Quantified Analysis for Video Recordings of Seizure

Jen-Cheng Hou

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Analyse Quantifiée des Vidéos de Crises Epileptiques

Jen-Cheng HOU

INRIA Sophia Antipolis, STARS

Présentée en vue de l'obtention du grade de docteur en Informatique **d'**Université Côte d'Azur
Dirigée par : Monique THONNAT
Co-dirigée par : Fabrice BARTOLOMEI
Co-encadrée par : Aileen MCGONIGAL
Soutenue le : 13 Décembre 2021

Devant le jury, composé de :
Président du jury :
Philippe ROBERT, Université Côte d'Azur
Rapporteurs :
Pau-Choo CHUNG, Université Nationale Cheng Kung
Séverine DUBUISSON, Aix-Marseille Université
Examinateur :
Aileen MCGONIGAL, Aix-Marseille Université
Fabrice BARTOLOMEI, Aix-Marseille Université
Philippe KAHANE, Université Grenoble Alpes
Monique THONNAT, INRIA

Inria

Analyse Quantifiée des Vidéos de Crises Epileptiques

Quantified Analysis for Video Recordings of Seizure

Jury :

Président du jury :

Philippe ROBERT	- PU-PH, Université Côté d'Azur, France

Rapporteurs :

Pau-Choo CHUNG	- Professeur, Université Nationale Cheng Kung, Taïwan
Séverine DUBUISSON	- MdC, HDR, Aix-Marseille Université, France

Examinateurs :

Aileen MCGONIGAL	- MD, HDR, Aix-Marseille Université, France
Fabrice BARTOLOMEI	- PU-PH, Aix-Marseille Université, France
Monique THONNAT	- DR0, Inria Sophia Antipolis, France
Philippe KAHANE	- PU-PH, Université Grenoble Alpes, France

ANALYSE QUANTIFIÉE DES VIDÉOS DE CRISES EPILEPTIQUES

Jen-Cheng Hou
Directeur de thèse: Monique Thonnat
Co-Directeur de thèse: Fabrice Bartolomei
STARS, Inria Sophia Antipolis, France

RÉSUMÉ

L'épilepsie est un trouble neurologique causé par une activité neuronale anormale dans le cerveau. Environ 1% de la population mondiale en est affectée. De nombreuses manifestations motrices (incluant des convulsions, des modifiations toniques, cloniques, hyperkinétiques) peuvent s'observer et sont une source de handicap majeur pour les patients. La motivation de cette recherche est de développer des méthodes basées sur des techniques récentes d'apprentissage automatique pour fournir une analyse objective des vidéos de crises cliniques.

Dans cette thèse, nous proposons trois contributions principales á l'analyse automatisée des vidéos de crises. Dans la première contribution, nous explorons des crises d'épilepsie hyperkinétiques en analysant les trajectoires du mouvement de la tête des patients. Les résultats fournissent une base pour étudier la corrélation entre le modifications spectrales de l'EEG et la fréquence des mouvements de la tête.

Néanmoins, l'épilepsie n'est pas la seule cause qui donne lieu á des crises. Par exemple, les crises psychogènes non épileptiques (PNES) en font partie. Ce sont des événements ressemblant á une crise d'épilepsie (ES), mais sans les décharges électriques caractéristiques associées á l'épilepsie. Bien les distinguer est donc important pour un diagnostic précis et des traitements de suivi. Les signes cliniques ou sémiologiques, sont évalués par les neurologues, mais leur interprétation subjective est susceptible de variabilité inter-observateur. Par conséquent, il est urgent de créer un système automatisé pour analyser les vidéos de crises. Dans cette recherche, nous proposons deux autres contributions pour classer ES et PNES uniquement sur la base des vidéos. Notre deuxième contribution utilise des informations issues de l'apparence et de points clés du corps et du visage des patients. En introduisant aussi un mécanisme de distillation des connaissances, les performances du score F1 et la précision sont de 0,85 et 0,82.

Puis sur la base de cette approche, nous menons une expérience parallèle pour distinguer ES avec émotion/non-émotion et dystonie/non-dystonie en fonction des composantes visage ou corps de la méthode. La validation LOSO donne des résultats satisfaisants, indiquant que notre modèle peut capturer des caractéristiques spatio-temporelles efficaces pour le visage et le corps pour l'analyse des crises. Dans notre troisième contribution, nous proposons un modèle en deux étapes qui est d'abord pré-entraîné sur de grandes vidéos contextuelles puis ce modèle est affiné pour la classification des types de crises.

Le modèle est basé sur l'encodeur du modèle Transformer. Étant donné qu'il est coûteux d'obtenir de grandes bases de données étiquetés par des médecins, nous cherchons á exploiter des données volumineuses non étiquetées pour initialiser les poids du modèle, puis le modèle est affiné sur la tâche cible en aval. Ce modèle traite uniquement les caractéristiques d'apparence par contre il implique plus de cas que ceux de la première étude. Le score F1 et la précision de la validation LOSO sont de 0,82 et 0,75. Grâce aux résultats très encourageants de cette recherche, nous proposons une base pour une direction de recherche prometteuse dans le domaine de l'analyse vidéo automatisée des crises.

Mots clés: Analyse vidéo de crise, apprentissage profond, apprentissage auto-supervisé

QUANTIFIED ANALYSIS FOR VIDEO RECORDINGS OF SEIZURE

by
Jen-Cheng Hou
Supervisor: Monique Thonnat
Co-Supervisor: Fabrice Bartolomei
STARS, Inria Sophia Antipolis, France

ABSTRACT

Epilepsy is a neurological disorder caused by abnormal neuron activity in the brain. Around 1% of the population worldwide is affected by it. Numerous motor manifestations (including convulsions, tonic, clonic, hyperkinetic changes) can be observed and are a source of major disability for patients. The motivation of this research is to develop methods based on recent machine learning techniques to provide objective analysis for clinical seizure videos.

In this thesis, we propose three main contributions towards automated vision-based seizure analysis. In the first contribution, we explore some hyperkinetic epileptic seizures by analyzing the head movement trajectories of the patients. The results provide a basis for studying the correlation between the spectrum of EEG and the head movement frequency. Nevertheless, epilepsy is not the only cause that gives rise to a seizure event. For example, psychogenic non-epileptic seizures (PNES) are one of them. They are events resembling an epileptic seizure (ES), but without the characteristic electrical discharges associated with epilepsy. How to distinguish them is important for accurate diagnosis and follow-up treatments. The clinical signs or semiology are evaluated by neurologists, but the subjective interpretation is liable for inter-observer variability. Hence, there is an urgent need to build an automated system to analyze seizure videos with the latest computer vision progress. In this research, we propose two other contributions for classifying ES and PNES solely based on the videos. Our second contribution utilizes multi-stream information from appearance and key-points for both the bodies and faces of the patients. In addition by introducing the knowledge distillation mechanism, the performance of the F1-score and the accuracy are 0.85 and 0.82. Furthermore, based on this approach, we conduct a side experiment for distinguishing ES with emotion/non-emotion and dystonia/non-dystonia based on the face and body streams in the method. The LOSO validation gives satisfactory results, indicating our model can capture effective spatio-temporal features for face and body for seizure analysis. In our third contribution, we propose a two-step model which is first pre-trained on large contextual videos then this model is fine-tuned for seizure type classification.

The model is based on the encoder of the Transformer model. Given that it is expensive to get large datasets labeled by doctors, we try to leverage large unlabeled data for a good weight initialization point for the model, and then fine-tune it on the target downstream task. This model only processes the appearance features but more cases than those in the first study are involved. The F1-score and accuracy of the LOSO validation are 0.82 and 0.75. With the very encouraging results in this research, we demonstrate a basis for a promising research direction in the field of automated seizure video analysis.

Keywords: seizure video analysis, deep learning, self-supervised learning

ACKNOWLEDGMENTS

I would like to thank Prof. Pau-Choo Chung and Prof. Severine Dubuisson for accepting to review my PhD thesis. I want to thank them for their very constructive advices and remarks. I also want to thank Prof. Philippe Kahane for the participation of my jury and Prof. Philippe Robert for accepting the charge as the president of my jury.

I would like to express my gratitude to my two thesis supervisors, Dr. Monique Thonnat and Prof. Fabrice Bartolomei. I want to thank them both for offering me this topic and for having the trust on me during the thesis. I am grateful for all the discussions, questions, advice, numerous and constructive criticisms from them, and their moral support in difficult times. In addition, I am grateful to Dr. Aileen McGonigal for the help in collecting video recording data and useful advices on how to improve my academic writing.

I also thank the entire STARS team (present and past) for having made these three years an unforgettable experience in an always stimulating and joyful atmosphere. I would particularly like to thank Francois Bremond for very constructive discussions. I want to thank Sandrine Boute, the project assistant who always provides kind help for all the members in the team.

Finally, I am thankful to my family and friends in Taiwan for supporting me spiritually throughout my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Epilepsy is one of the most prevalent neurological disorders, affecting nearly 1% of the population worldwide. It is characterized by recurrent seizures, which are caused by abnormal, excessive neuronal activity in the brain [1]. Globally, there are estimated five million people being diagnosed with epilepsy each year. People with epilepsy often experience negative impacts on their quality of life, such as less mobility, social interactions, learning or school attendance. Thus, how to perform an effective diagnosis of epilepsy and its monitoring are crucial towards a better quality of life for the patients.

Epilepsy is also known as a seizure disorder. It is usually diagnosed after a person has had two seizures, or one seizure with the tendency to have more. Seizures happen when the brain nerve cells fire more rapidly with less control than usual, affecting how a person feels or acts. Nevertheless, not all seizures are epileptic in origin. Some are caused by psychological reasons, and such type of seizures are called psychogenic non-epileptic seizures (PNES), which are not associated with an epileptic discharge. To determine if a seizure is caused by epileptic discharges, Video-EEG/Video-SEEG monitoring is used to check the existence of simultaneous culprit brain EEG rhythms during the seizure. Fig. 1.1 shows examples of how Video-EEG/Video-SEEG monitoring records the semiology and the real-time neuron activities for assessment. Despite the different cause of epileptic seizures (ES) and PNES, these two types of seizure could be similar in terms of the semiology, i.e. the clinical signs. Even for experienced neurologists, it could be challenging sometimes for them to correctly distinguish them. In addition, the evaluation could be subject to inter-observer variability. Hence, a computer-aided diagnosis is naturally considered as a way to improve the quality of the assessment.

Semiological signs play an important role in analyzing the clinical symptoms regarding a seizure event. It relates to multiple informative sources from the patients in a tempo-

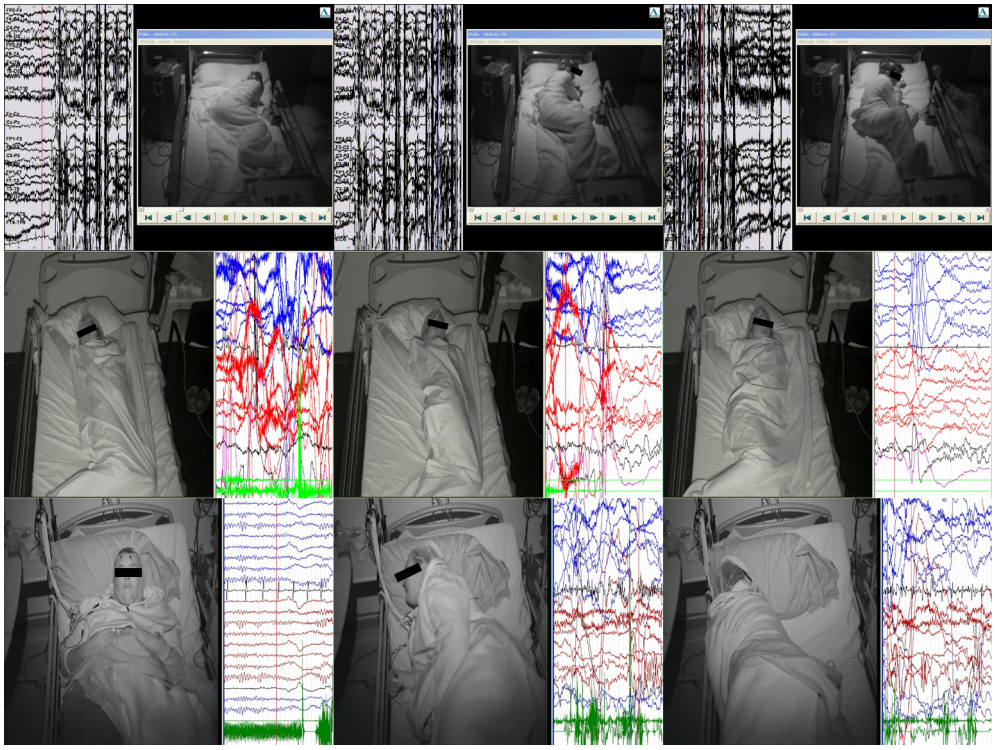


Figure 1.1: Samples of the Video-EEG/Video-SEEG recordings during seizures. The real-time neuron activities are used to determine if a seizure is caused by an epileptic discharge.

ral progression, like the evolving of one's facial expression, emotion, and body gestures. Fig. 1.2 and Fig. 1.3 show some examples. The motivation of this research is to develop methods based on recent machine learning progress to provide objective analysis for clinical seizure videos. This research aims to step towards the development of computer-based methodologies for seizure diagnosis considering the semiological information. The proposed approaches could be useful for developing related methodologies for monitoring other disease, such as dementia.

1.2 Challenges

As opposed to conditions within a highly controlled environment, the automated vision-based seizure analysis system is aimed at dealing with real-world seizure videos. To build such a system, there are several challenges we could encounter:

The complicated conditions in videos

For seizure videos, there is usually a complicated condition involved. Patients might be partially occluded by bed sheet, clinical staff, or even themselves due to hyperkinetic

Figure 1.2: Selected video sequences of seizure events. The semiological signs might include the time-evolving changes of facial expressions.

behavior. The occlusion could affect the performance of some automatic detection process, like region or key-points detection. In addition, low camera resolution or inadequate light could also incur poor performance. Fig. 1.4 shows some challenges in real conditions.

Insufficent data

For AI applications in the medical domains, doctor's annotation is usually time-consuming. So one of the major challenges for machine learning for medical applications is the fact that the scale of the labeled data are small compared to other problem domains. Seizure video analysis is no exception. Nevertheless, on the bright side of this challenge, recently some approaches are proposed to address this issue by self-supervised learning (SSL). In SSL, a learning machine captures the dependencies between variables, and learns representations of the data without requiring human-provided labels. The methodology usually first pre-trains on large volume of unlabeled data, and then fine-tunes the pre-trained model on downstream tasks, which usually have smaller labeled dataset. The SSL-based methods have revolutionized natural language processing (NLP) and is making very fast progress for speech and image recognition.

Model explainability

Model explainability means being able to explain model's predictions. As like the insufficient labeled data issue, the model explainability is another important topic for AI in medical applications. Doctors would favor an explainable model more than a blackbox model. Nevertheless, how to better leverage the trade-off between performance and ex-

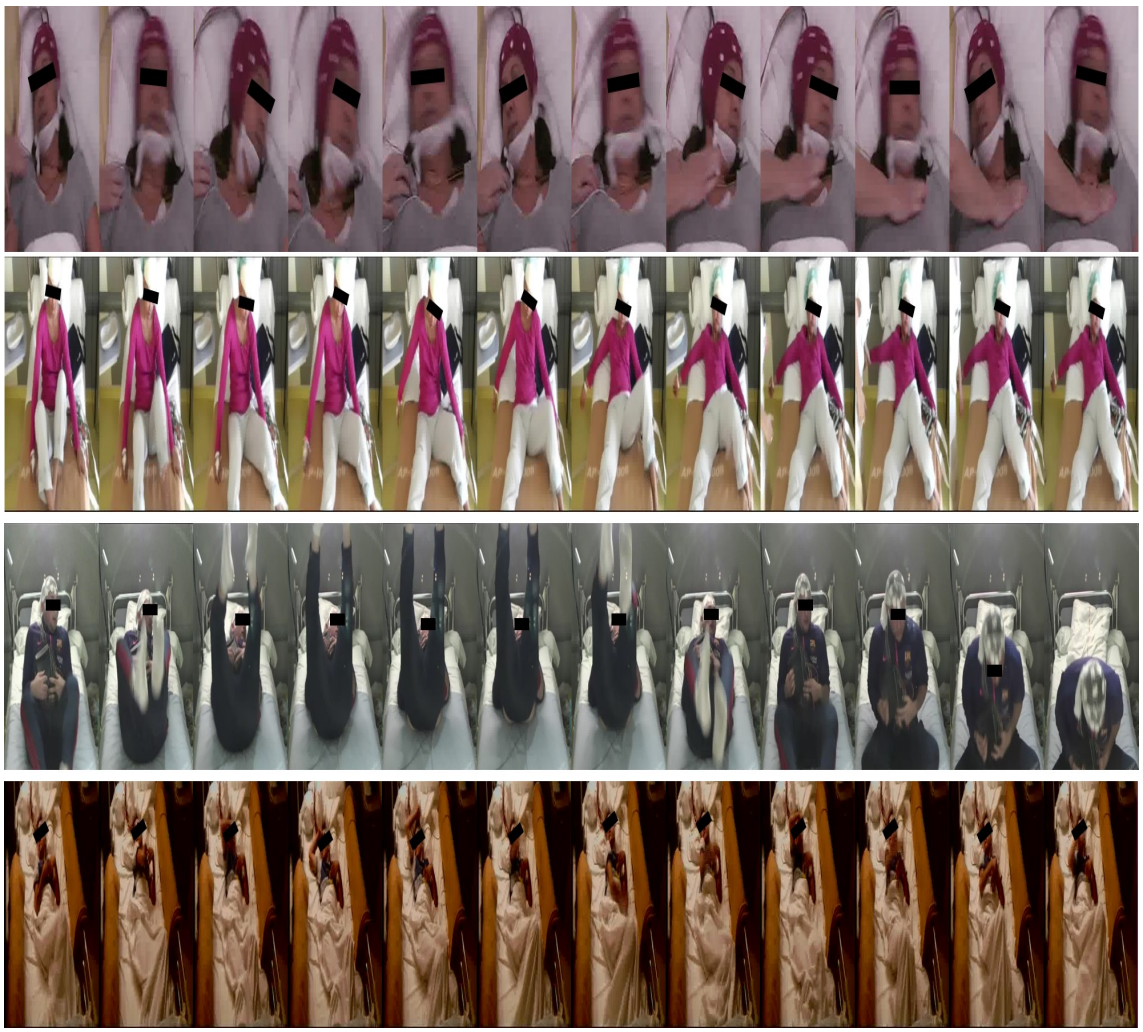


Figure 1.3: Selected video sequences of seizure events. The semiological signs might include the repetitive lateral head turing movement, limb rigidity, anterior-posterior rocking movement, and irregular upper-limb postures.

Figure 1.4: In real-world seizure videos, occlusions of the patients are often observed, either by (a) medical staff, (b) bed sheets, or (c) themselves due to hyperkinetic behavior. In addition, (d) illumination changes could also happen and affect the performance of some pre-processing procedures.

plainability is still an unsolved research topic. Related to our study, some video prediction models visualize the attention map as an indicator for "what the model sees". Yet as more and more research on the combination of vision and language, perhaps it is another good way towards explainable AI. So the model could provide both a prediction and an explanation. These could provide inspirations for models in medical applications.

1.3 Contribution of this study

The innovation of this thesis is the investigation of how deep learning can be exploited in the presence of limited data and complex conditions towards automated seizure video analysis. In this thesis, we propose several contributions as follows:

Head movement analysis for hyperkinetic seizures

For hyperkinetic seizures, in which high amplitude and/or rapid movements are involved in the ictal phase, we propose a method to analyze the trajectories of the head movements of the patients. This analysis provides a basis for investigating the correlation between the frequency of head movement and that of the EEG signals.

A multi-stream framework for seizure classification

We propose a multi-stream deep learning architecture in order to characterize semiological patterns from epileptic seizures (ES) and psychogenic non-epileptic seizures (PNES), based on the semiological signs from patients' motor manifestations and facial expressions.

A self-supervised learning framework for seizure classification

We propose a self-supervised learning framework to classify ES and PNES. The proposed transformer-based method is pre-trained on large unlabeled clinical contextual videos. Then the pre-trained model is fine-tuned on labeled datasets for the seizure type classification task.

1.4 Structure of the thesis

The chapters of this thesis are structured as follows:

Chapter 2 will conduct a literature survey on recent vision-based seizure analysis works.

Chapter 3 elaborates the data collection process and specification.

Chapter 4 shows the study on head movement analysis for hyperkinetic seizures.

Chapter 5 presents our multi-stream framework for seizure classification.

Chapter 6 illustrates the proposed self-supervised framework for seizure classification.

Chapter 7 summarizes this research and provides perspectives for future works.

Chapter 2

Related Work

This chapter presents a survey of the development of automated methods for analyzing seizure motions. Some of the methods are based on traditional machine learning techniques and some are built on recent deep learning models. We will first have a quick review of the idea of machine learning and deep learning, and then dive into the related works on seizure motion analysis.

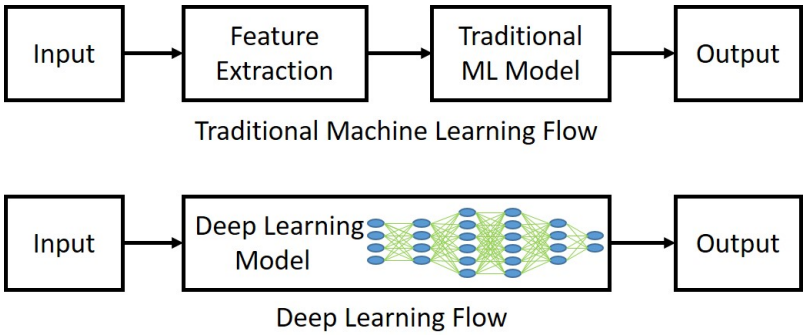


Figure 2.1: A comparison of the flow of machine learning and deep learning methods.

2.1 Traditional machine learning and deep learning methods

Machine learning is a field of studies that focus on using data and trying to imitate how human learns, and gradually improve the task performance. Deep learning can be seen as a sub-category of machine learning. It mimics the function and structure of neurons in the brain. Usually the deep learning models consist of multiple layers of non-linear transformation. Artifical Neural Networks (ANN) is used in general to refer to this type

of models. One major difference between machine learning and deep learning is in the feature engineering. As shown in Fig. 2.1, machine learning usually adopts hand-crafted features, which may need domain expertise for different tasks. Then these features are fed into traditional machine learning models, like Support Vector Machines (SVM), Decision Trees, or Random Forest, for the target puropose. On the other hand, feature engineering in deep learning is done automatically. It uses an end-to-end learning framework, and hence requires less human intervention. The following introduces some widely-used deep learning models or concepts.

Convolutional Neural Networks (CNN)

The convolutional neural network is a type of ANN, and is especially suitable for tasks related to image processing. It involves a kernel filter to conduct convolution computation across the whole input data, as like in Fig 2.2. A CNN typically consists of stacked convolutional layers, and then ends with several fully connected layers. The CNN is then learning to optimize the output by achieving the possible minimum loss defined by a cost fuction. The weights of the CNN are then updated by back-propagation [2]. Fig 2.3 shows one of the earliest work to utilize CNN for image-based digit recognition. Until now, there have been numerous succesful works using deep CNN to deal with image processing problems. For example, the VGG-19 model [3] consisting of 19 layers achieved the state-of-the-art in image recognition and at the time. A year later, He et al. [4] proposed Resnet, which adds a skip connection for residual learning in CNNs, and thus allows CNNs to be trained up to 152 layers. The idea of the skip connection has become a standard building blocks in many deep learning works now.

