technologies will significantly help people with disorders or attention-related symptoms.

Appendix B

Publications related to this thesis

B.1 Regular Papers in Journals

1. **Delvigne** V., Tits, N., La Fisca, L., Hubens, N., Maiorca, A., Wannous, H., Dutoit, T. and Vandeborre, J-P."*Where Is My Mind (Looking at)? A Study of the EEG–Visual Attention Relationship*", Informatics. Vol. 9. No. 1. MDPI, 2022.
2. **Delvigne** V., Wannous H., Dutoit, T., Ris, L. and Vandeborre, J-P. "*PhyDAA: Physiological Dataset Assessing Attention*", IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 5, pp. 2612-2623, 2022, doi: 10.1109/TCSVT.2021.3061719.
3. **Delvigne** V., Wannous H., Dutoit, T., Ris, L. and Vandeborre, J-P. "*Deep Learning Approaches to Multimodal EEG Signal Processing for Brain-Computer Interfaces*", IEEE Transactions on Affective Computing [under review]

B.2 Papers in International Conference with Peer Review

1. **Delvigne** V., Wannous H., Vandeborre, J.P., Ris L. and Dutoit T. "*Spatio-Temporal Analysis of Transformer based Architecture for Attention Estimation from EEG*", IEEE International Conference on Pattern Recognition (ICPR), Montreal, CA, 2022.
2. **Delvigne** V., Facchini A., Wannous, H., Dutoit T., Ris L. and Vandeborre, J.P. "*A Saliency based Feature Fusion Model for EEG Emotion Estimation*", IEEE International Engineering in Medicine and Biology Conference (EMBC), Glasgow, UK, 2022.
3. **Delvigne** V., Wannous H., Vandeborre J.P., Ris L. and Dutoit T. "*Attention Estimation in Virtual Reality with EEG based Image Regression*", IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), Utrecht, NL, 2020.

4. **Delvigne** V., Ris L., Dutoit T., Wannous, H. and Vandeborre, J.P. "*VERA: Virtual Environments Recording Attention*", IEEE International Conference on Serious Games and Applications for Health (SeGAH), Vancouver, CA, 2020.

B.3 Absracts in International Conference with Peer Review

1. **Delvigne** V. and Wannous H. "*Towards the use of self-supervised learning for EEG analysis*", Meeting in Signal and Image Processing in Neuroscience, GdR ISIS, Paris, FR, 2022.
2. **Delvigne** V., Ris L., Dutoit T., Wannous, H. and Vandeborre, J.P. "*An innovative Neurofeedback for children with ADHD using Virtual Reality*", HBP Student Conference on Interdisciplinary Brain Researc, Pisa, IT, 2020.

List of Figures

List of Tables

Bibliography

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Advances in neural information processing systems*, vol. 25, 2012.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge", *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[4] O. D. T. Catalá, I. S. Igual, F. J. Pérez-Benito, D. M. Escrivá, V. O. Castelló, R. Llobet, and J.-C. Peréz-Cortés, "Bias analysis on public x-ray image datasets of pneumonia and covid-19 patients", *IEEE Access*, vol. 9, pp. 42 370–42 383, 2021.

[5] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review", *Journal of neural engineering*, vol. 16, no. 5, p. 051001, 2019.

[6] D. American Psychiatric Association, A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association Washington, DC, 2013, vol. 5.

[7] E. G. Willcutt, "The prevalence of dsm-iv attention-deficit/hyperactivity disorder: a meta-analytic review", *Neurotherapeutics*, vol. 9, no. 3, pp. 490–499, 2012.

[8] S. R. Raman, K. K. Man, S. Bahmanyar, A. Berard, S. Bilder, T. Boukhris, G. Bushnell, S. Crystal, K. Furu, Y.-H. KaoYang *et al.*, "Trends in attention-deficit hyperactivity disorder medication use: a retrospective observational study using population-based databases", *The Lancet Psychiatry*, vol. 5, no. 10, pp. 824–835, 2018.