Recurrent Neural Networks (RNN)

The recurrent neural network is another type of ANN. As shown on the left side of Fig. 2.4, there is a loop within recurrent neural networks, which make the time-evolving information of the sequential data be kept. Specifically, in the Fig. 2.4, we have a main neural network block A, dealing with some input x_t, and outputs h_t. If we unroll RNN as like the right side Fig. 2.4, we can see the module A is copied and reused for different timesteps. This chain-like architecture turns out to be effective in handling sequential data. Based on the concept of RNN, there are several variants of RNN, like Long Short-Term Memory (LSTM) [6] and gated recurrent units (GRU) [7]. They have achieved great success in applications like speech recognition [8].

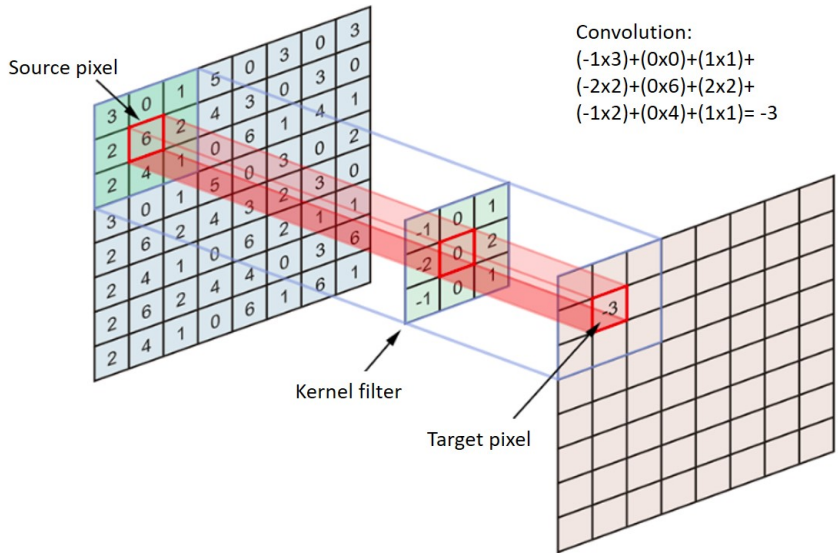


Figure 2.2: The convolution operation in CNN.

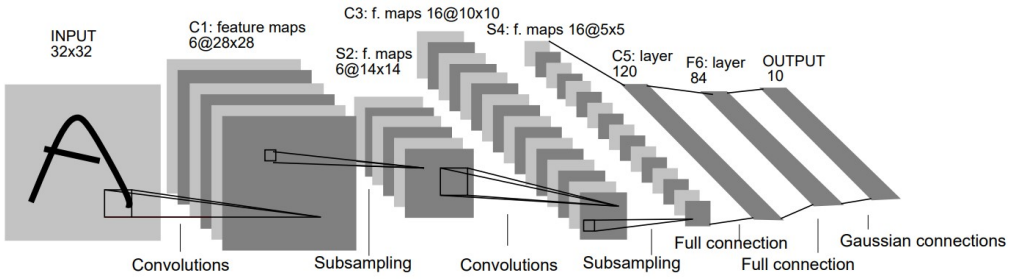


Figure 2.3: The first proposed CNN for digit recognition. The image is adapted from [5].

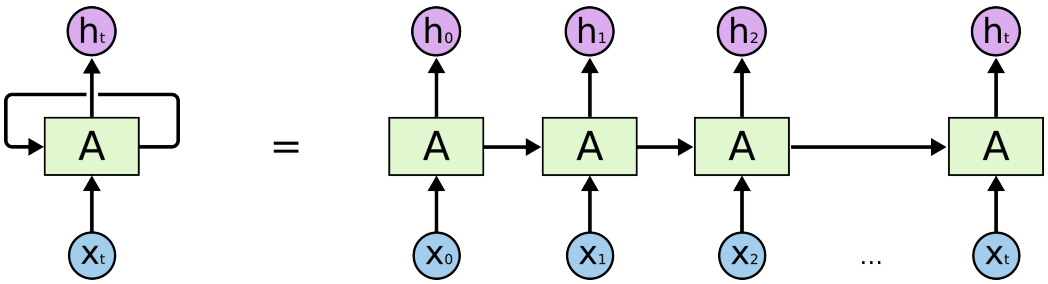


Figure 2.4: An unrolled recurrent neural network. The image is adapted from [9]

The Encoder-Decoder Architecture

In deep learning, the encoder-decoder architecture is widely used. It usually consists of an encoder and a decoder. The encoder is responsible for encoding the input into a

representation, and by feeding it into the decoder, we can decode the representation and so have the desired output. One popular implementation is the Denoising Autoencoder, as shown in Fig. 2.5. It encodes a noisy image, and then reconstructs a clean image through the decoder. Besides, combined with RNNs, the encoder-decoder architecture can be powerful for tasks like language translation. As shown in Fig. 2.6, it encodes the input sentence from source language by RNNs, and then decodes the representation by the decoder into target sentence in the desired language.

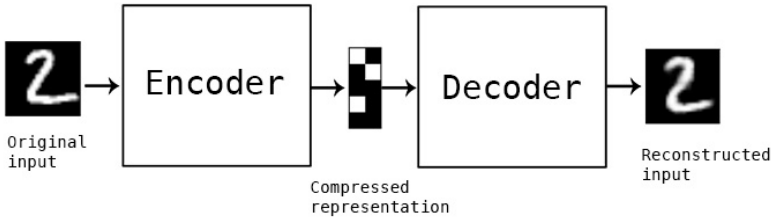


Figure 2.5: A denoising autoencoder encodes a noisy image, and then reconstructs a clean image through the decoder. The image is adapted from [10]

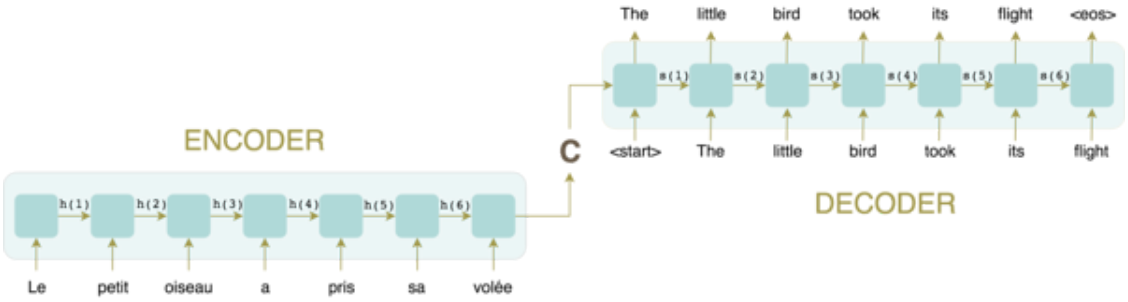


Figure 2.6: An encoder-decoder architecture can be suitable for machine translation tasks. The image is adapted from [11]

2.2 Seizure motion analysis with traditional machine learning methods

Compared to using EEG signals, there are relatively limited studies working on seizure motion analysis. An early review of seizure motion analysis with traditional machine learning methods is done by Pediaditis et al. [12]. In general, previous methods are

proposed under one of the two conditions: it is either a marker-based or a marker-free system. A marker-based system usually requires patients to attach sensors, such as inertial sensors [13] or reflective markers [14], on their bodies or wear customized outfits to effectively detect patients' limbs [15]. The recorded limb motion trajectories can be transformed into metrics like velocity, acceleration, and angular speed for motor seizure analysis. Nevertheless, these approaches may be inconvenient for the patients and subject to sensor detachment when patients are having violent behaviors due to an onset seizure. On the other hand, a marker-free system requires no body-attached sensors, and video cameras are usually used as sensors to record patients' behaviors. Different types of camera have been used, such as single, stereo, and depth cameras. Single camera system is the most widely used one.

In the previous pioneering studies of marker-based systems, Li et al. [14] proposed a system to analyze the motion trajectories of human body joints in videos, with the help of infrared reflective markers attached on patients bodies, as shown in Fig. 2.7 a. Lu et.al [15] developed a color-based system to track and analyze the limb trajectories, as shown in Fig. 2.7 b.

As for marker-free systems, Pediaditis et al. [16] proposed a method for vision-based seizure detection. The proposed work detect patients's faces and then extract the facial features with dense optical flow [17] and discrete Fourier transform. Based on those features, five hand-crafted values are designed to discriminate facial expressions for seizure detection via a decision tree algorithm [18]. Maurel et al. [19] developed a 3D head model given a face image, and utilize the model for facial expression analysis for patients with epileptic seizures. For vision-based body motion analysis, some studies use optical flow and clustering analysis [20, 21, 22]. Some combined spatio-temporal interest point detectors (STIPs) and histograms-of-flow features [23, 24]. Fig. 2.8 shows examples of marker-free systems.

2.3 Vision-based seizure video analysis with deep learning methods

Deep learning has excelled in many computer vision tasks. More and more healthcare applications are introducing deep learning into their systems for better performances [25]. Nevertheless, in the context of automated seizure video analysis, there are still relatively limited studies and few datasets are dedicated for the topic. Here we present some deep learning-based seizure video analysis studies.

Achilles et al. [26] proposed a system to detect seizures by using CNNs to learn features

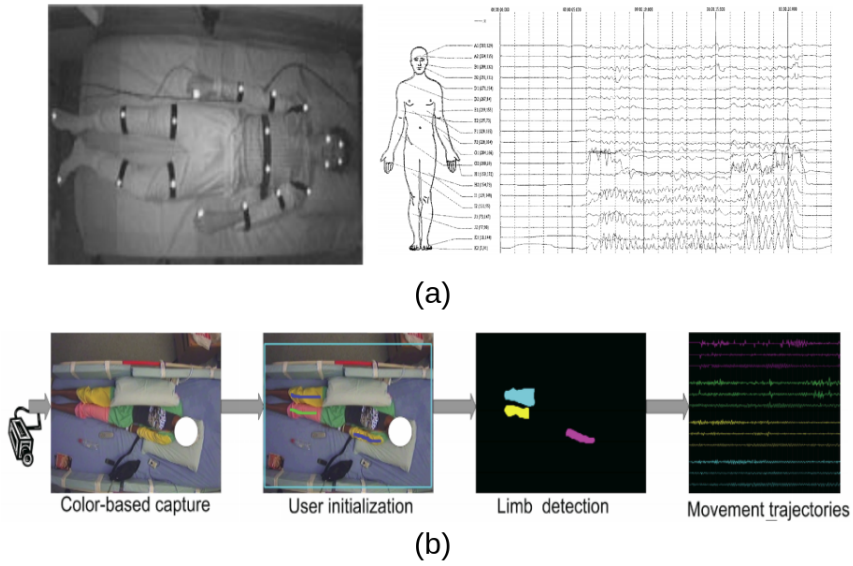


Figure 2.7: Illustration of two examples of the marker-based seizure analytic systems. (a) The system attaches reflective markers on patients keypoints for seizure motion analysis, and in (b), a color-based limb detection is applied with customized outfits. The images are adapted from [14] and [15].

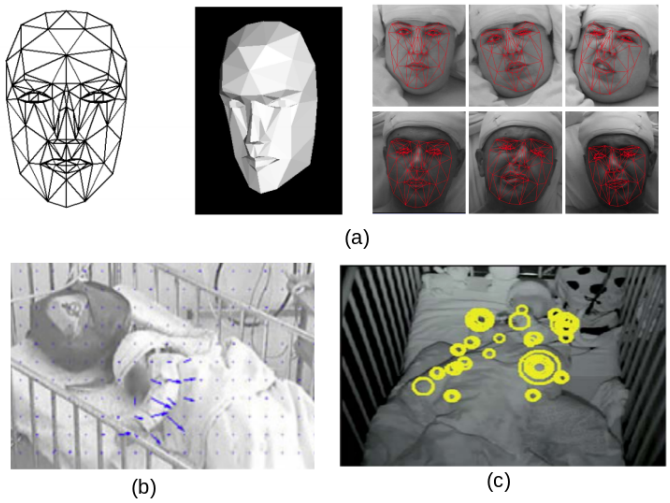


Figure 2.8: Illustration of examples of the marker-free seizure analytic systems. (a) Given a face image, the developed 3D face model is used to conduct facial expression analysis on patients with epileptic seizures. As for vision-based body motion analysis, (b) the optical flow features and (c) spatio-temporal interest point detectors (STIPs) are used. The images are adapted from [19], [22], and [23].

on streams from a combined depth and infrated (IR) sensor. The model can show better results than traditional methods like the combination of Histogram of Oriented Gradients (HOG) and SVM. Fig. 2.9 demonstrates the proposed framework.
Ahmedt-Aristizabal et al. [27] proposed a deep learning model to classify facial semiology from patients with mesial temporal lobe epilepsy (MTLE). As shown in Fig. 2.10, the model crops the face region and feed it to CNNs for generating features. Then a LSTM network handles temporal relations through the whole video.

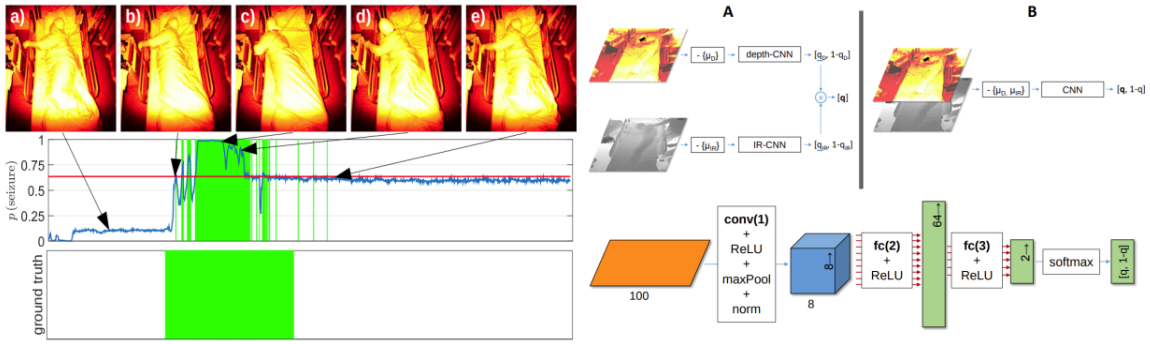


Figure 2.9: The task and the proposed architecture in [26]. The model uses CNN to learn features on depth and IR images for seizure detection. The image is adapted from [26].

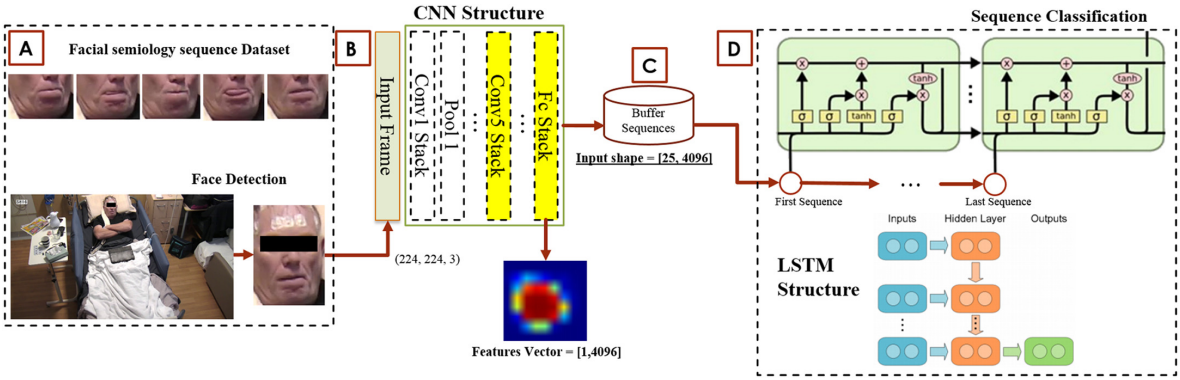


Figure 2.10: A deep facial analysis work proposed in [27]. After cropping the face region, the model uses CNNs to learn spatial features and a LSTM network to learn the temporal relation. The image is adapted from [27].

The information from patient's pose is typically used for seizure motion analysis. Nevertheless, unlike marker-based systems, marker-free or vision-based systems need an algorithm to predict patient's pose in the bed. Patient pose estimation could be challenging

due to various occlusion conditions, e.g. occlusion from blankets, medical staff, or patients themselves. Achilles et al. [28] proposed a CNN-RNN model trained on depth video to predict joint positions even under blanket occlusion, as shown in Fig. 2.11.

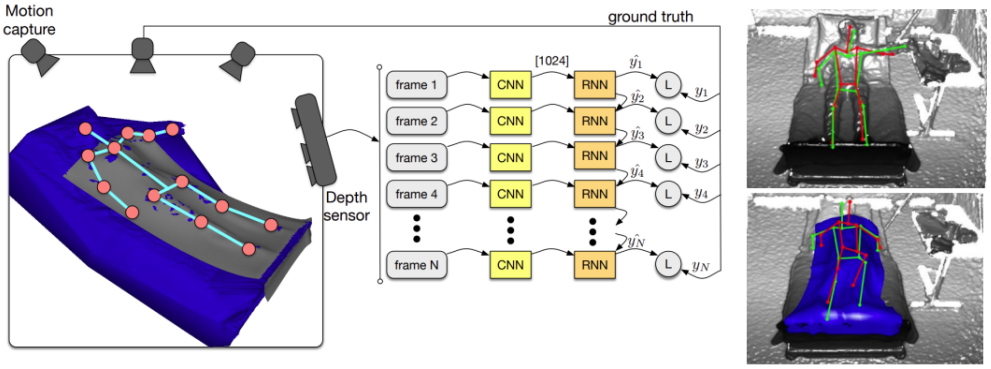


Figure 2.11: Achilles et al. collected videos via depth sensors and a motion capture system. The data is used to train a pose estimation model based on a CNN-RNN framework. On the right side is the pose estimation without and with blanket occlusion. Green/red skeletons denote the ground-truth/prediction. The image is adapted from [28].

One of the main reasons that the use of deep learning can be hindered in medical applications is the lack of enough sample data to train a neural network. To overcome this issue for the in-bed pose estimation research community, Liu et al. [29, 30] released a large scale dataset on in-bed poses, as shown in Fig. 2.12. The dataset features:

- Two data collection settings: (a) Hospital setting: 7 participants (3 females), and (b) Home setting: 102 participants (28 females, age range: 20-40).
- Four imaging modalities: RGB (regular webcam), long-wave infrared (FLIR LWIR camera), Pressure Map (Tekscan Pressure Sensing Map), and depth sensor (Kinect v2).
- Three cover conditions: No cover, bed sheet (cov1), and blanket (cov2).
- Fully labeled poses with 14 joints.

2.4 Concluding remarks and discussion

According to the literature reviewed, research on seizure motion anlaysis can be divided into three main directions: marker-based systems, marker-free systems, and deep learning based systems. The first two kind of systems usually apply traditional machine learning

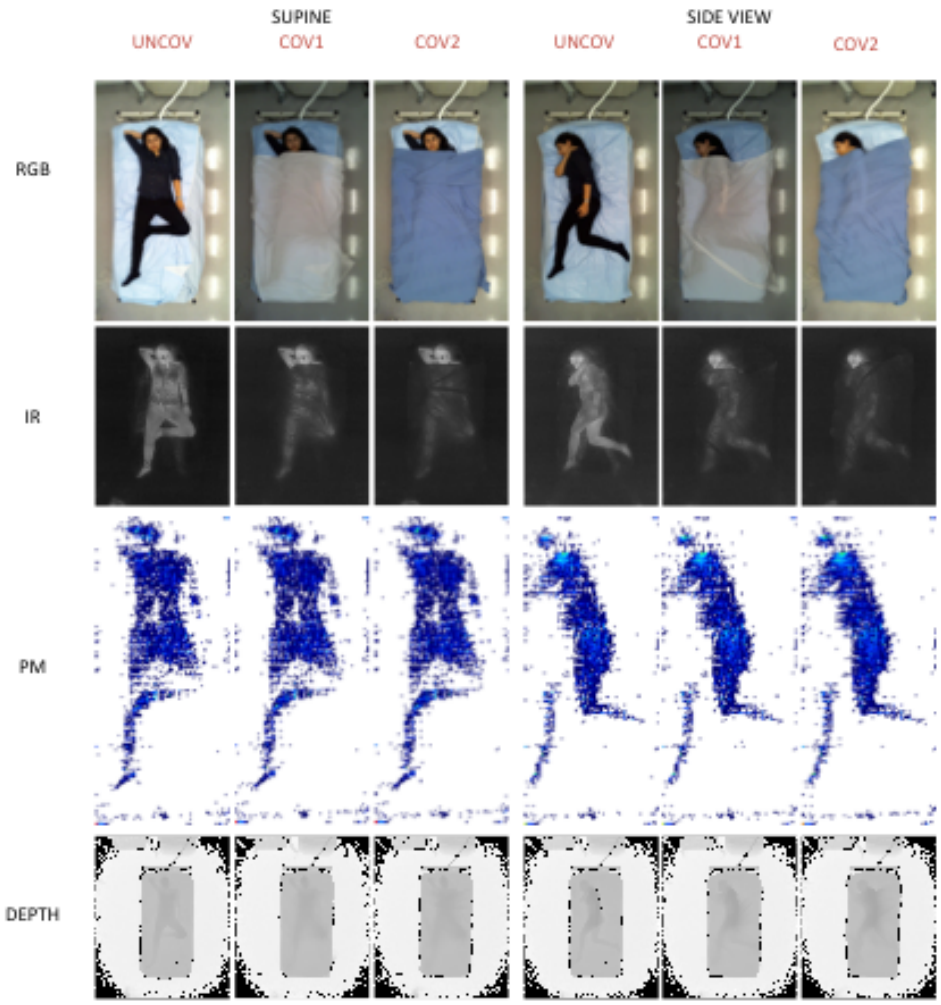


Figure 2.12: Sample images from a large-scale in-bed pose collection dataset. The image is adapted from [30].

techniques and need domain expertise for desigining a good feature for the target tasks. On the other hand, deep learning provides a data-driven option to learn features end-to-end and thus needs less human intervention. With deep learning, the choices of video sensors for vision-based seizure analysis can be more flexible, from single cameras to infrared and depth cameras. In spite of the motivations, using deep learning for seizure video analysis is still much under-explored, not only exhibited in the limited numbers of studies in this field, but also the model architectures used. Current architecures used in this field are mostly combinations of CNNs and RNNs. Indeed, those two neural nets are among the most famous building blocks in deep learning, and they seem able to give

promising results in the literature. However, deep learning field is evolving extremely fast, and since the success of CNNs and RNNs, there have been several powerful models proposed, such as Graph Neural Networks (GNNs) [31] and the Transformer model [32]. Given that, in this thesis, we include the GNNs and the Transformer model into our seizure video analysis, aiming to provide some concrete results with these state-of-the-art models. Besides, deep learning usually needs large volumes of labeled data to train, and it is usually difficult and expensive to get large volume of labeled seizure videos. This could be a reason hindering deep learning from being developed in seizure video analysis. As a silver lining, recently deep learning researchers proposed a work [33] showing how large volume of unlabeled data can be helpful for target tasks whose labeled data scale is much smaller. In this research, we also introduce this idea into our seizure video analysis. In short, after giving a short summary of the literature, we present some complementary viewpoints and implement them in this thesis. We hope this can provide some inspirations for the research community working on automated seizure video analysis.

Chapter 3

Data Collection

3.1 Overview

In this section, we introduce how we collect the seizure video dataset for this research. This part will cover the patient selection criteria, the context about video recording, ethical approval process and codification of the data. In addition, we will show how we use some data labelling tools and detectors to localize important body parts for better semiology analysis. In addition to the videos containing seizure events, we also collected a large volume of videos without seizures. They are called as 'contextual videos', aiming to provide contexutal information of where seizure is occurring. Finally, we compare our dataset to the ones collected by other related works, which shows that our dataset is moderately large in the field of video-based seizure motion analysis.