[9] H. Caci, M. Doepfner, P. Asherson, R. Donfrancesco, S. Faraone, A. Hervas, and M. Fitzgerald, "Daily life impairments associated with self-reported childhood/adolescent attention-deficit/hyperactivity disorder and experiences of diagnosis and treatment: Results from the european lifetime impairment survey", *European Psychiatry*, vol. 29, no. 5, pp. 316–323, 2014.

[10] M. Cueli, C. Rodríguez, P. Cabaleiro, T. García, and P. González-Castro, "Differential efficacy of neurofeedback in children with adhd presentations", *Journal of clinical medicine*, vol. 8, no. 2, p. 204, 2019.

[11] A. Zilverstand, B. Sorger, P. Sarkheil, and R. Goebel, "fmri neurofeedback facilitates anxiety regulation in females with spider phobia", *Frontiers in behavioral neuroscience*, p. 148, 2015.

[12] R. Coben, M. Linden, and T. E. Myers, "Neurofeedback for autistic spectrum disorder: a review of the literature", *Applied psychophysiology and biofeedback*, vol. 35, no. 1, pp. 83–105, 2010.

[13] E. A. Schoneveld, M. Malmberg, A. Lichtwarck-Aschoff, G. P. Verheijen, R. C. Engels, and I. Granic, "A neurofeedback video game (mindlight) to prevent anxiety in children: A randomized controlled trial", *Computers in Human Behavior*, vol. 63, pp. 321–333, 2016.

[14] F. Blume, J. Hudak, T. Dresler, A.-C. Ehlis, J. Kühnhausen, T. J. Ren-

ner, and C. Gawrilow, "Nirs-based neurofeedback training in a virtual reality classroom for children with attention-deficit/hyperactivity disorder: study protocol for a randomized controlled trial", *Trials*, vol. 18, no. 1, pp. 1–16, 2017.

[15] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review", *sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.

[16] K. Kurzhals, B. Fisher, M. Burch, and D. Weiskopf, "Evaluating visual analytics with eye tracking", in *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, 2014, pp. 61–69.

[17] F. Lotte, "A tutorial on eeg signal-processing techniques for mental-state recognition in brain–computer interfaces", *Guide to brain-computer music interfacing*, pp. 133–161, 2014.

[18] O. Bazanova and L. Aftanas, "Individual eeg alpha activity analysis for enhancement neurofeedback efficiency: two case studies", *Journal of Neurotherapy*, vol. 14, no. 3, pp. 244–253, 2010.

[19] C. Pernet, M. I. Garrido, A. Gramfort, N. Maurits, C. M. Michel, E. Pang, R. Salmelin, J. M. Schoffelen, P. A. Valdes-Sosa, and A. Puce, "Issues and recommendations from the OHBM COBIDAS MEEG committee for reproducible EEG and MEG research", *Nature Neuroscience*, vol. 23, no. 12, pp. 1473–1483, 2020.

[20] R. A. Fisher, "The use of multiple measurements in taxonomic problems", *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation", California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches", *arXiv preprint arXiv:1409.1259*, 2014.

[24] A. LeNail, "Nn-svg: Publication-ready neural network architecture schematics." *J. Open Source Softw.*, vol. 4, no. 33, p. 747, 2019.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] D. G. Dillon and D. A. Pizzagalli, "Inhibition of action, thought, and emotion: a selective neurobiological review", *Applied and Preventive Psychology*, vol. 12, no. 3, pp. 99–114, 2007.

[28] C. K. Conners, M. Staff, V. Connelly, S. Campbell, M. MacLean, and J. Barnes, "Conners' continuous performance test ii (cpt ii v. 5)", *Multi-Health Syst Inc*, vol. 29, pp. 175–96, 2000.

[29] J. C. Mullane and P. V. Corkum, "The relationship between working memory, inhibition, and performance on the wisconsin card sorting test in children with and without adhd", *Journal of Psychoeducational Assessment*, vol. 25, no. 3, pp. 211–221, 2007.

[30] A. Bashiri, M. Ghazisaeedi, and L. Shahmoradi, "The opportunities of virtual reality in the rehabilitation of children with attention deficit hyperactivity disorder: a literature review", *Korean journal of pediatrics*, vol. 60, no. 11, p. 337, 2017.

[31] Y. Pollak, P. L. Weiss, A. A. Rizzo, M. Weizer, L. Shriki, R. S. Shalev, and V. Gross-Tsur, "The utility of a continuous performance test embedded in virtual reality in measuring adhd-related deficits", *Journal of Developmental & Behavioral Pediatrics*, vol. 30, no. 1, pp. 2–6, 2009.

[32] Y. Tan, D. Zhu, H. Gao, T.-W. Lin, H.-K. Wu, S.-C. Yeh, and T.-Y. Hsu, "Virtual classroom: An adhd assessment and diagnosis system based on virtual reality", in *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*. IEEE, 2019, pp. 203–208.

[33] H. Eom, K. Kim, S. Lee, Y.-J. Hong, J. Heo, J.-J. Kim, and E. Kim, "Development of virtual reality continuous performance test utilizing social cues for children and adolescents with attention-deficit/hyperactivity disorder", *Cyberpsychology, Behavior, and Social Networking*, vol. 22, no. 3, pp. 198–204, 2019.

[34] W.-L. Zheng and B.-L. Lu, "A multimodal approach to estimating vigilance using eeg and forehead eog", *Journal of neural engineering*, vol. 14, no. 2, p. 026017, 2017.

[35] Z. Cao, C.-H. Chuang, J.-K. King, and C.-T. Lin, "Multi-channel eeg recordings during a sustained-attention driving task", *Scientific data*, vol. 6, no. 1, pp. 1–8, 2019.

[36] N. Kosmyna and P. Maes, "Attentivu: an eeg-based closed-loop biofeedback system for real-time monitoring and improvement of engagement for personalized learning", *Sensors*, vol. 19, no. 23, p. 5200, 2019.

[37] D. Szafir and B. Mutlu, "Pay attention! designing adaptive agents that monitor and improve user engagement", in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2012, pp. 11–20.

[38] A. García-Baos, D. Tomas, I. Oliveira, P. Collins, C. Echevarria, L. P. Zapata, E. Liddle, H. Super *et al.*, "Novel interactive eye-tracking game for training attention in children with attention-deficit/hyperactivity disorder", *The primary care companion for CNS disorders*, vol. 21, no. 4, pp. 0–0, 2019.

[39] P. Varela Casal, F. Lorena Esposito, I. Morata Martínez, A. Capdevila, M. Solé Puig, N. de la Osa, L. Ezpeleta, A. Perera i Lluna, S. V. Faraone,

J. A. Ramos-Quiroga *et al.*, "Clinical validation of eye vergence as an objective marker for diagnosis of adhd in children", *Journal of attention disorders*, vol. 23, no. 6, pp. 599–614, 2019.

[40] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "Mped: A multi-modal physiological emotion database for discrete emotion recognition", *IEEE Access*, vol. 7, pp. 12 177–12 191, 2019.

[41] D. F. Dinges and R. Grace, "Perclos: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance", *US Department of Transportation, Federal Highway Administration, Publication Number FHWA-MCRT-98-006*, 1998.

[42] A. Harati, S. Lopez, I. Obeid, J. Picone, M. Jacobson, and S. Tobochnik, "The tuh eeg corpus: A big data resource for automated eeg interpretation", in *2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE, 2014, pp. 1–5.

[43] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals", *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[44] 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, 2011.

[45] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks", *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[46] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter:

A multimodal framework for recognizing human emotions", *IEEE transactions on cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2018.

[47] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "Bci competition 2008–graz data set a", *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, vol. 16, pp. 1–6, 2008.

[48] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg", *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.

[49] V. Delvigne, H. Wannous, T. Dutoit, L. Ris, and J.-P. Vandeborre, "PhyDAA: Physiological Dataset Assessing Attention", *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, iEEE Transactions on Circuits and Systems for Video Technology.

[50] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy", *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[51] S. Davis, K. Nesbitt, and E. Nalivaiko, "A systematic review of cybersickness", in *Proceedings of the 2014 conference on interactive entertainment*, 2014, pp. 1–9.

[52] R. Oostenveld and P. Praamstra, "The five percent electrode system for high-resolution eeg and erp measurements", *Clinical neurophysiology*, vol. 112, no. 4, pp. 713–719, 2001.