3.2 Introduction of the curated seizure video dataset

It is important to have a seizure video dataset containing semiology for vision-based seizure motion analysis. To our best knowledge, there is no such dataset that is publicly available. To conduct experiments and evaluate our proposed methods, we curated a seizure video dataset, which contains epileptic seizures (ES) and psychogenic non-epileptic seizures (PNES). The participated patients are selected by Prof. Fabrice Bartolomei and Dr. Aileen McGonigal, from the Epileptology department in the Marseille University Hospital (a.k.a. the Timone Hospital), France. The research associated with the collected data is approved by the Institutional Review Board (IRB) in the Marseille University Hospital, and the patients involved in the dataset have provided the informed consent statements.

For epileptic seizures in our dataset, video recordings were carried out in the Epilepsy Monitoring Units (EMUs). The patients are with drug-resistant epilepsy, indicating the

medication does not work well for them, and instead they need a brain surgery to remove the brain regions that cause the seizures for an improved cure. Before undergoing a brain surgery, the patients need to go through a pre-surgical evaluation, where clinicians will identify the culprit brain regions that cause the seizures. The evaluation procedures usually involve a Video-SEEG monitoring in the EMUs. Stereo-EEG (SEEG) is an invasive approach for monitoring the epileptic discharge within the brain, as shown in Fig. 3.1.

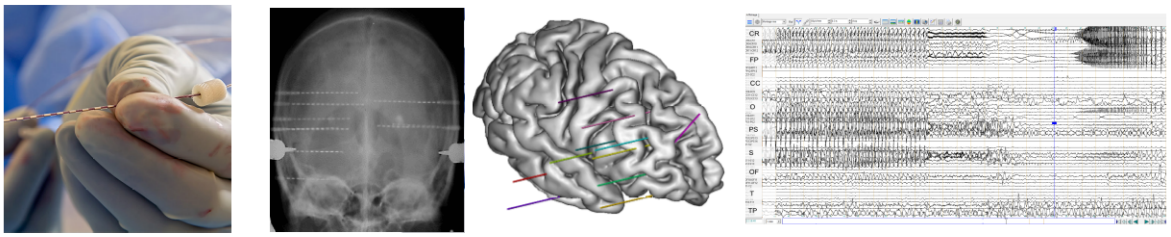


Figure 3.1: (a) Electrodes used in SEEG monitoring. (b) and (c) represent how multiple depth electrodes sample distributed neural systems in the brain. (d) is an example of the recorded multi-channel SEEG signals.

Figure 3.2: Video monitoring of epileptic patients. (a). Epilepsy monitoring units at Epileptology department in the Marseille University Hospital. (b). Samples of video recordings of patients under SEEG monitoring.

Fig. 3.2 shows how the clinicians monitor the patients in the EMUs, and some video samples from patients under SEEG monitoring. Besides, for the collected ES videos, we divide them into subgroups based on the presence of hyperkinetic motor movements. Neurologists are especially interested in seizures with hyperkinetic motor movements, as they usually have more complicated and characteristic semiology. The assessment of the existing of hyperkinetic motor movements are determined by trained clinicians. Some

examples of hyperkinetic seizures are shown in Fig. 3.3.

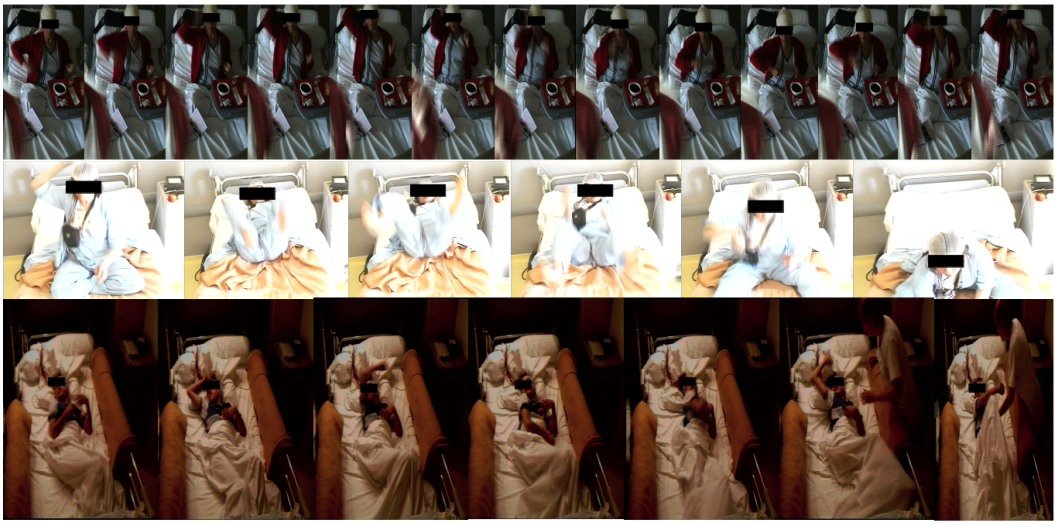


Figure 3.3: Selected samples of epileptic seizure featuring hyperkinetic motor movements.

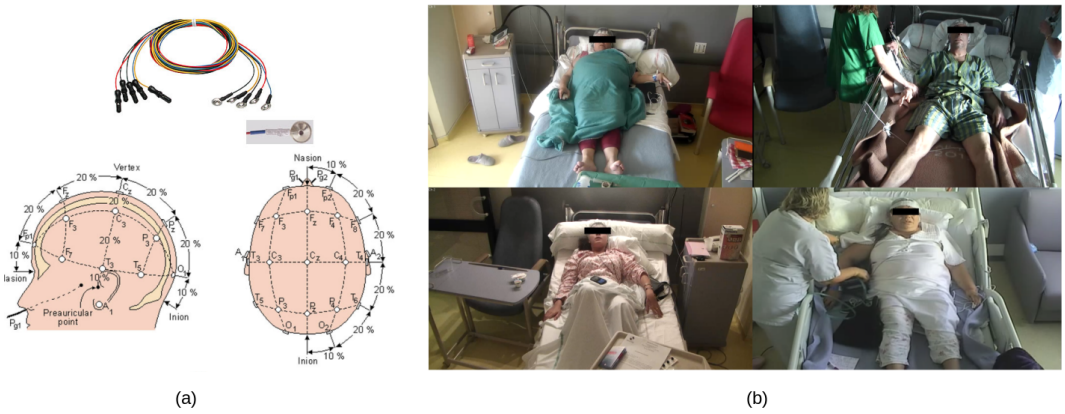


Figure 3.4: (a) Electrodes used and the placement for the scalp EEG monitoring. (b) Selected samples of patients under EEG monitoring.

For PNES videos, video recordings are also carried out in the EMUs, while the patients are under Video-EEG monitoring procedures, unlike the invasive SEEG used in the ES videos. The main context of receiving these patients is to check whether the seizure is epileptic or non-epileptic. The video appearance gives clinicians clues to this, but the important factor is the simultaneous recording of brain EEG rhythms, since epileptic seizures have an epileptic discharge visible on the EEG but PNES do not. Fig. 3.4 demonstrates some selected samples of the EEG monitoring procedure. All the patients in the PNES

video are diagnoised with PNES, indicating the seizures are caused by psychological reasons rather than epileptic discharges in the brain. In particular, the collected PNES videos are with the hyperkinetic motor movement features.
The seizure video clips are segmented from the untrimmed video recordings during a Video-SEEG or Video-EEG monitoring procedure. The video segmentation is purely based on the seizure semiology, where we keep the main expression of semiology of a seizure event. For example, for a tonic-clonic seizure, there are usually three phases involved: a tonic phase (stiffening), a clonic phase (jerking), and a post-ictal confusional fatigue phase (relaxation). In this case, we only look to the first two phases.
For patients who participated in this research, there were no external or additional activities involved other than the regular monitoring procedures. We aim to collect as much data as possible. In addition to the clinical data from patients under active treatments, we also analyze the data in a retrospective way, as to include videos dating back to September, 2000. Besides, all the videos are in real conditions. There is no specific device or settings catering to this research. With the considered points, we are able to develop a challenging and representative dataset at a moderately large scale in the field.

3.3 More details about the seizure video dataset

The developed seizure video dataset aims to support and evaluate the proposed methods in this research. The main task in this research is to distinguish ES from PNES, simply based on the visual information. The motivation is from the importance of the correct diagnosis of whether a seizure belongs to ES or not, which makes considerable difference for the follow-up treatments. In this section, we present more details about the curated seizure video dataset, including some specification, statistics, and codification of the data. In addition, we show how we use semi-automated tools to annotate ground-truth for better region-of-interest (ROI) detection.

The collected seizure videos
As mentioned, our seizure video clips are segmented from the video recordings in the EMUs. Since the time span within our database is across more than 20 years, there were several camera system changes. As shown in Fig. 3.5, before 2006, the camera system provided a focus on patient's face overlapped on the main content of the video frame. For the system after 2006 and before 2012, the location of the face focus had been parallel to the whole view. As for videos recorded after 2012, there is no more focus on the frame. Although zooming in gave a better visualization for clinicians to assess the semiological features, it might increase the difficulty for preprocessing the videos as clean data for

(a) (b)

Figure 3.5: Changes of camera view settings in the EMUs at the Marseille University Hospital. Before 2006, there was a zoom-in overlapped with the main frame, as in (a). After 2006 and before 2012, the focus was located parallel to the main frame, as in (b).

developing automated methods. Later in this section, we will show how we utilize semi-automated labelling tools to improve the ROI detection in our dataset.

As could be expected due to different recording systems, there is no unified video format and specification for the untrimmed video recordings in the EMUs. After segmenting out each seizure for a new video clip, we saved the new clip as its original format, including the MPG, MP4, AVI, and ASF. We then converted the trimmed clips into image sequences for each clip at a frame rate of 25 frames per second, while keeping the resolution unchanged, inlcuding 352×576, 352×288, 704×576, 720×576, and 1280×720. We resize the aspect ratio until the frames were fed into the developed models. Table. 3.1 shows some main statistics of our seizure video dataset.

Class name	ES	PNES
Number of patients	52	29
Number of seizures	235	48
Average seizure duration [sec]	45	52
Min. seizure duration [sec]	7	12
Max. seizure duration [sec]	150	119
Earliest date recorded	Jan. 2000	Oct. 2007
Latest date recorded	Oct. 2020	Feb. 2021
Number of HKNS	HKNS:101, non-HKNS:134	all

Table 3.1: Some statistics about the seizure video dataset. HKNS denotes hyperkinetic seizures.

De-identify patient data via semi-anonymization

To protect the personal information of patients from being identified, we carried out a semi-anonymization on the collected video data. The naming rule for each seizure video is $AaBbXXM\_YYYY$, where

- Aa: The first two characters of the surname of the patient.
- Bb: The first two characters of the first-name of the patient.
- XX: Two digit number indicating the ordinal of the seizure.
- M: This is either a S or E. A S in this position denotes the seizure is from a SEEG monitoring (ES), while a E indicates a scalp EEG monitoring (PNES).
- $YYYY$: Year when the seizure was recorded.

For example, if a patient is named Timothy Roberts (a hypothetical name), and he had his first ES in 2015. Then the semi-anonymized name for the seizure video clip would be $RoTi01S\_2015$. This simple nomenclature not only covers enough private information for developers, but also provides an easier way for clinicians to decode the target patient for clinical discussion in case.

Semi-automated labelling tools for ROI detection

Our dataset is challenging because it involves different camera systems, changes of illumination, occlusion issues, etc. To more effectively detect the ROI for semiology analysis, we utilized some graphical image annotation tools to manually label some ROI in our dataset for better detection. As shown in Fig. 3.6, we used LabelImg [34] to annotate the bounding box of face and full body of the target patient. As for keypoint annotation, as shown in Fig. 3.7, we utilized Visipedia [35] to help label the joints on our patients, which are exported as a COCO [36] style format. With the annotated images, we fine-tuned some pre-trained detectors to fit better on our dataset. Fig. 3.8 demonstrates some selected samples of the detection results. More details about the ROI detection can be found in Chapter 5.

3.4 The contextual video dataset

After a routine Video-SEEG or Video-EEG monitoring, there could be several video recordings at hours for the entire session. If seizures occurs during the session, the medical staff

Figure 3.6: A demonstration of using LabelImg, a graphical image annotation tool, to annotate the bounding box of ROI, i.e. face and body.

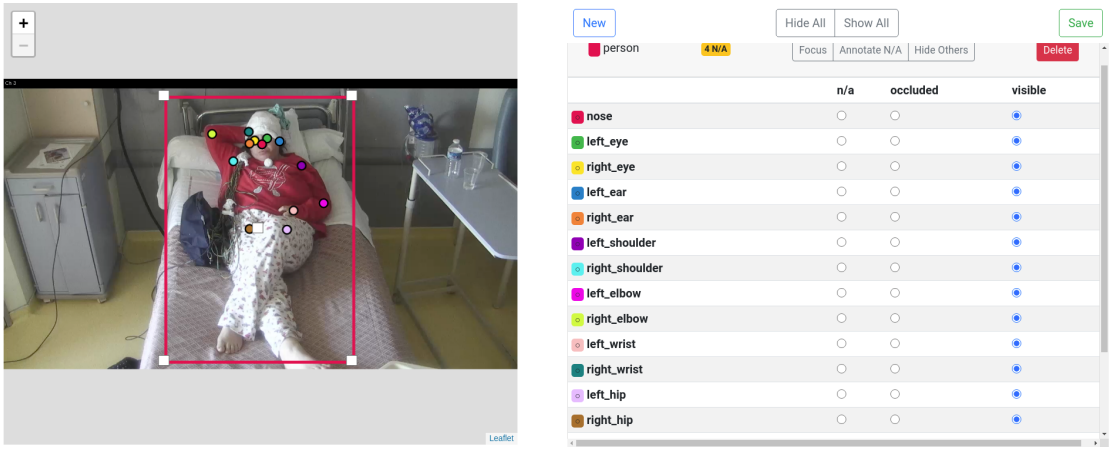


Figure 3.7: A demonstration of using Visipedia, a graphical image annotation tool, to annotate the keypoints of the patient. We labelled 11 keypoints, including nose, eyes, ears, shoulders, elbows, wrists, and hips, for selected frames.

will identify and extract the video segments afterward, and then save them in the database of the hospital. As for parts where no seizure events are involved, these recordings will be erased weeks later, because they could be bulky for storage yet not informative in terms of medical viewpoints.

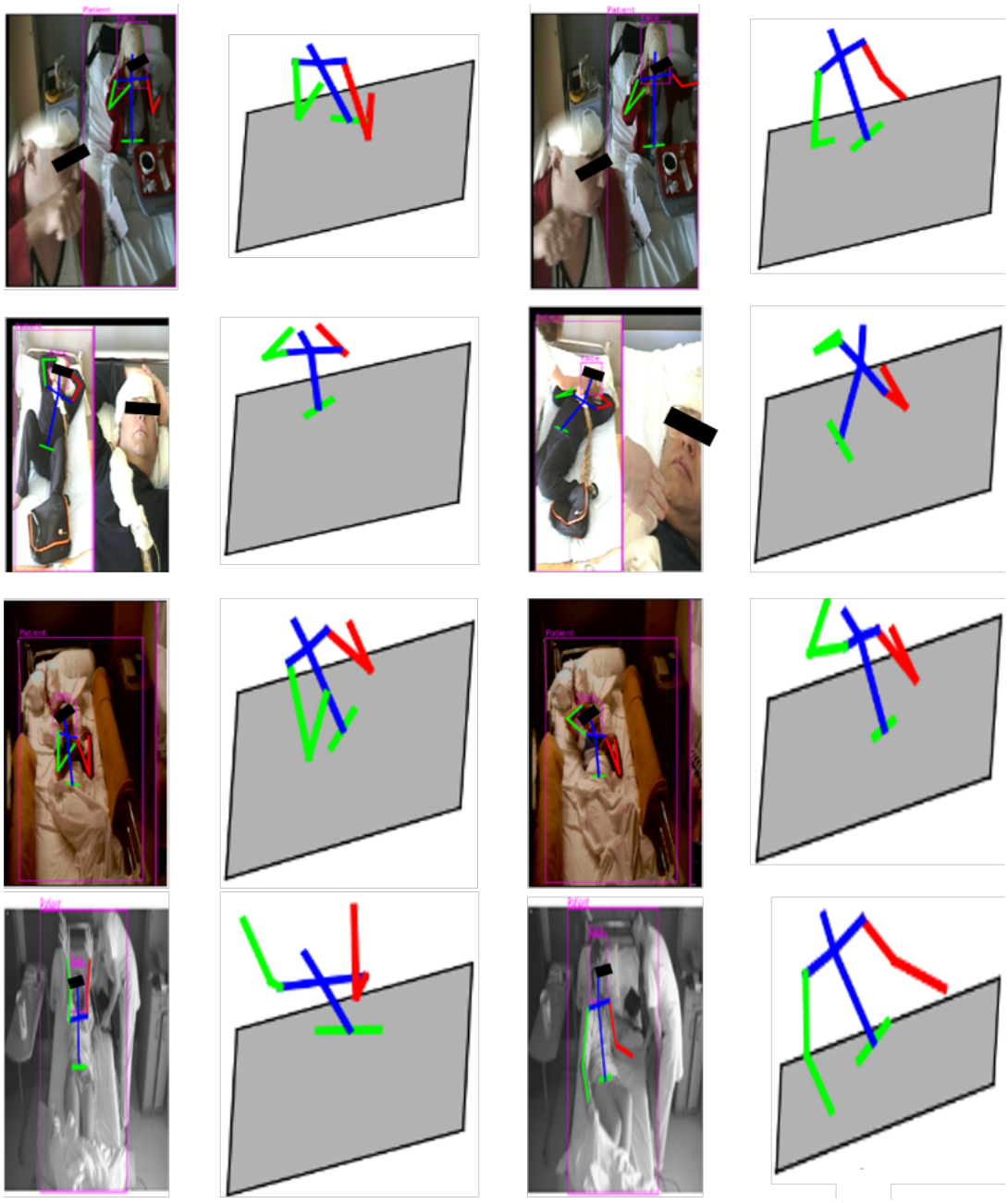


Figure 3.8: A demonstration of face region, body region, and (2D/3D) upper-limb keypoint estimation on selected hyperkinetic seizures recorded in different illumination conditions and camera systems.

Nevertheless, from the side of machine learning, these 'meaningless' seizure-free videos might be useful. The reason is that they can provide the visual information of the surroundings/environments of how seizures are captured. In addition, we can have a large

quantity of them, and the mainstream deep learning/machine learning models usually favors big data. Given that, in this research, we intentionally collected more than 1000 hours seizure-free videos, and we call them as 'contextual videos' in this research. The behavior in these contextual videos can be as diverse and natural as those in daily routines, such as eating, sleeping, chatting with their families, and interaction with clinicians. The recording conditions include both daytime and night. Some selected samples are shown in Fig. 3.9. With these seizure-free data, we can utilize some modern algorithms benefiting from training large unlabeled data [33]. More details about how we exploit these videos for seizure type classification can be found in Chapter 6.

3.5 Concluding remarks

In this section, we present the details about the definition and the collection process of our seizure video dataset. To better analyze the semiology, we demonstrate some ROI detection results with the help of semi-automated tools. In our dataset, there are 283 seizure events and 81 patients involved in total. The numbers make it a moderately large scale compared to the ones used in other vision-based seizure anlaysis works, as shown in Table. 3.2. In addition, to allow novel learning algorithms better adapt to our research problems, we also gather a large amount of non-seizure videos. As can be seen, the research and the dataset scale in this field are still limited. We hope our dataset can provide a good basis for successors to develop a larger and more comprehensive one for better facilitating vision-based seizure analysis.

Research	Number of seizure videos	Number of patients
Achilles [26]	52	10
Ahmedt-Aristizabal [37]	52 (MTLE:40, ETLE:12)	18 (MTLE:12, ETLE:6)
Karácsony [38]	126 (FLE:85, TLE:41)	35 (FLE:20, TLE:15)
Maia [39]	143 (ETLE:107, TLE:36)	31
Ahmedt-Aristizabal [40]	161 (MTLE:90, ETLE:71)	34 (MTLE:17, ETLE:17)
Ours	283 (ES:235, PNES:48)	81 (ES:52, PNES:29)

Table 3.2: Comparison of the seizure video datasets used in the literature. MTLE, ETLE, FLE denotes mesial temporal lobe epilepsy, extra-temporal lobe epilepsy, and frontal lobe epilepsy, respectively.

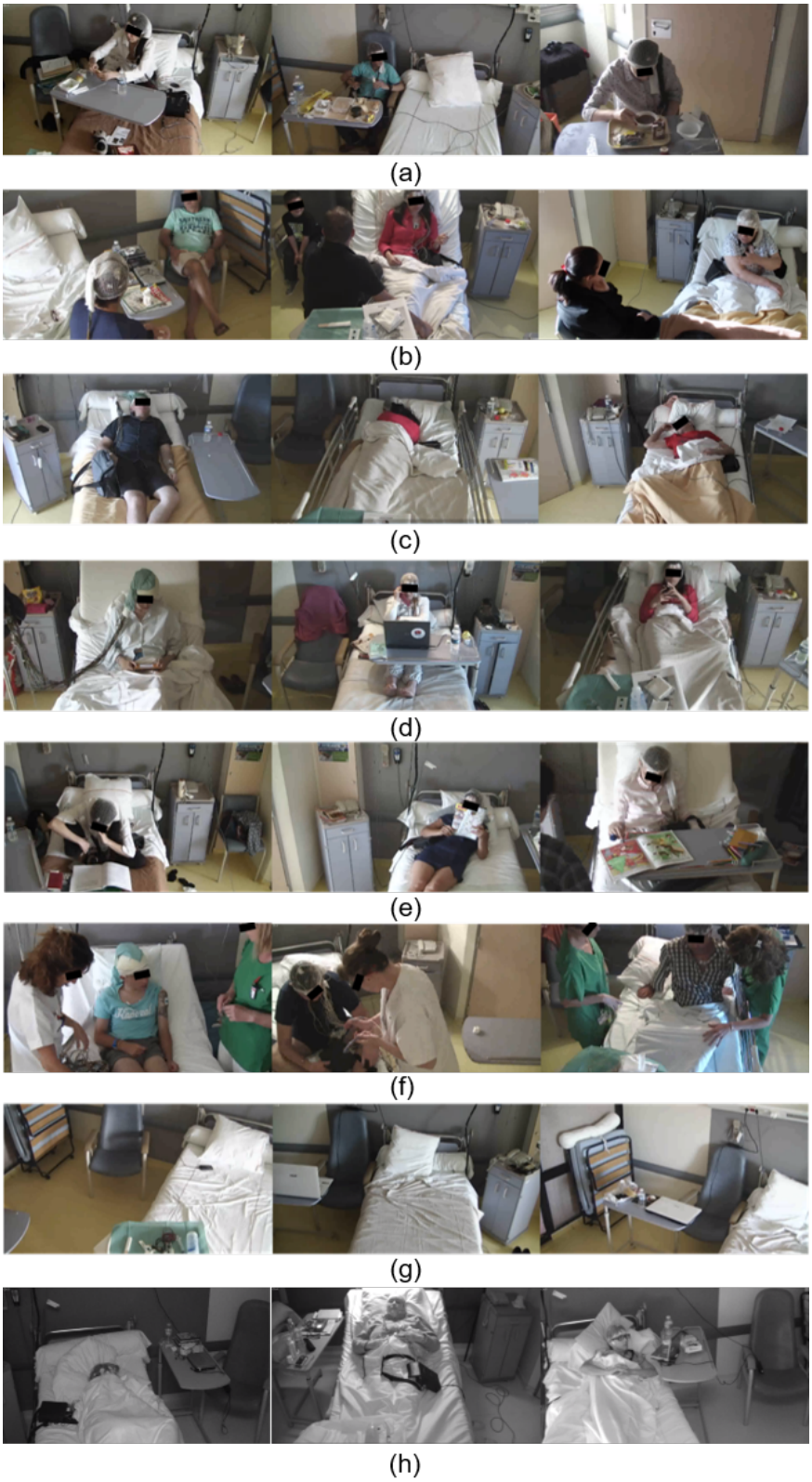


Figure 3.9: The contextual videos cover the daily behaviors of patients in the Video-SEEG/Video-EEG monitoring unit, except for the onset seizure events. They include (a) eating food, (b) interacting with their family, (c) sleeping, (d) using laptops/smartphones, (e) reading books, (f) being checked by the clinical staff. The empty settings are possibly recorded if patients leave the room, as like (g). (h) shows some night conditions.

Chapter 4

Head Movement Analysis for Hyperkinetic Seizures

This study investigates the time-evolving frequency of head movement of an charateristic hyperkinetic behavior during epileptic seizures. Two research journal papers have been published based on this study [41, 42].

4.1 Introduction

Rhythmic movement patterns constitute typical functional motor behaviors across species and across the lifespan [43], and are considered to arise from subcortical central pattern generators [44, 45]. Stereotyped non-functional rhythmic movements are observed in sleep disorders [46], movement disorders [47, 48], and epilepsy [49]. To date, very little data exist quantifying the time-evolving frequencies of these movement sequences, which is unfortunate as these may help in elucidating the mechanisms underlying such pathological behaviors. Semiological "fingerprints" of similar rhythmic movements occurring in both epileptic seizures and sleep disorders have led to speculation of possible shared mechanisms [49]. However, understanding of these is limited, notably in terms of how higher cortical circuits might interact with subcortical components of the motor system to produce similar clinical expressions in conditions with different physiopathologies [48, 50, 49], even though the neural circuitry involved in repetitive behaviors is increasingly well characterized from animal models, notably in terms of defining the corticostriatal circuitry involved. Methods allowing more precise documentation of clinical and physiological phenomena involving complex motor patterns might facilitate further investigation and understanding of this relatively unknown domain [51, 50]. Rhythmic movements are a common feature of epileptic seizures, the best-known example being clonic jerk movements in the context of generalized tonic-clonic seizures [52, 53]. Other

Figure 4.1: Samples of the characteristic antero-posterior rocking movement from the selected 3 patients. In this study, the patients from top to down are called 'patient 1', 'patient 2', and 'patient 3'.

seizure-related rhythmic movements involve multi-segmental motor behaviors, which can involve the axial segment (e.g. rocking movements of the trunk) [54] or upper or lower limbs (e.g. bicycling-like movements of the lower limbs, hand tapping) [55]. Oro-alimentary automatisms may also occur rhythmically [56, 57]. Amongst possible methods of movement quantification in neurological disorders (for comprehensive review, see [58]), there is increasing interest in video analysis techniques, including those based on deep learning or machine learning, for automated analysis of movements in epileptic seizures [59, 60, 61, 62, 63] and for motor stereotypies (e.g., in autism [64]). However, such studies have tended to focus on detection and categorization of movement patterns; quantification of multi-segmental rhythmic behaviors in terms of time-evolving movement frequencies has not yet been reported. Here, we describe a series of prefrontal seizures characterized by highly stereo-typed rhythmic antero-posterior body rocking movements, analyzed using quantitative video methods as well as electroencephalography (EEG).

4.2 Methods

Clinical data

Videos recorded in the context of presurgical epilepsy evaluation in Timone University Hospital, Marseille, France were studied. All patients provided written informed consent for use of data. From a series of 220 cases of frontal epilepsy, 3 patients demonstrated a

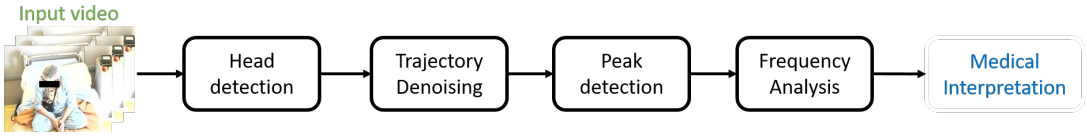


Figure 4.2: Workflow of the proposed approach for head movement trajectory analysis.

characteristic pattern of antero-posterior rocking during epileptic seizures, as can be seen in Fig. 4.1. Two patients had intracerebral electrode exploration with stereoelectroencephalography (SEEG) and one had surface video-EEG. All had full presurgical work-up including neuroimaging.

Video method

Videos of seizures with rocking movements were analyzed using a head tracking method. Recorded seizures in which rocking movements were not clearly visible were excluded. Each seizure video was converted into an image sequence under 25 frame-per-second and resized to 512×512 pixels dimension.

As shown in Fig. 4.2, video processing consisted of four parts: head detection, trajectory denoising, peak detection, and frequency analysis. First, the head of the patient was detected in each video by utilizing a robust detector: the Single Shot Multibox Detector (SSD) network [65], a deep learning-based model. We opted for this approach as partial occlusions and changing environmental lighting conditions, could have jeopardized the performance of simpler head detection approaches. The location of the head was manually annotated for 10-15% of the image samples per video, selected randomly, then SSD was used for head detection. The SSD network is pre-trained on ImageNet [66], a large-scaled image recognition dataset. The pre-trained weights were used as the initialization weights for the SSD network while the network was retrained with the manually labelled samples for fine-tuning the network. After head detection throughout the whole video, the head movement trajectory was computed in the horizontal and vertical direction. Fig 4.3 demonstrates some head detection results and the head movement trajectory. We can observe some cyclic patterns in the vertical direction, as the antero-posterior movements are mainly perpendicular to the camera. The trajectory was then normalized between 0 and 1 for further processing.

To remove the jitter in the trajectory caused by the detector, the trajectory was denoised in both directions by filtering with Empirical Mode Decomposition (EMD) [67]. EMD breaks down signals into different components without leaving the time domain. The components are called Intrinsic Mode Functions (IMFs) and need to satisfy certain conditions as follows:

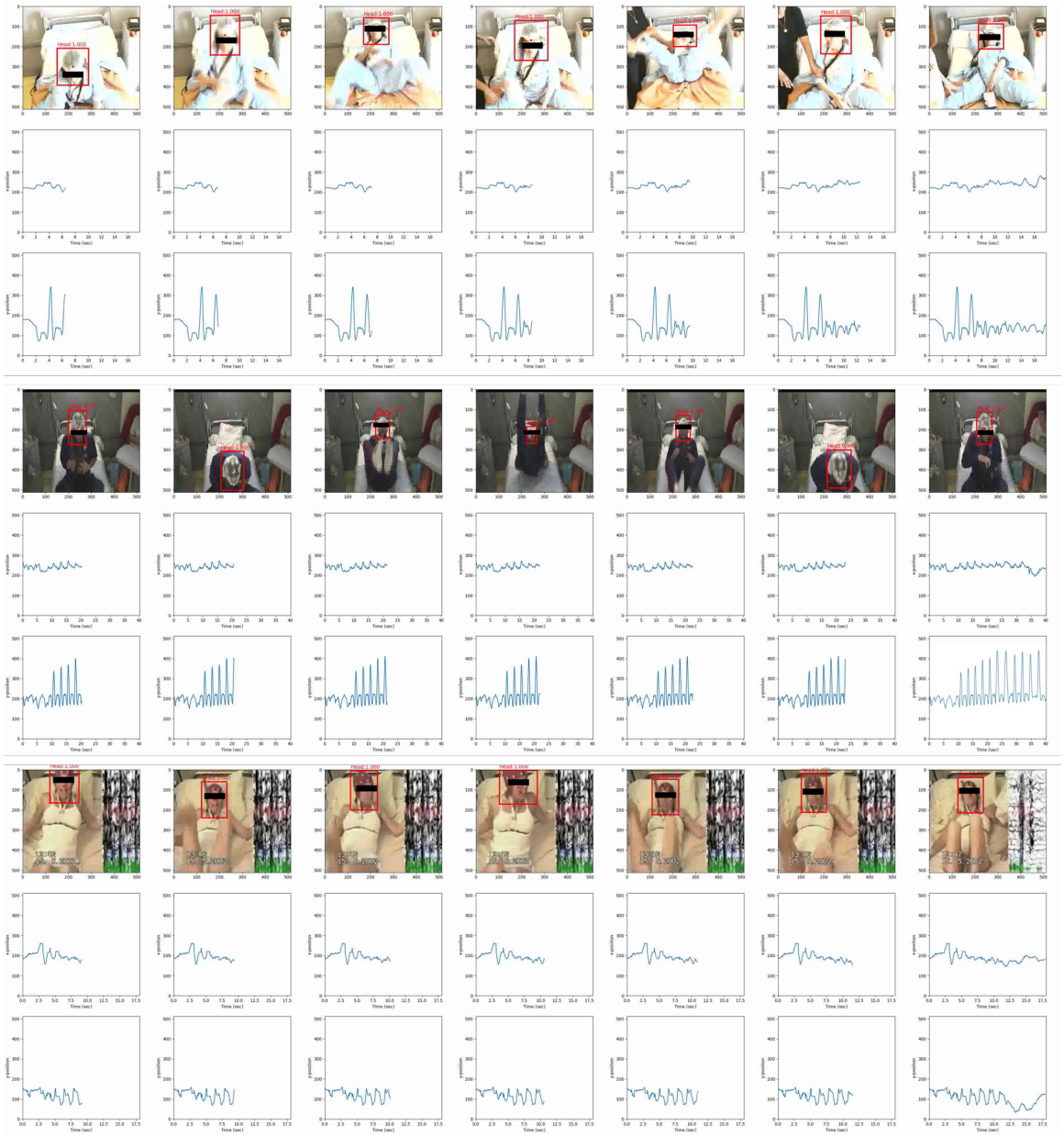


Figure 4.3: Selected samples of the head detection and the head movement trajectory from the cases in Fig. 4.1. For each demonstration, the first row is the image sequence of the seizure video with head detected. The second/third rows represent the horizontal/vertical coordinates of the center of the detected bounding box throughout the whole seizure event. Cyclic patterns are more obvious in the vertical directions, as the antero-posterior rocking movements are mainly perpendicular to the camera.

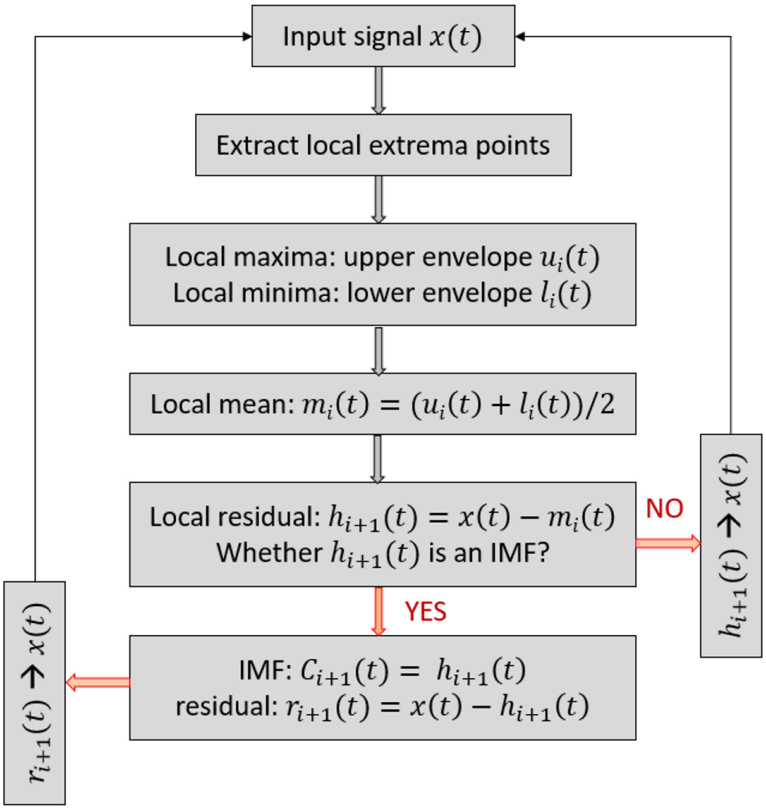
- The number of extrema and the number of zero-crossings must either be equal or differ at most by one.

- The mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

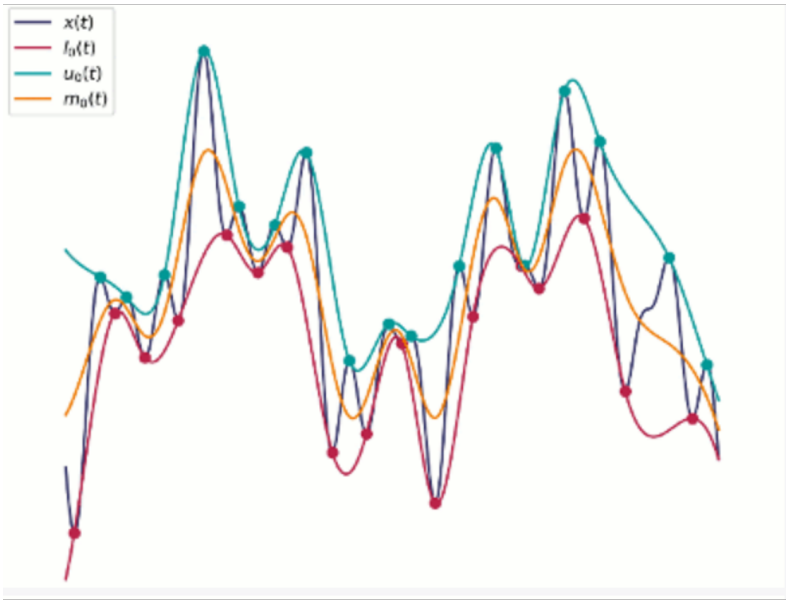
Fig. 4.4 shows the procedure of extracting an IMF. The method is purely data-driven and suitable for non-stationary and nonlinear signals, which corresponds to the conditions for our movement trajectories. An application of EMD is denoising. It can be achieved by dropping the high frequency IMFs. Inspired by this, the head movement trajectory was decomposed into several IMFs. By adding the IMFs with lower frequency components, we can reconstruct and obtain the denoised trajectory signal, as can be seen in Fig. 4.5.
Peaks of the trajectories in both directions were next detected for calculating the cyclic head movement frequencies. To identify peaks corresponding to real antero-posterior movements, we referred to the trajectories in both directions. We defined two functions f and g, such that $f(n_x)$ and $g(n_y)$ represent the trajectory value in the horizontal and vertical directions at time-sampled points n_x and n_y. In addition, n_x and n_y are denoted as n_{xp} and n_{yp} respectively, once $f(n_x)$ and $g(n_y)$ are viewed as peaks. If $|n_{xp} - n_{yp}| < T$ where T is a threshold for deciding how close the peaks in the trajectories from both directions are to be considered as real peaks associated with antero-posterior head movement. The valid peaks are then used to calculate a moving average frequency based on the reciprocal of the peak-peak duration, in order to inspect the time-evolving frequency properties of the seizures. The results can be seen in Fig. 4.6. Specifically, take the seizure 5 in Fig. 4.6 as example, the 0-th detected peaks with a peak-peak frequency around at 0.35 Hz represents the reciprocal of the mean peak-peak duration of the next five antero-posterior rocking movement episodes from the first detected valid peak. Our medical interpretation of the results is in the following discussion section.

4.3 Results

Clinical and neurophysiological results
Localization by SEEG showed low voltage fast epileptic discharge in a widespread right dorsolateral prefrontal distribution for patient 1, and focally in the left orbitofrontal cortex for patient 2. Patient 3 did not require SEEG for presurgical work-up since non-invasive investigations including high resolution scalp EEG and positron emission tomography confirmed focal right prefrontal epilepsy organisation (right intermediate frontal sulcus). All 3 patients had normal neuroimaging. All 3 underwent subsequent cortectomy, with cure of epileptic seizures and disappearance of stereotyped behavior, with minimally 2 years'



(a)



(b)

Figure 4.4: (a) The procedure of extracting an IMF in EMD. (b) An illustrative signal $x(t)$ for (a), and its upper/lower envelope and local mean in the first iteration of extracting an IMF.

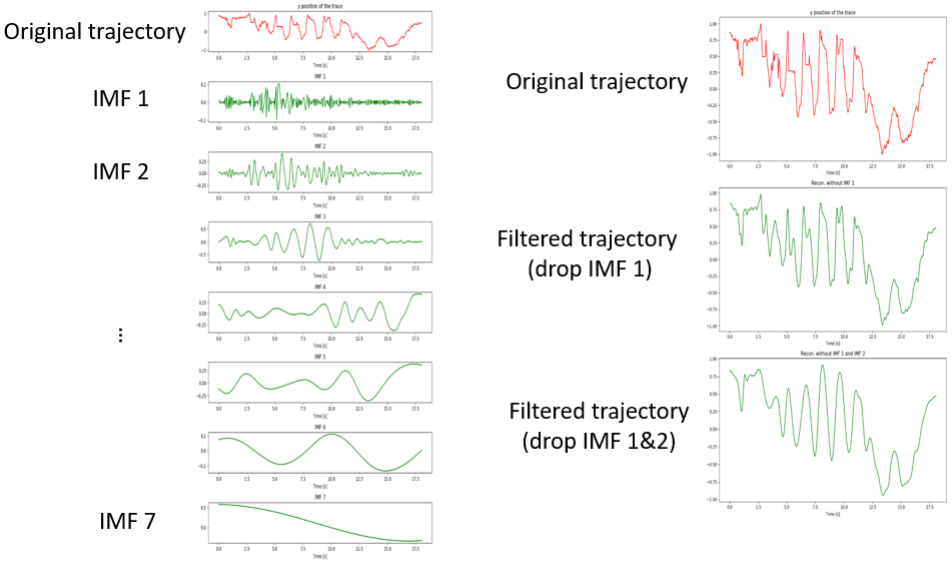


Figure 4.5: On the left, a head movement trajectory (in red) and its seven derived intrinsic mode functions (IMFs) (in green). On the right, the same head movement trajectory (in red) and the two denoised trajectories by selecting different IMFs for reconstruction.

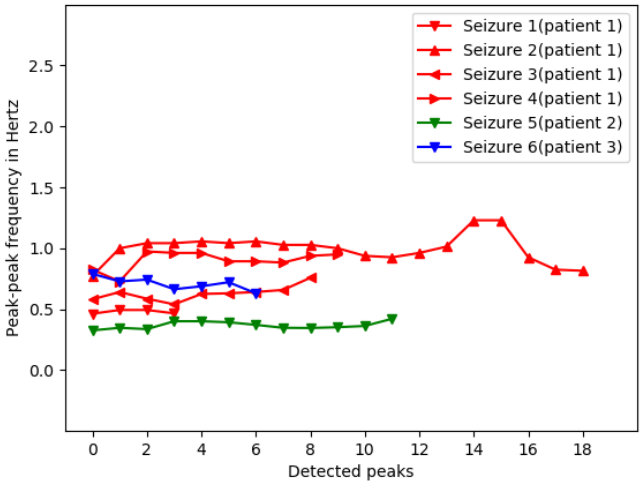


Figure 4.6: Peak-peak frequency in hertz for each detected peak in each seizure video. The color represents individual patients.

Table 4.1: Data for each video on frequency of rocking movements, as calculated using automated head tracking.

	Seizure 1 (patient 1)	Seizure 2 (patient 1)	Seizure 3 (patient 1)	Seizure 4 (patient 1)	Seizure 5 (patient 2)	Seizure 6 (patient 3)
Duration of seizure-related rocking (sec)	16	20	17	14	40	15
Maximum (Hz)	0.49	1.23	0.76	0.97	0.42	0.79
Minimum (Hz)	0.46	0.77	0.54	0.73	0.33	0.63
Median (Hz)	0.48	1.01	0.63	0.92	0.36	0.72
Standard deviation (Hz)	0.02	0.12	0.06	0.08	0.03	0.05
Mean (Hz)	0.48	1.00	0.63	0.90	0.37	0.71
Coefficient of variation (%)	4.17	12	9.52	8.89	8.11	7.04

follow-up. Etiology of epilepsy was cryptogenic in patient 1 and due to focal cortical dysplasia in patients 2 and 3.

Video analysis

Six seizure videos with antero-posterior rocking movement from the 3 patients were included. Duration of rocking within seizures lasted 15-40 sec (mean 16.5). Mean frequency of rocking movements was 0.37-1.00 Hz (median 0.67) (Table 4.1). Each seizure was characterized by a stable frequency of rocking throughout its duration, with a mean coefficient of variation 8.3% (Table 4.1, Fig. 4.7, and Fig. 4.6).

4.4 Discussion

The directionality and regularity of rhythmic axial rocking movements were highly stereotyped across seizures and patients, without goal-directed or habitual behavior [68]. Body rocking also occurs in sleep-related movement disorder [69], or "self-stimulatory behaviors" [70] e.g., in autism, as well as in normal infant development [71]. Anteroposterior spinal rocking is relatively uncommon in epileptic seizures and has only been seen by us in a context of prefrontal epilepsy; such a pattern occurs in only a small proportion of seizures involving prefrontal cortex (around 1% in our series). The 3 cases reported here represent the only ones with this characteristic semiological pattern seen in our Epilepsy Unit over decades of recording.

The repetition, rhythmicity, cyclicity and topographical similarity suggest a pathophysiological role for a temporal assembly of neural structures acting as an oscillator [72, 73], with clinical expression reflecting interaction between nervous system activity and biomechanical dynamics of the musculoskeletal system [74]. Rocking frequency varied between individuals and between seizures, within a mean range (0.37-1.0 Hz) overlapping with but slightly lower than that associated with most physiological rhythmic behaviors (0.8-2 Hz) [75] and those occurring in sleep disorders (0.5-2 Hz) [46]. Despite the short durations available for calculation, each sequence showed a stable frequency throughout, in keep-

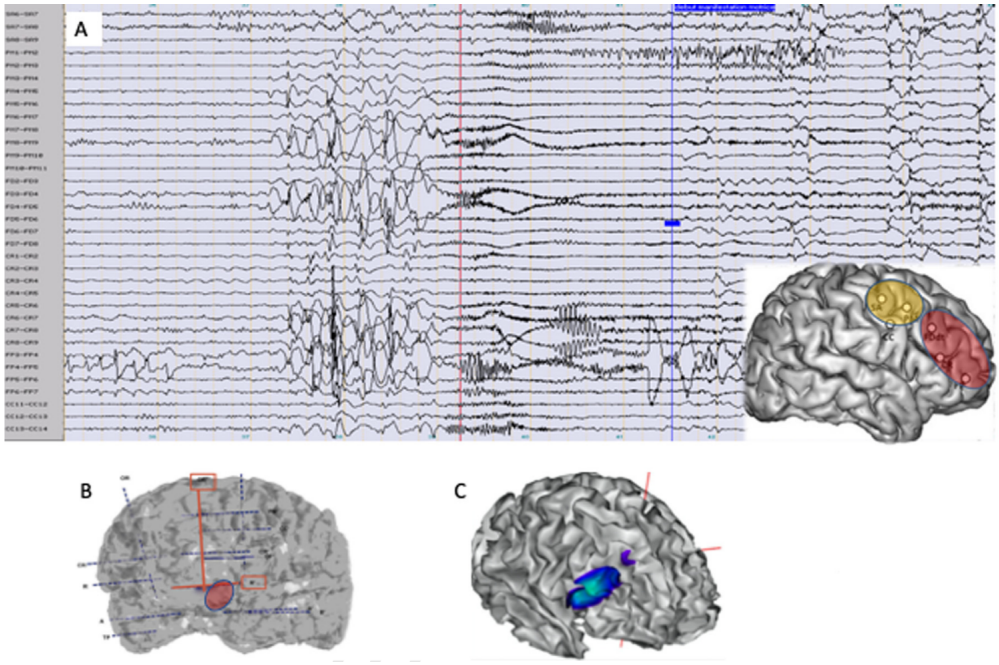


Figure 4.7: A. Ictal SEEG trace from patient 1 (10 seconds per page, 50 microV/mm). Note preictal spiking across a widespread right predominantly dorsolateral prefrontal distribution, followed by abrupt transition to a low voltage fast discharge in the gamma band (vertical red line), showing similar distribution as the preictal spikes. A less tonic and slightly later discharge is seen in electrodes exploring right premotor cortex (top of SEEG trace). The first semiological sign (sudden onset of antero-posterior rhythmic rocking and altered contact; vertical blue line) occurs approximately 3 seconds after electrical seizure onset, at which point slower diffuse rhythmic activity is seen on SEEG. Inset to panel A: schematic illustration of epileptogenic zone of Patient 1, with right dorsolateral prefrontal organisation projecting to premotor areas. B. Patient 2: focal left orbitofrontal organisation of epileptogenic zone, based on SEEG exploration. C. Patient 3: focal right intermediate frontal sulcus organisation of epileptogenic zone, based on non-invasive presurgical evaluation; here, source localisation of HR-EEG interictal data is shown.

ing with the reduced variability typically associated with stereotyped movements [76]. The coefficient of variation was similar to that reported for healthy gait [77]. It is of interest that patient 1, who presented 4 seizures with rocking, showed slight differences in rocking frequency between seizures, suggesting that these were not caused by rigidly predetermined rhythmic generators. The observation that the rocking frequency differed somewhat between seizures may reflect individual differences, as for spontaneous natural rocking [78]. Seizure-specific factors may also have played a role, for example, variations in frequency of cortical seizure discharge, and degree of synchrony between key structures, may contribute to differences in clinical output. Higher rocking frequencies were related to smaller amplitudes (e.g. upper trunk rocking rather than whole body), in keeping with known biomechanical effects of inertia [55]. Antero-posterior directionality remained the same throughout each seizure.