[53] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces", *Journal of Neural Engineering*, vol. 15, no. 5, 2018, iOP Publishing.

[54] A. Delorme and S. Makeig, "Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis", *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.

[55] N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, and N. Langer, "Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading", *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.

[56] L.-W. Ko, O. Komarov, W. D. Hairston, T.-P. Jung, and C.-T. Lin, "Sustained attention in real classroom settings: An eeg study", *Frontiers in human neuroscience*, vol. 11, p. 388, 2017.

[57] S. Bioulac, D. Purper-Ouakil, T. Ros, H. Blasco-Fontecilla, M. Prats, L. Mayaud, and D. Brandeis, "Personalized at-home neurofeedback compared with long-acting methylphenidate in an european non-inferiority randomized trial in children with adhd", *BMC psychiatry*, vol. 19, no. 1, pp. 1–13, 2019.

[58] O. M. Bazanova, T. Auer, and E. A. Sapina, "On the efficiency of individualized theta/beta ratio neurofeedback combined with forehead emg training in adhd children", *Frontiers in human neuroscience*, vol. 12, p. 3, 2018.

[59] Y. Li, L. Wang, W. Zheng, Y. Zong, L. Qi, Z. Cui, T. Zhang, and T. Song, "A Novel Bi-hemispheric Discrepancy Model for EEG Emotion Recognition", *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[60] B. P. Harne, "Higuchi fractal dimension analysis of eeg signal before and after om chanting to observe overall effect on brain", *International*

Journal of Electrical and Computer Engineering, vol. 4, no. 4, p. 585, 2014.

[61] R. Baravalle, O. A. Rosso, and F. Montani, "Causal shannon–fisher characterization of motor/imagery movements in eeg", *Entropy*, vol. 20, no. 9, p. 660, 2018.

[62] M. Martin, F. Pennini, and A. Plastino, "Fisher's information and the analysis of complex signals", *Physics Letters A*, vol. 256, no. 2-3, pp. 173–180, 1999.

[63] F. S. Bao, X. Liu, and C. Zhang, "Pyeeg: an open source python module for eeg/meg feature extraction", *Computational intelligence and neuroscience*, vol. 2011, 2011.

[64] Z. Mardi, S. N. M. Ashtiani, and M. Mikaili, "Eeg-based drowsiness detection for safe driving using chaotic features and statistical tests", *Journal of medical signals and sensors*, vol. 1, no. 2, p. 130, 2011.

[65] R. Hamila, J. Astola, F. A. Cheikh, M. Gabbouj, and M. Renfors, "Teager energy and the ambiguity function", *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 260–262, 1999.

[66] A. Erdamar, F. Duman, and S. Yetkin, "A wavelet and teager energy operator based method for automatic detection of k-complex in sleep eeg", *Expert Systems with Applications*, vol. 39, no. 1, pp. 1284–1290, 2012.

[67] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, "Optimizing spatial filters for robust eeg single-trial analysis", *IEEE Signal processing magazine*, vol. 25, no. 1, pp. 41–56, 2007.

[68] M. J. Shensa *et al.*, "The discrete wavelet transform: wedding the a trous and mallat algorithms", *IEEE Transactions on signal processing*, vol. 40, no. 10, pp. 2464–2482, 1992.

[69] G. Rilling, P. Flandrin, P. Goncalves *et al.*, "On empirical mode decomposition and its algorithms", in *IEEE-EURASIP workshop on nonlinear signal and image processing*, vol. 3. Citeseer, 2003, pp. 8–11.

[70] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory", *Physica D: Nonlinear Phenomena*, vol. 31, no. 2, pp. 277–283, 1988.

[71] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, and J. Wang, "SST-EmotionNet: Spatial-Spectral-Temporal Based Attention 3D Dense Network for EEG Emotion Recognition", in *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, 2020, pp. 2909–2917.

[72] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks", in *4th International Conference on Learning Representations, ICLR*, 2016.

[73] V. Delvigne, H. Wannous, J.-P. Vandeborre, L. Ris, and T. Dutoit, "Attention estimation in virtual reality with eeg based image regression", in *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2020, pp. 10–16.