Repetitive, rhythmic movement patterns in frontal lobe seizures may be characterized as stereotypies [55, 48], whose segmental distribution was previously shown to be correlated with localization of the epileptogenic zone along a rostro-caudal axis: axial/proximal motor stereotypies were associated with more posterior frontal regions and distal stereotypies with anterior prefrontal regions [50]. All seizures here showed prefrontal cortex epileptic discharge, but with different sublobar localization across patients. Thus, the here observed movement patterns were not directly related to epileptic activity within a single specific cortical region [55], but suggested an effect involving associative motor regions that might project to a “final common pathway” underlying the repetitive movements. Clinical expression would likely depend on subcortical circuits triggered by different possible cortical localizations of epileptic activity [79] and probably specific temporal (frequency, synchrony) conditions of discharge [56, 51].

From a hierarchical perspective of nervous system organization, abnormal triggering of innate movement patterns may occur by top-down “release” due to transiently altered dynamics within topographically organized cortico-subcortical motor control circuits [80], as has been suggested for some other seizure patterns involving “programmed” behaviors (e.g. rhythmic movements related to locomotion or mastication). From an ontogenetic perspective, fetal somersaults around the transverse axis occur from around 12 weeks’ gestation [81]; in addition, rhythmic stereotypies seem to play a specific developmental role in normal infants, with rhythmic trunk movements occurring mainly between 6 and 12 months of age [71]. Phylogenetically, rhythmic spinal flexion underlies rectilinear locomotion in some limbless vertebrates [82].

In previous SEEG work on seizure-related oroalimentary automatisms, the authors suggested that functional coupling between cortical structures during seizures may be responsible for a top-down effect on outflow pathways from masticatory cortex [56]. Similarly, specific synchronization dynamics created during certain prefrontal seizures might allow

expression of subcortical generators of regular, stereo-typed, rhythmic movement, in this case involving spinal musculature.
Following the research interests on the intersection of neuronal activity and ictal behavior, we extend this study with the inclusion of SEEG. In particular, the four seizures with the corresponding SEEG signals of the patient 1 are analyzed. A finding of spectral correlation between head movement trajectory and SEEG singals is discovered, suggesting a neural signature during expression of motor semiology incorporating both temporal features associated with rhythmic movements and spatial features of seizure discharge [42].
Limitations of our work include the use of a single camera, which allowed 2-D video analysis. More data allowing more detailed characterization of movements could have been achieved using a 3-D video approach, through recording with multiple cameras placed at different angles to the subject. A novel 3-D method, NeuroKinect, has recently been used to successfully record and quantify movements in epileptic seizures [61]. One specific advantage of a multi-camera approach in the present series would have been a lateral view of these movements, which could have allowed assessment of their amplitude. However, since the present data came from a retrospective series recorded in the conventional way in our videotelemetry unit, we were obliged to work with the available video data. The other major limitation is the small number of cases (due to the rarity of this specific semiological pattern), and the possibility that including several seizures from the same patient was a source of bias in determining the mean rocking frequency of the whole group.

4.5 Conclusion

Automated video analysis confirmed stable frequency throughout rocking sequences in the prefrontal seizures, suggesting a mechanism involving intrinsic oscillatory generators. Since localization of seizure onset varied within prefrontal cortex across patients, altered dynamics within a “final common pathway” involving cortico-subcortical movement circuits is hypothesized. The results provide a basis for studying the correlation between the spectrum of EEG and the head movement frequency. Further work on time-evolving frequencies of stereotyped movements across a range of pathologies could help shed light on possible shared pathophysiological mechanisms; to this end, documentation of kinematic properties of stereotypies using automated video analysis could be a useful tool. Future studies could focus on a larger series of seizures with complex motor behaviors, aiming to identify clinical subgroups based on automated video analysis (including a control group), and to correlate these with intracerebral EEG signal analysis.

Chapter 5

A Multi-Stream Framework for Seizure Classification

In this study, we investigate semiology-based seizure classification problems with a deep learning-based method. The proposed method utillzes information from keypoints and appearance, from both face and body pose. Knowledge distillation is introduced for regulating the model learning. Two tasks are explored: epileptic/non-epileptic seizure classification, and recognition of limb dystonia and emotion in epileptic seizures.

5.1 Introduction

As stated in Chapter 1, seizures can be categorized as epileptic seizures (ES) or psychogenic non-epileptic seizures (PNES), based on the presence of epileptic discharges in the brain. The clinical management of ES and PNES is different and as such, accurate diagnosis is crucial to avoid therapeutic errors. To diagnose the type of seizure, one important information comes from semiology [83], i.e., the clinical signs that occur during the seizure, independently from auxiliary information such as EEG or neuroimaging. The gold standard diagnostic method is to record habitual events on video-EEG, with simple visual analysis by an expert in epileptology. Nevertheless, distinguishing between ES and PNES may be challenging, with low accuracy rates for less experienced clinicians, especially when seizures of either type involve complex hyperkinetic motor behavior [83]. There have been many works trying to deal with seizure classification problems with machine learning based on either EEG signals [84, 85] or visually observed semiology [86, 19]. However, to our knowledge, none so far have specifically focused on distinguishing ES from PNES.

In this study, we take advantage of recent deep learning frameworks in computer vision for directly analyzing patients' semiology, focusing particularly on the body pose and face

regions. Several related works have been proposed recently [40, 37, 38]. In [40], the authors use semiological signs from face, body, and hands to classify epilepsy with convolution neural networks (CNNs) and recurrent neural networks (RNNs). The work in [37] also utilized similar strategy with pre-trained CNN features combined with RNNs for analyzing and fusing the information from face and body pose. The method proposed in [38] used a I3D [87] backbone to extract spatio-temporal features followed by RNNs as the classifier.

Rather than using the standard combination framework like CNN-RNN architectures, in this work, we propose to leverage the recent powerful graph convolutional networks (GCNs) for seizure classification. The GCN model [88], which operates convolution on graphs, have been adopted in various tasks, such as skeleton-based human action recognition [89, 90] and facial landmark-based emotion recognition [91, 92]. In this study, we apply a novel, adaptive GCN (AGCN) [89] in which the topology of the graph can be learned, on the detected body joints and facial landmarks for seizure classification. In addition, inspired by [90], we introduce a knowledge distillation (KD) mechanism from the complementary appearance stream for regulating the keypoint features leaned by AGCN. To obtain further improvement, we combined the prediction from each AGCN separately trained on body pose keypoints and facial landmarks with the knowledge distillation mechanism. To our best knowledge, this work is the first attempt to utilize GCNs for seizure type classification (ES versus PNES) based on semiological information. The next section will describe the proposed methodology, followed by experimentation and conclusion.

5.2 Methodology

5.2.1 Overview

In this section, we describe our proposed multi-stream framework for classifying two types of seizures, i.e. ES and PNES. The overall architecture is shown in Fig. 5.1. After converting the seizure video into an image sequence, we detected and cropped the region of patient's body and face, followed by keypoint detectors for joint and facial landmark localization. The detected keypoints were then fed into separated AGCN for classification, which are viewed as Keypoint Streams. The cropped detected region of patient and face were fed into their corresponding feature extractor, and adopted temporal convolutional networks (TCNs) for temporal reasoning. The outputs of these streams, termed as Appearance Streams, were then used to transfer the learned knowledge to the Keypoint Streams. The predictions by AGCN from the pose and face streams were further combined for better performance. The following are the details for each stream.

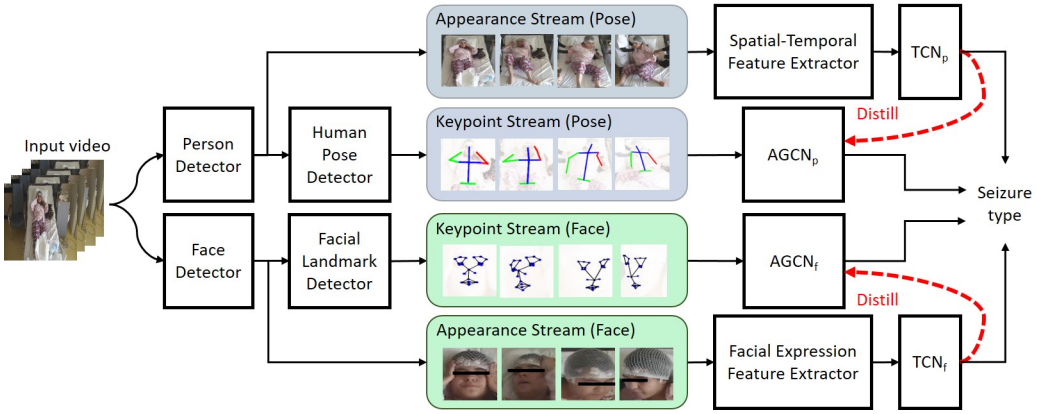


Figure 5.1: Overview of the proposed framework.

5.2.2 Region of interest and keypoint detection

We adopted a fast SSD network [93] with MobileNet [94] backbone for region of interest (ROI) detection, i.e. detecting patients and their faces. The SSD model was pretrained on Imagenet dataset [95] and fine-tuned on our dataset. For body joint localization, we detected the 2D keypoints of upper-limb on the detected patient with Keypoint-RCNN [96, 97], which is pretrained on MS COCO [36] and fine-tuned on our dataset. The 11 detected points include head, neck, left/right shoulders, left/right elbows, left/right wrists, left/right hips, and bottom of the spine. The detected 2D keypoints were fed into a 3D estimator [98] for 3D pose estimation. For face stream, we used a toolbox for extracting 2D facial landmarks with the detected face. There are 23 keypoints detected for each face, focusing on eyebrows, eyes, nose, and mouth. The toolbox was not optimized for our dataset. Fig. 5.2 and Fig. 5.3 show some illustrations and detection results.

5.2.3 The appearance stream

After the ROI detection on a video with T frames, we have the detected cropped region for patient as $R_P = \{r_{p1}, r_{p2}, \cdots, r_{pT}\}$ and for detected face as $R_F = \{r_{f1}, r_{f2}, \cdots, r_{fT}\}$, with $r_{pt} \in \mathbb{R}^{W_p \times H_p \times 3}$ and $r_{ft} \in \mathbb{R}^{W_f \times H_f \times 3}$. W_p and W_f are normalized width, and H_p and H_f are normalized height for detected regions for pose and face streams respectively. We leverage pretrained models for feature extraction followed by a temporal convolution layer. For pose stream, we used R(2+1)D model [99] pretrained on Kinetics [100] with the last classification layer removed as backbone to extract spatio-temporal features on a L-frame snippet, by

$$v_t = Model_{R(2+1)D}(r_{pt}, r_{p(t+1)}, \cdots, r_{p(t+L-1)}) \tag{5.1}$$

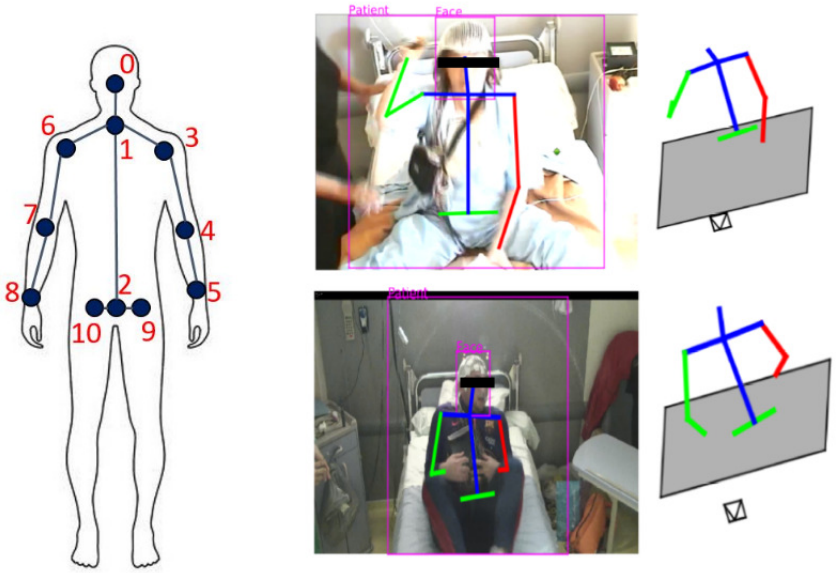


Figure 5.2: (a) Illustration of detected upper-limb joints. (b) Samples of ROI detection and (2D/3D) upper-limb keypoints detection.

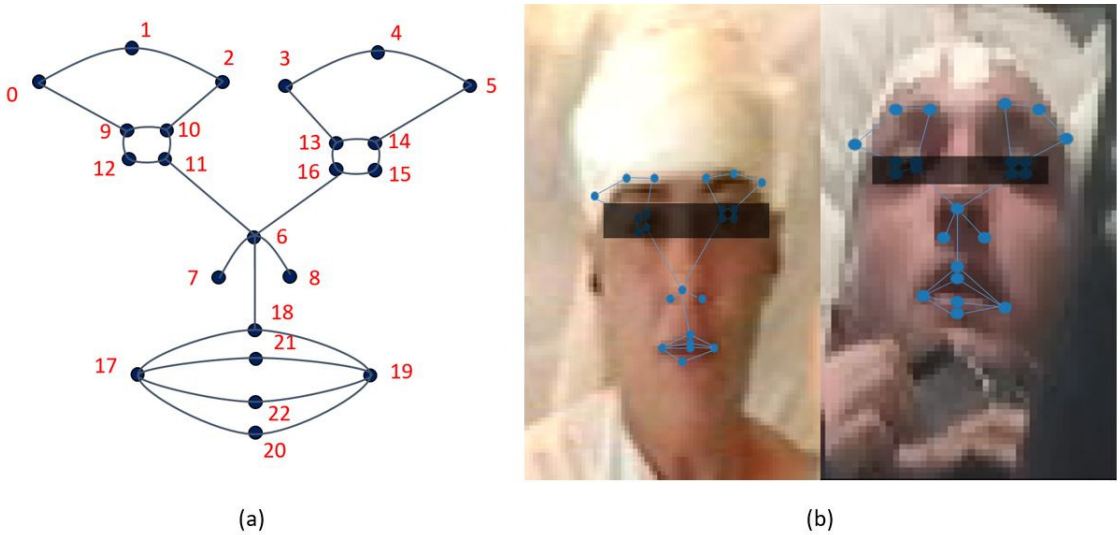


Figure 5.3: (a) Illustration of detected facial landmarks. (b) Samples of facial keypoint detection on our dataset.

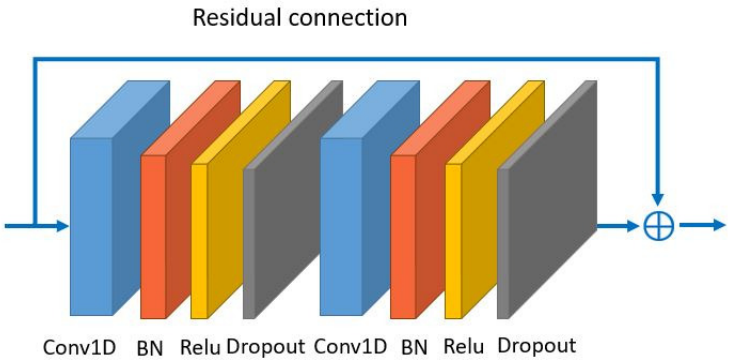


Figure 5.4: Illustration of the temporal convolutional block. Conv1D represents the 1D convolution on the temporal axis, followed by a batch normalization (BN) layer, a ReLU layer, and a Dropout layer. Moreover, a residual connection was added for each block.

Hence for each time step t, the feature represents spatio-temporal information from a video snippet rather than a still image. As for facial feature extraction, we use the last layer output before classification layer of a VGG-19 model [101] pretrained on a public facial expression recognition dataset [102] as

$$u_t = VGG(r_{ft}) \tag{5.2}$$

With the extracted spatio-temporal feature sequence $V = \{v_1, v_2, \cdots, v_T\}$ and facial feature sequence $U = \{u_1, u_2, \cdots, u_T\}$, we feed them into respective temporal convolutional networks for video-level temporal reasoning as the following,

$$c_p = softmax(TCN_p(V)) \tag{5.3}$$

$$c_f = softmax(TCN_f(U)) \tag{5.4}$$

TCN_p and TCN_f represent the TCNs used for the pose and face streams respectively. Both of them are composed by stacks of the temporal convolutional block, as shown in Fig. 5.4, followed by a linear layer as the classification layer. TCN_p and TCN_f are trained separately with standard cross-entropy loss for seizure classification. Later we used these pretrained models as teacher models to distill the learnt knowledge to the Keypoint Streams.

5.2.4 The keypoint stream

In the keypoint streams, we processed the spatio-temporal dynamics of detected keypoints for pose and face with their respective AGCN. The used AGCN is the one proposed in [89],

in which the topology of the graph can be optimized while training for specific tasks. This property hence increases the flexibility of the model for graph construction and brings more generality to adapt to various data samples, such as the highly complex behavioral patterns in our case. For pose stream, we have detected upper-limb keypoint sequence $K_P = \{k_{p1}, k_{p2}, \cdots, k_{pT}\}$, with $k_{pt} \in \mathbb{R}^{C_p \times V_p}$ where C_p and V_p represent the number of channels and joints respectively. With pre-defined adjacency matrix $A_p \in \mathbb{R}^{V_p \times V_p}$, describing the connection relation between the keypoints, we have output logits after $softmax$ operation as

$$o_p = AGCN_p(K_P, A_p) \tag{5.5}$$

Likewise for face stream, we have a facial landmark sequence $K_F = \{k_{f1}, k_{f2}, \cdots, k_{fT}\}$, with $k_{ft} \in \mathbb{R}^{C_f \times V_f}$ where C_f and V_f represent the number of channels and facial landmarks respectively. With adjacency matrix $A_f \in \mathbb{R}^{V_f \times V_f}$, we can have its output likewise by,

$$o_f = AGCN_f(K_f, A_f) \tag{5.6}$$

Instead of computing the cross-entropy for o_p and o_f, we introduced the learned knowledge in the Appearance Streams as addressed in the following part.

5.2.5 Knowledge distillation and ensemble

We have demonstrated how to process the appearance and keypoint information for both pose and face streams. For many multi-stream video analysis cases, it is usual to explicitly combine the learned knowledge from appearance and keypoint sources for a performance boost. Nevertheless, in this work we argue the keypoints should be the main information source for distinguishing seizures. First, we have decent fidelity of the keypoint detection throughout the whole videos. For the appearance stream, on the other hand, occlusions often occur in our dataset and so make the information less reliable. Besides, in medical scenarios like our study cases, privacy and confidentiality are important issues. To align these concepts, the strategy we adopted was to utilize both the appearance and keypoint information while training and only use keypoint information during testing. In addition to the cross-entropy loss, we introduced a standard knowledge distillation mechanism (KD) [103] while training the keypoint streams. It was implemented by minimizing the KL divergence between the probability distributions from the pretrained appearance streams and the keypoint streams. The overall objective losses for pose and face keypoint branches

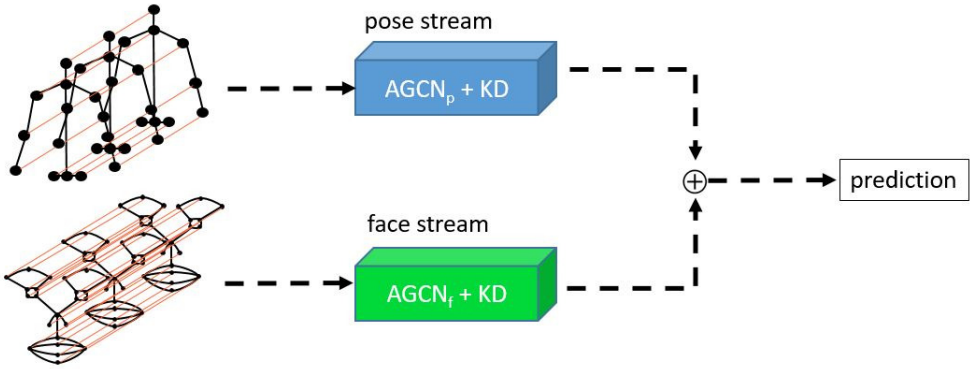


Figure 5.5: Illustration of the ensemble of the prediction from the pose and face streams in the testing phase, with the respective spatio-temporal graphs. The orange line denotes the temporal edges. AGCN+KD denotes AGCN network trained with additional knowledge distillation loss with the global context branches as teachers.

are hence as follows:

$$\begin{aligned} L_{CE,pose} = & -\frac{1}{N}\sum_{i=1}^{N} y^i \cdot log(AGCN_p(K_p^i, A_p)) \\ & + (1-y^i) \cdot log(1 - AGCN_p(K_p^i, A_p)) \end{aligned} \tag{5.7}$$

$$L_{KD,pose} = \frac{1}{N}\sum_{i=1}^{N} D_{KL}(TCN_p(V^i)||AGCN_p(K_p^i, A_p)) \tag{5.8}$$

$$\begin{aligned} L_{CE,face} = & -\frac{1}{N}\sum_{i=1}^{N} y^i \cdot log(AGCN_f(K_f^i, A_f)) \\ & + (1-y^i) \cdot log(1 - AGCN_f(K_f^i, A_f)) \end{aligned} \tag{5.9}$$

$$L_{KD,face} = \frac{1}{N}\sum_{i=1}^{N} D_{KL}(TCN_f(V^i)||AGCN_f(K_f^i, A_f)) \tag{5.10}$$

$$L_{Total,pose} = L_{CE,pose} + \lambda_p L_{KD,pose} \tag{5.11}$$

$$L_{Total,face} = L_{CE,face} + \lambda_f L_{KD,face} \tag{5.12}$$

where $D_{KL}(P||Q) = \sum_j P_j log\frac{P_j}{Q_j}$, denoting the KL divergence. The λ_p and λ_f are trade-off hyper-parameters, and y^i is the label for the i-th example. We train the $AGCN_p$ and $AGCN_f$ separately. For the final prediction, we combined the prediction from pose and face streams for performance improvement, as shown in Fig. 5.5.

Figure 5.6: Seizure examples in a real-world setting during daytime and night.