[74] P. Hauri and D. R. Hawkins, "Alpha-delta sleep", *Electroencephalography and Clinical Neurophysiology*, vol. 34, no. 3, pp. 233–237, 1973.

[75] C. W. Anderson and Z. Sijercic, "Classification of eeg signals from four subjects during five mental tasks", in *Solving engineering problems with neural networks: proceedings of the conference on engineering applications in neural networks (EANN'96)*. Turkey, 1996, pp. 407–414.

[76] E. Haselsteiner and G. Pfurtscheller, "Using time-dependent neural networks for eeg classification", *IEEE transactions on rehabilitation engineering*, vol. 8, no. 4, pp. 457–463, 2000.

[77] R. Palaniappan, "Brain computer interface design using band powers extracted during mental tasks", in *Conference Proceedings. 2nd International IEEE EMBS Conference on Neural Engineering, 2005.* IEEE, 2005, pp. 321–324.

[78] B. Zou, M. Shen, X. Li, Y. Zheng, and L. Zhang, "Eeg-based driving fatigue detection during operating the steering wheel data section", in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).* IEEE, 2020, pp. 248–251.

[79] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, "Mlp-mixer: An all-mlp architecture for vision", *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[80] Y. Yuan, K. Jia, F. Ma, G. Xun, Y. Wang, L. Su, and A. Zhang, "A hybrid self-attention deep learning framework for multivariate sleep stage classification", *BMC Bioinformatics*, vol. 20, no. 16, p. 586, 2019.

[81] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization", *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[82] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, 2018.

[83] N. Zhang, W.-L. Zheng, W. Liu, and B.-L. Lu, "Continuous vigilance estimation using lstm neural networks", in *International Conference on Neural Information Processing.* Springer, 2016, pp. 530–537.

[84] G. Zhang and A. Etemad, "Capsule attention for multimodal eeg-eog representation learning with application to driver vigilance estimation", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1138–1149, 2021.

[85] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber *et al.*, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies", 2001.

[86] W. Wu, W. Sun, Q. J. Wu, Y. Yang, H. Zhang, W.-L. Zheng, and B.-L. Lu, "Multimodal vigilance estimation using deep learning", *IEEE Transactions on Cybernetics*, 2020.

[87] T. Wen and Z. Zhang, "Deep Convolution Neural Network and Autoencoders-Based Unsupervised Feature Learning of EEG Signals", *IEEE Access*, vol. 6, pp. 25 399–25 410, 2018, iEEE Access.

[88] G. Honke, I. Higgins, N. Thigpen, V. Miskovic, K. Link, S. Duan, P. Gupta, J. Klawohn, and G. Hajcak, "Representation learning for improved interpretability and classification accuracy of clinical factors from eeg", *arXiv preprint arXiv:2010.15274*, 2020.

[89] P. Zhong, D. Wang, and C. Miao, "Eeg-based emotion recognition using regularized graph neural networks", *IEEE Transactions on Affective Computing*, 2020.

[90] G. Cisotto, A. Zanga, J. Chlebus, I. Zoppis, S. Manzoni, and U. Markowska-Kaczmar, "Comparison of attention-based deep learning models for eeg classification", *arXiv preprint arXiv:2012.01074*, 2020.

[91] A. Demir, T. Koike-Akino, Y. Wang, M. Haruna, and D. Erdogmus, "Eeg-gnn: Graph neural networks for classification of electroencephalogram (eeg) signals", in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 1061–1067.

[92] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks", *arXiv preprint arXiv:1701.00160*, 2016.

[93] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, and M. Shah, "Generative adversarial networks conditioned by brain signals", in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3410–3418.

[94] P. Tirupattur, Y. S. Rawat, C. Spampinato, and M. Shah, "Thoughtviz: Visualizing human thoughts using generative adversarial network", in *Proceedings of the 26th ACM international conference on multimedia*, 2018, pp. 950–958.

[95] R. Li, J. S. Johansen, H. Ahmed, T. V. Ilyevsky, R. B. Wilbur, H. M. Bharadwaj, and J. M. Siskind, "The perils and pitfalls of block design for eeg classification experiments", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 316–333, 2020.