5.3 Experimentation

5.3.1 Dataset

In this work, we aimed to differentiate between ES and PNES, and tackle the problem in a real-world setting, as in Fig. 5.6, rather than a highly controlled environment. We collected 38 ES videos from 19 patients and 23 PNES videos from 15 patients, resulting in total 61 seizures and 34 patients. These selected seizures are a subset of a larger curated video dataset, as depicted in Chapter 3 and Table 3.1. All patients have been recorded in the Video-EEG Epilepsy unit of the Epileptology department of the Marseille University Hospital. Both ES and PNES were selected according to presence of hyperkinetic motor behavior [104], which involve large amplitude, often explosive whole body movements. Due to the clinical challenges of localizing hyperkinetic ES seizures, and the challenges of discriminating between ES and PNES, this type of semiology is of great interest to neurologists [105, 106, 79]. The duration of the seizures ranged from 15 seconds to 120 seconds. Each patient had at least one and at most 6 recorded seizures. Both day and night conditions were included. All the seizure videos were collected from the video-EEG monitoring unit in the hospital. All patients had a firm diagnosis of either ES or PNES, established by expert epileptologists based on their video-EEG data. Patients gave informed consent for use of video-EEG data. Examples in Fig. 5.6 are from this dataset.

5.3.2 Data preprocessing

All seizure videos were converted to image sequence by 25 fps, and for each video, T frames were equally sampled for analysis. For video frame length shorter than T, the video

itself was concatenated to enough frame length for sampling. In this study, T is set to 300. For image preprocessing, pixel values were normalized to 0 to 1.0, and normalized image size W_p, H_p, W_f, H_f are 112, 112, 48, and 48, respectively. For the 2D spatial coordinates of the detected keypoints, the values of the coordinates were normalized between -1.0 to 1.0 w.r.t the width and height of the cropped region. As for the third dimension in 3D pose estimation, the values were normalized with regards to the maximum and minimum values at the third axis across the video.

5.3.3 Quality of ROI and keypoint detection

As mentioned in section 5.2.2, we fine-tuned the ROI and keypoint detection with manually labeled data in our dataset. For the ROI detection, the intersection-over-union (IoU) is used for quantitative evaluation. The definition of IoU is as formula 5.13, where B_{gt} and B_{pd} represent the bounding box of ground-truth and prediction, respectively. The detection model used reached an average IoU of 0.89 for face detection and 0.94 for patient detection. As for the 2D body joint detection, the keypoint evaluation metric for MS COCO dataset is used. The mean average precision (mAP) at IoU of 0.50 is 0.82. As for facial landmark detection, the model used was not fine-tuned and we visually checked the quality of the results.

$$IoU = \frac{area(B_{gt} \cap B_{pd})}{area(B_{gt} \cup B_{pd})} \tag{5.13}$$

5.3.4 Experimental setup

We conducted both seizure-wise 10-fold cross validation and leave-one-subject-out validation on our datasets. Stochastic gradient descent (SGD) was applied as the learning optimizer. The initial learning rate for either of the four streams was 0.001, with linear learning rate decay scheduling used. The training epochs were set at 50, and we choose the weights at the epoch where the test sets had the highest accuracy for evaluation. The batch size was 4. The hyperparameters λ_p and λ_f are both set as 0.5, and the video snippet length L is 32. The configuration of $AGCN_p$ and $AGCN_f$ were the same as [89]. The kernel size and the dropout rate for both TCN_p and TCN_f are 4 and 0.4.

5.3.5 Experimental results

Table 5.1 shows the F1-score and accuracy of the 10-fold cross validation experiment, where

$$precision = \frac{TP}{TP + FP} \tag{5.14}$$

model	F1-score	accuracy
$AGCN_p$	0.79	0.74
$AGCN_f$	0.78	0.70
TCN_p	0.75	0.69
TCN_f	0.80	0.74
$AGCN_p + KD$	0.86	0.84
$AGCN_f + KD$	0.84	0.82
Ensemble	**0.89**	**0.87**

Table 5.1: The 10-fold cross validation result: comparison of F1-score and accuracy between different models. AGCN+KD denotes AGCN network trained with additional knowledge distillation loss with the appearance streams as teachers.

$$recall = \frac{TP}{TP + FN} \tag{5.15}$$

$$F1\text{–}score = \frac{2 \times precision \times recall}{precision + recall} \tag{5.16}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{5.17}$$

and TP, TN, FP, and FN denote true positive, true negative, false postive, and false negative, respectively. We take ES as a positive case. As shown in Table 5.1, we can see that $AGCN_P$ performs better than TCN_p, indicating that keypoint-based feature is more informative than appearance when correlating body pose to seizure classification. On the other hand, TCN_f slightly outperforms $AGCN_f$, inferring that for seizure analysis based on face, the appearance could provide more characteristic information than facial landmarks. Besides, for both the pose and face streams, we can have significant performance gain by introducing the knowledge distillation on the keypoint branch. This indicates the importance of utilizing complementary information (i.e. from keypoints and appearance) for seizure analysis. Lastly, combining the prediction from pose and face stream with our proposed ensemble method, the performances of the F1-score and the accuracy are 0.89 and 0.87, respectively. This performance improvement shows the effectiveness of integrating multi-stream information. Fig. 5.7 is the receiver operating characteristic (ROC) curve for different models in the 10-fold validation experiment. The ensemble model has the highest value of area under the ROC curve (AUC), indicating the best performance among the models. After the inclusion of knowledge distillation, AUCs of the keypoint branches can gain a significant boost.

Table 5.2 shows the F1-score and accuracy of the leave-one-subject-out validation experiment. We can observe a performance drop compared to the 10-fold validation experiment, possibly due to that the inter-subject variance is considered in the setting and makes the

model	F1-score	accuracy
$AGCN_p$	0.68	0.62
$AGCN_f$	0.68	0.59
TCN_p	0.53	0.56
TCN_f	0.68	0.61
$AGCN_p + KD$	0.74	0.67
$AGCN_f + KD$	0.72	0.66
Ensemble	**0.76**	**0.72**

Table 5.2: The leave-one-subject-out validation result: comparison of F1-score and accuracy between different models. AGCN+KD denotes AGCN network trained with additional knowledge distillation loss with the appearance streams as teachers.

model	F1-score (10-fold)	accuracy (10-fold)
[38]	0.80	0.71
[40](pose)	0.82	0.79
[40](face)	0.75	0.72
model	F1-score (LOSO)	accuracy (LOSO)
[38]	0.64	0.58
[40](pose)	0.70	0.62
[40](face)	0.66	0.61

Table 5.3: We implement the methods in Karácsony et al. [38] and Ahmedt-Aristizabal et al. [40], and test the model in our task. The table shows the results of 10-fold cross validation and leave-one-subject-out (LOSO) validation.

task harder. Otherwise the overall result in Table 5.2 basically indicates the same trend and conclusion as that in the 10-fold cross validation. Besides, we also compare some deep learning based seizure classification studies with ours, as shown in Table 5.3 and Table 5.4. Table 5.3 shows how the methods in the related works performed in our task. Table 5.4 present the results on their own work.

5.4 Recognition of limb dystonia and emotion in epileptic seizures

With our developed method, here we conduct a pilot test on recognizing the presence of limb dystonia and emotion in epileptic seizures. We divided the 19 patients with epileptic seizures in this study into sub-groups based on the presence of limb dystonia or emotion. For patients with limb dystonia, they usually have seizures featuring involuntary and prolonged muscle contractions that result in abnormal postures. The other patients

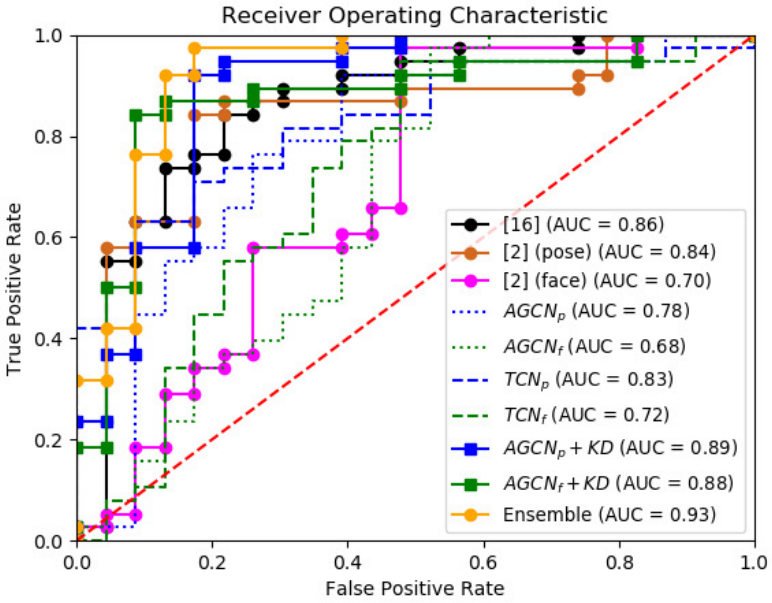


Figure 5.7: The 10-fold cross validation result: the ROC curve for the binary seizure classification task. AGCN+KD denotes AGCN network trained with additional knowledge distillation loss with the appearance streams as teachers.

Method	Classes	Result
A.-Aristizaba et al. (2018) [40]	MTLE ETLE	Average ACC: 0.53-0.56
Maia et al. (2019) [39]	TLE ETLE	AUC: 0.65
Karácsony et al. (2020) [38]	TLE FLE	F1-score: 0.84 AUC: 0.90
Ours	ES PNES	F1-score: 0.89 ACC: 0.87 AUC: 0.93

Table 5.4: Comparison of deep learning-based seizure classification studies. The results shown are based on N-fold cross validation. MTLE, ETLE, and FLE denote mesial temporal lobe epilepsy, extra temporal lobe epilepsy, and frontal lobe epilepsy, respectively.

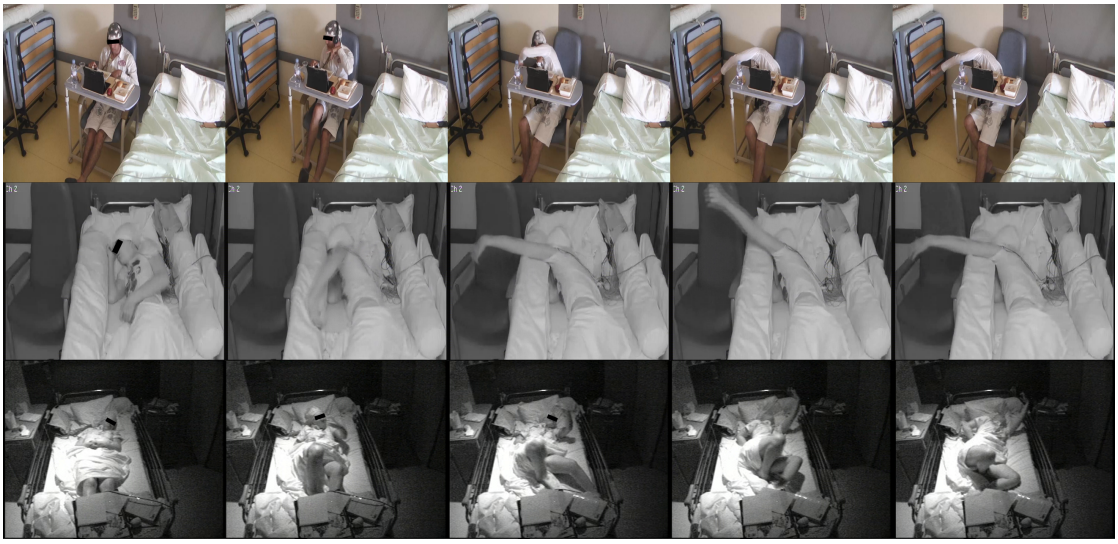


Figure 5.8: Samples of the selected seizures with limb dystonia. Seizures with limb dystonia usually features involuntary and prolonged muscle contractions that result in abnormal postures.

Figure 5.9: Samples of seizures without limb dystonia. The ictal behaviors tend to be more kinetic than those with limb dystonia in general.

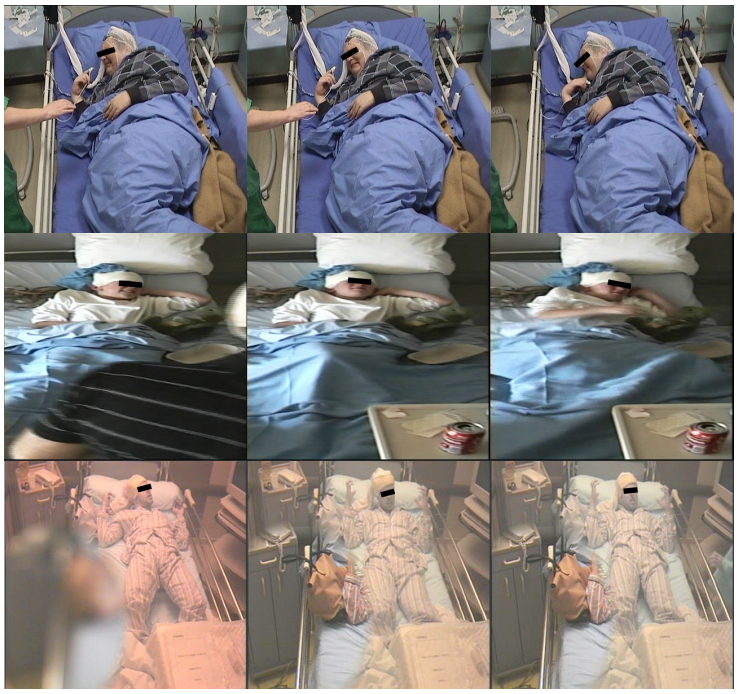


Figure 5.10: Samples of the selected seizures with emotion. Seizures with emotion come with prominent facial expressions, and usually accompany vocal sounds.

Figure 5.11: Samples of the selected seizures without emotion. The selection criteria is based on the presence of less notable facial expressions, or cases where faces are mostly invisible during the seizures.

without the features are then catergorized as a non-dystonia group, where ictal behaviors tend to be more kinetic than those with limb dystonia in general in our cases. Evaluated by experienced clinicians, 9 patients (20 seizures) are viewed as with limb dystonia and the remaining 10 patients (18 seizures) belong to the non-dystonia group. Fig. 5.8 and Fig. 5.9 show some samples of the seizures in the two conditions.

As for the selection critera for seizures with emotion, patients who exhibit prominent facial expressions during the ictal phase are considered as such cases. Those seizures are often accompanied by vocal sounds from the patients as well. Among the 19 patients, 9 patients (15 seizures) are considered as having seizures with emotion, while the other 10 patients (23 seizures) are seen as a non-emotion group. Seizures in the non-emotion group here do not necessarily indicate the absence of emotion, but seizures with milder-to-none emotion expressions, or the insufficiency/unavailability of face observation during the seizures. Fig. 5.10 and Fig. 5.11 show some samples of the seizures in the two conditions.

As shown in Fig. 5.1, our proposed method consists of streams regarding pose and face. To recognize if a seizure has limb dystonia, we adopt the pose stream of the proposed method to conduct a classification task, i.e. *dystonia group v.s. non-dystonia group*. On the other hand, the streams regarding face are used for recognizing if a seizure has emotion involved, i.e. *emotion group v.s. non-emotion group*. Previous deep-learning work on identifying dystonia use neuroimaging data (structural and functional MRI) [107], while we attempt to solely utilize the semiological signs. Besides, in spite of popularity in the research of automated facial emotion recognition [108], its use is aimed at normal conditions and still relatively under-investigated for medical applications, particularly when it comes to vision-based seizure analysis. Given that, we explore our proposed model for semiology-based dystonia and emotion recognition in the same dataset.

We conduct a LOSO validation and use the same experimental setup and configuration as stated in 5.3.4. The results can be seen in Table 5.5 and Table 5.6. For limb dystonia recognition, $AGCN_p$ outperfoms TCN_p, suggesting the pose keypoint can be more informative than the appearance regarding analyzing limb dystonia. With the knowledge distillation from the appearance stream, $AGCN_p$ can get a further boost in the performance, showing the effectiveness of knowledge distillation in this task. As for recognizing the presence of emotion in the epileptic seizures, we can observe TCN_f has the best result, indicating the spatio-temporal appearance features of the face can be more crucial than the facial landmarks for the task.

The automated recognition of limb dystonia or emotion in a seizure event could be of great interest for neurologists, as the presence of these iconic subtype behavior is important for the clinical evalutation. Further we may include more subtle ictal behavior for fine-grained recognition. Besides, our method may suit other clinical applications where behavior-based assessment is crucial, such as attention deficit hyperactivity disor-

der (ADHD) for children or Alzheimer's disease for the elder.

model	F1-score (LOSO)	accuracy (LOSO)
$AGCN_p$	0.78	0.76
TCN_p	0.76	0.73
$AGCN_p + KD$	0.83	0.81

Table 5.5: Results on classifying if seizures have limb dystonia based on our method regarding the body/pose stream.

model	F1-score (LOSO)	accuracy (LOSO)
$AGCN_f$	0.74	0.74
TCN_f	0.84	0.80
$AGCN_f + KD$	0.80	0.78

Table 5.6: Results on classifying if seizures have emotion involved based on our method regarding the face stream.

5.5 Conclusion

In this work, we propose a novel multi-stream framework with knowledge distillation for seizure classification, specifically for distinguishing between ES and PNES with hyperkinetic motor behavior. The contributions are twofold. First, we utilized multi-stream information from keypoint and appearance for both body pose and face streams. From experimental results, we give hints about which type of information should be used based on which stream information is being dealt with for seizure analysis, that is, for analysis based on body pose, keypoint-based features should be considered and for those based on face, appearance information seems more crucial. Second, by introducing a knowledge distillation mechanism, we show the importance of utilizing complementary information for keypoint-based seizure analysis. The performance obtained on real-world data for the challenging task of discriminating epileptic seizures from psychogenic non-epileptic seizures improve the state-of-the-art and are very encouraging with respective F1-score/accuracy 0.89/0.87 for seizure-wise cross validation and 0.75/0.72 for leave-one-subject-out validation.

In addition, we conduct a pilot test on recognizing the presence of limb dystonia and emotion in the same dataset based on our propopsed method. The F1-score/accuracy for limb dystonia and emotion presence recognition can be 0.83/0.81 and 0.84/0.80, respectively. We hope the pilot test can be inspiring for more semiology recognition

research in vision-based seizure analysis.

Chapter 6

A Self-supervised pre-training framework for Seizure Classification

6.1 Introduction

Deep learning has shown its effectiveness in various applications and domains, spanning from computer vision, speech recognition to natural language processing (NLP). Nevertheless, one of the notorious traits in deep learning is the need for large labeled data to train. Yet practically, not every field can acquire large labeled data with ease. For example, for medical applications, it is often expensive to get doctor's data annotation. Thus training a supervised model for medical applications with large data could be very difficult. Recently a popular learning diagram called self-supervised learning (SSL) has come to many research scientists' view scope. SSL uses data itself as its own supervision, and thus needs no external labels for learning. The mainstream SSL framework, i.e. the pretraining-finetuning paradigm, aims at learning effective representations with a large volume of unlabeled data, and take advantages of the learnt knowledge in the pre-trained model for fine-tuning downstream tasks. BERT [33] is one of the iconic SSL model achieving success in NLP, as shown in Fig. 6.1. The BERT model is pre-trained on large unlabeled corpus data, i.e. the wikipedia articles, with several learning objectives, and by simply fine-tuning the pre-trained BERT model, it achieves state-of-the-art results on 11 NLP tasks. Besides, there are also SSL-based research works gaining improvement in computer vision [109] and speech processing domains [110]. In spite of the success of SSL model, it is still an under-explored area for SSL-based model to show its power in medical domains, in particular for the vision-based seizure video analysis. Our work makes the first attempt introducing deep learning-based SSL into the research topic in the literature.

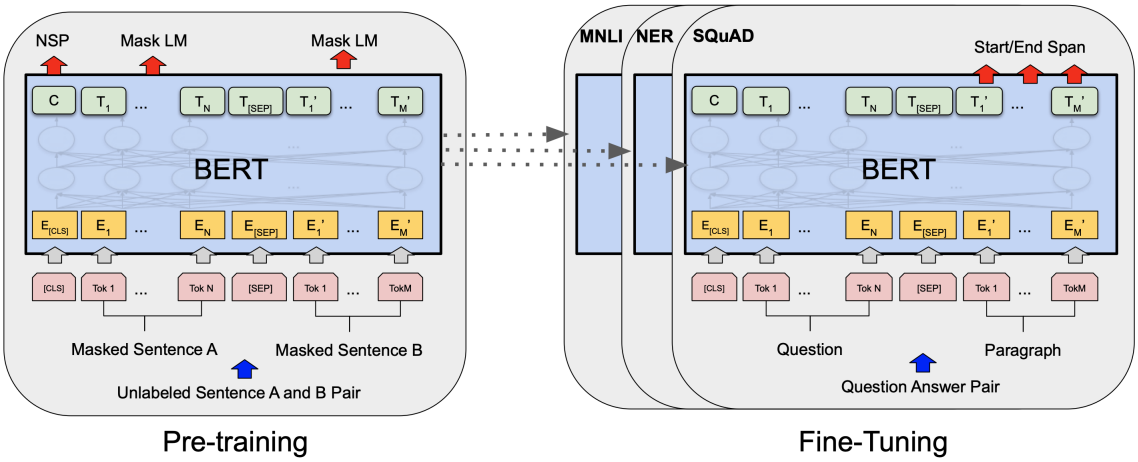


Figure 6.1: The BERT model pre-trains on large unlabeled corpus data with several learning objectives (left), and the pre-trained model is fine-tuned on several NLP downstream tasks (right). The image is adapted from [33].

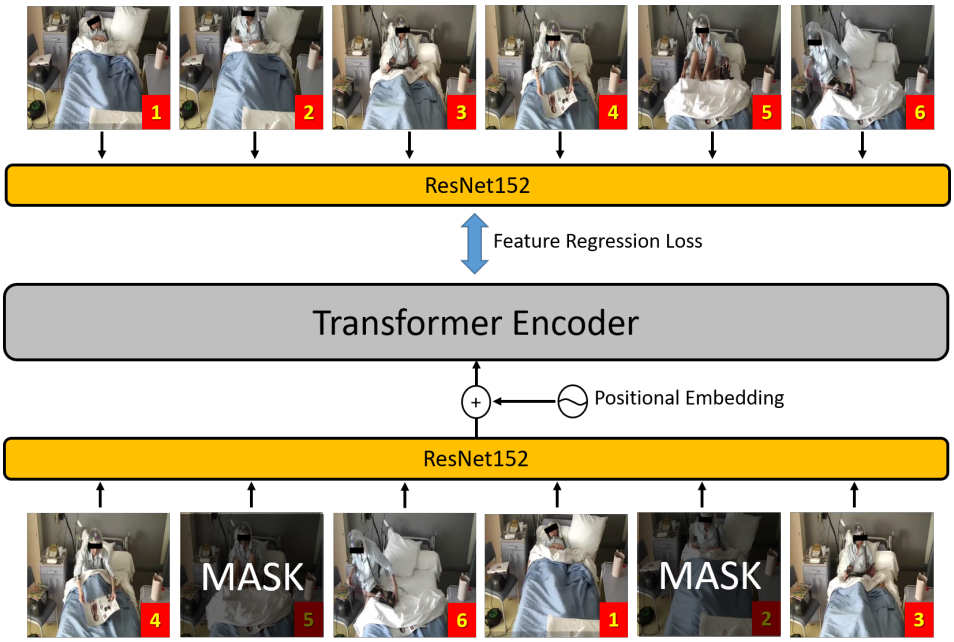


Figure 6.2: SSL-based pretraining on contextual videos: The input sequence is the "noised" version of the target sequence, where random frames are masked out and permutation is applied. We pretrain the encoder of Transformer to reconstruct the corresponding visual features.

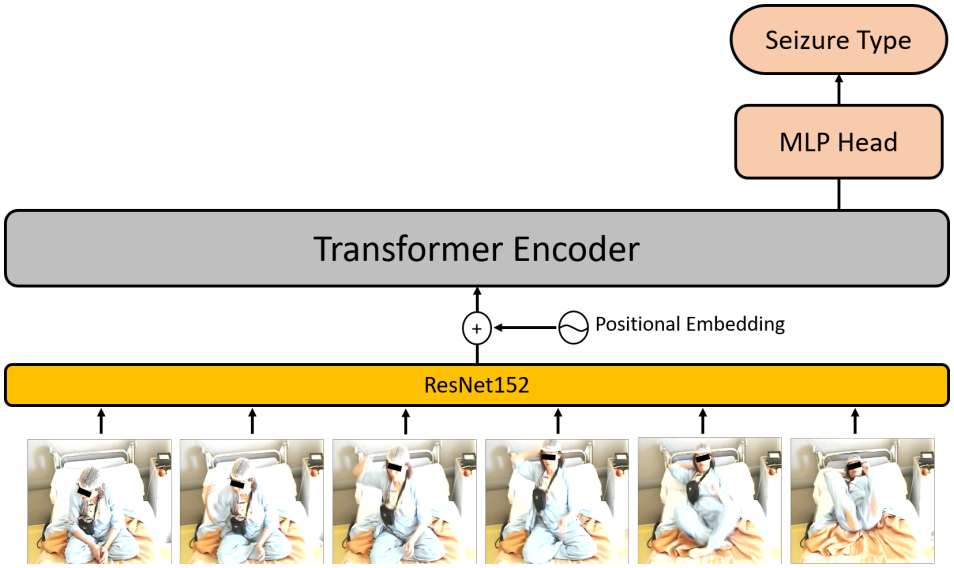


Figure 6.3: Finetuning phase for seizure type classification: In the fine-tuning phase, an uncorrupted seizure video sequence is fed into the pretrained model. A classification layer, i.e. multi-layer perceptron (MLP), is added on top of the pretrained model for the classification task.

6.2 Methodology

Overview of the proposed model

As mentioned in Section 6.1, acquiring large labeled medical data for training neural nets could be practically difficult. The proposed SSL-based method aims at leveraging large volume of unlabeled medical data for learning good representations. Particularly, as shown in Fig. 6.2, our model is pre-trained on voluminous contextual videos without labels. A classification head is added on top of the pre-trained model and the whole model is fine-tuned on the labeled videos for seizure type classification, as shown in Fig. 6.3.

Similar to the BERT model, we adopt the encoder of the Transformer model [32] as our main model architecture. The Transformer model can process sequential data in a parallel form by introducing positional embeddings. In addition, the proposed multi-head attention in the model is another reason that makes the model effective. We will elaborate more about the Transformer model in the following part.

As for the voluminous unlabeled data we use for pre-training, we call them as "contextual videos". Like the labeled seizure videos, the contextual videos are recorded in the EEG-Video monitoring unit. They record daily behaviors of patients and possibly other associated people in the unit. As mentioned in Chapter. 3, the recorded content can be as diverse and natural as those in daily routines, except that seizure onset events are

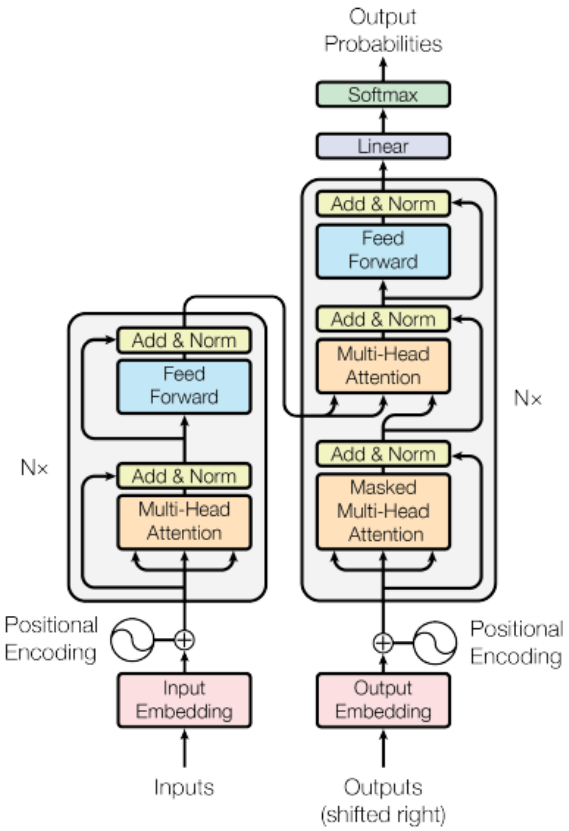


Figure 6.4: Architecture of the Transformer model. The image is adapted from [32].

not included. It is worth mentioning that the large amount of unlabelled data was not patient-specific, i.e. we did not compare non-seizure activities with seizure appearance in individual patients. The idea is to get readily available and abundant data that can provide contextual information in the video-recording environment, and the model can learn good contextual representations from them in a self-supervised way. Then, the learnt representations in the pretrained model are utilized for the downstream task, i.e. the seizure type classification task.

The method needs no complex task-specific design and with just minimal modification of the model, our method can present promising results. In addition, we show an entrypoint of how to take advantage of large, cheap, and unlabeled videos into medical video analysis. To our knowledge, this is the first work utilizing large unlabeled videos to faciliate vision-based seizure video anlaysis. The following parts detail the proposed framework.

The Transformer model and the multi-head attention

Here we give a brief introduction of the Transformer model. The Transformer model

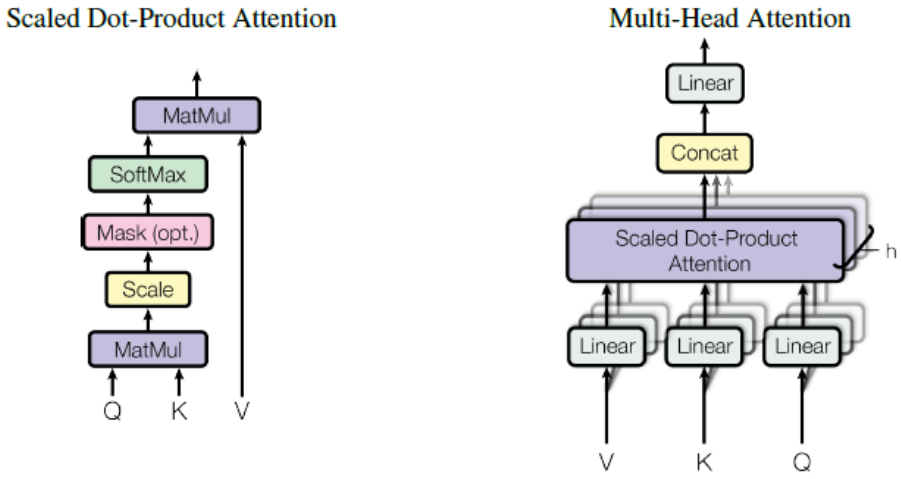


Figure 6.5: (Left) The proposed attention in the Transformer model, and its multi-head version (right). The image is adapted from [32].

shows state-of-the-art results on language translation tasks. The model introduces positional embeddings to avoid the recurrent procedures while dealing with sequential data, and thus be more parallelizable to train. The model is with an encoder-decoder architecture, as shown in Fig. 6.4. The encoder of the Transformer is adopted as the main architecture in our model. Besides, the proposed multi-head attention in the Transformer is another main contribution of the paper. As shown in Fig. 6.5, the proposed scaled dot-product attention is defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6.1}$$

The matrix Q, V, and V denotes packed queries, keys and values, respectively. The input consists of queries and keys of dimension d_k, and values of dimension d_v.
Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions.

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O$$
$$\text{where } head_i = Attention(QW_i{}^Q, KW_i{}^K, VW_i{}^V)$$

Where the projections are parameter matrices $W_i{}^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i{}^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i{}^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$. The encoder and decoder of the Transformer have the same layers of the attention layer. One difference is masked attention is used to prevent leftward information flow in the decoder. The effectiveness of the Transformer

model makes it one of the most widely used models in the mainstream deep learning research. In the following part, we elaborate how we utilize it for our target application.

Pre-training and Fine-tuning the model

Inspired by BART [111], another Transformer-based SSL model for NLP, which corrupts input text with an arbitrary noising function and makes Transformer to reconstruct the original text, we include this concept of denoising objective into our model in the pre-training phase. From the contextual video dataset $\mathbb{D}_c$, for each video $V_c \in \mathbb{D}_c$, we have ordered image sequence as $V_c = (m_c^1, \ldots, m_c^K)$. Two noising functions are applied on V_c. We change the sequence ordering by permuting V_c, and then randomly mask out some frames, resulting in a noised version of V_c, denoted as $\widetilde{V_c}$. The pre-training objective is to regress the Transformer output of each frame in $\widetilde{V_c}$ to the visual features of V_c. The L2 regression loss is formulated as:

$$L(\theta) = E_{v_c \sim \mathbb{D}_c} \sum_{i=1}^{K} \| h_\theta(\widetilde{v_c}^{(i)}) - r(v_c^{(i)}) \|_2^2 \tag{6.2}$$

Where θ is the trainable parameters of the Transformer, and the Transformer output is expressed as h_θ. We take ResNet152 [4] as our CNN backbone to generate visual features. The ResNet152 is pre-trained on ImageNet [66], and we remove the last classification layer to generate a 2048-d feature. We denote it as r as the function for frame descriptor. After pre-training the Transformer on the contextual video dataset $\mathbb{D}_c$ with the defined objective loss as equation 6.2, we add a multilayer perceptron (MLP), i.e. fully-connected (FC) layer, on top of our pre-trained Transformer for classification. Then fine-tune the whole model on the target dataset $\mathbb{D}_s$, which contains seizure videos for seizure type classification. For each seizure video $V_s \in \mathbb{D}_s$, we have an uncorrupted image sequence as input to Transformer as $V_s = (m_s^1, \ldots, m_s^N)$, with the corresponding binary seizure type labels $y_s \in \mathbb{L}$. In the fine-tuning phase, the seizure classification task is optimized based on the standard binary cross-entropy loss as

$$L_{CE} = E_{v_s \sim \mathbb{D}_s}(y_s \cdot log(Softmax(FC(h_\theta(v_s)))) + (1 - y_s) \cdot log(1 - Softmax(FC(h_\theta(v_s))))) \tag{6.3}$$

6.3 Experimentation

In this section, we give the details of the implementation of the experimentation.

Dataset and pre-processing

For pre-training the Transformer model, there are about 13k 10-second clips in the

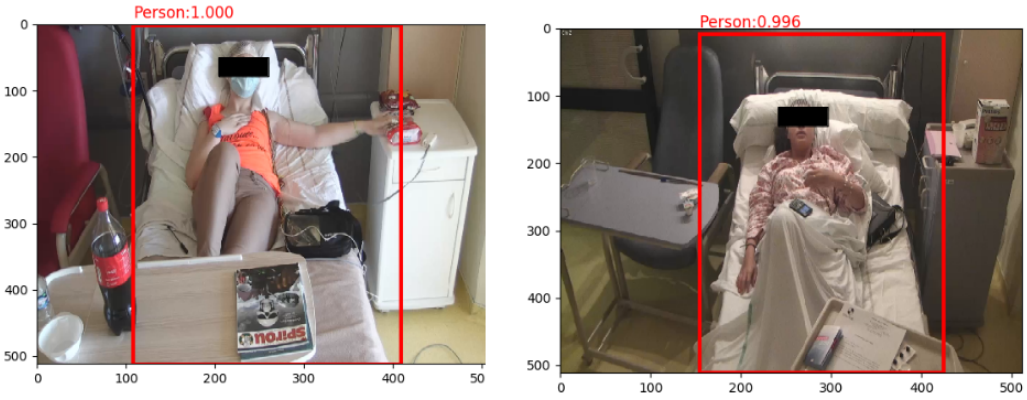


Figure 6.6: Patients in image frames are detected and cropped before feeding into the pre-trained Transformer model for the downstream seizure classification task.

contextual video dataset $\mathbb{D}_c$, resulting in a total 36 hours of clip duration. We convert the clip into image sequence at 25 fps. We resize the frame into a 128 $\times$ 171 dimension, and while generating the training mini-batch, a random crop of 112 $\times$ 112 is applied on the frames.

In the fine-tuning step for seizure type classification, seizure dataset $\mathbb{D}_s$ is used. $\mathbb{D}_s$ covers all the seizure videos specified in Chapter 3. In other words, $\mathbb{D}_s$ contains 283 trimmed seizure videos, and among them, 235 videos belong to ES, and 48 videos are PNES. A total of 81 patients are involved, in which the ES and PNES class has 52 and 29 patients, respectively. The length of seizure videos ranges from 7 seconds to 150 seconds. After converting the seizure videos into image sequence, we detect the target patient with a SSD detector [93] pre-trained on our seizure video dataset, as shown in Fig 6.6. The Intersection over Union (IoU) is 0.89. The cropped region is then resized to a 128 $\times$ 171 dimension, and a center crop of 112 $\times$ 112 is applied on the frames. Normalization of image tensors are implemented by subtracting the mean and divided by the standard deviation across each channel.

Specification of the Transformer

Regarding the specification of the Transformer used in this work, the number of attention head h is 8. Model dimension d_{model} is set as 1024. The maximum position is set as 256. The number of encoder layer is 6. The number of total trainable parameters is about 78M.

Corrupt the input for pre-training

The video sequence as input for the Transformer while pre-training is corrupted, in terms of frame ordering and information masking. The input length of the model is set as

256. A span of consecutive 30 frames are randomly selected and relocated for frame permutation. As for frame masking, following BERT [33], we replace 15% of frames with visual `MASK` tokens. The visual `MASK` and `PAD` tokens are tensors with the shape of image tensors but filled with fixed values of -1.0 and 0.0, respectively.

Training details

We pre-train the Transformer for 60 epochs. The initial learning rate is 0.01, with a linearly decreased scheduler. Weight decay for pre-training is set as 0.0001. The pre-training process takes 670 gpu-hours (roughly 14 hours on 48 V100 gpus across 6 nodes). We adopt AdamW [112] as the optimizer. As for fine-tuning the Transformer, we train it for 50 epochs. Batch size is 16. Except for setting the initial learning rate as 0.005, other training settings are the same as those in the pre-training phase. We test the whole videos by temporally averaging the predictive results. A dropout rate of 0.5 in the final classifier layer is set. In addition, to mitigate the imbalanced dataset, a class weight (reciprocal of the number of class clips) is added in the cross entropy loss. The implementation of the Transformer model is based on the Huggingface library [113].

Experimental results

We perform a leave-one-subject-out (LOSO) validation. The F1-score and the accuracy are 0.82 and 0.75, respectively. As shown in Table 6.1, our results are comparable to other state-of-the-art seizure classification tasks given different class targets. For ES and PNES classificaiton, our method outperforms the best results proposed by the approach in Chapter 5. It is worth mentioning the dataset used in Chapter 5 is the subset of the one used in this section, and the method involves information from multi-streams. This indicates our proposed Transformer-based pre-training approach can learn robust and generalizable features for the downstream task. The video-wise confusion matrix is shown in Table 6.2.

6.4 Conclusion

In this study, we propose a Transformer-based self-supervised pre-training framework for learning features suitable for the downstream task, i.e. classifying ES and PNES videos. The paradigm aligns with the research direction of self-supervised pre-training that takes advantage of large unannotated data and learns useful representations from it for downstream tasks. This may be especially favored for medical applications where data annotations are usually costly. In our work, a Transformer-based model is pre-trained on a large volume of contextual videos with denoising pre-training objectives. By simply fine-tuning the pre-trained model with a minimum model modification, the experimental classification results can compete with methods from other state-of-the-art works for

Method	Classes	Result
A.-Aristizaba et al. (2018) [40]	MTLE ETLE	Average ACC: 0.53-0.56
Maia et al. (2019) [39]	TLE ETLE	ACC: 0.83
Karácsony et al. (2020) [38]	TLE FLE	F1-score: 0.84
Methods in §5	ES PNES	F1-score: 0.76 ACC: 0.72
Ours	ES PNES	F1-score: 0.82 ACC: 0.75

Table 6.1: Comparison of deep learning-based seizure classification studies. Our results can compete to other state-of-the-art seizure classification tasks with different class targets. For ES and PNES classification, our method outperforms the best results proposed by the approach in Chapter 5. MTLE, ETLE, and FLE denote mesial temporal lobe epilepsy, extra temporal lobe epilepsy, and frontal lobe epilepsy, respectively.

	(predict) ES	(predict) PNES
(actual) ES	181	54
(actual) PNES	15	33

Table 6.2: Confusion matrix of the video-wise classification results by leave-one-subject-out validation.

similar tasks. To our knowledge, this is the first deep learning work exploiting large unlabeled data for facilitating vision-based seizure analysis. We hope our study can inspire the research community regarding seizure video analysis to rethink how we can benefit from large unannotated data.

Chapter 7

Conclusion and Future Work

In this thesis we have proposed and evaluated several methods for video-based seizure analysis. The proposed methods utilize some of the latest concepts and architectures in the current deep learning research community. Our experiments demonstrated encouraging results compared to existing methods applied on vision-based seizure video analysis. We conclude our work by pointing out key contributions (section 7.1) and discuss short and long-term perspectives of our work (section 7.2).

7.1 Key Contributions

- **Curation of a large-scale seizure video dataset** - In collaboration with the Epileptology department in the Marseille University Hospital, we managed to build a large seizure video dataset aiming to automate semiology analysis. In our dataset, there are 283 seizure events in total and 81 patients involved. We have 235 epileptic seizures (52 patients) and 48 psychogenic non-epileptic seizures (29 patients). We are particulaly interested in hyperkinetic seizures, as the semiology is often complex yet charateristic. Among the 235 epileptic seizures, 101 seizures are regarded as hyperkinetic ones. As for psychogenic non-epileptic seizures, all of the collected ones are hyperkinetic. The video recording conditions are unconstrained and thus suitable for developing automated methods for **analyzing real-world seizure cases**. To our knowledge, our curated video dataset is the largest one by far in the vision-based seizure video analysis literature. In addition, we have high-quality detection results for all patients' body and head, and upper-body limbs for some of the seizures. The detection information can faciliate further research based on this dataset.

- **Head trajectory analysis for hyperkinetic seizures** - We proposed a simple workflow to analyze the head trajectory of 5 hyperkinetic epileptic seizures. The head of

the patients in each frame of the video was first detected with the Single Shot Multibox Detector (SSD) network [65], and the trajectory of the head motion can be obtained. In particular, the analyzed seizures exhibit charateristic antero-posterior rocking movement, resulting in cyclic partterns in the head trajectory. The trajectories are then denoised by Empirical Mode Decomposition (EMD) through dropping high frequency intrinsic mode functions (IMFs). Time-evolving frequency are obtained with the moving-averaged reciprocal of peak-peak duration in the trajectories. Our results confirmed stable frequency throughout rocking sequences in the prefrontal seizures, suggesting a mechanism involving intrinsic oscillatory generators. The results can be a basis for further spectral investigation along with EEG signal [41, 42].

- **A Multi-Stream Framework for Seizure Classification** - We proposed a multi-stream framework for semiology-based seizure analysis. The proposed deep-learning method utillzes information from keypoints and appearance, from both face and body pose. We use Graphical Neural Networks (GNN) to handle the keypoint features, by treating the detected keypoints as a graph. As for the appearance stream, CNN-based spatio-temporal and facial features are utilized for representing the body and facial parts. The features are then fed into its Temporal Convolutional Networks (TCN) for classification. Besides, knowledge distillation from appearance to keypoint stream is introduced for regulating the model learning. Two tasks are explored for the proposed method: epileptic/non-epileptic seizure classification, and recognition of limb dystonia and emotion in epileptic seizures. For the first task, experimental results show our best method can outperform other existing methods used in the literature regarding seizure video anlaysis. As for the second task, it demonstrated encouraging results for subtype semiology recognition, indicating the generalizability of our method for the related tasks.

- **A Self-supervised framework for Seizure Classification** - We proposed a pretraining-finetuning paradigm for seizure video analysis based on the widely used Transformer model. The Transformer was pre-trained with a denoising objective to reconstruct the correct feature on the pre-training data. This aims at learning contextual representation that is generalizable for downstream tasks. In this study, the downstream task is to discriminate between ES and PNES seizures on the full scale of data we collected. An additional classification layer is added on top of the pre-trained Transformer for classification. By simply fine-tuning the pre-trained model, the experimental results show promising results compared to related vision-based seizure classification works.

  The pre-training data we used is the contexutal videos recorded in the EMU where

no seizure events are involved. Utilizing these unlabelled, easily-accessible data is important for medical applications, as a large amount of annotation from doctors are usually costly or unavailable. This research direction aligns with the theme of unsupervised and self-supervised learning that use data itself as its own supervision. To our knowledge, in the context of video-based seizure analysis research, our work is the first attempt to take advantage of large-scale unlabelled data with self-supervised pre-training.

7.2 Future Work

7.2.1 Short-term Perspectives

- **Generalizing the pre-trained model for other seizure-related tasks** - We have pre-trained a Transformer model with a large amount of unlabeled video data. It has shown that the learned contextual representations can be used for seizure classification task. A simple and expected extension of the model usage could be utilization for other vision-based seizure-related tasks, such as seizure detection or fine-grained semiology recognition.

- **Improve the video dataset** - Data is the fuel for deep learning, and there is never too much of them. A continual improvement of our dataset, in terms of quality and quantity, can be important for the seizure analysis research community. Thanks to the extensive experiences and patient cases with SEEG monitoring in the Marseille University Hospital, a conceivable next step is to add the brain region category for epileptic seizures. It may allow a new research aspect. Besides, more new and retrospective cases will be considered for collection once the semiology of the seizure are of interest.

7.2.2 Long-term Perspectives

- **Improve model interpretability via vision-and-language learning** - For AI in medical applications, doctors would like to know the reason why a model gives such predictions. Nevertheless, highly explainable models, such as Decision Trees, usually give poorer performance than those with low interpretability, such as deep neural network models. The trade of between performance and interpretability of machine learning models in healthcare applications is still an open topic. Given the recent success of vision-and-language research with Transformer [114, 115, 116], it makes us rethink if we can improve the model interpretability via text/language. Language accounts for a large portion of how humans communicate. It would be interesting

if we can have a model that can provide both a prediction (e.g. seizure type) and an explanation (e.g. clinical description of a seizure). The model then would not be a pure black-box for doctors, and doctors could have an idea of how data are perceived and how predictions are made by the model.

- **Multi-modal self-supervised pre-training with EEG/neuroimaging data** - In this thesis we propose a self-supervised pre-training paradigm solely based on video data for vision-based seizure video analysis. It is natural to think if we can include other accessible medical data as a multi-modal framework. Data from different modalities usually contain complementary information, and thus have potential to improve the overall performance. In the context of seizure analysis, along with the video data, it would be interesting to inlcude auxiliary data such as EEG or neuroimages for multi-modal self-supervised pre-training. The learned contextual multi-modal features could be useful for downstream tasks where multi-modal data is involved, e.g. predicting seizure type with both video and EEG/neuroimages.

- **Applications for other behavioral disorders** - One advantage of our proposed methods in this thesis is that they are not limited for seizure anlaysis. They also can be suitable for other video-based behavioral disorder analysis, such as Alzheimer's disease for the elder or ADHD for the children. These disorders may contain characteristic behaviors, which form a basis for correct diagnosis by clinicians. Yet subtle changes or features may be hard to recognize sometimes, and we think our proposed methods have potentials to be helpful regarding general behavioral disorder analysis based on vision information.