[96] C. Brett, "8 crazy dreams for neuralink", 2020.

[97] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[98] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need", *Advances in neural information processing systems*, vol. 30, 2017.

[99] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186.

[100] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis,

L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach", *arXiv preprint arXiv:1907.11692*, 2019.

[101] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale", *arXiv preprint arXiv:2010.11929*, 2020.

[102] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations", *arXiv preprint arXiv:2006.11477*, 2020.

[103] R. Hu, A. Singh, T. Darrell, and M. Rohrbach, "Iterative answer prediction with pointer-augmented multimodal transformers for textvqa", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9992–10 002.

[104] D. Merkx and S. L. Frank, "Comparing Transformers and RNNs on predicting human sentence processing data", *arXiv:2005.09471 [cs]*, 2020.

[105] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics", in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1153–1160.

[106] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols", in *Proceedings of the 2000 symposium on Eye tracking research & applications*, 2000, pp. 71–78.

[107] Y. Liu, Z. Lan, J. Cui, O. Sourina, and W. Müller-Wittig, "EEG-Based Cross-Subject Mental Fatigue Recognition", in *2019 International Conference on Cyberworlds (CW)*, 2019, pp. 247–252, iSSN: 2642-3596.

[108] ——, "Inter-subject transfer learning for EEG-based mental fatigue recognition", *Advanced Engineering Informatics*, vol. 46, 2020.

[109] M. Arns, C. K. Conners, and H. C. Kraemer, "A decade of EEG Theta/Beta Ratio Research in ADHD: a meta-analysis", *Journal of Attention Disorders*, vol. 17, no. 5, pp. 374–383, Jul. 2013.

[110] W.-C. L. Lew, D. Wang, K. Shylouskaya, Z. Zhang, J.-H. Lim, K. K. Ang, and A.-H. Tan, "Eeg-based emotion recognition using spatial-temporal representation via bi-gru", in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 116–119.

[111] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics", in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1153–1160.

[112] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images", *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[113] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 740–757, 2018.

[114] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion", *IEEE Access*, vol. 8, pp. 168 865–168 878, 2020.

[115] H. Li, W.-L. Zheng, and B.-L. Lu, "Multimodal vigilance estimation with adversarial domain adaptation networks", in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–6.

[116] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning", in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[117] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules", *Advances in neural information processing systems*, vol. 30, 2017.

[118] L.-H. Du, W. Liu, W.-L. Zheng, and B.-L. Lu, "Detecting driving fatigue with multimodal deep learning", in *2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2017, pp. 74–77.

[119] A. B. Said, A. Mohamed, T. Elfouly, K. Harras, and Z. J. Wang, "Multimodal deep learning approach for joint eeg-emg data compression and classification", in *2017 IEEE wireless communications and networking conference (WCNC)*. IEEE, 2017, pp. 1–6.

[120] X. Zhang, J. Liu, J. Shen, S. Li, K. Hou, B. Hu, J. Gao, and T. Zhang, "Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine", *IEEE transactions on cybernetics*, vol. 51, no. 9, pp. 4386–4399, 2020.

[121] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep canonical correlation analysis", *arXiv preprint arXiv:1908.05349*, 2019.

[122] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and eeg to enhance emotion recognition", in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[123] M. Karnati, A. Seal, A. Yazidi, and O. Krejcar, "Lienet: A deep convolution neural networks framework for detecting deception", *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[124] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 317–326.

[125] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps", in *Proceedings of the 19th ACM SIGKDD international*

conference on Knowledge discovery and data mining, 2013, pp. 239–247.

[126] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[127] M. Brousmiche, J. Rouat, and S. Dupont, "Audio-visual fusion and conditioning with neural networks for event recognition", in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2019, pp. 1–6.

[128] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units", *arXiv preprint arXiv:2106.07447*, 2021.

[129] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language", *arXiv preprint arXiv:2202.03555*, 2022.

[130] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7077–7087.

[131] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis", *arXiv preprint arXiv:2006.15955*, 2020.

[132] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition", in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3507–3511.

[133] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-

based recurrent neural network", *IEEE Access*, vol. 8, pp. 61 672–61 686, 2020.