Bibliography

[1] R. S. Fisher, W. van Emde Boas, W. Blume, C. Elger, P. Genton, P. Lee, and J. Engel, "Epileptic seizures and epilepsy: Definitions proposed by the international league against epilepsy (ILAE) and the international bureau for epilepsy (IBE)," *Epilepsia*, vol. 46, pp. 470–472, Apr. 2005. (Cited on page 1.)

[2] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986. (Cited on page 10.)

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015. (Cited on page 10.)

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, June 2016. (Cited on pages 10 and 64.)

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, pp. 2278–2324, 1998. (Cited on pages ix and 11.)

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, Nov. 1997. (Cited on page 10.)

[7] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014. cite arxiv:1412.3555Comment: Presented in NIPS 2014 Deep Learning and Representation Learning Workshop. (Cited on page 10.)

[8] A. Graves, A. rahman Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, May 2013. (Cited on page 10.)

[9] "Understanding lstm networks." https://colah.github.io/posts/2015-08-Understanding-LSTMs/. (Cited on pages ix and 11.)

[10] "Building autoencoders in keras." https://blog.keras.io/building-autoencoders-in-keras.html. (Cited on pages ix and 12.)

[11] "Understanding the hype around transformer nlp models." https://blog.dataiku.com/decoding-nlp-attention-mechanisms-to-understand-transformer-modelsl. (Cited on pages ix and 12.)

[12] M. Pediaditis, M. Tsiknakis, and N. Leitgeb, "Vision-based motion detection, analysis and recognition of epileptic seizures—a systematic review," *Computer Methods and Programs in Biomedicine*, vol. 108, pp. 1133–1148, Dec. 2012. (Cited on page 12.)

[13] H. Joo, S.-H. Han, J. Lee, D. Jang, J. Kang, and J. Woo, "Spectral analysis of acceleration data for detection of generalized tonic-clonic seizures," *Sensors*, vol. 17, p. 481, Feb. 2017. (Cited on page 13.)

[14] Z. Li, A. da Silva, and J. Cunha, "Movement quantification in epileptic seizures: a new approach to video-EEG analysis," *IEEE Transactions on Biomedical Engineering*, vol. 49, pp. 565–573, June 2002. (Cited on pages ix, 13 and 14.)

[15] H. Lu, Y. Pan, B. Mandal, H.-L. Eng, C. Guan, and D. W. S. Chan, "Quantifying limb movements in epileptic seizures through color-based video analysis," *IEEE Transactions on Biomedical Engineering*, vol. 60, pp. 461–469, Feb. 2013. (Cited on pages ix, 13 and 14.)

[16] M. Pediaditis, M. Tsiknakis, L. Koumakis, M. Karachaliou, S. Voutoufianakis, and P. Vorgia, "Vision-based absence seizure detection," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, Aug. 2012. (Cited on page 13.)

[17] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, pp. 363–370, Springer Berlin Heidelberg, 2003. (Cited on page 13.)

[18] S. L. Salzberg, "C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993," *Machine Learning*, vol. 16, pp. 235–240, Sept. 1994. (Cited on page 13.)

[19] P. Maurel, A. McGonigal, R. Keriven, and P. Chauvel, "3d model fitting for facial expression analysis under uncontrolled imaging conditions," in *2008 19th Interna-*

tional Conference on Pattern Recognition, IEEE, Dec. 2008. (Cited on pages ix, 13, 14 and 41.)

[20] N. Karayiannis and G. Tao, "Extraction of temporal motion velocity signals from video recordings of neonatal seizures by optical flow methods," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)*, IEEE. (Cited on page 13.)

[21] N. B. Karayiannis, G. Tao, J. D. Frost, M. S. Wise, R. A. Hrachovy, and E. M. Mizrahi, "Automated detection of videotaped neonatal seizures based on motion segmentation methods," *Clinical Neurophysiology*, vol. 117, pp. 1585–1594, July 2006. (Cited on page 13.)

[22] K. Cuppens, L. Lagae, B. Ceulemans, S. V. Huffel, and B. Vanrumste, "Automatic video detection of body movement during sleep based on optical flow in pediatric patients with epilepsy," *Medical & Biological Engineering & Computing*, vol. 48, pp. 923–931, June 2010. (Cited on pages ix, 13 and 14.)

[23] K. Cuppens, C.-W. Chen, K. B.-Y. Wong, A. V. de Vel, L. Lagae, B. Ceulemans, T. Tuytelaars, S. V. Huffel, B. Vanrumste, and H. Aghajan, "Using spatio-temporal interest points (STIP) for myoclonic jerk detection in nocturnal video," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, Aug. 2012. (Cited on pages ix, 13 and 14.)

[24] B. Mandal, H.-L. Eng, H. Lu, D. W. S. Chan, and Y.-L. Ng, "Non-intrusive head movement analysis of videotaped seizures of epileptic origin," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, Aug. 2012. (Cited on page 13.)

[25] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, pp. 24–29, Jan. 2019. (Cited on page 13.)

[26] F. Achilles, F. Tombari, V. Belagiannis, A. M. Loesch, S. Noachtar, and N. Navab, "Convolutional neural networks for real-time epileptic seizure detection," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, pp. 264–269, July 2016. (Cited on pages ix, 13, 15 and 27.)

[27] D. Ahmedt-Aristizabal, C. Fookes, K. Nguyen, S. Denman, S. Sridharan, and S. Dionisio, "Deep facial analysis: A new phase i epilepsy evaluation using computer vision," *Epilepsy & Behavior*, vol. 82, pp. 17–24, May 2018. (Cited on pages x and 15.)

[28] F. Achilles, A.-E. Ichim, H. Coskun, F. Tombari, S. Noachtar, and N. Navab, "Patient MoCap: Human pose estimation under blanket occlusion for hospital monitoring applications," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pp. 491–499, Springer International Publishing, 2016. (Cited on pages x and 16.)

[29] S. Liu and S. Ostadabbas, "Seeing under the cover: A physics guided learning approach for in-bed pose estimation," in *Lecture Notes in Computer Science*, pp. 236–245, Springer International Publishing, 2019. (Cited on page 16.)

[30] "SLP dataset for multimodal in-bed pose estimation." https://web.northeastern.edu/ostadabbas/2019/06/27/multimodal-in-bed-pose-estimation/. (Cited on pages x, 16 and 17.)

[31] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020. (Cited on page 18.)

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017. (Cited on pages xii, 18, 61, 62 and 63.)

[33] 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, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019. (Cited on pages xii, 18, 27, 59, 60 and 66.)

[34] "Labelimg: a graphical image annotation tool." https://github.com/tzutalin/labelImg. (Cited on page 24.)

[35] "Visipedia: Annotation toolkit for editing keypoints and bounding boxes.." https://github.com/tzutalin/labelImg. (Cited on page 24.)

[36] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. DollÃir, "Microsoft coco: Common objects in context," in *2014 ECCV*, 2014. (Cited on pages 24 and 43.)

[37] D. Ahmedt-Aristizabal, K. Nguyen, S. Denman, S. Sridharan, S. Dionisio, and C. Fookes, "Deep motion analysis for epileptic seizure classification," in *2018 40th*

Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, July 2018. (Cited on pages 27, 41 and 42.)

[38] T. Karacsony, A. M. Loesch-Biffar, C. Vollmar, S. Noachtar, and J. P. S. Cunha, "A deep learning architecture for epileptic seizure classification based on object and action recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2020. (Cited on pages xiii, 27, 41, 42, 51, 52 and 67.)

[39] P. Maia, E. Hartl, C. Vollmar, S. Noachtar, and J. P. S. Cunha, "Epileptic seizure classification using the NeuroMov database," in *2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*, IEEE, Feb. 2019. (Cited on pages 27, 52 and 67.)

[40] D. Ahmedt-Aristizabal, C. Fookes, S. Denman, K. Nguyen, T. Fernando, S. Sridharan, and S. Dionisio, "A hierarchical multimodal system for motion analysis in patients with epilepsy," *Epilepsy & Behavior*, vol. 87, pp. 46–58, Oct. 2018. (Cited on pages xiii, 27, 41, 51, 52 and 67.)

[41] J.-C. Hou, M. Thonnat, R. Huys, F. Bartolomei, and A. McGonigal, "Rhythmic rocking stereotypies in frontal lobe seizures: A quantified video study," *Neurophysiologie Clinique*, vol. 50, pp. 75–80, Apr. 2020. (Cited on pages 29 and 70.)

[42] A. Zalta, J.-C. Hou, M. Thonnat, F. Bartolomei, B. Morillon, and A. McGonigal, "Neural correlates of rhythmic rocking in prefrontal seizures," *Neurophysiologie Clinique*, vol. 50, pp. 331–338, Oct. 2020. (Cited on pages 29, 39 and 70.)

[43] C. R. Gallistel, *The organization of action : a new synthesis*. Hillsdale, N.J. New York: L. Erlbaum Associates Distributed by Halsted Press, 1980. (Cited on page 29.)

[44] S. Grillner, "Biological pattern generation: The cellular and computational logic of networks in motion," *Neuron*, vol. 52, pp. 751–766, Dec. 2006. (Cited on page 29.)

[45] E. Marder and D. Bucher, "Central pattern generators and the control of rhythmic movements," *Current Biology*, vol. 11, pp. R986–R996, Nov. 2001. (Cited on page 29.)

[46] R. Manni and M. Terzaghi, "Rhythmic movements during sleep: a physiological and pathological profile," *Neurological Sciences*, vol. 26, pp. s181–s185, Dec. 2005. (Cited on pages 29 and 36.)

[47] G. Deuschl, P. Bain, and M. Brin, "Consensus statement of the movement disorder society on tremor," *Movement Disorders*, vol. 13, pp. 2–23, Oct. 2008. (Cited on page 29.)

[48] M. J. Edwards, A. E. Lang, and K. P. Bhatia, "Stereotypies: A critical appraisal and suggestion of a clinically useful definition," *Movement Disorders*, vol. 27, pp. 179–185, Dec. 2011. (Cited on pages 29 and 38.)

[49] C. Tassinari, G. Cantalupo, B. Högl, P. Cortelli, L. Tassi, S. Francione, L. Nobili, S. Meletti, G. Rubboli, and E. Gardella, "Neuroethological approach to frontolimbic epileptic seizures and parasomnias: The same central pattern generators for the same behaviours," *Revue Neurologique*, vol. 165, pp. 762–768, Oct. 2009. (Cited on page 29.)

[50] A. McGonigal and P. Chauvel, "Prefrontal seizures manifesting as motor stereotypies," *Movement Disorders*, vol. 29, pp. 1181–1185, Oct. 2013. (Cited on pages 29 and 38.)

[51] P. Chauvel and A. McGonigal, "Emergence of semiology in epileptic seizures," *Epilepsy & Behavior*, vol. 38, pp. 94–103, Sept. 2014. (Cited on pages 29 and 38.)

[52] H. Gastaut, *Epileptic seizures; clinical and electrographic features, diagnosis and treatment*. Springfield, Ill: Thomas, 1972. (Cited on page 29.)

[53] A. Marchi, B. Giusiano, M. King, S. Lagarde, A. Trébuchon-Dafonseca, C. Bernard, S. Rheims, F. Bartolomei, and A. McGonigal, "Postictal electroencephalographic (EEG) suppression: A stereo-EEG study of 100 focal to bilateral tonic-clonic seizures," *Epilepsia*, vol. 60, pp. 63–73, Nov. 2018. (Cited on page 29.)

[54] S. Rheims, P. Ryvlin, C. Scherer, L. Minotti, D. Hoffmann, M. Guenot, F. Mauguière, A.-L. Benabid, and P. Kahane, "Analysis of clinical patterns and underlying epileptogenic zones of hypermotor seizures," *Epilepsia*, vol. 49, pp. 2030–2040, Dec. 2008. (Cited on page 30.)

[55] F. Bonini, A. McGonigal, A. Trébuchon, M. Gavaret, F. Bartolomei, B. Giusiano, and P. Chauvel, "Frontal lobe seizures: From clinical semiology to localization," *Epilepsia*, vol. 55, pp. 264–277, Dec. 2013. (Cited on pages 30 and 38.)

[56] J. Aupy, I. Noviawaty, B. Krishnan, P. Suwankpakdee, J. Bulacio, J. Gonzalez-Martinez, I. Najm, and P. Chauvel, "Insulo-opercular cortex generates oroalimentary automatisms in temporal seizures," *Epilepsia*, vol. 59, pp. 583–594, Feb. 2018. (Cited on pages 30 and 38.)

[57] S. Meletti, G. Cantalupo, L. Volpi, G. Rubboli, A. Magaudda, and C. A. Tassinari, "Rhythmic teeth grinding induced by temporal lobe seizures," *Neurology*, vol. 62, pp. 2306–2309, June 2004. (Cited on page 30.)

[58] M. do Carmo Vilas-Boas and J. P. S. Cunha, "Movement quantification in neurological diseases: Methods and applications," *IEEE Reviews in Biomedical Engineering*, vol. 9, pp. 15–31, 2016. (Cited on page 30.)

[59] B. Abbasi and D. M. Goldenholz, "Machine learning applications in epilepsy," *Epilepsia*, vol. 60, pp. 2037–2047, Sept. 2019. (Cited on page 30.)

[60] D. Ahmedt-Aristizabal, C. Fookes, S. Dionisio, K. Nguyen, J. P. S. Cunha, and S. Sridharan, "Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: A focused survey," *Epilepsia*, vol. 58, pp. 1817–1831, Oct. 2017. (Cited on page 30.)

[61] J. P. S. Cunha, H. M. P. Choupina, A. P. Rocha, J. M. Fernandes, F. Achilles, A. M. Loesch, C. Vollmar, E. Hartl, and S. Noachtar, "NeuroKinect: A novel low-cost 3dvideo-EEG system for epileptic seizure motion quantification," *PLOS ONE*, vol. 11, p. e0145669, Jan. 2016. (Cited on pages 30 and 39.)

[62] J. P. S. Cunha, L. M. Paula, V. F. Bento, C. Bilgin, E. Dias, and S. Noachtar, "Movement quantification in epileptic seizures: A feasibility study for a new 3d approach," *Medical Engineering & Physics*, vol. 34, pp. 938–945, Sept. 2012. (Cited on page 30.)

[63] J. Rémi, J. P. S. Cunha, C. Vollmar, Özgür Bilgin Topçuoğlu, A. Meier, S. Ulowetz, P. Beleza, and S. Noachtar, "Quantitative movement analysis differentiates focal seizures characterized by automatisms," *Epilepsy & Behavior*, vol. 20, pp. 642–647, Apr. 2011. (Cited on page 30.)

[64] U. Großekathöfer, N. V. Manyakov, V. Mihajlović, G. Pandina, A. Skalkin, S. Ness, A. Bangerter, and M. S. Goodwin, "Automated detection of stereotypical motor movements in autism spectrum disorder using recurrence quantification analysis," *Frontiers in Neuroinformatics*, vol. 11, Feb. 2017. (Cited on page 30.)

[65] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision – ECCV 2016*, pp. 21–37, Springer International Publishing, 2016. (Cited on pages 31 and 70.)

[66] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009. (Cited on pages 31 and 64.)

[67] A. Zeiler, R. Faltermeier, I. R. Keck, A. M. Tome, C. G. Puntonet, and E. W. Lang, "Empirical mode decomposition - an introduction," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, IEEE, July 2010. (Cited on page 31.)

[68] H. S. Singer, "Motor control, habits, complex motor stereotypies, and tourette syndrome," *Annals of the New York Academy of Sciences*, vol. 1304, pp. 22–31, Oct. 2013. (Cited on page 36.)

[69] G. MAYER, J. WILDE-FRENZ, and B. KURELLA, "Sleep related rhythmic movement disorder revisited," *Journal of Sleep Research*, vol. 16, pp. 110–116, Mar. 2007. (Cited on page 36.)

[70] I. Lovaas, C. Newsom, and C. Hickman, "SELF-STIMULATORY BEHAVIOR AND PERCEPTUAL REINFORCEMENT," *Journal of Applied Behavior Analysis*, vol. 20, pp. 45–68, Mar. 1987. (Cited on page 36.)

[71] E. Thelen, "Kicking, rocking, and waving: Contextual analysis of rhythmical stereotypies in normal human infants," *Animal Behaviour*, vol. 29, pp. 3–11, Feb. 1981. (Cited on pages 36 and 38.)

[72] A. M. Graybiel, "Habits, rituals, and the evaluative brain," *Annual Review of Neuroscience*, vol. 31, pp. 359–387, July 2008. (Cited on page 36.)

[73] M. H. Lewis, A. A. Baumeister, and R. B. Mailman, "A NEUROBIOLOGICAL ALTERNATIVE TO THE PERCEPTUAL REINFORCEMENT HYPOTHESIS OF STEREOTYPED BEHAVIOR: A COMMENTARY ON "SELF-STIMULATORY BEHAVIOR AND PERCEPTUAL REINFORCEMENT"," *Journal of Applied Behavior Analysis*, vol. 20, pp. 253–258, Sept. 1987. (Cited on page 36.)

[74] N. G. Hatsopoulos, "Coupling the neural and physical dynamics in rhythmic movements," *Neural Computation*, vol. 8, pp. 567–581, Apr. 1996. (Cited on page 36.)

[75] B. Morillon, L. H. Arnal, C. E. Schroeder, and A. Keitel, "Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception," *Neuroscience & Biobehavioral Reviews*, vol. 107, pp. 136–142, Dec. 2019. (Cited on page 36.)

[76] N. Yamada, "Nature of variability in rhythmical movement," *Human Movement Science*, vol. 14, pp. 371–384, Oct. 1995. (Cited on page 38.)

[77] J. H. Hollman, F. M. Kovash, J. J. Kubik, and R. A. Linbo, "Age-related differences in spatiotemporal markers of gait stability during dual task walking," *Gait & Posture*, vol. 26, pp. 113–119, June 2007. (Cited on page 38.)

[78] M. J. Richardson, K. L. Marsh, R. W. Isenhower, J. R. Goodman, and R. Schmidt, "Rocking together: Dynamics of intentional and unintentional interpersonal coor-

dination," *Human Movement Science*, vol. 26, pp. 867–891, Dec. 2007. (Cited on page 38.)

[79] L. Vaugier, A. McGonigal, S. Lagarde, A. Trébuchon, W. Szurhaj, P. Derambure, and F. Bartolomei, "Hyperkinetic motor seizures: a common semiology generated by two different cortical seizure origins," *Epileptic Disorders*, vol. 19, pp. 362–366, Sept. 2017. (Cited on pages 38 and 48.)

[80] J. SCHMAHMANN and D. PANDYA, "Disconnection syndromes of basal ganglia, thalamus, and cerebrocerebellar systems," *Cortex*, vol. 44, pp. 1037–1066, Sept. 2008. (Cited on page 38.)

[81] J. de Vries, G. Visser, and H. Prechtl, "The emergence of fetal behaviour. i. qualitative aspects," *Early Human Development*, vol. 7, pp. 301–322, Dec. 1982. (Cited on page 38.)

[82] R. GAYMER, "New method of locomotion in limbless terrestrial vertebrates," *Nature*, vol. 234, pp. 150–151, Nov. 1971. (Cited on page 38.)

[83] U. Seneviratne, D. Rajendran, M. Brusco, and T. G. Phan, "How good are we at diagnosing seizures based on semiology?," *Epilepsia*, vol. 53, pp. e63–e66, Jan. 2012. (Cited on page 41.)

[84] Y. Kaya, M. Uyar, R. Tekin, and S. Yıldırım, "1d-local binary pattern based feature extraction for classification of epileptic EEG signals," *Applied Mathematics and Computation*, vol. 243, pp. 209–219, Sept. 2014. (Cited on page 41.)

[85] K. Samiee, P. Kovacs, and M. Gabbouj, "Epileptic seizure classification of EEG time-series using rational discrete short-time fourier transform," *IEEE Transactions on Biomedical Engineering*, vol. 62, pp. 541–552, Feb. 2015. (Cited on page 41.)

[86] C. Hubsch, C. Baumann, C. Hingray, N. Gospodaru, J.-P. Vignal, H. Vespignani, and L. Maillard, "Clinical classification of psychogenic non-epileptic seizures based on video-EEG analysis and automatic clustering," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 82, pp. 955–960, May 2011. (Cited on page 41.)

[87] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. (Cited on page 42.)

[88] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017. (Cited on page 42.)

[89] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019. (Cited on pages 42, 45 and 49.)

[90] B. Pan, H. Cai, D. Huang, K. Lee, A. Gaidon, E. Adeli, and J. C. Niebles, "Spatio-temporal graph for video captioning with knowledge distillation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 10867–10876, IEEE, 2020. (Cited on page 42.)

[91] Q. T. Ngoc, S. Lee, and B. C. Song, "Facial landmark-based emotion recognition via directed graph neural network," *Electronics*, vol. 9, p. 764, May 2020. (Cited on page 42.)

[92] X. Xu, Z. Ruan, and L. Yang, "Facial expression recognition based on graph neural network," in *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)*, IEEE, July 2020. (Cited on page 42.)

[93] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," *Lecture Notes in Computer Science*, pp. 21–37, 2016. (Cited on pages 43 and 65.)

[94] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017. (Cited on page 43.)

[95] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009. (Cited on page 43.)

[96] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2017. (Cited on page 43.)

[97] F. Massa, "New person keypoint detection models in pytorch domain libraries," 2019. (Cited on page 43.)

[98] C.-H. Chen and D. Ramanan, "3d human pose estimation = 2d pose estimation + matching," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, July 2017. (Cited on page 43.)

[99] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, June 2018. (Cited on page 43.)

[100] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, July 2017. (Cited on page 43.)

[101] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. (Cited on page 45.)

[102] "Challenges in representation learning: Facial expression recognition challenge," 2013. (Cited on page 45.)

[103] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015. (Cited on page 46.)

[104] W. T. Blume, H. O. Lüders, E. Mizrahi, C. Tassinari, W. V. E. Boas, and J. Engel, "Glossary of descriptive terminology for ictal semiology: Report of the ILAE task force on classification and terminology," *Epilepsia*, vol. 42, pp. 1212–1218, Jan. 2002. (Cited on page 48.)

[105] J. Fayerstein, A. McGonigal, F. Pizzo, F. Bonini, S. Lagarde, A. Braquet, A. Trébuchon, R. Carron, D. Scavarda, S. Julia, I. Lambert, B. Giusiano, and F. Bartolomei, "Quantitative analysis of hyperkinetic seizures and correlation with seizure onset zone," *Epilepsia*, vol. 61, pp. 1019–1026, May 2020. (Cited on page 48.)

[106] A. Kheder, U. Thome, T. Aung, B. Krishnan, A. Alexopoulos, G. Wu, I. Wang, and P. Kotagal, "Investigation of networks underlying hyperkinetic seizures utilizing ictal SPECT," *Neurology*, vol. 95, pp. e637–e642, July 2020. (Cited on page 48.)

[107] D. Valeriani and K. Simonyan, "A microstructural neural network biomarker for dystonia diagnosis identified by a DystoniaNet deep learning platform," *Proceedings of the National Academy of Sciences*, vol. 117, pp. 26398–26405, Oct. 2020. (Cited on page 55.)

[108] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," *Procedia Computer Science*, vol. 175, pp. 689–694, 2020. (Cited on page 55.)

[109] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, June 2020. (Cited on page 59.)

[110] W.-C. Huang, C.-H. Wu, S.-B. Luo, K.-Y. Chen, H.-M. Wang, and T. Toda, "Speech recognition by simply fine-tuning bert," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, June 2021. (Cited on page 59.)

[111] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 7871–7880, Association for Computational Linguistics, July 2020. (Cited on page 64.)

[112] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019. (Cited on page 66.)

[113] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2019. (Cited on page 66.)

[114] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems*, 2019. (Cited on page 71.)

[115] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. (Cited on page 71.)

[116] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *ECCV*, 2020. (Cited on page 71.)