[134] J. Chen, B. Hu, Y. Wang, P. Moore, Y. Dai, L. Feng, and Z. Ding, "Subject-independent emotion recognition based on physiological signals: a three-stage decision method", *BMC medical informatics and decision making*, vol. 17, no. 3, pp. 45–57, 2017.

[135] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness", *arXiv preprint arXiv:1811.12231*, 2018.

[136] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles", in *European conference on computer vision*. Springer, 2016, pp. 69–84.

[137] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1920–1929.

[138] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 132–149.

[139] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations", in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[140] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments", *arXiv preprint arXiv:2006.09882*, 2020.

[141] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2495–2504.

[142] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort, "Uncovering the structure of clinical eeg signals with self-supervised learning", *Journal of Neural Engineering*, vol. 18, no. 4, p. 046020, 2021.

[143] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances", *Neural computation*, vol. 14, no. 4, pp. 715–770, 2002.

[144] I. Obeid and J. Picone, "The temple university hospital eeg data corpus", *Frontiers in neuroscience*, vol. 10, p. 196, 2016.

[145] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data", *arXiv preprint arXiv:2101.12037*, 2021.

[146] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system", *IEEE Transactions on biomedical engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[147] V. Delvigne, L. Ris, T. Dutoit, H. Wannous, and J.-P. Vandeborre, "Vera: Virtual environments recording attention", in *2020 IEEE 8th International Conference on Serious Games and Applications for Health (SeGAH)*. IEEE, 2020, pp. 1–7.

[148] V. Delvigne, A. Facchini, H. Wannous, T. Dutoit, L. Ris, and J.-P. Vandeborre, "Emotion estimation from eeg–a dual deep learning approach combined with saliency", in *2020 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 116–119.

[149] V. Delvigne, H. Wannous, J.-P. Vandeborre, L. Ris, and T. Dutoit, "Spatio-temporal analysis of transformer based architecture for attention estimation from eeg", *arXiv preprint arXiv:2204.07162*, 2022.

[150] V. Delvigne, N. Tits, L. La Fisca, N. Hubens, A. Maiorca, H. Wannous,

T. Dutoit, and J.-P. Vandeborre, "Where is my mind (looking at)? a study of the eeg–visual attention relationship", in *Informatics*, vol. 9, no. 1. MDPI, 2022, p. 26.

[151] Apple, "Breathe - apple store", 2022.

[152] S. Yan, G. Ding, H. Li, N. Sun, Y. Wu, Z. Guan, L. Zhang, and T. Huang, "Enhancing audience engagement in performing arts through an adaptive virtual environment with a brain-computer interface", in *Proceedings of the 21st international conference on intelligent user interfaces*, 2016, pp. 306–316.

[153] J. Huang, C. Yu, Y. Wang, Y. Zhao, S. Liu, C. Mo, J. Liu, L. Zhang, and Y. Shi, "Focus: enhancing children's engagement in reading by using contextual bci training sessions", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 1905–1908.

[154] T. Pham, D. Tezza, and M. Andujar, "Enhancing drone pilots' engagement through a brain-computer interface", in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 706–718.

[155] A. Libert and M. M. Van Hulle, "Predicting premature video skipping and viewer interest from eeg recordings", *Entropy*, vol. 21, no. 10, p. 1014, 2019.

[156] Y. Renard, F. Lotte, G. Gibert, M. Congedo, E. Maby, V. Delannoy, O. Bertrand, and A. Lécuyer, "Openvibe: An open-source software platform to design, test, and use brain–computer interfaces in real and virtual environments", *Presence*, vol. 19, no. 1, pp. 35–53, 2010.

[157] C. Kothe, "pylsl", Jun. 2022, original-date: 2015-04-25T04:00:38Z. [Online]. Available: https://github.com/chkothe/pylsl

[158] Siliconifier, "Siliconifier/Python-Unity-Socket-Communication", Jun. 2022, original-date: 2020-12-25T02:59:15Z. [Online]. Available: https://github.com/Siliconifier/Python-Unity-Socket-Communication

This thesis was made using a customized version of “hepthesis”.